B.D. Singh · A.K. Singh

Marker-Assisted Plant Breeding: Principles and Practices

Springer

Marker-Assisted Plant Breeding: Principles and Practices

B.D. Singh • A.K. Singh

Marker-Assisted Plant Breeding: Principles and Practices

Springer

B.D. Singh
School of Biotechnology
Banaras Hindu University
Varanasi, UP, India

A.K. Singh
Division of Genetics
Indian Agricultural Research Institute
New Delhi, Delhi, India

ISBN 978-81-322-2315-3 ISBN 978-81-322-2316-0 (eBook)
DOI 10.1007/978-81-322-2316-0

Library of Congress Control Number: 2015943502

Springer New Delhi Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer (India) Pvt. Ltd is part of Springer Science+Business Media (www.springer.com)

To
Prof. M.S. Swaminathan,
who reshaped agricultural research in India
and inspired a whole generation of plant scientists

Foreword

Plant breeding is the discipline that fashioned our crop plants out of the wild weedy species and continues its endeavor to modify their genotypes to enhance their performance and usefulness to the changing human needs and climate conditions. In the past, the new genotypes developed by plant breeders have been considerably successful in keeping pace with the growing global food needs and consumer preferences. For example, the evolution of hybrid varieties and semi-dwarf cereal genotypes has contributed to quantum jumps in crop productivity, and the latter was responsible for the 'green revolution' that made countries like India virtually self-sufficient in their food grain requirements within a short span of a few years.

The world population is increasing at a rapid rate and is expected to go past the nine billion mark by the year 2042. In addition, the nature and the relevance of both abiotic and biotic stresses are undergoing unrelenting changes in the wake of the environmental alterations engendered by climate change and global warming. In view of these, it is necessary not only to continue to evolve crop genotypes with higher yield potential and tolerance to the various prevailing stresses but also to develop them at a much faster pace. The plant breeders thus face unprecedented challenges of harnessing the reservoirs of genetic variability present in the unadapted germplasm with the minimum investment of time and in a highly precise and predictable manner.

Traditional plant breeding methods rely on phenotype-based selection, but phenotypic evaluation of many traits is problematic, unreliable or expensive. Also, the usefulness of trait phenotypes of individuals/lines in predicting the performance of their progeny is questionable. In addition, the conventional breeding methods do not allow the use of desirable genes from related species in an efficient manner, and there is always the risk of linkage drag. Plant breeders have always been trying to develop breeding strategies that would make their selections more effective and reliable and that would facilitate the utilization of unadapted germplasm with the minimum risk of linkage drag. One of the options that was pursued with some success was the use of simply inherited traits for an indirect selection for complex traits. This effort led to the discovery of protein-based markers and, eventually, the DNA-based markers.

Since the deployment of RFLPs in biological studies, several user-friendly DNA markers like SSRs and SNPs have been developed. The current

emphasis is on technologies that permit low-cost, high-throughput genotyping using molecular markers. Markers are being increasingly used for marker-assisted selection to facilitate gene introgression and for accelerated recurrent selection with the use of off-season nurseries and greenhouse facilities. In addition, markers have found applications in many other plant breeding activities like diversity analysis, germplasm characterization, hybrid seed lot genetic purity determination, elucidation of heterosis loci, etc. In view of the increasing integration of markers in plant breeding programs, many universities have introduced courses on marker-based plant breeding. There is, therefore, an urgent need for a book covering the various aspects relevant to the use of markers in plant breeding.

The book 'Marker-Assisted Plant Breeding' is designed to provide up-to-date information on molecular markers and their applications. The authors have attempted to provide sufficient basic information in an easily understandable narrative so that even the beginners have little difficulty in following the subject. This book will also be useful to teachers, breeders and research workers since it makes available at one place the current information on the various aspects of the subject. The development of different molecular markers and their various applications are described in a simple language, and in a clear and easily comprehendible manner. In the first chapter, the field of marker-assisted plant breeding is introduced and placed in the proper perspective in relation to plant breeding. The next three chapters describe the various molecular marker systems, while mapping populations and mapping procedures, including high-throughput genotyping and association mapping, are discussed in the subsequent five chapters. Four chapters are devoted to various applications of markers, while the last two chapters provide information about relevant bioinformatics tools and phenomics.

The authors deserve compliments for conceiving this book and for developing this concept into a useful and informative book. I am confident that the students, teachers and the professional plant breeders will find this book to be of considerable usefulness as it provides a wealth of information at one place. The book assumes contemporary relevance and importance, since varieties breed with the help of marker-assisted selection are eligible for certification under organic farming.

M S Swaminathan
Research Foundation
Third Cross Street
Taramani Institutional Area
Chennai 600 113, India

Prof. M.S. Swaminathan

Preface

Improved genotypes developed by plant breeding remain pivotal to global food security. In the wake of ever-increasing human population, declining agricultural resources and the stresses generated by climate change, plant breeding is expected to make larger contributions in increasingly shorter time frames. Therefore, plant breeding methods and schemes would have to be made more efficient and capable of accelerated variety development, say, by making efficient use of off-season nurseries, greenhouse facilities and innovative breeding methods. One of the chief limitations of plant breeding is the low effectiveness of phenotypic selection for many traits, particularly the quantitative traits. Further, selection for many other traits is tedious, problematic, time consuming and/or poorly reliable due to threshold requirements, difficulties in assay procedures and phenotype measurement, etc. Breeders have long been searching for tools that would permit effective indirect selection for such traits. Oligogenic phenotypic traits were the first to be used for this purpose, followed by protein-based/isozyme markers. However, the chief limitation of the above marker systems was the limited availability of good informative markers closely linked to the traits of interest.

In 1980, Botstein and coworkers proposed the use of restriction fragment length polymorphism (RFLP) for linkage mapping in humans. RFLP soon emerged as the first DNA-based molecular marker system, and it was used for the preparation of marker linkage maps and for the mapping of several traits of interest in many crops. The greater abundance and other desirable features of RFLPs as compared to phenotypic and protein markers, prompted the development of other relatively more convenient DNA marker systems like random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), simple sequence repeat (SSR), etc. Single nucleotide polymorphism (SNP) has emerged as the most abundant molecular marker that is amenable to high-throughput genotyping. Each of these marker systems offers some advantages and suffers from certain limitations.

Molecular markers provide a tool for identifying genomic regions involved in the control of traits of interest. They also facilitate selection for the target genomic regions on the basis of marker genotype rather than the phenotype of the concerned trait. The reliability of such indirect selection depends mainly on the strength of linkage between the marker and the genomic region of interest. Therefore, markers located within the genes of interest, particularly those associated with allelic differences with respect to

trait phenotype (the functional nucleotide polymorphism, FNP, being the ultimate), would be the most informative and useful. However, the practical usefulness of MAS will be primarily determined by the relative cost of marker development, identification of trait-linked markers and marker genotyping in the breeding populations as compared to the direct trait phenotype-based selection.

The first step in the use of markers for MAS is the identification of markers tightly linked to the traits of interest. Ordinarily, a suitable mapping population needs to be constructed to identify the linked markers by linkage mapping. Several different types of mapping populations, ranging from simple F_2 through recombinant inbred lines to multi-parent advanced generation intercross (MAGIC) and interconnected populations can be used for linkage analyses. Alternatively, a collection of germplasm lines/individuals from natural populations can be used for linkage disequilibrium-based association mapping. In addition, the rich genomic resources that are now becoming available for most crops of interest can be analysed for marker identification.

Molecular mapping of oligogenes is relatively simple, while that of quantitative trait loci (QTLs) poses many problems, and the results from mapping studies are affected by a variety of factors, including the genetic model and the statistical algorithm used for QTL analysis. Generally, different QTLs governing the same trait are identified from different studies and consensus QTLs need to be identified by QTL meta-analysis. In addition, for a reliable detection of association/linkage between markers and traits, the trait phenotypes have to be measured precisely and reliably; a discipline called phenomics devoted to large-scale precision phenotyping is currently the area of intensive research activity.

Molecular markers tightly linked to the desired traits can be used for MAS to select for the genes governing the concerned traits, recover the recurrent parent genotype in backcross programs as well as to eliminate linkage drag, wherever required. Innovative breeding schemes are being designed to facilitate an efficient utilization of resources and to maximize gains from the marker technology. For example, marker-assisted recurrent selection (MARS) is being used for the improvement of quantitative traits, including yield, and up to three generations can be raised in a single season using off-season nurseries or a phytotron. The comprehensive scheme of genomic selection has been proposed for the selection of all genomic regions influencing the traits of interest, whether or not they show significant association with the trait phenotype. An ambitious breeding scheme, breeding by design, has been proposed to accumulate all the positive alleles for all the relevant traits into a single genotype that may be expected to have an outstanding performance. Similarly, a reverse breeding scheme for isolation of complementing inbred pairs from any heterotic hybrid combination has been patented.

Molecular markers have found a variety of other applications, including genetic diversity analysis, phylogenetic studies and construction of heterotic pools. Markers enable unambiguous identification of lines/varieties and facilitate seed certification and PBR (plant breeder rights) implementation. Tightly linked markers provide the basis for fine mapping and positional cloning of genes, which enables generation of information on gene function and regulation, as well as production of transgenic lines expressing the traits of interest.

A successful integration of molecular marker technology in plant breeding would require a low-cost, user-friendly marker systems amenable to high-throughput marker genotyping. Considerable effort is currently focused on the development of low-cost marker identification and genotyping platforms, including genotyping strategies that reduce the volume of genotyping work and/or combine marker discovery with marker genotyping without greatly sacrificing the amount of information obtained. The exciting developments in the above areas are generating new information and concepts/ideas with concomitant creation of specialized terms and phrases that together constitute the discipline of 'marker-assisted plant breeding'.

The chief constraints that limit the integration of molecular markers as a common tool in plant breeding are relatively higher cost of marker genotyping, and the fact that marker technology may appear unfamiliar to those trained in conventional plant breeding. There is continuous generation of new information, concepts/ideas and, inevitably, terminology related to molecular markers and their applications for achieving plant breeding objectives. Further, the marker technology has triggered innovations in breeding strategies and methods and has necessitated the creation of statistical and bioinformatics tools for data processing to facilitate their use for timely decision making. Plant breeding students need to be exposed to the various concepts, procedures and techniques relevant to the field in order to be able to appreciate the opportunities and the limitations of various options offered by the marker technology. It is encouraging that most educational institutions are introducing courses devoted either fully or partly to molecular markers.

The book 'Marker-Assisted Plant Breeding, Principles to Practice' is designed for such students who have had little or no exposure to molecular markers, but have a basic knowledge of genetics and plant breeding, and some exposure to molecular biology. This book will also be useful for teachers, research workers and practicing plant breeders. We have attempted to explain the basic principles, procedures and techniques of marker technology and provide, in brief, the up-to-date information on various aspects in a clear and easily comprehendible manner. Figures and line drawings are provided to highlight the chief features of important procedures/schemes/concepts with a view to facilitate their understanding by the students. In the first chapter, the field of marker-assisted plant breeding is introduced and placed in the proper perspective in relation to plant breeding. The next three chapters describe the various molecular marker systems, while mapping populations and procedures, including high-throughput genotyping and phenotyping, are discussed in the following five chapters. Four chapters are devoted to various applications of molecular markers, including MAS, diversity analysis, positional cloning, etc. The last two chapters provide information about relevant bioinformatics tools and phenomics.

Varanasi, UP, India — Brahma Deo Singh
New Delhi, Delhi, India — Ashok Kumar Singh
November 25, 2014

Acknowledgements

We wish to acknowledge the valuable help received from several of our colleagues and many of our research scholars. Prof. Umesh Singh and Prof. A.M. Kayastha, Varanasi, Dr. Anjana Jajoo, Indore, Dr. Sanjeev Kumar and Dr. Kusum Yadav Lucknow, Dr. K.K. Vinod, Aduthurai, Dr. Vinay, Kumar, Hyderabad, Dr. J.K. Joy, Mohali, Dr. Gopala Krishnan, Dr. Shailesh Tripathi, Dr. Neelu Jain, Dr. M. Vignesh, Dr. Shailendra K. Jha, Dr. Ramya Kurian, Dr. Renu Singh and Ms. Prachi Yadav, New Delhi, and Dr R.K. Sharma, Palampur read parts of the manuscript and/or proof, and suggested very useful additions and alterations. Our research scholars Mallikarjuna, Haritha, Ranjith, Sateesh, Fiyaz, Niranjana, and Naresh, New Delhi, and Vinay K. Singh and Reena Deshmukh, Varanasi, assisted us in a variety of ways, including finding of a suitable flavour for the text, checking for format, style spelling, grammatical errors, etc.

Dr. Kusum Yadav, Lucknow, provided PC scatter plots and the gel image, Dr. Balram Marathi provided the LOD score curves for SIM and CIM, and Reena Deshmukh provided images of multiple sequence alignment and rooted and unrooted trees; these contributions have been suitably acknowledged at the appropriate places. The efforts of Ranjith deserve a special mention as he has meticulously generated and later corrected the line drawings for the entire book.

Several of our colleagues and numerous students suggested that we should develop a book on marker-assisted plant breeding; this was one of the reasons for our decision to undertake this effort. We appreciate that Prof. Kole encouraged us to contact Springer for the publication of this book. We would also like to record our happiness for Springer's decision to publish this book and are thankful to all those involved in this process.

In the end, we are highly appreciative of the affection, support and encouragement we have always received from our family members, including our wives, sons and daughters.

Contents

About the Authors

Brahma Deo Singh Brahma Deo Singh is currently Emeritus Professor at School of Biotechnology, Banaras Hindu University, Varanasi, India. He obtained his bachelor's degree in agriculture from Allahabad Agricultural Institute, Allahabad, India, and master's degree in Agricultural Botany from Government Agricultural College, Kanpur, India, with first position in the university and was awarded the University Gold Medal. He earned his Ph.D. degree from University of Saskatchewan, Saskatoon, Canada. Prof. Singh has 40 years of teaching and research experience in the areas of genetics and breeding of pulse crops, plant tissue culture, biological nitrogen fixation and molecular markers. He has published over 150 research papers in reputed journals and authored several books in genetics, plant breeding and biotechnology. He was awarded the First Prize of the Dr. Rajendra Prasad Puraskar in 1987 and 1990 by the Indian Council of Agricultural Research, New Delhi, for the books *PadapPrajanan* and *Anuvanshiki*, respectively.

Ashok Kumar Singh Ashok Kumar Singh, Fellow of National Academy of Agricultural Sciences, India, is the present Head, Division of Genetics at the prestigious Indian Agricultural Research Institute, New Delhi. He completed his bachelor's and master's degrees from Banaras Hindu University, Varanasi, India, and earned his Ph.D. degree from the institute where he is currently working as a dedicated teacher and rice breeder. He has been associated with the development of 11 Basmati rice varieties, including the first superfine grain aromatic rice hybrid Pusa RH 10, which combine earliness with higher yield and higher per day productivity with excellent grain and cooking quality. He has successfully integrated marker-assisted selection for incorporating resistances to bacterial blight, blast, brown plant hopper drought, salinity and submergence in rice varieties. His current research interests include TILLING, bio-prospecting for genes and novel alleles and marker-assisted breeding in rice. He is well recognized for his contributions to Basmati rice breeding and marker-assisted breeding. He has

over 70 research publications in journals of international repute, and he has been honoured by several awards, including Borlaug Award 2012, Rafi Ahmad Kidwai Award 2013 for research contributions and Bharat Ratna Dr. C. Subramaniam Award 2013 for contribution to teaching.

Abbreviations

AB-QTL analysis	Advanced backcross QTL analysis
AceDB	A *C. elegans* database
AFLP	Amplified fragment length polymorphism
AGR	Absolute growth rate
AIL	Advanced intercross line
ALP	Amplicon length polymorphism
AM	Association mapping
AMP-PCR	Anchored microsatellite-primed PCR
API	Application program interface
AP-PCR	Arbitrary primed PCR
APS	Adenosine 5′-phosphosulfate
ArMet	Architecture for metabolomics
AS-PCR	Allele-specific PCR
ASAP	Allele-specific associated primers
ASO	Allele-specific oligo
ASSRs	Anchored simple sequence repeats
ATP	Adenosine triphosphate
BAC	Bacterial artificial chromosome
BB	Bacterial blight
BC	Backcross
BCF	Binary variant call format
BIL	Backcross inbred line
BLAST	Basic local alignment search tool
BLOSUM	Blocks substitution matrices
BLR	Bayesian linear regression
bp	Base pairs
BSA	Bulked segregant analysis
BSR-Seq	Bulked segregant RNA-Seq
Bt	Bacillus thuringiensis
BV	Breeding value
CAPS	Cleaved amplified polymorphic sequence
CBDP	CAAT box-derived polymorphism
CCD	Charge-coupled device; conserved domains database
CDDP	Conserved DNA-derived polymorphism
cDNA SNPs	SNPs discovered in cDNAs
cDNA-SSCP	SSCP analysis of cDNA
cDNA	Copy DNA or complementary DNA

CDS	Coding sequence
CGH	Comparative genomic hybridization
ChD	Discordant chastity
CID	Carbon isotope discrimination
CIM	Composite interval mapping
cM	CentiMorgans
CMLM	Compressed MLM
CMS	Cytoplasmic male sterility
CNV	Copy number variation
Combined MAS	Combined marker-assisted selection
CoRAP	Conserved region amplification polymorphism
COS	Conserved orthologous set (of genes)
CRoPS	Complexity reduction of polymorphic sequences
Cry protein	Crystal protein
cSNPs	Copy SNPs
CSSL	Chromosomal segment substitution line
CTAB	Cetyltrimethylammonium bromide
CTD	Canopy temperature depression
D/TGGE	Denaturing/temperature gradient gel electrophoresis (also, DGGE/TGGE)
DAF	DNA amplification fingerprinting
DArT	Diversity array technology
DAS	Distributed annotation server
DASH	Dynamic allele-specific hybridization
dATP	Deoxyadenosine triphosphate
dATPS	Deoxyadenosine -thiophosphate
Dbfetch	Database fetch
DBMS	Database management systems
dbSNP	SNP Database
dCAPS	Derived cleaved amplified polymorphic sequence
dCTP	Deoxycytidineine triphosphate
DDBJ	DNA data bank of Japan
ddNTP	Dideoxynucleotide triphosphate
ddRAD-Seq	Double digest restriction-site-associated DNA sequencing
DG_{ij}	Gower's measure of distance or the average taxonomic distance
dGTP	Deoxyguanonosine triphosphate
DH	Doubled haploid
dHPLC	Denaturing high-performance liquid chromatography
DNA	Deoxyribose nucleic acid
dNTP	Deoxynucleotide triphosphate
DP	Donor parent
DP	Donor parent
dRAMP	Digested random amplified microsatellite polymorphism
dsDNA	Double-stranded DNA
dTTP	Deoxythymidine triphosphate
DUS	Distinctness, uniformity and stability

e-PCR	Electronic PCR
EBI	European Bioinformatics Institute
EBL	Expected bin length
EBV	Estimated breeding value
EMAIL	Endonucleotic mutation analysis by internal labeling
EMBL	European Molecular Biology Laboratory
EMF	Expectation maximization algorithm (under fixed effect model)
EMMA	Efficient mixed model association
EMMAX	EMMA expedited
epiRIL	Epigenetic recombinant inbred line
eQTLs	Expression QTLs
eSNPs	Electronic SNPs
eSSRs	Expressed SSRs
ESTs	Expressed sequence tags
EVA	Evaluation of automatic protein structure prediction
F_2:F_3	F_2-derived F_3
FaST-LMM	Factored spectrally transformed linear mixed model
FBP	Field-based phenomics
FDR	False discovery rate
FEN-1	Flap endonuclease 1
FISH	Fluorescence in situ hybridization
FM	Fitch-Margoliash
FPD	Fractioned-pool design
FRET	Fluorescence resonance energy transfer
FTP	File transfer protocol
G × E interaction	Genotype × environment interaction
Gb	Gigabase
GBS	Genotyping by sequencing
GBA	Genome bit analysis
GCA	General combining ability
GD	Genetic distance
GD_J	Jaccard's coefficient
GD_{MR}	Modified Rogers' distance
GD_{NL}	Nei and Li's coefficient
GD_{SM}	Simple matching coefficient
GEBV	Genomic estimated breeding value
GEI	Genotype × environment interaction
GEMMA	Genome-wide efficient mixed model association
GEO	Gene expression omnibus
GFP	Green fluorescent protein
GLD	Gametic linkage disequilibrium
GLM	General linear model
GLM	General linear model
GO	Gene ontology
GOC	Gene ontology consortium
GPD	Gametic phase disequilibrium

GPS	Geographical positioning system
GRAMMAR	Genome-wide rapid association using mixed model and regression
Gy	Gray
GS	Genomic selection
GSDB	Genome sequence databases
GSFLX	Genome sequencer FLX
gSSR	Genomic SSR
GUI	Graphical user interface
GWAS	Genome-wide association study
H_1	Alternative hypothesis
HLs	Heterosis loci
hME	Homogeneous MassEXTEND
HMM	Hidden Markov model
HMPR	Hypomethylated partial restriction
H_0	Null hypothesis
HPGA	High-throughput plant growth analysis
hQTLs	Heterosis QTLs
HRR	Haplotype relative risk
HTG	High throughput genome
HTML	Hypertext-markup language
htSNPs	Haplotype tagging SNPs
IBD	Identity-by-descent
ICIM	Inclusive composite interval mapping
IF_2	Immortalized F_2
ILL	Introgression line library
IM	Interval mapping
iMAS	Integrated marker-assisted selection
IMP	Inter-MITE polymorphism
InDels	Insertions and deletions
INSDC	International nucleotide sequence database collaboration inter-SSR PCR
IP	Intellectual property
IPR	Intellectual property rights
IR	Infrared
IRAP	Inter-retrotransposon amplified polymorphism
IRILs	Intermated recombinant inbred lines
ISA	Inter-SSR amplification
isSNPs	In silico SNPs
ISSR	Inter-simple sequence repeat
ITP	Intron-targeting polymorphism
JICIM	Joint inclusive CIM
JLAM	Joint linkage and association mapping
KASP	Kompetitive allele-specific PCR
kb	Kilo base pairs
KEGG	Kyoto encyclopedia of genes and genomes
LAD	Leaf area duration

LASSO	Least absolute shrinkage and selection operator
LCR	Ligase chain reaction
LD	Linkage disequilibrium
LDU	Linkage disequilibrium units
LIDAR	Light detection and ranging
LIFT	Laser-induced fluorescence transient
LMM	Linear mixed model
LOD	Logarithm of odds
LRT	Likelihood-ratio test
LSO	Locus-specific oligo
LTR	Long terminal repeat
M	Morgan
MABC	Marker-assisted backcrossing
MADAM	Microarray data manager
MAGE-ML	Microarray gene expression markup language
MAGIC	Multiparent advanced generation inter-cross
MaizeGDB	Maize genetics and genomics Database
MALDI-TOF MS	Matrix-assisted laser desorption ionization time of flight mass spectrometry
MANOVA	Multivariate analysis of variance
MARS	Marker-assisted recurrent selection
MAS	Marker-assisted selection
Mb	Mega base pairs
MBDC	Molecular biology database collection
MBL	Maximum bin length
McFISH	Multicolor FISH
MCMC	Markov chain Monte Carlo
MDS	Multidimensional scaling
ME	Minimum evolution
MEGA-AFLP	Multiplex-endonuclease genotyping approach AFLP
MES	Marker-evaluated selection
MeV	Multi-experiment viewer
Mha	Million hectares
MIAME	Minimal information about a microarray experiment
MIDAS	Microarray Data Analysis System
MIM	Multiple interval mapping
MIP	Molecular inversion probe
MIPS	Martinsried Institute of Protein
MITE	Miniature inverted-repeat transposable element
MLE	Maximum likelihood estimate
MLM	Mixed linear model, modified location model
MLs	MAGIC lines
MMAPPR	Mutation mapping analysis pipeline for pooled RNA-Seq
MP-PCR	Microsatellite-primed PCR
MPS	Massively parallel sequencing
MPSS	Massively parallel signature sequencing

MQM	Multiple QTL mapping
mQTLs	Metabolic QTLs
MRI	Magnetic resonance imaging
MS	Mass spectrometry
MSG	Multiplexed shotgun genotyping
MTA	Marker-trait association
MTI	Multitrait index
MTMM	Multi-trait mixed model
MuSICA	Multiclone shotgun integrated cDNA assembler
mvLMM	Multivariate linear mixed models
NAM	Nested association mapping
NCBI	National Centre for Biotechnology Information
ncSNPs	Non-coding SNPs
NDVI	Normalized difference vegetative index
Ne	Effective population size
ng	Nanogram
NGM	Next generation mapping
NGS	Next generation DNA sequencing
NIH	National Institutes of Health
NIL	Near-isogenic line
NIR	Near infrared
NJ	Neighbor joining
nL	Nanolitre
NLM	United States national library of medicine
NMR	Nuclear magnetic resonance
NR	Nonredundant Database
nsSNP	Nonsynonymous SNP
nt	Nucleotide
OLA	Oligonucleotide ligation assay
OMIM	Online Mendelian inheritance in man
ORF Finder	Open reading frame finder
ORF	Open reading frame
PAC	P_1-derived artificial chromosome
PAGE	Polyacrylamide gel electrophoresis
PAM	Point accepted mutation, pulse amplitude modulated
PBR	Plant breeder's rights
PC	Principal components
PCA	Principal component analysis
PCoA	Principal coordinate analysis
PCR	Polymerase chain reaction
PDB	The protein data bank
PEA	Plant efficiency analyzer
PERL	Practice extraction and reporting language
pFDR	Positive false discovery rate
PHI-BLAST	Pattern hit iterated BLAST
PHYLIP	Phylogeny inference package
PIC	Polymorphic information content

PIR	Protein information resource
PlantGDB	Plant genome database
PlaRoM	Root-monitoring platform
PO	Plant ontology
pQTLs	Protein quantity QTLs
PPi	Pyrophosphate
PS II	Photosystem II
PSI-BLAST	Position specific iterated BLAST
pSNPs	Promoter SNPs
QPM	Quality protein maize
QTL	Quantitative trait locus
QTL-NIL	QTL-near isogenic lines
QTNs	Quantitative trait nucleotides
RAD	Restriction-site-associated DNA
RADSeq	Restriction-site-associated DNA sequencing
RAHM	Random amplified hybridization microsatellites
RAM	Randomly amplified microsatellites
RAMP	Random amplified microsatellite polymorphism (also, RAMPO)
RAP-PCR	RNA fingerprinting by arbitrarily primed PCR
RAPD	Random amplified polymorphic DNA
RBFNN	Radial basis function neural network
RBIP	Retrotransposon-based insertion polymorphism
RD	Roger's distance
RDBSM	Relational database management system
REBASE	Restriction enzyme database
refSNP	Reference SNP
REMAP	Retrotransposon-microsatellite amplified polymorphism
REML	Restricted maximum likelihood
RFLP	Restriction fragment length polymorphism
RFP	Red fluorescent protein
RGB	Red, green and blue
RGR	Relative growth rate
RIL	Recombinant inbred line
RIX	Recombinant inbred intercross
RKHS	Reproducing kernel Hilbert spaces
RNA-Seq	RNA sequencing
RNAi	RNA interference
RP	Recurrent parent
RQM	Reverse QTL mapping
RR-Seq	Reduced representation sequencing
RRLs	Reduced representation libraries
RSB	Recurrent selection backcross
RSBI	Recurrent selection backcross *inter se* intercross
RT-PCR	Reverse transcriptase PCR
RWC	Relative water content
S-SAP	Sequence-specific amplification polymorphism

SA model	Structured association model
SAMPL	Selective amplification of microsatellite polymorphic loci
SBE	Single base extension
SBS	Sequencing by synthesis
SCA	Specific combining ability
SCAR	Sequence characterized amplified regions
SCN	Soybean cyst nematode
SCoT	Start codon-targeted
SDP	Selective DNA pooling
SE	Standard error
SFP	Single feature polymorphism
SIM	Simple interval mapping
SLS-MAS	Single large scale MAS
SMRT	Single-molecule real-time
SMS	Single molecule sequencing
SNP	Single nucleotide polymorphism
SNP2CAPS	SNP-to-CAPS
SOAP	Simple object access protocol
SOLiD	Sequencing by oligonucleotide ligation detection
SPARs	Single primer amplification reactions
SPP	Single position polymorphism
SQL	Structured query language
SRAP	Sequence-related amplified polymorphism
SRS	Sequence retrieval system
SSAP	Sequence-specific amplification polymorphism
SSCP	Single strand conformation polymorphism/profile
SSD	Single seed descent
ssDNA	Single-stranded DNA
SSLP	Simple sequence length polymorphism
SSR	Simple sequence repeat
STMP	Sequence-tagged microsatellite profiling
STMS	Sequence-tagged microsatellite site
STS	Sequence tagged site
synSNP	Synonymous SNP
TAC	Transformation competent artificial chromosome
TASSEL	Trait Analysis by aSSociation, Evolution and Linkage
TD	Transposon display
TDR	Time domain reflectometry
TDT	Transmission disequilibrium test
TE-AFLP	Three-endonuclease AFLP
TE	Transposable element
TGS	Third generation DNA sequencing
TILLING	Targeting induced local lesions in genomes
TMV	Tobacco mosaic virus
TO	Trait ontology
T_m	Melting temperature

TPA	Third party annotation
TRAP	Target region amplification polymorphism
TrEMBL	Translated EMBL
tSNPs	tagSNPs
TSR	Target-specific region
TUS	Tentative unique sequences
UGMs	Unigene-derived microsatellites
UML	Universal modeling language
UniProtKB	UniProt Knowledgebase
UPGMA	Unweighted paired group method using arithmetic average
UPGMC	Unweighted paired group method using centroids
UPOV	Union Internationale pour la Protection des Obstentions Vegetales
UV	Ultraviolet
VNTRs	Variable number of tandem repeats
WGRS	Whole genome resequencing
WUE	Water-use efficiency
XML	Extended HTML
YAC	Yeast artificial chromosome
YFP	Yellow fluorescent protein
ZLD	Zygotic linkage disequilibrium

Part I

General

Introduction to Marker-Assisted Crop Improvement

1

1.1 Introduction

The development of agriculture, i.e., the cultivation of some selected plant species, was perhaps the most important turning point in human history. As food availability from crop cultivation would have increased, the need for hunting and gathering would have declined. As a result, there would have been a shift from the nomadic to the settled lifestyle. As the uncertainty of food availability would have declined, more and more time would have become available for activities that form the basis of our civilization and culture. Agriculture would also have encouraged innovation in both tools and methods to support the crop cultivation and associated activities. There would have been conscious, albeit unplanned, effort to select plants with desirable features for planting in the next season. Planned and systematic selection efforts began during the nineteenth century, and plant breeding activities began to acquire a scientific framework with the rediscovery of the Mendel's laws of inheritance. Plant breeding has become a highly organized activity and is credited with dramatic increases in agricultural production. In many countries, plant breeding has developed into a highly successful industry. However, the gains in agricultural production have been negated mainly by population growth and, of late, climate change, making it necessary to develop improved crop varieties more efficiently and rapidly.

One of the chief limitations of the breeding methods is that the decision about the worth of different lines and even individual plants has to be based on their phenotypes. This has been recognized for a long time to reduce the efficiency of breeding methods and, in some situations, to delay the development of improved varieties. Therefore, a systematic and sustained search for easily scorable markers that could be used for a reliable indirect selection for target traits was initiated. This search began with morphological traits and eventually led to the development of DNA-based molecular markers. These markers allow the identification and mapping of the desired genes and an efficient indirect selection for the target traits. They have also motivated the development of novel breeding approaches for fully exploiting the potential of marker technology. Thus, the integration of molecular markers in plant breeding activities has given rise to the new discipline described as "smart breeding," "molecular breeding," or "marker-assisted breeding." The crop varieties developed by marker-assisted breeding are often referred to as 'Super Organics'.

1.2 Domestication: The Evolution of Crop Plants

The present-day crop species have evolved from wild weedy species. This evolution began about 11,000 years ago when humans chose a relatively small number of wild species for growing them under human management, i.e., domestication.

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_1, © Author(s) 2015

The cereals, legumes, and plant species used for their fruits and roots were the first to be domesticated. The exact series of events during domestication is not well known, but a strong selection pressure seems to have been exerted by the humans in the domesticated populations. As a result, rapid and radical changes occurred in these populations causing them to diverge from their progenitor wild species and, ultimately, to evolve as our present-day crop species. The crop plants seem to have been domesticated in six different regions of the world. Yet the sets of traits selected for in the various domesticated plant species, i.e., the domestication syndrome traits, are largely similar. Almost all the important crop species were domesticated early in the history of agriculture, and only few crops have been domesticated during the recorded history. However, domestication continues to be relevant, and the current focus is on species having desirable features for biofuel production. Selection would have led to a progressive decline in the genetic variability present in the domesticated populations. This trend has continued with the modern plant breeding schemes, which usually exploit crosses among a small number of related elite lines (Singh 2012a).

1.3 Plant Breeding

Plant breeding aims at changing the genetic constitution of crop plants to make them more useful to humans. In view of this, plant breeding has been often described as "plant evolution directed by man." The main outcome of plant breeding are improved cultivars or varieties having superior agronomic features, higher yield potential and/or better produce quality. The improved crop varieties have been the chief contributors to the large increases in agricultural production during the past several decades. However, hunger and malnutrition continue to afflict about 10 % of the human population; they claim more human lives each year than AIDS, tuberculosis, and malaria combined together. The world human population has been increasing at a rapid rate since the industrial revolution around early 19th century, and it crossed the 7.2 billion mark in April 2014. It is projected to reach eight billion by the year 2024 and to cross nine billion by the year 2042 (http://www.worldometers.info/world-population/#pastfuture). About 90 % of the population increase will take place in the developing countries where food and water are already in short supply. Cereals are grown on more than half of the global cropped area. The annual growth rate of global cereal production was over 3 % during the 1970s. It declined to 1 % per year during the 1990s and to zero between 2000 and 2003, but it rose to over 2 % toward the end of the first decade of the twenty-first century. The rate of growth in global yields of pulses and root crops has been well below 1 % per year during the last five decades (http://www.fao.org/docrep/018/i3107e/i3107e03.pdf). Therefore, continued growth in agricultural production at a sufficient rate to avert large-scale food shortage is regarded as one of the greatest challenges for plant scientists during the twenty-first century.

1.3.1 Major Developments in Plant Breeding

Plant breeding may be considered to have begun with the domestication of a small number of promising wild species. Planned selection for superior plant types began around the second decade of the nineteenth century, and several excellent varieties of cereals were evolved. Plant hybridization dates back to 1717 when Thomas Fairchild crossed carnation with sweet William. Hybridization has been extensively used for creating genetic variation, and it continues to be the dominant method for this purpose. Distant hybridization, including somatic hybridization, has been used to access genes/alleles present in the wild relatives but not available in the cultivated germplasm (Singh 2012a). It has been argued that many crops would have lost their commercial status without the support from their wild relatives. Genetic variation has also been created by mutagenesis and through somaclonal variation generated by tissue culture. Both auto- and allopolyploidy

have been evaluated for their usefulness, and a new allopolyploid crop species, triticale, has been developed. The recombinant DNA technology enables transfer and expression of specific genes into plants (Sect. 1.5). The above activities create genetic variation to be utilized by selection for developing genotypes with superior characteristics. The selected genotypes are evaluated in replicated trials, preferably, over locations and years to ascertain their superiority over the existing varieties. A new superior genotype is finally multiplied and distributed for commercial cultivation.

1.3.2 The Genotype and Phenotype

Genotype describes the allelic constitution of an individual at one or more loci, while phenotype is the observable expression of one or more traits. The concept of genotype and phenotype was articulated by Johannsen in 1909, but it was reflected in the progeny test used by the breeders and elaborated by Vilmorin during the 1850s. The findings of Mendel showed that the phenotypes of qualitative traits are good indicators of the genotypes at the concerned loci. But some of the individuals with the dominant phenotype for a trait breed true, i.e., are homozygous, while others behave as hybrid, i.e., are heterozygous, and produce both dominant and recessive phenotypes in their progeny. The progeny test reveals the homozygous or heterozygous state of the concerned loci as well as the extent of the environmental effect on the phenotype (Singh 2012a). The latter is particularly relevant for quantitative traits because their phenotypic expression is affected by the environment and, often, an interaction between the genotype and the environment. Thus, the phenotype can be expressed by the following equation:

$$P = \mu + G + E + (G \times E) \qquad (1.1)$$

where P is the phenotype of a quantitative trait, μ is the population mean, G is the effect of genotype of the concerned individual, E is the effect of environment on expression of the trait, and $(G \times E)$ is the interaction component between the genotype and the environment. A precise estimation of G, E and $G \times E$ components of the phenotypic variation for different quantitative traits is one of the continuing quests of plant breeding.

1.3.3 Genetic Variation: Qualitative and Quantitative Inheritance

The genetic variation present in a population is observable as differences in trait phenotypes of the individuals in a given population. The various traits of an organism can be grouped into the following two categories: (1) qualitative and (2) quantitative traits. In general, each *qualitative trait* is governed by one or few *major genes* or *oligogenes*, each of which produces a large effect on the trait phenotype (Singh 2012a). Mendel analyzed the inheritance of qualitative traits to formulate the laws of segregation and independent assortment. When a trait is governed by two or more genes, they may interact in various ways to produce different phenotypic ratios in F_2. The phenotypic expression of oligogenes is generally little affected by the environment. Therefore, the individuals can be grouped into distinct classes on the basis of trait phenotype, which often serves as good indicator of the genotype at the concerned locus (Table 1.1). However, some oligogenes require a specific environment, i.e., a threshold environment, for their expression; such traits are called *threshold characters*. For example, the phenotypic expression of a disease resistance gene can be assessed only when the concerned pathogen comes in contact with the plants, and the environmental conditions are favorable for disease development. It was once believed that each oligogene affects a single trait, but many of them are known to affect multiple traits; this is known as *pleiotropy*. In general, the expression of the wild-type allele is not affected by the normal range of variation in the environment and the genetic background. However, the expression of mutant alleles is often influenced

Table 1.1 The classical view of qualitative and quantitative traits

Feature	Qualitative trait	Quantitative trait
Number of genes involved	One or few	Many
Genes are known as	Oligogenes or major genes	Polygenes
Effect of each gene on trait phenotype	Large	Small
Effect of other genes affecting the trait	Various types of gene interaction	Cumulative and epistatic gene effects
Effect of environment on trait phenotype	Usually, little[a]	Small to large
Effect of the genetic background	Generally, little[b]	Generally, considerable
Trait variation	Discontinuous	Continuous
Individuals grouped into distinct phenotypic classes	Almost always[c]	Never

[a]Some traits have threshold requirement
[b]Modifying genes are known to affect the expression of, generally, mutant alleles of oligogenes. The expressions of the wild-type alleles of oligogenes are generally well adjusted to the normal range of variation in the genetic background
[c]Some mutant alleles exhibit variable expressivity in response to the environment and the genetic background. This generates almost continuous variation in the phenotype of the concerned qualitative trait

Table 1.2 The current view of qualitative and quantitative traits

Qualitative/ quantitative trait	Distribution in F_2	Phenotypic variation explained (%)	Example	Governed by
Qualitative	Discrete	100	Blast resistance in rice, opaque kernels in maize	Major gene or oligogene
Semi-quantitative	Discrete	100	Semi-dwarfism (*sd1*) in rice	Oligogene
Quantitative	Continuous	>50	Submergence tolerance (*sub1*) in rice; most biochemical traits	Oligogene
Quantitative	Continuous	25–50	Stem rot resistance in rice	Large effect QTL
Quantitative	Continuous	<25	Most agronomic and physiological traits	QTLs

Based on Mackill and Junjian (2001) and Babu et al. (2004)

by changes in the environment as well as the genetic background.

Most of the traits of biological as well as economic significance, however, show continuous variation, and the individuals cannot be grouped into distinct phenotypic classes; these traits are called *quantitative* or *metric traits*. In 1908, Nilsson-Ehle proposed the multiple factor hypothesis, which provided the basis for polygenic inheritance of quantitative traits. It is presumed that each *polygene* has a small effect on trait phenotype, and the effects of all polygenes affecting a trait are cumulative (Table 1.1). It is now recognized that polygenes show dominance and epistatic gene actions in addition to their additive effects. The continuous variation of quantitative traits is explained to be the result of polygenic control, and the effects of the environment and the genetic background on the expression of polygenes. In fact, even monogenic traits tend to show continuous variation due to the environmental effects on their expression (Singh 2012a). The current view of qualitative and quantitative traits is based almost exclusively on the extent of phenotypic variation explained by the concerned gene(s) (Table 1.2).

The most important outcome of the differential environmental effects on qualitative and quantitative traits concerns their response to phenotypic selection. In case of qualitative traits, phenotypic selection is relatively simple and highly effective. But in case of quantitative traits, the effectiveness of phenotypic selection depends primarily on heritability of the trait. *Heritability* is the proportion of genetic variance for a trait to its phenotypic variance. But in case of segregating generations, the appropriate measure of heritability will be the ratio of additive

genetic variance to the phenotypic variance for the trait. The heritability is reflected in the effect size of the QTL (quantitative trait locus) governing the trait. A *QTL*, in simple terms, is the genomic region that is involved in the control of a quantitative trait. However, a single QTL may have one or more genes affecting the concerned trait. Selection is generally effective for quantitative traits governed by one or few QTLs with large effects. But selection is nearly ineffective for traits that have moderate to low heritability and are governed by several small effect QTLs. Most of the traits of economic interest, including yield, seem to belong to the latter category (Singh 2012a).

In spite of the limitations of phenotypic selection, conventional plant breeding has been remarkably successful in improving our crop species. A large number of varieties with improved characteristics, including yield and produce quality, have been developed in each crop species. The breeding strategies and the genetic makeup of the improved varieties depend primarily on the modes of reproduction and pollination of the concerned crop. In self-pollinated crops, the varieties are generally pure lines, hybrid varieties are common in cross-pollinated crop species, and clones are used for commercial cultivation in asexually propagated crops.

1.3.4 Contributions: Pure Line Varieties

A *pure line* is the self-pollinated progeny of a single homozygous plant of a self-pollinated species. As a result, all the individuals within a pure line have identical genotype, and the phenotypic variation observable in a pure line is nonheritable. The pure line concept was developed by Johannsen in 1903 on the basis of the results of selection for seed size in common bean (*Phaseolus vulgaris*), a self-pollinated species. Self-pollinated species show <5 % natural cross-pollination. The most important outcome of self-pollination is the progressive increase in homozygosity with the generation. As a result, populations of self-pollinated species eventually become mixtures of pure lines. Initially, self-pollinated crops were improved by individual plant or pure line selection that utilized the genetic variation existing in their populations. Subsequently, hybridization between selected lines/genotypes was used to create the desired genetic variation, and the pedigree method was the most widely used selection scheme for isolating superior pure lines. In this scheme, individual plants are selected till they become homozygous by, say, F_5 or F_6. But in the bulk method, the population is grown in bulk till F_6 or later; individual plants are then selected to isolate pure lines (Singh 2012a). The bulk scheme has been used to a limited extent, but it is receiving renewed interest as marker-evaluated selection (Sect. 9.11.8). A variant of the bulk scheme, the single-seed descent (SSD) method, is widely used for isolating a mixture of pure lines, e.g., recombinant inbred lines (Sect. 5.8), from appropriate crosses. The above breeding schemes allow selection for new genotypes that may be superior to the parents of the cross (transgressive breeding). The backcross breeding scheme, however, is designed for transferring one or a few genes from an otherwise undesirable genotype (the donor parent) into a popular variety deficient in the concerned traits (the recurrent parent). Thus, the end product of a backcross program is the recurrent parent without the defect(s) that were corrected by the introgressed gene(s). The above breeding schemes have supported around 1–3 % annual increase in the yields of the three main cereals, viz., rice, maize, and wheat.

1.3.5 Contributions: Hybrid Varieties

The cross-pollinated crops are essentially random mating and may show up to 100 % cross-pollination. As a result, they are highly heterozygous and show loss in vigor and fertility as a result of inbreeding (inbreeding depression). The genetic constitution of such populations is described in terms of gene and genotype frequencies. When such a population is at equilibrium for a gene with two alleles, the frequencies of the three genotypes at this locus

are p^2, $2pq$, and q^2. Selection in such a population is expected to increase the frequencies of selected alleles. Therefore, the mean of concerned traits would change in the direction of selection. But mass selection and line-breeding schemes like ear-to-row method were ineffective in increasing the yields of maize populations. It was realized that this ineffectiveness was mainly due to the low heritability of traits like yield and an increase in the level of inbreeding of the selected populations. Therefore, schemes using progeny test as the basis of selection and a mating scheme designed to minimize inbreeding were developed. The three recurrent selection schemes are the most elaborate and perhaps the most effective of the various selection schemes. These schemes involve selfing as well as crossing of the phenotypically superior plants to a tester, evaluation of the test-cross progeny in replicated trials and final selection of the plants on the basis of progeny test. The selfed seeds from the selected plants are grown separately, and their progeny are intermated in all possible combinations to generate the selected version of the population. Thus, each selection cycle of these schemes requires 2–3 years. In general, selection has been used to increase the frequencies of desirable alleles in open-pollinated populations, from which superior inbreds have been isolated (Singh 2012a).

The ineffectiveness of the early selection schemes prompted the use of F_1 hybrids for commercial cultivation. Initially, crosses between open-pollinated populations were used as hybrid varieties. But in 1909, Shull suggested the use of inbreds, isolated from open-pollinated populations, for the production of hybrid varieties. The first commercial hybrid variety of maize, Burr Leaming Dent, was released in the USA in 1922. Maize hybrids were slow to gain popularity, but by 1950s they had completely replaced the open-pollinated varieties in the USA. Hybrid varieties are the best means of exploiting *heterosis*, which signifies the superiority of a F_1 hybrid over its parents. Self-pollinated crop species also show heterosis, and hybrid varieties are in commercial cultivation wherever hybrid seed production is not a constraint. It may be pointed out that the genetic and biochemical/physiological bases of heterosis are far from clear, but its commercial exploitation has been quite rewarding (Singh 2012a).

Several of our crops are often cross-pollinated as they show more than 5 % cross-pollination. The genetic makeup of these crops is regarded as intermediate between those of self- and cross-pollinated species. Therefore, both pure line and hybrid varieties of these crops are used for commercial cultivation.

1.3.6 Contributions: Clones

Many of our crops like potato, sweet potato, sugarcane, etc. are asexually propagated. These crops, often called clonal crops, are not exposed to segregation and recombination, which are the inevitable consequences of sexual reproduction. As a result, clonal crops are highly heterozygous and show severe inbreeding depression. These crops offer a unique advantage as any plant with desirable features can be asexually propagated to obtain a superior clone. A *clone* is asexual progeny of a single asexually reproducing plant. Therefore, all the plants in a clone have the same genotype, and the phenotypic variation within a clone is nonheritable. The improvement of clonal crops usually involves selection of individual plants from a variable population (clonal selection), hybridization followed by individual plant selection and/or mutagenesis coupled with selection. The chief problems in breeding of clonal crops are reduced flowering and fertility, perennial life cycle (in many cases) and difficulties in genetic analyses (Singh 2012a).

1.3.7 Limitations of Phenotype-Based Plant Breeding

The decisions in conventional plant breeding are based on phenotypic evaluation for the target traits. The value of a quantitative trait phenotype for selection depends on the heritability of the trait. Therefore, quantitative traits have to be

evaluated in replicated trials preferably conducted under different environments. This increases the evaluation costs and limits the trials to such locations and seasons that allow meaningful expression of the concerned traits. Therefore, off-season nursery and greenhouse facilities cannot be used for selection for traits like yield (Singh 2012a). Further, traits like fruit/seed characteristics and yield can be evaluated only at maturity. As a result, the selected plants cannot be used for hybridization in the same generation/season. The phenotypic evaluation for many traits may require specific environments, including inoculation with a specific race of the concerned pathogen. The creation of some environments may be difficult or demanding. In addition, phenotypic evaluation for some traits may take time, may be tedious or may be expensive. In some cases, the results from phenotypic evaluation may not be reliable due to the environmental effects.

One of the chief limitations of phenotype-based breeding is the nonavailability of an effective selection scheme during the early segregating (F_2–F_4) generations from crosses. Since individual plants are selected in these generations, selection is effective only for highly heritable traits. Another major limitation relates to the selection of parents for hybridization for the improvement of quantitative traits. A variety of approaches based on performance of the parents themselves or of the progeny (F_1 or a later generation) from their crosses have been proposed, but none of them is effective in all the cases (Singh 2012a).

1.4 The Growing Food Needs

The world human population has been increasing at a rapid rate. It crossed 7.2 billion in April 2014 and is expected to reach nine billion by 2042 (http://www.worldometers.info/world-population/#pastfuture). It has been projected that yields of rice, maize, and wheat must increase by at least 70 % before 2050 to feed the increasing human population (Furbank and Tester 2011). In the past, agricultural production increased due to the combined effects of improved crop varieties; increased use of inputs like fertilizers, pesticides, and irrigation water; and increased area of cultivation. But the increase in agronomic inputs cannot continue for long as freshwater reserves and petroleum resources (used for fertilizer and pesticide production) are steadily declining, and the costs of both fuel and fertilizer are on the rise. According to an estimate, nearly 70 % of the world's fresh water extracted for human use is utilized for agriculture, and the demand for water is expected to increase substantially with time. At the same time, the high-input agricultural practices are known to cause environmental pollution.

It may be added that a part of the existing farmland is being used for urban expansion and other developmental activities. The increased use of irrigation water and chemical fertilizers is causing salinization of the cultivated areas: about 30 % of the arable land may become salinized by 2025, and this figure may rise to 50 % by 2050. In addition, water stress and desertification are also reducing the area of arable land. Therefore, the total cultivated area can be increased mainly by using forest land for cultivation, which is not desirable. The global climate change is expected to cause a steady rise in temperature and unpredictable precipitation leading to moisture stress and reduced crop yields (Reynolds et al. 2009). There is some evidence that climate change is leading to altered prevalence of plant diseases, evolution of new pathotypes, and increased activities of insect pests; these changes would reduce crop yields. In addition, many insect pests are becoming resistant to insecticides, and many of the effective insecticides are now banned. The modern agricultural practices are perceived to encourage soil erosion, loss of fertility, and reduced biodiversity. It is feared that their continued use would lead to a serious degradation of the environment (http://www.foodsecurity.ac.uk, Collard and Mackill 2008). Another factor affecting food availability is the diversion of food grains for biofuel production. In 2007–2008, about 10 % of the global coarse grain production was used for making ethanol (Sticklen 2007). In addition, increasing crop areas may be expected to be

diverted for cultivation of designer crops for producing specific biochemicals. Finally, there is an increasing demand for meat and meat products, especially in the developing world. The increased meat production will further strain food availability as increasingly greater quantities of food grains will be used as feed for the meat-producing animals (http://www.foodsecurity.ac.uk).

Thus, the genetic improvement of crops seems to be the most viable approach to enhance agricultural production. It will be necessary to develop new high-yielding genotypes that combine high yield potential with yield stability under abiotic and biotic stresses. A possible positive impact of climate change on crop yields might be through the beneficial effects of enhanced CO_2 levels on photosynthesis. But the modern crop varieties show poor responses to elevated CO_2 levels, and there is large within and between species variation for this response. Therefore, efforts should be made to develop "smart" crop genotypes capable of taking full advantage of the environment generated by the climate change (Zisca and Bunce 2007; Tester and Langridge 2010; Furbank and Tester 2011). But the annual increase in cereal yields as a result of conventional breeding has declined considerably and has reached a plateau. During 1997–2010, the annual increase in cereal yields was almost one third of that between 1960 and 1980 (Fischer et al. 2014). Thus, the traditional breeding programs do not seem to be capable of meeting the projected demands for agricultural production. It has been suggested that exploitation of the genomics resources by the transgenic and the molecular marker technologies might offer solutions to the current challenges in plant breeding.

1.5 The Transgenic Technology: Lukewarm Social Response

A gene introduced into an organism by recombinant DNA technology is known as *transgene*, and a plant expressing such gene(s) is called *transgenic plant*. The transgene is integrated into a suitable plant expression vector and then introduced into the plant cells using a suitable transformation technique like *Agrobacterium* coculture or particle gun acceleration. The expression vector has all the regulatory sequences required for efficient gene expression in plants. It also has a selectable reporter gene for the selection of the transformed plant cells. The putative transgenic plants obtained from the transformed cells are intensively evaluated for the expression of the transgene and for agronomic performance. All transgenic plants are subjected to the required biosafety assays, including toxicity and allergenicity tests. The findings from these evaluations are considered by the regulatory authorities before the transgenic plants are approved for commercial cultivation. The transgene may encode a protein that itself is the desired product, or it may by itself generate a desirable phenotype, e.g., the Cry protein specifying insect resistance in Bt crops. The transgene-encoded protein may participate in a biosynthetic pathway and modify it in various ways, and a group of transgenes may be expressed in concert to introduce a novel pathway and generate a novel product. Finally, the expression of an endogenous gene may be blocked to produce a desirable phenotype, e.g., suppression of the polygalacturonase gene in Flavr Savr transgenic tomato. The transgenic plants are modified for the target traits in a highly specific and efficient manner. Therefore, the transgenic technology is regarded as a clean technology for directed genetic modifications (Singh 2012b).

Transgenic plants for plant breeding use are being developed since 1980s. The "Flavr Savr" tomato was the first transgenic to be approved in 1994 for commercial cultivation. The cultivation of transgenic varieties began in 1996 on 1.7 million hectares (Mha), of which >88 % was in the USA. In 2012, transgenic crops were grown in 28 countries in 170.3 Mha (>11 % of the global cropped area). But >91 % of this area was located in merely five countries (the USA, Brazil, Argentina, India and Canada), and USA alone accounted for 40.8 % of the area

under transgenic varieties. Similarly, transgenic varieties belonging to merely four crops, viz., soybean, maize, cotton, and canola, were grown in >99 % of the area under the transgenic crops. In addition, only two modified traits, viz., herbicide tolerance and insect resistance, together accounted for ~98 % of the area under transgenic genotypes. Although 124 new transformation events are likely to be commercialized by 2015 as compared to only 40 events released so far, they will expand the range of modified traits and crop species only slightly. In general, cultivation of transgenic crops is associated with a small but significant increase in yield, and a reduction in the use of all types of inputs like capital, labor, energy, pesticides, etc. The transgenic crops are estimated to generate large economic benefits that are distributed among the farmers, the processors/consumers and the concerned biotechnology companies (Lusser et al. 2012a).

A variety of safety concerns have been raised against the transgenic crops, including (1) risks to human health due to toxicity and allergenicity of the transgene products and transfer of the antibiotic resistance markers to gut microflora, (2) transgene transfers to the wild weedy relatives of crops making them more persistent and noxious, (3) persistence of the transgenics themselves as weeds, (4) transgene transfers to the microflora, (5) detrimental effects on the nontarget flora and fauna, and (6) contamination of non-GM (non-genetically modified) food. In order to address these and other concerns, many countries enacted new legislation during the 1980s and 1990s. These legislations regulate the experimental evaluation and commercial release of the transgenic plants and the import and marketing of their seed. They also include rules governing comprehensive risk assessments for environmental and food and feed safety. More recently, studies on the socioeconomic impact of the transgenics have been added as an additional requirement to help the policymakers. While environmental and health risk assessments are mandatory, the socioeconomic assessment is optional. The consumer acceptance of GM food has been one of the major issues in the adoption of transgenic crops. For example, the consumers in European Union prefer non-GM food, and the level of this preference rises with income and education of the consumers. However, when GM food products are kept on the shelves, the European Union consumers tend to buy them. The general policy of European food industry is to avoid GM raw materials. In the USA and Canada, the segregation of GM and non-GM crops/foods is not explicitly regulated, and the issues related to contamination by GM food are settled between the involved parties (Lusser et al. 2012a).

It may be pointed out that new technologies are being devised for use as tools in breeding programs; these technologies facilitate the creation of the desired genetic variation and the realization of breeding objectives. In these technologies, transgenics constitute only an intermediate step and are not represented in the end product. Therefore, it is hoped that products from these technologies will find much greater consumer acceptance than transgenic crops. Some of the new technologies already being used by the commercial breeding programs are as follows: zinc-finger nuclease technology, oligonucleotide-directed mutagenesis, cisgenesis and intragenesis, RNA-dependent DNA methylation, grafting on transgenic rootstocks, reverse breeding, etc. The traits targeted for improvement include both simply inherited agronomic traits as well as complex traits. In many breeding programs, varieties developed by the new technologies are expected to reach commercial cultivation around 2015. Further, many newer methods like targeted mutagenesis using, say, engineered meganucleases are being developed (Lusser et al. 2012b).

Thus, the contributions of transgenic technology have been remarkable, and its potential for generating novel useful genotypes is immense. But its overall impact on the global agriculture remains limited to few crops and traits, and the transgenic varieties are cultivated in a small number of countries on only 10 % of the world cropped area. In view of this, the molecular marker technology remains the only widely acceptable approach for supplementing the plant breeding efforts in meeting the global food needs.

1.6 Molecular Markers: Selection Made Easy and More Reliable

The efforts for detection and localization of polygenes began soon after the mechanism of inheritance of quantitative traits was established. In 1923, Sax reported linkage between seed coat color (a qualitative trait) and seed size (a quantitative trait) in common bean. Subsequently, Thoday (1961) used elaborate cytogenetic techniques coupled with genetic analyses to map QTLs for several quantitative traits in *Drosophila melanogaster*. In these studies, association between oligogenic characters, used as markers, and quantitative traits was used to detect and localize the concerned polygenes. Subsequently, protein-based markers like isozymes were developed, and they generated considerable interest and expectations. But the number of useful protein-based markers, like that of morphological markers, in a given species is rather limited. In 1980, Botstein and associates described the concept underlying the use of restriction fragment length polymorphism (RFLP) for linkage mapping in humans. The RFLP technique was soon used in a variety of biological investigations, including linkage mapping. In fact, RFLP became the marker of choice, and it remains the standard reference marker even today. The development of polymerase chain reaction (PCR) technique in 1985 triggered the search for more convenient PCR-based markers, and several such markers were discovered (Table 1.3). The simple sequence repeat (SSR) markers offered several advantages over RFLPs and soon became the most widely used marker system. The genome-sequencing projects revealed the existence of single nucleotide variation, i.e., the single nucleotide polymorphism (SNP), in the genome sequences of different individuals. The SNPs have been found to be the most abundant DNA sequence polymorphism. The SSR markers have now been replaced by SNP marker system, which is the current marker of choice due to its abundance and amenability to high-throughput genotyping.

Molecular markers have been used to construct high-density linkage maps and for the detection of QTLs and their mapping to specific genomic regions. Linkage mapping is generally based on genotypic and phenotypic analyses of a suitable mapping population constructed by crossing two lines differing for the target trait. In addition, collections of germplasm lines/breeding lines and samples drawn from natural populations can be analyzed to detect marker-trait associations. Once close linkage between a marker and a trait of interest is established, the marker genotype can be used as the basis for indirect selection for the target gene/QTL, i.e., marker-assisted selection (MAS). MAS has limited the role of phenotypic evaluation to the establishment of marker-trait linkages and the evaluation of products of MAS before their release for cultivation. The determination of molecular marker genotypes is virtually error-free and independent of the developmental stage of the plants, and the prevailing environment. The analysis of marker genotype is generally much easier than phenotypic evaluation for many target traits and is amenable to moderate to high/very high throughput. Thus, DNA markers provide a much easier and highly reliable means of indirect selection for such traits that are affected by the environment and/or whose phenotypic evaluation is tedious/time consuming or expensive.

Molecular markers have been generally used to facilitate target gene introgression using the backcross scheme (marker-assisted backcrossing, MABC). MABC also facilitates the recovery of recurrent parent genotype and the elimination of donor parent genome flanking the target gene for minimizing linkage drag. MABC is well suited for introgression of oligogenes and large effect QTLs for defect correction of an otherwise superior variety that is used as the recurrent parent. Marker technology has prompted the development of ingenious breeding schemes designed to accumulate QTLs for various traits. For example, the marker-assisted recurrent selection (MARS) scheme is designed for the accumulation of QTLs with significant effect on the target trait, while genomic selection (GS) aims to accumulate all the QTLs affecting the trait irrespective of whether their effects are significant or not. The two major advantages of MAS are as follows: (1) the selected plants can

Table 1.3 A chronology of the development of molecular markers and their major applications

Year	Development
1980	Botstein and coworkers: described the approach for using RFLP (restriction fragment length polymorphism) for the preparation of human linkage map
1985	Saiki and colleagues: first demonstration of PCR (polymerase chain reaction)
1989	The SSCP (single-strand conformation profile/polymorphism) technique described by Orita and coworkers
1989	Sequence-tagged site (STS) markers reported by Olson and coworkers
1990	Development of AP-PCR (arbitrary primed PCR) technique by Welsh and McClelland
1990	Microsatellite markers
1990	Development of RAPD (randomly amplified polymorphic DNA) technique by Williams JGK and coworkers
1991	Development of DAF (DNA amplification fingerprinting) technique by Caetano-Anolles and associates
1991	Allele-specific technique CAPSs (cleaved amplified polymorphic sequences) by Williams MNV and coworkers and Konieczny and Ausubel; named as CAPS by Konieczny and Ausubel (1993)
1993	SCAR (sequence-characterized amplified regions) markers developed by Paran and Michelmore
1993	Development of AFLP (amplified fragment length polymorphism) technique by Zabeau and Vos
1994	Inter-simple sequence repeat (ISSR) markers described by Zietkiewicz and coworkers (Zietkiewicz et al. 1994)
1994	RAMPO (random amplified microsatellite polymorphism) technique for the detection of minor amplicons of AP-PCR by Wu KS and coworkers
1996	cDNA-AFLP (copy DNA-AFLP) developed by Bachem and associates (Bachem et al. 1996)
1997	Retrotransposon-based techniques like S-SAP (sequence-specific amplification polymorphism), IRAP (inter-retrotransposon amplified polymorphism), and REMAP (retroposon-microsatellite amplified polymorphism) to detect genome-wide polymorphism
1998	Allele-specific technique dCAPS (derived cleaved amplified polymorphic sequence) by Michaels and Amasino
1998	Resistance gene analogue markers by Chen and associates
2001	SRAP (sequence-related amplified polymorphism) technique for the detection of polymorphism in ORFs (open reading frames) by Li and Quiros
2002	COS (conserved orthologous set) markers by Fulton and coworkers
2003	TRAP (target region amplification polymorphism) technique for the detection of polymorphism in regions surrounding the targeted exons reported by Hu and Vick
2004	Intron-targeting polymorphism (ITP) markers developed by Choi and coworkers
2009	CDDP (conserved DNA-derived polymorphism) markers devised by Collard and Mackill
2009	SCoT (start codon targeted polymorphism) markers developed by Collard and Mackill
2009	CoRAP (conserved region amplification polymorphism) markers reported by Wang and coworkers
2014	CAAT box-derived polymorphism (CBDP) marker described by Singh AK and coworkers

be used for hybridization in the same season, and (2) the populations grown in off-season nurseries/greenhouses can be subjected to selection for even traits like yield, which should otherwise be assessed only in the target environment. These features of MAS have speeded up the development of varieties by a factor of 2–3.

Molecular markers are being used for genetic diversity analyses and to aid an effective conservation and utilization of genetic resources. Marker-gene/QTL linkage has been exploited for cloning and characterization of the concerned genes. It is expected that molecular marker-based genetic analyses would generate further insights into the developmental regulation of quantitative traits. In addition, markers may facilitate the unraveling of genetic basis of heterosis and, eventually, the prediction of heterotic cross combinations.

1.7 Designer Crops

Crop varieties developed to express a specified desirable monogenic, oligogenic, or polygenic trait are often referred to as "designer crops." The term "designer crops" is not restricted to transgenic plants alone. In fact, varieties developed by any methodology, including MAS, are

called "designer crops" provided they exhibit a specified phenotype. For example, insect-resistant varieties of maize and cotton are designer crops; they are produced by the integration of the *cry* gene from *Bacillus thuringiensis* into the genomes of these crops and an efficient expression of this gene to generate resistance to the target insects. Similarly, pyramiding of multiple QTLs/genes from different donors into a recurrent parent through MABC would yield "designer crops," e.g., Improved Pusa Basmati 1 variety of rice. The transgenic and molecular marker technologies are the two potent approaches for a planned and precise transfer of genes to generate plant varieties having the specified set of, often novel, features.

1.8 Some Notable Achievements of Marker-Assisted Plant Breeding

The molecular markers have been used for MAS during backcross programs designed primarily for the introgression of disease resistance genes. In most such cases, two or more genes have been pyramided to achieve durable resistance to the concerned pathogens. The first variety developed by MAS was a maize hybrid released in the USA for commercial cultivation in 2006 by Monsanto, USA. Since then, several varieties developed by MAS, often improved versions of popular varieties produced through MABC, have been released for commercial cultivation. Some examples of varieties developed by MAS are Cadet and Jacinto rice varieties (USA); Indonesian rice varieties Angke and Conde; barley varieties Sloop SA and Sloop Vic developed in Australia; rice varieties Improved Pusa Basmati 1, Improved Samba Mahsuri and Swarna Sub-1, and maize hybrid Vivek QPM 9 released in India. However, many varieties developed through MAS remain unknown because they are not included in scientific publications (Collard and Mackill 2008).

The molecular markers have enabled detection and mapping of QTLs, which was not possible with morphological or even protein-based markers. The markers have made selection independent of the phenotype, as a result of which selection for the desired traits, including yield, can be effectively practiced in off-season nurseries/greenhouses. Further, the desirable plants can be selected in the seedling stage, and the selected plants can be used for hybridization in the same season. These features of MAS enable the completion of up to three recurrent selection cycles per year by utilizing off-season nursery/greenhouse facilities (Eathington et al. 2007). In addition, MAS allows easy pyramiding of oligogenic resistance and combining of horizontal resistance with vertical resistance. It may be pointed out that pyramiding is considerably difficult on the basis of disease tests, while combining of vertical and horizontal resistances is not possible.

Molecular markers allow definitive identification of different cultivars/varieties and germplasm lines and can be used for testing the purity of inbred parents of hybrids and that of seed lots. Closely linked molecular markers have been used for positional cloning of a number of plant genes. Molecular markers have stimulated the development of novel breeding schemes like GS and the highly ambitious "breeding by design" schemes. Another breeding scheme, the marker-evaluated selection, is designed for the identification of genomic regions associated with adaptation to specific agro-ecological conditions and accumulation of these genomic regions using MAS to develop varieties with superior adaptation.

1.9 Future Prospects of Marker-Assisted Plant Breeding

In a moderate-size plant breeding program, thousands of plants have to be evaluated for a number of markers within a relatively short period of time. Therefore, the molecular marker systems should be amenable to high throughput, and their genotyping cost should be reasonably low. None of the currently available marker systems meets the requirements of low cost with high throughput, especially in the context

of developing countries. In fact, the high cost of marker genotyping is considered as one of the main factors limiting the widespread global adoption of marker technology in routine breeding programs (Collard and Mackill 2008). Therefore, increasing the throughput and reducing the cost of marker genotyping are two of the major future challenges. The current sequencing costs are not an issue in the developed world particularly in the private sector, which is routinely using molecular markers in breeding programs. However, the sequencing costs are still high for routine use by most public sector breeding programs, particularly in the developing countries. It is expected that the sequencing costs will continue to decline, and newer approaches like reduced representation sequencing and low coverage sequencing for genome-wide SNP development will further reduce the marker genotyping costs (Mir and Varshney 2013). Further, schemes like selective genotyping and targeted marker discoveries would help reduce the genotyping work and, thereby, the total genotyping cost of a breeding program.

The bulk of available markers reveal the allelic state of the target gene due to linkage between the marker and the gene. However, the linkage relationship between the marker and the gene might change due to recombination. In view of this, generally two markers flanking the target locus are used for MAS; this increases the total number of markers to be genotyped. Therefore, it will be highly desirable to develop markers located within the target genes. Further, the alleles of such markers should be based on sequence differences between the respective alleles of the concerned gene(s), and this sequence should be involved in the differential function of the alleles. The term "functional markers" is used to describe molecular markers of this type. The use of functional markers for MAS will eliminate the risk of change in the linkage relationship and drastically reduce the number of markers to be genotyped. In addition, functional markers may provide biologically more meaningful estimates of genetic diversity/distance than random markers, especially when such markers also include the regulatory regions of genes/QTLs. But the development of functional markers is a demanding and expensive task.

The molecular markers facilitated the detection and mapping of QTLs, but certain issues still remain to be satisfactorily resolved. One important issue relates to the identification of QTLs involved in epistatic interactions since the earlier methods of QTL mapping, such as simple interval mapping and composite interval mapping, did not have the provision for the estimation of epistatic interactions. The mapping methods like multiple interval mapping and Bayesian approaches have been developed to address this issue, but further improvements are required in these methods. Another limitation of the QTL mapping methods is that they detect only such QTLs that produce statistically significant effects on the target traits. However, the significant effect QTLs detected for a trait are not able to account for the total genetic variance for the trait. The GS scheme was designed to select for even such QTLs that do not produce a significant effect on the trait, but considerable work is needed to establish the usefulness of this demanding breeding scheme. The QTL mapping is generally based on biparental populations, and each QTL detected in them usually represents a large genomic region (~10–20 cM). It may be pointed out that MAS for such large genomic regions would be problematic. Association mapping would allow identification of much shorter genomic regions representing QTLs and also offer some unique opportunities for the detection of marker-trait associations. However, association studies in plants present several difficulties for which suitable statistical tools need to be developed. Finally, many different QTLs affecting a single trait have been identified in different studies, which represent slightly different genomic regions. QTL meta-analysis has been used to identify and map the "true" or "meta" QTLs, but the biological significance of the "meta" QTLs remains to be elucidated.

The MABC for oligogenic traits is relatively straightforward but that for QTLs may or may not produce the expected phenotype (Hospital

2005; Collard and Mackill 2008). Since most of the economically important traits of crop species are governed by QTLs, introgression of QTLs using MABC will be highly desirable. Thus far, there is no way of predicting the outcome of QTL introgressions, except that QTLs with large effect size and stable expression over environments may be expected to generate the desired phenotype in the recurrent parent. But large effect QTLs can be successfully introgressed on the basis of phenotype, and the real usefulness of MABC will be with reference to relatively small effect QTLs. In this context, breeding schemes that combine QTL discovery and MAS would be preferable to performing these activities separately. The breeding scheme MARS successfully accumulates QTLs with significant effects on the trait phenotype, while GS is designed for accumulation of all QTLs affecting the trait. Further, novel breeding schemes need to be designed to take full advantage of the potentials of the marker technology. Another area where marker technology might ultimately prove useful is the prediction of heterotic cross combinations, which remains a nagging issue. It was hoped that marker-based estimates of genetic diversity between the parents would predict heterosis more precisely than those of phenotypic diversity, but this expectation has not been realized. Molecular markers permit the assignment of inbred lines to appropriate heterotic groups and the identification of heterosis loci. A detailed analysis of these loci may provide a better insight into the genetic basis of heterosis and afford a more reliable heterosis prediction.

The use of molecular markers depends on establishing a reliable and predictable relationship between a trait phenotype and a marker genotype. This would require a precise phenotyping of the individuals of a test population. Precision phenotyping is regarded as one of the most challenging tasks, particularly when a relatively large population is to be evaluated for a large number of traits. The discipline of phenomics is devoted to large-scale phenotyping and is currently an area of intense research and development. The integration of marker technology into breeding programs leads to the following two consequences: (1) generation of a large amounts of data and (2) the need for quick decisions based on these data. In view of the above, appropriate statistical tools and a strong bioinformatics support for acquisition, handling, storage, and management of the huge amounts of data need to be developed.

Questions

1. "Plant breeding has been remarkably successful in improving the performance of crop plants". Evaluate the validity of this statement with the help of various contributions of plant breeding.
2. "The chief limitation of the breeding methods is that the decision about the worth of different lines/plants has to be based on their phenotype". Discuss this statement in the light of available information.
3. Why do we need to integrate molecular markers in plant breeding activities?
4. Why do we need to accelerate the development of improved crop varieties?
5. Briefly discuss the future prospects of marker-assisted plant breeding.

Part II

Genetic Markers

Hybridization-Based Markers

2

2.1 Introduction

There will be no need for molecular markers if all the traits of various organisms had the following three features: (1) the traits were easily scored, (2) the individuals were reliably classified into few distinct phenotypic classes, and, more particularly, (3) there were complete correspondence between trait phenotypes and genotypes at the concerned loci. In reality, however, only some of the traits of any organism exhibit the above features, and they are called *qualitative traits*. Gregor J. Mendel consciously selected qualitative traits for his classical experiments in plant hybridization. The findings from these experiments, published in 1866, enabled the discovery of the fundamental laws of inheritance, which laid the foundation of the discipline of genetics. Mendel had carefully selected such pea varieties that differed for one or more of seven different qualitative traits. Each of these traits had two easily identifiable contrasting forms and showed stable expression over 2 years of evaluation. In the subsequent years, many qualitative traits were extensively used in genetic studies in a variety of organisms to obtain information on a range of issues. But it was soon recognized that many traits of economic importance showed continuous variation and, as a result, could not be used for classical inheritance studies. These traits, known as *quantitative traits*, exhibit continuous variation mainly due to the environmental influences on their phenotype. Therefore, the phenotype of such traits is not a reliable indicator of genotype, and phenotype-based selection for them is often disappointing. The genetics of these traits has been extensively investigated since this knowledge could help devise effective selection schemes for them. Efforts were also made to identify qualitative traits linked to various quantitative traits (Sax 1923; Thoday 1961). It was expected that linkage relationships might help unravel the genetic basis of quantitative traits and permit indirect selection for them. Therefore, a search for other easily detectable and stable characteristics was initiated, leading to the discovery and development of protein-based and, finally, DNA-based markers.

2.2 Genetic Markers

A trait that is polymorphic, easily and reliably identified, and readily followed in segregating generations and indicates the genotype of the individuals that exhibit the trait is known as *genetic marker*. This trait could be visible to the naked eye, or a biochemical feature, including that of protein. Thus, a *genetic marker locus* would be a specific location in the genome of an organism that can be identified by a genetic marker of the organism. An "*ideal*" *genetic marker* should be polymorphic and multiallelic to permit classification of individuals into more than two groups. It should be codominant to

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_2, © Author(s) 2015

enable discrimination between heterozygotes and homozygotes. It should not be epistatic, i.e., should not show inter-locus interactions, so that identification of marker alleles at one locus does not interfere with that at other marker loci. The marker loci should be neutral so that marker alleles by themselves do not affect fitness of the individuals. It should be abundant and distributed almost evenly over the entire genome, and should not be pleiotropic. Finally, environmental variation should not affect the marker trait so that marker phenotype accurately reflects the genotype at the marker locus irrespective of the prevailing environment (de Vienne 2003).

Genetic markers can be grouped as follows: (1) visible/morphological markers, (2) protein markers, and (3) DNA markers. In addition, structural features of chromosomes and the chromosome banding patterns generated by specific staining techniques are used as markers to identify linkage groups and to support physical mapping, especially in the case of animal species. Sometimes, morphological, cytological, and protein markers together are called *classical markers*. Further, a variety of biomolecules are fast developing as promising biomarkers useful in the identification of genotypes expected to generate desirable phenotypes (Sect. 15.15). Strictly speaking, the term *molecular markers* includes both protein and DNA markers as well as the metabolite-based biomarkers, *but the current usage of this term is limited to DNA markers*.

2.2.1 Visible/Morphological Markers

Morphological traits were the earliest genetic markers used in scientific studies. Some examples of such markers are shape and color of flowers, color and shape of fruits and seeds, etc. Since these traits are scorable by the naked eye, they are also termed as *naked eye polymorphisms*. These traits represent the actual phenotypes of plants that are relevant to plant breeders. In contrast, protein and DNA markers ordinarily represent arbitrary locations in genomes and may or may not directly correspond to specific phenotypes. Generally, assays for morphological markers require neither sophisticated equipment nor preparatory procedures. Therefore, scoring of these markers is simple, rapid, and inexpensive, and often they can be scored even from preserved specimens (Stussey 1990). The chief limitations of morphological markers are as follows: (1) The number of good visible/morphological markers in a species is rather limited. (2) Typically, only a few of these markers can be analyzed in a single cross/mapping population mainly due to difficulties in determining phenotypes of different traits in a single plant. (3) Generally, they can be scored only on whole plants and that too during specific developmental stages. (4) Many traits, e.g., disease resistance, may have a threshold requirement for their expression. (5) Some genes governing the marker traits may have pleiotropic effect on the trait of interest, i.e., the trait with which marker association is to be tested. This would distort the segregation ratio and cause error in gene mapping. (6) Finally, maintenance of suitable genetic stocks expressing the various marker traits would be necessary.

In the early days of linkage mapping, crosses between parents differing for two or three traits were widely used, and data from several such crosses were pooled to construct linkage maps. As long as one or more loci are common between the crosses, the gene order and the distances between genes can be integrated with some degree of confidence. Some morphological markers are known to be associated with important agronomic traits, e.g., leaf-tip burning is associated with leaf rust resistance gene *Lr34* in wheat, and pigmented seedling/black chaff is associated with stem rust resistance gene *Sr2*. Some markers are useful in identification of crop varieties, e.g., brown glumes in the case of wheat variety HD-2329 and crooked neck (peduncle) in the case of Kalyan Sona wheat. These markers will continue to be used wherever they are available.

2.2.2 Protein-Based Markers

Protein-based markers are detected as electrophoretic variants of proteins, including enzymes.

These markers are generated by such small changes in the coding sequences of the concerned genes that alter the amino acid sequences of the concerned proteins. As a result, the variant protein molecules differ from the wild-type molecules in electrical charge detected as differential electrophoretic mobility. *Isozymes* are different forms of an enzyme that have the same catalytic function and are present in the same individual. The differences in electrophoretic mobility result from differences in their net charge or conformation. Strictly speaking, *isozymes* are closely related variants of an enzyme encoded by different genes, which may have arisen by gene duplication or polyploidization. In contrast, the variants of an enzyme encoded by different alleles of the same gene are called *allozymes*. Therefore, only allozymes will behave as alleles of a marker locus and will be useful in linkage analyses. In contrast, strict isozymes will be inherited as separate loci and may show independent segregation. In practice, a decision about which of the electrophoretic variants of an enzyme are isozymes (or allozymes) would require genetic analysis using suitable crosses. *Generally, the terms isozymes and allozymes are used as synonyms and, in the context of genetic markers, the term allozyme is seldom used*. Most protein-based markers are isozymes, but molecular weight or isoelectric variants of nonenzymatic proteins also serve as markers. Isozymes and seed storage proteins have been widely used as markers (Tanksley and Orton 1983).

A functional enzyme molecule may comprise one (monomer), two (dimer), or more (multimer) identical (homomultimer) or distinct (heteromultimer) polypeptides. A monomeric enzyme will always yield two bands in the F_1. But a homomultimeric enzyme will give rise to $n + 1$ bands in F_1, where n is the number of copies of the polypeptide present in the enzyme molecule (Fig. 2.1). Suppose alleles *A* and *a* of a gene encoding an enzyme produce slow- and fast-moving polypeptides A and a, respectively. If this enzyme were monomeric, each of the two homozygotes, *AA* and *aa*, will exhibit one slow and one fast band, respectively. But the F_1 (*Aa*) from the two homozygotes will show both the parental bands. However, if the enzyme molecule were dimeric, e.g., AA and aa, the F_1 will show three bands. Two of these bands will be the two parental bands, and the additional band with intermediate mobility will have both the polypeptides (Aa). The band patterns will be more complex and their interpretation will be

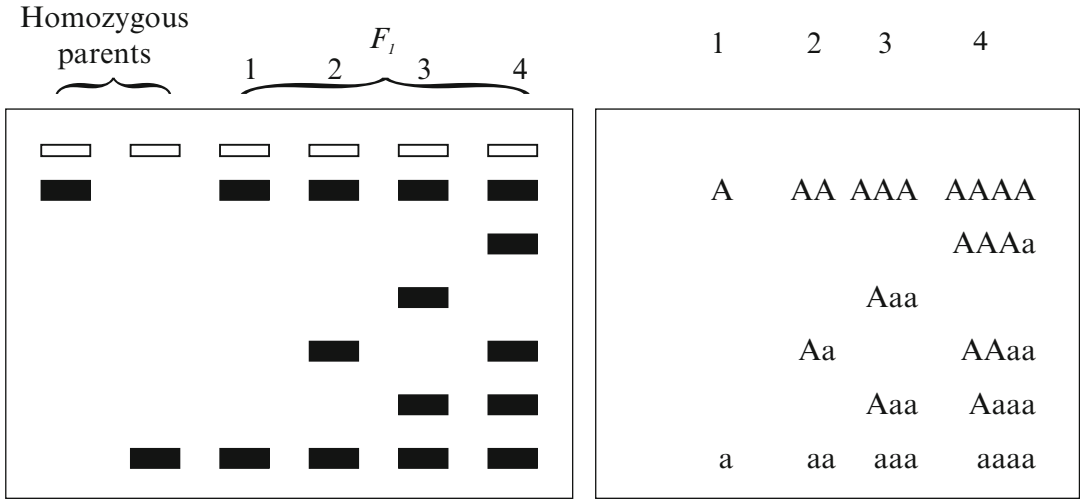


Fig. 2.1 A schematic representation of the isozyme banding patterns seen in F_1 generation and their interpretations. The enzyme molecule may be (*1*) monomer, (*2*) homodimer, (*3*) homotrimer, or (*4*) homotetramer. A and a are the polypeptides encoded by the alleles *A* and *a*

more difficult in the case of heteromultimeric enzymes.

Protein-based markers offer the following advantages over visible/morphological markers: (1) They reflect differences in gene sequences more directly than visible/morphological markers, (2) only a small amount of tissue is needed for their detection, (3) they can often be detected at seedling stage or even from seeds, and (4) analysis of one marker usually does not interfere with that for other protein-based markers. (5) Therefore, many different marker loci can be analyzed in the same cross. (6) In addition, isozymes are usually codominant, (7) their analysis is relatively easy, and (8) data interpretation is facilitated by numerous reference data (Muller-Starck 1998). The major limitation of this marker system is that (1) any two parents may be polymorphic for only a relatively small number of protein-based markers. Some other limitations are that (2) isozymes represent only a small, nonrandom sample of the structural genes of an organism. (3) They detect only such mutations that produce a functional enzyme with changed electrophoretic mobility. (4) In addition, a single band may represent two different isozymes having identical mobility. (5) Finally, they may vary with the tissue, the developmental stage, and the environment (Beckmann and Soller 1983, 1986; Muller-Starck 1998).

Isozyme markers linked to genomic regions/genes involved in the development of several traits of interest have been identified. Some of these markers have also been used for indirect selection for the concerned traits, i.e., marker-assisted selection (MAS). For example, acid phosphatase locus *Asp1*, closely linked with nematode resistance in tomato, has been used for MAS. But isozyme markers have been almost completely replaced by DNA-based markers.

2.2.3 DNA Markers

The *DNA-based markers* represent variation in genomic DNA sequences of different individuals. They are detected as differential mobility of fragments in a gel, hybridization with an array or PCR amplification, or as DNA sequence differences. The development of these markers began in 1974, when analysis of fragments generated by restriction enzyme digestion of DNA was used for physical mapping of a gene in adenovirus. In 1980, human geneticists observed that digestion of genomic DNA with restriction enzymes generated DNA fragments of different lengths from the same genomic regions of different individuals. This variation in fragment length could be detected by gel electrophoresis of the DNA digests, followed by hybridization with a suitable probe representing the concerned genomic region. The pattern of bands generated in this way differed among the different individuals. This variation was called restriction fragment length polymorphism (RFLP) and was used as the first DNA-based marker. With time, a variety of different DNA-based marker systems were developed to satisfy one or more of the following needs: (1) increased throughput, i.e., number of assays per unit time, (2) lower cost, (3) higher reproducibility, (4) greater abundance, and (5) more user-friendliness. These marker systems detect the following three types of DNA sequence polymorphisms: (1) variation at single nucleotides, (2) insertion/deletion (InDel) of one to several bases, and (3) variation in the number of tandem repeats of few to several nucleotides.

2.2.3.1 Types of DNA Markers

There are several marker systems, some of which are as follows: (1) restriction fragment length polymorphism (RFLP), (2) randomly amplified polymorphic DNAs (RAPDs), (3) arbitrary-primed PCR, (4) DNA amplification fingerprinting (DAF), (5) amplified fragment length polymorphism (AFLP), (6) sequence-characterized amplified regions (SCAR), (7) sequence-tagged sites (STS), (8) allele-specific associated primers (ASAP), (9) single primer amplification reactions (SPARs), (10) simple sequence repeat (SSR) polymorphisms, (11) SSR-anchored PCR, (12) cleaved amplified polymorphic sequences (CAPSs), (13) allele-specific PCR, (14) allele-specific ligation, (15) single-strand conformation

polymorphism (SSCP), (16) diversity array technology (DArT), (17) inter-SSR (ISSR) markers, (18) amplicon length polymorphism (ALP), (19) sequence-related amplified polymorphism (SRAP), (20) target region amplification polymorphism (TRAP), (21) transposable element-based markers, and (22) single-nucleotide polymorphism (SNP).

2.2.3.2 An Ideal DNA Marker

An ideal DNA marker system should have the following features. It should generate a very large number of single-copy neutral effect markers that are polymorphic and, preferably, evenly distributed throughout the genome. It should be codominant and have multiple alleles to provide adequate resolution of genetic differences among individuals/lines. The detection of marker alleles, i.e., genotyping, should be simple, easy, quick, inexpensive, reproducible, and amenable to automation and have high throughput. Further, only small amount of DNA should be needed for genotyping, and the error in genotyping should be near zero. Finally, the marker system should not require prior information about the genome of an organism. However, none of the available marker systems meets all the above criteria, but SNPs are the closest to being ideal molecular markers (Xu 2010; Jiang 2013).

2.2.3.3 Features/Advantages of DNA Markers

DNA markers have the following useful features/advantages (Helentjaris 1992): (1) They represent polymorphism in the actual base sequence of DNA distributed over the entire genome. (2) The number of different marker loci is very large so that all the genomic regions can be mapped at very high marker densities. (3) The sequence variation detected by these markers is generally neutral, except when it is located in coding sequences and affects the functions of concerned genes. (4) Scoring for one DNA marker usually has no effect on that of the others, so that multiple markers can be evaluated simultaneously. (5) Molecular markers show simple Mendelian inheritance. (6) Marker genotyping is independent of the prevailing environment and (7) the developmental stage of the plant, and (8) the marker assays are nondestructive. Therefore, MAS can be effectively used under any environment and at any stage of development since the trait phenotypes are not evaluated. (9) MAS can also be used for an allele that is not expressed in the available genotypes, e.g., a recessive allele in heterozygotes. (10) The DNA samples can be stored for future use, and (11) specific marker stocks are not required.

Since the number of polymorphic markers in a single cross/mapping population is very large (several thousands in the case of some markers), construction of molecular marker linkage maps is very rapid. As a result, high-density DNA marker maps have been developed in several crop species, e.g., rice, maize, wheat, common bean, lettuce, potato, etc. However, the amount of DNA available and the overall level of DNA polymorphism between the two parents of a cross affect the numbers of markers that can be reasonably scored in a single mapping population.

2.2.3.4 Applications of DNA Markers

Molecular markers have a variety of applications, including (1) fingerprinting of strains/varieties for unequivocal identification; (2) mapping of genes and quantitative trait loci (QTLs); (3) efficient MAS for tightly linked QTLs and such oligogenes, direct selection for which may be costly or problematic; (4) positional cloning of genes/QTLs; (5) identification of chromosome segments that would contribute to improvements in the target traits; (6) establishing phylogenetic relationships among different strains/species; (7) selection of parents for hybridization; (8) assessing the basis of somaclonal variation; (9) identification of pathogen races and biotypes; (10) prediction of heterotic cross combinations; (11) identification of wide hybrids; (12) gene pyramiding; and (13) management and utilization of genetic resources. (14) Finally, MAS allows the use of off-season nursery and greenhouse facilities to reduce the time needed for variety development.

2.2.3.5 Categories of DNA Markers

The DNA marker systems have been classified on the basis of different criteria. In terms of the chronology of their development, markers are classified as (1) first-generation (RFLP, RAPD, and their modifications), (2) second-generation (SSRs, AFLPs, and their modifications), and (3) third-generation (ESTs and SNPs) markers. DNA markers have been classified as (1) PCR-based and (2) non-PCR-based markers depending on the use of PCR or as (1) SNPs (generated by variation in DNA sequence) and (2) non-SNPs (produced by variation in sequence length, e.g., SSRs) based on their molecular basis (Gupta et al. 2001). Another classification approach uses the location and the functional significance of markers. On this basis, the markers are grouped as (1) random, (2) gene-based, and (3) functional markers. But from the user's point of view, a more useful classification of markers would be the one based on the method of marker detection and genotyping since this would indicate the degree of their user-friendliness. On the basis of the above and the throughput criteria, the various marker systems can be grouped as (1) low-throughput hybridization-based markers, (2) medium-throughput PCR-based markers, and (3) high-throughput sequence-based markers. This grouping is often considered as (1) first-generation, (2) second-generation, and (3) third-generation molecular markers, respectively.

2.2.4 Concluding Remarks on Genetic Markers

In conclusion, few sufficiently polymorphic morphological markers are available in a given species. These markers are usually dominant, they often show epistasis, and their expression may be dependent on the developmental stage and also influenced by the environment. In any case, only few such markers are polymorphic in a single cross/population. The protein-based markers have most of the features of ideal markers, but the number of markers that can be scored in a single population is limited to ~40 or so, and the polymorphism pattern is largely tissue dependent. The DNA-based markers are abundant; the tissue, developmental stage, or the environment has no effect on the pattern of polymorphism; and many of them are codominant (de Vienne 2003). Therefore, DNA markers are the most relevant and are discussed in this and the next two chapters. The development of and genotyping for molecular markers involves DNA extraction and processing of this DNA as per specific protocols.

2.3 Random, Gene-Based, and Functional Markers

DNA markers can be divided into the following three broad groups: (1) random, (2) gene-based, and (3) functional markers. This classification is based on the location of the markers in the genome and their relationship with specific phenotypes of the relevant traits (Table 2.1). *Random DNA markers* are derived from polymorphic sequences located at random sites in the genome. These markers may or may not be located in genes, and their involvement in the development of a phenotype is not known. The *gene targeted*, *gene-specific*, or *gene-based markers* represent polymorphic sites within genes, but their relationships with the relevant trait phenotypes are not known. In contrast, *functional markers* (Table 2.2) are derived from such polymorphic sites within genes that have a causal relationship with specific phenotypes of the concerned traits. The functional markers are of two types, viz., direct and indirect functional markers. When the proof of allele function is based on either NIL comparison or genetic transformation, the markers are called *direct functional markers* or *allele-specific markers*. But when the proof of allele function is obtained by association studies, the markers are known as *indirect functional markers* (Anderson and Lubberstedt 2003).

Random DNA markers are the easiest to develop and were the first to be used. In contrast, the development of functional markers is much

Table 2.1 A comparison among random, gene-based, and functional DNA markers

Marker type	Origin of marker DNA sequence	Function of polymorphic site	Method of function characterization	Marker development cost	Quality of marker
Random	Not known	Not known	None	Low	Low[a]
Gene-based	Gene	Not known	None	Low	Medium
Indirect functional	Gene	Functional motif	Association studies	Medium	High
Direct functional	Gene	Functional motif	Isogenic lines, transformation	High	High

[a]Recombination may occur between marker and the linked gene/QTL

Table 2.2 Some examples of gene-based/functional markers

Crop species	Trait	Gene	Marker	Phenotype	Remarks
Rice	Blast resistance	*Pi-ta*	SNP		In the codon for the amino acid at position 918
			Allele *T*	Susceptible	
			Allele *G*	Resistant	
	Amylose content	*waxy*	SNP[a]		At the intron 1/exon 1 splice site
			Allele *G*	High	
			Allele *T*	Low	
	Wide compatibility	*S5*	PCR-based *S5*-MMS[b]		
Tomato	Spotted wilt resistance	*Sw5-b*	PCR products from primers Sw5-f2/r2	Resistance	Amplification in resistant plants
	Spotted wilt resistance	*Sw5-b*	Two SNP markers	Resistance	In amplicons from primers Sw5b-f1/r1
	Verticillium wilt resistance	*Ve2*	CAPS markers		
	Fusarium wilt resistance	*I-2*	InDel marker		
Wheat	Glutenin content	*Glu-1*			
	Grain hardness	*Pinb-D1*			
	Plant height	*Rht1*			
	Grain protein content	*Gpc-B1*			
	Starch quality	*GBSS1*			
	Leaf rust resistance	*Lr51*			

[a]Exon 6 SNP alleles (*C*/*A*) also affect amylose content; allele *A* further reduces the amylose content so that the *T* allele at exon 1/intron splice site and the *A* allele in exon 6 together produce the lowest amylose content
[b]A multiplex marker system of three primer pairs: one pair for InDel and two pairs for SNPs. Three alleles; the neutral allele sponsors wide compatibility

more recent and the most demanding. Their development requires knowledge of the functions of relevant genes and their alleles, the sequence differences among the alleles, and a direct proof that these differences are responsible for the concerned phenotypes of the relevant traits. The proof of function of different alleles of the marker (= gene) can also be obtained indirectly by association studies. In the case of random markers, QTL mapping of each population is necessary because different populations may segregate for different QTLs for the trait in question. Further, recombination between a marker and the QTL linked to it could change the phase of their linkage even in closely related lines/parents. In contrast, a functional marker will always be associated with the known QTL function/allele. As a result, different mapping populations need to be characterized only for the QTL alleles, and de novo QTL mapping is not required. Therefore, (1) functional markers do not require validation, and (2) they can be

applied directly to other populations. (3) They provide a better estimate of allelic diversity of genes/QTLs and (4) of genetic diversity of the species. (5) They would also generate knowledge about the nature and the physical location of sequences involved in phenotypic expression of the concerned traits (Anderson and Lubberstedt 2003). (6) Finally, the number of markers required for foreground selection will be reduced to the number of genes to be selected, and (7) there will be no recombination between a marker and the linked gene.

One limitation of functional marker development is that only a small fraction of the genes of different crop species have been functionally characterized. A more demanding task is to reliably characterize and distinguish among the phenotypic effects of the different alleles of a given gene/QTL and to develop suitable allele-specific markers. Once functional markers have been developed, they need to be evaluated in different genetic backgrounds in order to obtain more precise estimates of the phenotypic effects of different marker (= gene/QTL) alleles. Therefore, the initial focus should be on the development of functional markers for large effect QTLs (Anderson and Lubberstedt 2003).

2.4 Isolation and Purification of DNA from Plants

DNA is generally isolated from leaf, endosperm, or some other plant part collected from seedlings or plants growing in the field/greenhouse. DNA can be isolated from half-seeds lacking the embryo, while the half-seeds containing the embryo can be germinated in vitro to raise the next generation. This scheme would permit selection before planting and, thereby, greatly reduce the size of breeding population. The DNA isolation procedure has to tackle the problems posed by tough cellulosic cell wall, secondary metabolites, and other chemical compounds present in plant tissues. The various DNA isolation procedures can be grouped into the following three categories: (1) the standard CTAB (cetyltrimethylammonium bromide) method, (2) rapid DNA extraction methods, and (3) commercial DNA isolation kits. The salient features of these procedures are summarized in Table 2.3, and the details are given in Appendix 2.1.

The CTAB procedure is regarded as the standard method. It yields high-quality DNA that can be stored for long periods of time and is suitable

Table 2.3 A summary of relevant features of the chief DNA extraction strategies

	DNA extraction strategy		
Feature	DNA extraction kits	Rapid DNA extraction methods	The standard CTAB method
DNA quality	Good	Poor	Good
DNA yield	Low	Low	High (up to 5 times as much)
Suitability of DNA for long-term storage	Poor to good	Poor	Good
Flexibility in terms of sample size	High	High	Low to medium
Hands-on time per multiplex unit*	~30 min	>30 min	>2 h
Number of samples processed per day**	~1,000	>2,000	<400
Cost per sample (Euros)	~2.7	<0.0001	~0.70
Cost/μg of DNA[a]	~1.35	~0.00005	~0.70
Most suited for activities involving evaluation of:			
(a) Number of samples	Moderate to large	Large to very large	Small to moderate
(b) Number of markers	Small to moderate	Small	Large to very large

Based on Bagge and Lübberstedt (2008)
*The size of multiplex units for all the strategies is 96
**Including sample collection
[a]Excluding plastic materials and labor costs. The values are only indicative

Table 2.4 A generalized indication of the numbers of markers and samples (lines/individuals) evaluated in different activities related to plant breeding

Activity	Number of samples	Number of markers
Selection of parents	Small to moderate	Large
Marker-assisted selection:		
(a) Foreground selection	Small to moderate	Small
(b) Background selection	Small to moderate	Moderate to large
(c) Marker-assisted recurrent selection	Moderate to large	Small to moderate
(d) Genomic selection	Moderate to large	Large to very large
(e) Others	Moderate to large	Small to moderate
Fingerprinting:		
(a) Genetic characterization of germplasm and breeding materials	Moderate to very large	Moderate to large
(b) Variety identification and seed lot genetic purity test	Small to large	Small to moderate
(c) Product purity test	Small to very large	Small to moderate
Screening of transgenic materials	Moderate to large	Small
Diversity analysis	Moderate to large	Moderate to large
Association studies	Moderate to very large	Large to very large

Modified from Bagge and Lübberstedt (2008)

for polymerase chain reaction (PCR). But it is labor and time intensive, costs 1,000-fold higher than rapid DNA extraction procedures, and is not amenable to automation. The rapid DNA extraction methods allow processing of the largest number of samples at the lowest cost, but DNA yield, quality, and storability are poor. The commercially available kits are the costliest (~2–3 times as much as the CTAB method), process intermediate number of samples per day, and yield low amounts of good quality DNA. Thus, the CTAB method will be preferable when a small number of samples are to be evaluated for a large number of markers (Table 2.4). But the rapid DNA isolation methods are suited for evaluation of a large number of samples for a small number of markers. The commercial kits would be useful for evaluation of a small to moderate number of samples for a small (Mini kits) to moderate (Midi/Maxi kits) number of markers (Bagge and Lubberstedt 2008).

2.5 Restriction Fragment Length Polymorphism

Restriction fragment length polymorphism (*RFLP*) signifies that a single restriction enzyme generates fragments differing in lengths from the same genomic regions of different individuals/strains/lines of a given species or of different related species. The general procedure for RFLP detection (Fig. 2.2) is as follows: (1) To begin with, high-molecular-weight genomic DNA is isolated from several individuals/strains of a species/related species, (2) then each DNA sample is digested separately with the selected restriction enzyme, (3) the restriction fragments are separated (on the basis of size) by gel electrophoresis, (4) the fragments are denatured and transferred from the gel onto a suitable solid support in such a way that the relative positions of the fragments in the gel are preserved (*Southern blotting*), (5) the fragments are fixed and exposed to the labeled DNA probe under conditions favoring DNA–DNA hybridization (*Southern hybridization*), (6) the probe molecules not involved in hybridization are removed by washing, and (7) the fragments involved in hybridization with the probe are detected as distinct bands by autoradiography ("hot" probes) or by color development ("cold" probes). RFLPs are the *first generation of molecular markers*. They were first used in genetic analysis for determining the locations of temperature-sensitive mutations of adenovirus onto a physical map of the restriction fragments (Grodzicker et al. 1974). Later, Botstein

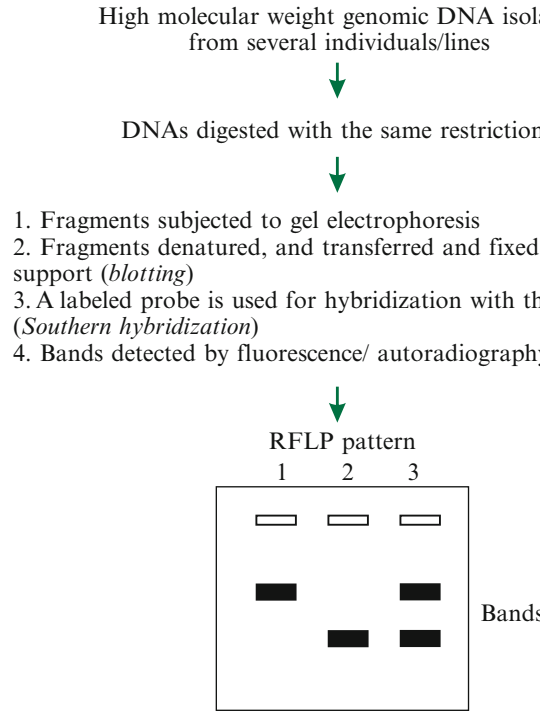


Fig. 2.2 A simplified representation of the RFLP procedure. The probe used for Southern hybridization was a unique sequence so that only a single band was produced in the homozygotes (*Lanes 1* and *2*), and two bands were observed in the heterozygotes (*Lane 3*) for the RFLP locus

et al. (1980) described in detail the principle and the procedure for use of RFLPs in construction of human linkage map. RFLPs have been extensively used for genetic mapping of animal and plant genomes.

2.5.1 Restriction Enzymes

Enzymes that produce internal cuts or cleavages in DNA molecules are known as *endonucleases*. A class of endonucleases cleaves DNA only within or near such sites that have specific base sequences. These enzymes are called *restriction endonucleases* or *restriction enzymes*, and the sites they recognize are termed as *recognition sequences*, *recognition sites*, or *restriction sites*. There are three different types (types I, II, and III) of restriction enzymes. *Type II restriction enzymes* are remarkably stable and cleave DNA either within or immediately outside the recognition sequences, which are palindromes with rotational symmetry (Fig. 2.3). The first type II restriction enzyme to be isolated, in 1970, was *Hind*II. Since then over 350 different enzymes

Fig. 2.3 A palindrome with rotational symmetry. The *arrow* represents the axis of symmetry

with over 100 different recognition sequences have been isolated/identified; some of these are listed in Table 2.5. The restriction enzymes used in genetic engineering and other genetic studies are exclusively type II enzymes, and the term "restriction enzyme" ordinarily signifies these enzymes. A restriction enzyme is highly specific for its recognition sequence, and a change of even a single base pair in the sequence is enough to prevent cleavage. This property of restriction enzymes is used to advantage in recombinant DNA technology and for detection of certain molecular markers, e.g., RFLP, AFLP, DArT, RAD, etc.

The recognition sequences of most type II enzymes have an even number, e.g., 4, 6, or 8, of base pairs (bp), which are predominantly GC-rich. If the four nucleotides, viz., A, T, G, and C, were distributed at random in a DNA molecule, a given nucleotide is expected to

Table 2.5 Some restriction enzymes and their recognition sequences

Restriction enzyme	Recognition sequence[a]
*Alu*I	5′ AG/CT 3′
	3′ TC/GA 5′
*Ape*KI	5′ G/CWGC 3′
	3′ CGWC/G 5′
*Apa*I	5′ GGGCC/C 3′
	3′ C/CCGGG 5′
*Bam*HI	5′ G/GATCC 3′
	3′ C/CTAG/G 5′
*Bgl*II	5′ A/GATCT 3′
	3′ TCTAG/A 5′
*Cla*I	5′ AT/CGAT 3′
	3′ TAGC/TA 5′
*Dra*I	5′ TTT/AAA 3′
	3′ AAA/TTT 5′
*Eco*RI	5′ G/AATTC 3′
	3′ CTTAA/G 5′
*Eco*RV	5′ GAT/ATC 3′
	3′ CTA/TAG 5′
*Hin*dIII	5′ A/AGCTT 3′
	3′ TTCGA/A 5′
*Hpa*I	5′ GTT/AAC3′
	3′ CAA/TTG 5′
*Hpa*II	5′ C/CGG 3′
	3′ GGC/C 5′
*Mse*I	5′ T/TAA 3′
	3′ AAT/T 5′
*Not*I	5′ GC/GGCCGC 3′
	3′ CGCCGG/CG 5′
*Pst*I[b]	5′ CTGCA/G 3′
	3′ G/ACGTC 5′
*Pvu*II	5′ CAG/CTG 3′
	3′ CTG/GAC 5′
*Sma*I	5′ CCC/GGG 3′
	3′ GGG/CCC 5′
*Sau*3A	5′/GATC 3′
	3′ CTAG/5′
*Taq*I	5′ T/CGA 3′
	3′ AGC/T 5′
*Xba*I	5′ TCT/AGA 3′
	3′ AGA/TCT 5′
*Xho*I	5′ GAG/CTC 3′
	3′ CTC/GAG 5′

[a]The "/" in the recognition sequence indicates the site of cleavage

[b]*Pst*I does not cleave a restriction site in which the 5′ C is methylated. This property of the restriction enzyme is exploited for construction of genomic libraries enriched in non-repeat sequences

occur, on an average, once after four nucleotides. Therefore, a restriction enzyme with recognition site of 4 bp would cleave DNA after, on an average, every 4^4 bp (=256 bp). Similarly, an enzyme with recognition site of 6 bp will cut DNA in fragments of, on an average, 4,096 bp (4^6 bp); therefore, it may be expected to cut a genome of 10^9 bp into about 250,000 fragments of different sizes. Most restriction enzymes cleave DNA molecules within their specific recognition sites, but some of them cut immediately outside their recognition sequences (Table 2.5). Most enzymes induce staggered cuts (the two strands are cleaved at different locations) to produce protruding ends (Fig. 2.4a). The protruding ends generated by a single restriction enzyme are complementary to each other due to the palindromic nature of the recognition site. Some restriction enzymes, on the other hand, cut both the strands at the same position so that they generate blunt or flush ends (Fig. 2.4b). Most enzymes do not cleave at such recognition sites that are methylated (methylation-sensitive restriction enzymes), while some enzymes recognize and cleave at both methylated and non-methylated recognition sites (methylation-insensitive restriction enzymes). In some cases, two restriction enzymes recognize the same target sequence, but one of them is methylation sensitive and the other is methylation insensitive; such enzymes are called *isoschizomers*. For example, enzymes *Hpa*II and *Msp*I are isoschizomers; they both recognize the sequence 5′CCGG3′ when it is non-methylated, but only *Msp*I recognizes the methylated (methylation at the second C) sequence.

2.5.2 Southern Hybridization

Southern hybridization is a DNA–DNA hybridization procedure named after E. M. Southern, who developed this method. In this procedure, either mechanical shearing or digestion with a restriction enzyme is used to fragment DNA samples (Singh 2012b). The mixture of fragments is separated by electrophoresis in either polyacrylamide or agarose gel (Fig. 2.5).

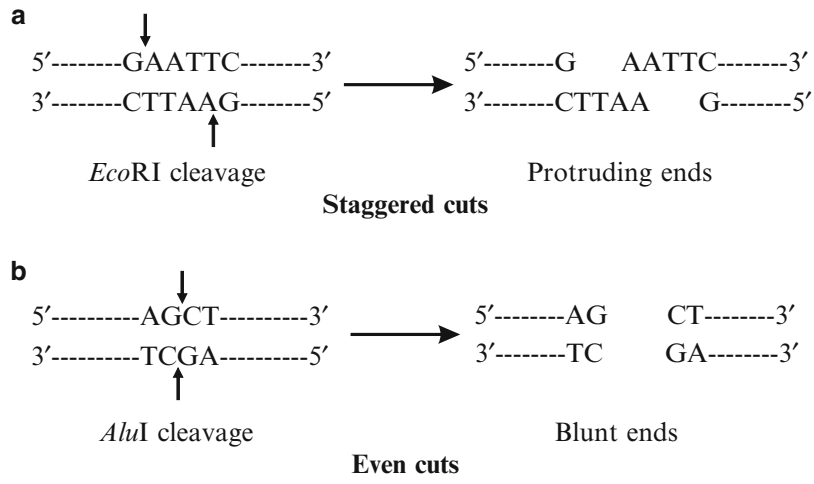


Fig. 2.4 DNA cleavage by restriction endonucleases. (**a**) Staggered cuts (*Eco*RI) produce protruding ends, and (**b**) even cuts (*Alu*I) generate blunt ends. The *vertical arrows* indicate the sites of cuts in the DNA strands

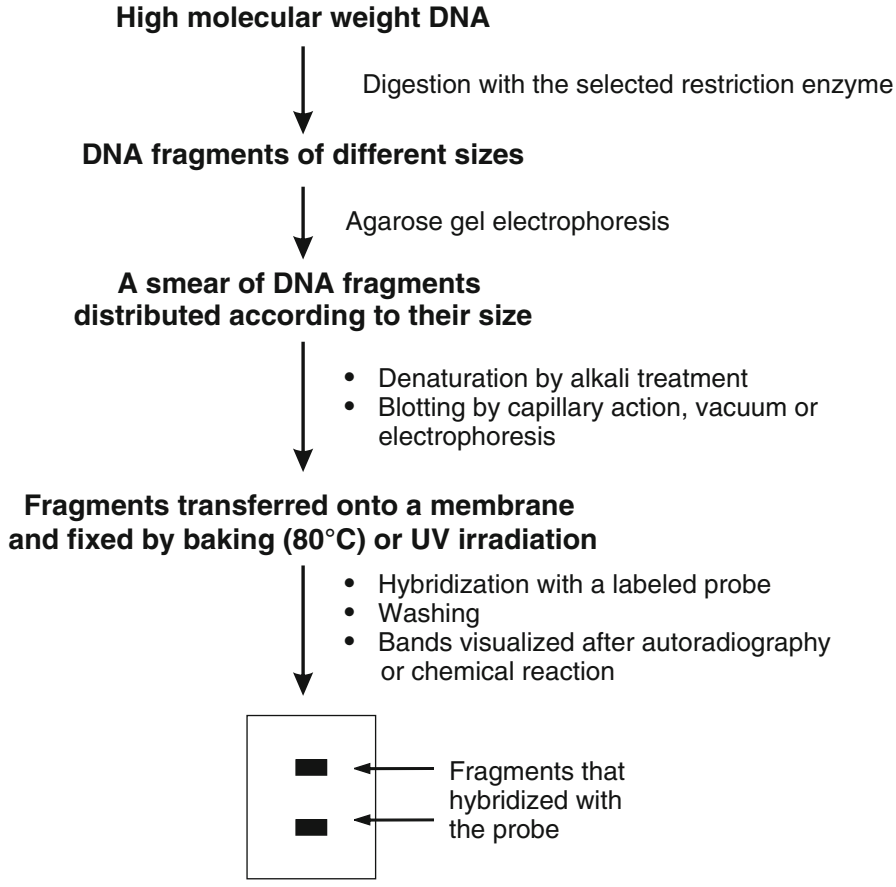


Fig. 2.5 A schematic representation of the Southern hybridization procedure. The relative positions of the fragments do not change during blotting. The probe may be labeled radioactively or chemically, the latter being preferable. A nitrocellulose filter or nylon membrane may be used as a solid support for DNA fragments

Agarose gel is used for separation of DNA fragments of few hundred to 20 kb (kilo base pairs), while smaller fragments are separated using polyacrylamide gel. The DNA fragments migrate through the gel depending on their size since the fragments are uniformly negatively charged. A mixture of DNA fragments of known size is used as marker; this mixture is run in a separate lane. This permits estimation of the size of an unknown DNA fragment present in other lanes of the gel. The fragments are now denatured by alkali treatment so that they become single-stranded. They are then transferred onto a solid support like nitrocellulose filter membrane by a process called *blotting*. Nitrocellulose filter membranes were used initially, but subsequently developed membranes like Hybond N+ have superior features and are preferred. Blotting was initially achieved by capillary action, which takes several hours to complete. The blotting based on vacuum or electrophoresis is much faster and is, therefore, preferred. The relative positions of the DNA fragments in the gel remain unaltered during blotting, and the loss in resolution (sharpness) of the bands is minimal.

The nitrocellulose membrane is removed and baked at 80 °C to permanently fix the DNA fragments onto the membrane. The baked nitrocellulose membrane is then pretreated with a specific solution. This pretreatment prevents nonspecific binding of the single-stranded probes (Sect. 2.5.3) used for hybridization. The probe represents the sequence that is to be detected from among the fragments fixed onto the membrane. After the pretreatment, the membrane is transferred into a hybridization solution containing the probe. The conditions maintained during the hybridization step are less stringent to allow a high rate of probe hybridization. After this step, the membrane is subjected to a series of washes of progressively increasing stringency to eliminate the free probe molecules as well as those paired to related sequences that are not completely homologous to the probe. The stringency of washes is increased by raising the temperature or, more commonly, lowering the ionic strength of the washing solution. The membrane is now placed in close contact with an X-ray film and incubated for the desired period of time. During this period, the images of the bands hybridized with the radioactive probes are formed on the film. The film is then developed and distinct bands are observed; these bands indicate the positions in the gel of those fragments that are complementary to the probe. Southern hybridization technique is highly precise and extremely sensitive. It is used for DNA fingerprinting, detection of RFLPs, detection and identification of the transferred transgenes in transgenic individuals, etc.

When sheared or restricted DNA fragments are subjected to gel electrophoresis, the fragments are distributed in a continuum leading to the formation of a smear, and there are no distinct bands. The bands become observable due to the hybridization of the selected probe with one or few fragments present in the gel. Further, some membranes like nylon membranes have become available, which are physically more robust than nitrocellulose membranes. DNA fragments become cross-linked to these new membranes after a brief exposure to UV light, which saves time. Further, the same membrane blot, i.e., the membrane along with the DNA fragments transferred from the gel and fixed onto it, can be reused for hybridization with another probe after the probe used earlier is removed by washing at high temperature or by some other suitable DNA denaturing procedure.

2.5.3 Probes

Probes are DNA or RNA fragments of typically 500–3,000 bp that are used for detecting specific fragments from among many different fragments present in a mixture. Probes are ordinarily derived from cloned DNA segments from either genomic or cDNA (copy DNA or complementary DNA) libraries (Appendix 2.2). Single-stranded copies of the desired DNA segments can also be generated by asymmetric PCR. In addition, synthetic oligonucleotides can also be used as probes. The genomic library may represent the entire genome of the organism or it may

be a chromosome-specific library obtained from addition/substitution lines or flow-sorted chromosomes. It may even be a library derived from a microdissected chromosome. The best probes are derived from single-copy sequences, which most likely represent structural genes. But DNA sequences with low number of copies or even multiple copies have also been used as probes. The genomic library may be enriched for unique sequences by using a methylation-sensitive restriction enzyme like *Pst*I for digestion of the genomic DNA. A cDNA library will contain genomic sequences representing the structural genes that are expressed in the tissue, from which the mRNA was isolated. In practice, DNA inserts from *Pst*I-generated genomic libraries and cDNA libraries are the most commonly used as probes (de Vienne 2003).

Generally, probes are prepared from the genomic sequences of the same species (*homologous probes*). But probes developed from sequences of other species are also used; such probes are called *heterologous probes*, but the term *heterospecific probes* would be more appropriate. The proportion of useful heterospecific probes declines with the taxonomic distance, and it is rare to have such probes from another family. cDNA probes are more likely to function as heterospecific probes because they are based on more conserved genomic sequences. Heterospecific probes allow mapping in a species without the development of homologous probes, a step that would require considerable effort. In addition, heterospecific probes permit comparative mapping of related species, which is useful in several ways, including isolation of genes of interest.

The probes are suitably labeled with either radioactivity (e.g., ^{32}P) or a chemical ligand using one of the several approaches. They can be *directly labeled* by providing a labeled nucleotide during production of the probe (using bacterial clone/PCR/chemical synthesis). The procedure of nick translation is widely used for labeling of double-stranded DNA probes. Single-stranded DNA probes can be labeled by a method called *random priming*. The single-stranded probe is added to a reaction mixture that supports DNA synthesis and contains the Klenow fragment of *E. coli* DNA polymerase I, the four dNTPs, and a combination of 6-bp-long arbitrary sequence primers. These arbitrary primers will anneal to the probe fragment at all those sites that have a sequence complementary to them and would enable the Klenow fragment to initiate DNA synthesis using the probe strand as template. One or more of the dNTPs are suitably labeled so that the newly produced strands are also labeled.

In the case of chemical labeling, a suitable ligand, e.g., digoxigenin (a plant-derived protein), biotin, an enzyme, or a fluorophore, is conjugated with the nucleotide to be labeled, and the nucleotide is used to label the desired probe. The labeled probe is used for hybridization, and the membrane carrying the hybridized probe molecules is incubated in a detection buffer. This buffer has the necessary reagents for color development after interaction with the chemical label. In the case of digoxigenin, the buffer has an antibody specific for digoxigenin (anti-digoxigenin) coupled with an enzyme, say, alkaline phosphatase. After some time the filter/membrane is washed and the locations of enzyme activity are detected by adding a suitable substrate to the buffer; the enzyme acts on the substrate to produce a colored insoluble precipitate. Several approaches are available for increasing the intensity of color generated by the chemical labels.

Probes are labeled either radioactively (*hot probes*; the label first to be used, but not favored any more) or chemically (*cold probes*; label of choice at present) to permit their easy and reliable detection. Hot probes should be used only in well-equipped and authorized laboratories. In contrast, cold probes can be used in any laboratory and are relatively safer, and some of them can be stored at $-20\ ^{\circ}C$ for long periods of time. But cold probes may not be cheaper than hot probes, and their preparation is not completely harmless; therefore, rigorous precaution should be taken during their preparation. The chief limitation of chemical labels is that the filters/membranes carrying DNA fragments cannot be reused for hybridization with other probes. This is because insoluble precipitates are formed during the detection process.

2.5.4 Polymorphisms Detected by RFLP Markers

The pattern of RFLP will mainly depend on the following: (1) sequence differences in the concerned DNA segments of the selected individuals/strains/species, (2) the particular restriction enzyme used for digestion of the genomic DNAs, and (3) the DNA probe used for Southern hybridization. Variations in restriction fragment lengths leading to detectable RFLP patterns are generated due to the following changes in the concerned genomic regions: (1) a change due to SNP in the base sequence of a recognition site for the restriction enzyme used for digestion of the DNAs (Fig. 2.6), (2) a relatively large (one to several hundred base pairs) deletion and/or insertion in the concerned stretch of the genomic DNA, and (3) a rearrangement (inversion and translocation) of large segments of DNA. SNPs either generate (a gain) or abolish (a loss) restriction sites, while insertions, deletions, inversions, and translocations change the location of one or more restriction sites for the concerned enzyme; all these changes generate RFLPs.

Whether a given RFLP is the result of a mutated restriction site or of deletion, insertion, or rearrangement can be determined by additional experiments. In case several different restriction enzymes are used with the same probe to generate RFLPs, polymorphism due to insertion/deletion/rearrangement should be detected in each case. In contrast, the RFLP due to mutated restriction site is likely to be absent in the case of some of the enzymes tested. Similarly, a comparison of the polymorphism data generated from two genotypes of a species

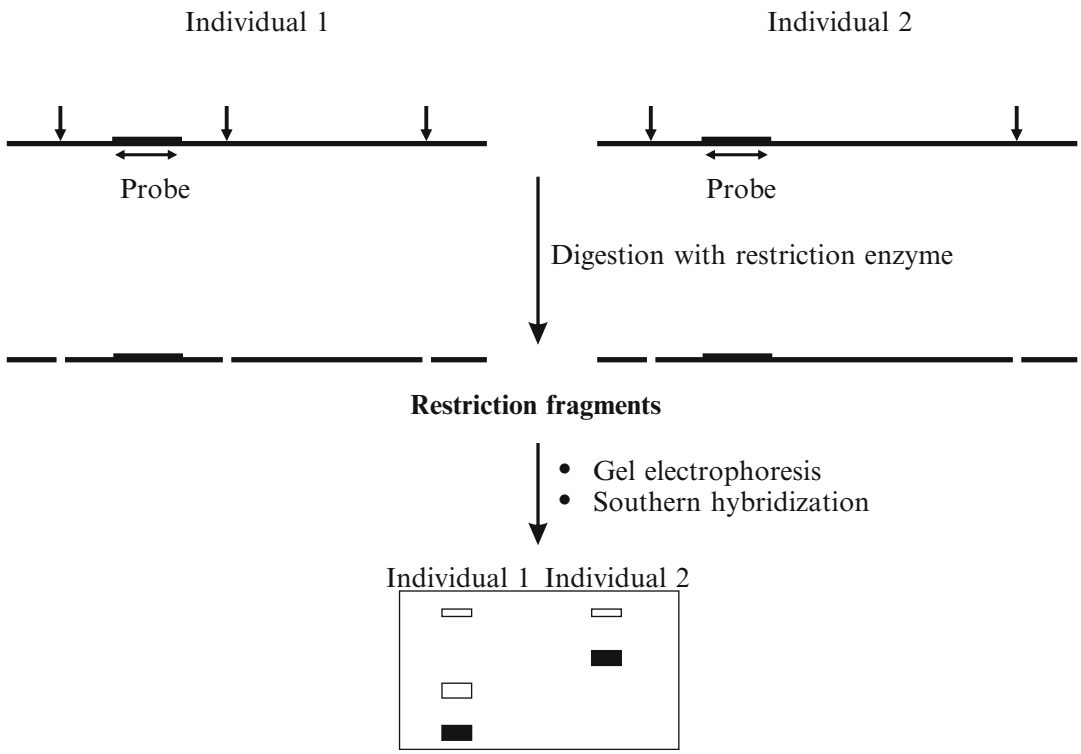


Fig. 2.6 Molecular basis of origin of RFLPs. *Arrows* indicate the recognition sites for the restriction enzyme used in the assay. The sequence marked as "probe" is used as probe for Southern hybridization. Only the solid bands are visualized by Southern hybridization. The open band cannot be visualized; it is indicated only to signify the location of the second restriction fragment from the relevant genomic region of individual 1. In individual 2, the restriction site located in the middle of the fragment has been lost due to a change in its base sequence (SNP)

using two restriction enzymes in combination with a large number of different probes would reveal the relative importance of the two sources of RFLPs. Such an analysis revealed a significant contribution of insertions/deletions to RFLPs in maize, but not in tomato (de Vienne 2003).

2.5.5 Genetic Aspects of RFLPs

An RFLP is detected as a differential movement of a band in the gel lanes of genomic DNAs from different individuals/strains/species digested with the same restriction enzyme. When a unique sequence is used as a probe, a total of two bands, i.e., one fast- and one slow-moving band, will be detected. These two bands represent the two alleles of the RFLP locus corresponding to the genomic region that hybridizes with the given probe. If this probe were used with other restriction enzymes, more alleles of this RFLP locus could be detected. The RFLP alleles are codominant because they represent fragments of different lengths that are easily separated and detected. Therefore, a homozygote will show a single band for an RFLP locus, while the heterozygote will exhibit two bands. But sometimes, a single probe may detect two bands even in homozygotes. There are two possible explanations for such a result. In case the genomic region to which the probe hybridizes contains a recognition site for the restriction enzyme used to generate the RFLP, the enzyme will cut the DNA molecule within this region. As a result, the probe will hybridize with two restriction fragments, and two bands will be observed in the homozygotes (Fig. 2.7). Alternatively, a duplication event may have generated two copies of the genomic region detected by the probe, which will generate two bands in the homozygotes. It should be kept in mind that in the first case, the two bands represent a single allele of a single locus, while in the second case they correspond to two different RFLP loci. The inheritance pattern of the two RFLP bands in appropriate crosses would easily discriminate between these two possibilities. When the two bands represent a single locus, they would always remain together in the

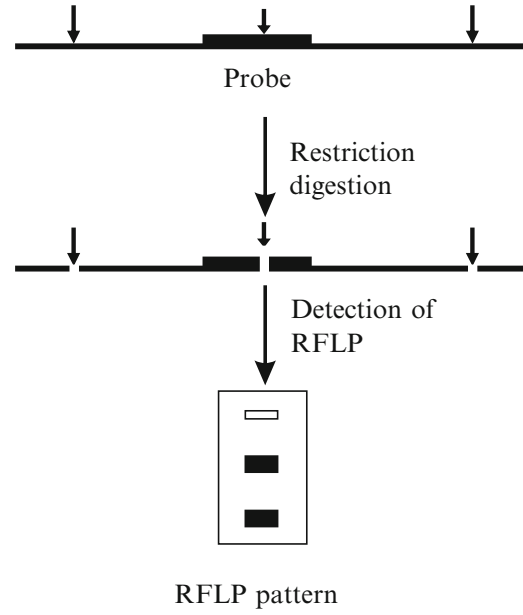


Fig. 2.7 Detection of two RFLP bands in homozygotes by a single probe representing a single locus. The two bands are generated due to cleavage by the restriction enzyme within the region that hybridizes with the probe (Based on de Vienne et al. 2003)

progeny and their patterns will not recombine. On the other hand, the patterns of the two bands will recombine in case they represented different RFLP loci. It may be pointed out that duplication events may generate more than two copies of a sequence resulting in multiple bands. So long as the banding patterns are not too complex for genetic analysis, inheritance studies would permit the determination of the number of different RFLP loci involved as well as identification of the alleles at the different loci. Further, genetic analysis would also allow the identification of allelic bands from nonallelic ones.

A suitable mapping population is analyzed to detect linkage between different RFLP loci as well as to assess whether an RFLP locus is linked with an oligogene/QTL. The RFLP loci can be mapped together to generate linkage maps, which are comparable to the conventional linkage maps. The RFLP maps can be successfully integrated with the conventional linkage maps. The RFLP maps can be placed onto specific chromosomes or even chromosome arms. This can be achieved by (1) detecting linkage between

RFLP markers in a linkage group with known genetic markers that are already mapped onto specific chromosomes. Further, one may utilize (2) addition/substitution lines, (3) appropriate translocation stocks, and/or (4) monosomic/trisomic lines for suitably designed studies for assigning RFLP markers to specific chromosomes. (5) Finally, RFLP probes may be used for in situ hybridization with the preparations of polytene chromosomes.

2.5.6 Advantages of RFLPs

RFLPs are well-accepted DNA markers and were widely used during the 1980s and 1990s for a variety of purposes, including preparation of linkage maps. They are still important as anchor markers in comparative mapping and synteny analyses. RFLP marker system offers several advantages that are as follows: (1) a very large number of RFLP loci can be scored and mapped in a mapping population so that even very small chromosome segments can be mapped; (2) the mapping of an RFLP marker does not require the associated gene to express itself; (3) they are highly reproducible, (4) are codominant in nature, and (5) allow mapping of even QTLs; and (6) construction of RFLP maps is very rapid as compared to that of conventional linkage maps. RFLP maps have been developed for several crop species, including maize, rice, wheat, etc.; the genome of *Arabidopsis thaliana* has been mapped to saturation. The RFLP analysis requires relatively large amount (5–10 μg) of DNA. However, a single Southern blot can be used for successive analyse with several (usually, eight to ten) probes, and they can be stored for a period of many years. Finally, several Southern blots, representing hundreds of individuals, can be analyzed simultaneously (Rafalski and Tingey 1993).

2.5.7 Limitations of RFLPs

The RFLP technique suffers from the following limitations: (1) The RFLP procedure is expensive and requires considerable labor and time. (2) The original method used radioactive probes that are hazardous to handle and require special disposal facilities. This difficulty can be resolved by using nonradioactive labels. (3) Considerable skill and effort is needed for the development of RFLPs, including the construction of genomic/cDNA libraries for the identification of suitable probes. (4) The DNA used for RFLP analysis must be of high purity to enable restriction digestion. (5) Further, scoring of RFLPs in different individuals/lines takes far greater time and effort than that for many other molecular markers like SSRs. (6) Finally, this marker system is not amenable to automation and high-throughput analysis. As a result, RFLPs are no more in common use.

2.5.8 Conversion of RFLP Markers into PCR-Based Markers

Once a useful RFLP marker has been identified, it can be converted into a more convenient and user-friendly PCR-based marker amenable to high-throughput procedures. This can be done by sequencing the two ends of the longer (slower-moving) RFLP fragment and designing a pair of primers using this sequence information. These primers are used for PCR amplification of the fragment from the genomic DNAs of the individuals/lines polymorphic for the concerned RFLP fragment; this approach is often called *PCR-RFLP*. In case the amplified fragment shows length polymorphism, we have a PCR-based sequence-tagged site (STS; Sect. 3.10) marker representing the RFLP locus. It may be pointed out that this fragment would exhibit fragment length polymorphism in case the RFLP was generated by either deletion or insertion of a sequence between the two primer-binding sites for the STS marker. However, the fragment will not exhibit length polymorphism if the RFLP were the result of a mutation in a recognition site located between the two primer-binding sites. In this situation, polymorphism can be detected by digesting the amplified fragment with the concerned restriction enzyme;

this would give rise to a cleaved amplified polymorphic sequence (CAPS; Sect. 3.14) marker. In addition, this type of polymorphism can also be detected by single-strand conformation polymorphism (SSCP; Sect. 3.15) or denaturing/temperature gradient gel electrophoresis (DGGE/TGGE; Sect. 3.16).

2.6 Diversity Array Technology

Diversity array technology (*DArT*) is a high-throughput, low-cost genotyping system. It is essentially similar to amplified fragment length polymorphism (AFLP; Sect. 3.9) procedure, except for the use of microarray-based nucleic acid hybridization in the place of gel electrophoresis for the detection of polymorphism (Jaccoud et al. 2001). DArT was initially developed for the assessment of genetic diversity present in a species, but it has found several other applications. DArT analysis consists of the following two steps: (1) construction of a microarray, called diversity array or genotyping array, and (2) genotyping of the test individuals/lines based on hybridization of their genomic fragments with the concerned genotyping array. A *genotyping array* contains such genomic DNA segments of a given species, which are found to be polymorphic across a range of germplasm of interest. These DNA fragments are obtained by a procedure involving the following steps: (1) isolation and purification of the genomic DNAs from several individuals/lines of a population/species representing the diversity to be studied, (2) pooling of ~5 ng DNA from each of these individuals/lines and digesting the pooled DNA with the selected restriction enzymes, (3) ligation of appropriate adapters to the restriction fragments, (4) reducing the complexity of fragments by 10–1,000-fold and PCR amplification of the selected fragments, and (5) cloning of the amplified fragments. (6) DNA insert from each of the clones is amplified individually using vector-specific primers, and (7) the amplification products are purified and spotted onto a solid support like a microscopic slide to prepare the microarray (Fig. 2.8). Thus, construction of the genotyping array does not require knowledge of either the sequence or the function of the DNA segments used for the purpose.

Complexity of a genome or DNA preparation represents the total number of different sequences present in it. Thus, a DNA preparation of low complexity will have a smaller number of different sequences than that of high complexity. One approach for reducing the complexity of DNA fragments is to use primers having one to three selection nucleotides at their 3′ ends for PCR amplification (Jaccoud et al. 2001). A *selection nucleotide* is an arbitrary nucleotide added to the 3′ end of the primer so that only such fragments that have the nucleotide complementary to this nucleotide at the corresponding position will be amplified. This will reduce the number of fragments amplified to one-fourth of the total number of different fragments for every selection nucleotide used in a primer. Another approach for complexity reduction is to digest the genomic DNA with a combination of two (one rare cutter enzyme like *Pst*I and one frequent cutter enzyme, such as *Taq*I or *Bst*NI) restriction enzymes. In this case, the enzyme combination has considerable effect on the level of polymorphism revealed, and the most successful combination in revealing polymorphism may depend on the plant species. For example, in barley, the enzyme combinations *Pst*I and *Taq*I and *Pst*I and *Bst*NI were equally effective in revealing polymorphism, and these combinations were superior to the other enzyme combinations tested. But in the case of wheat, the enzyme combination *Pst*I and *Taq*I was superior to the other enzyme combinations, including the combination *Pst*I and *Bst*NI (Wenzl et al. 2004; Akbari et al. 2006).

All the fragments amplified following the complexity reduction procedure are cloned (Fig. 2.8). DNA inserts from all the clones are amplified again using vector-specific primers, and the amplified fragments from each insert are spotted individually on a suitable solid support to generate a microarray (Appendix 2.3); this is called *discovery array*. It may be mentioned that only a small proportion (usually, around 4–10 %) of the fragments present in a

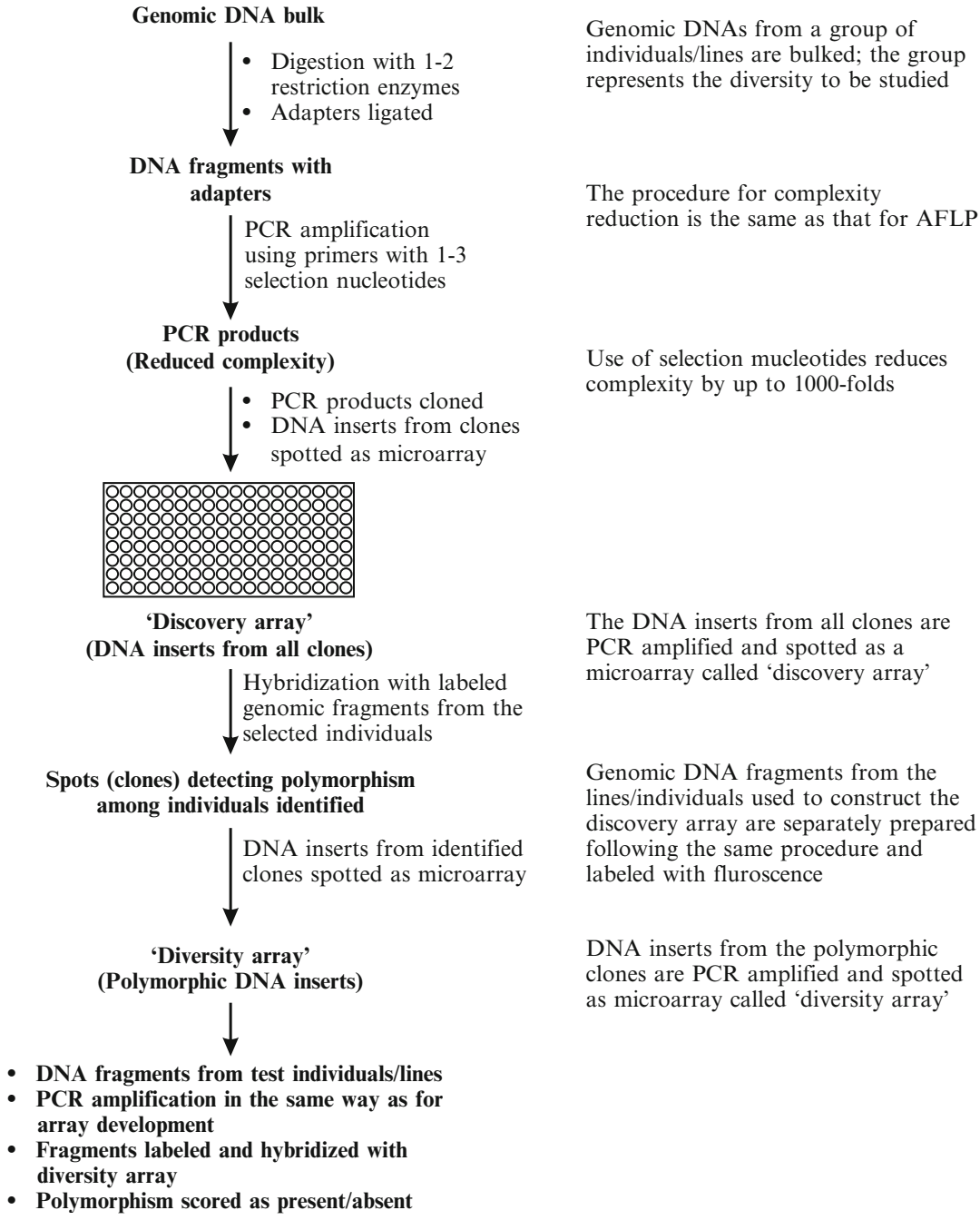


Fig. 2.8 A simplified schematic representation of diversity array technology (DArT) procedure (Based on Jaccoud et al. 2001)

discovery array would be polymorphic. The polymorphic fragments are identified by using fluorescence-labeled genomic DNA fragments from the same individuals/lines, whose genomic DNA was pooled to construct the discovery array, for hybridization with the fragments spotted onto the discovery array. The labeled fragments are prepared following the same

protocol of restriction digestion and PCR amplification that was used for construction of the discovery array. If genomic DNA fragments from two individuals/lines were labeled with two different fluorescent dyes and were used together for hybridization with the discovery array, most of the spots would hybridize with fragments from both the individuals/lines. These spots would produce a fluorescence color distinct from those generated individually by the two fluorescent dyes used for labeling. Some spots, however, will hybridize with the fragments from only one of the two lines/individuals. These spots would produce fluorescence color characteristic for the dye used for labeling of the concerned fragments, and they would contain fragments that differ between the two lines/individuals, i.e., are polymorphic. The spots containing polymorphic fragments are identified, and the DNA inserts are amplified from the corresponding clones and are finally spotted onto a solid support to develop the *genotyping array*.

For genotyping a line/an individual, the genomic DNA (50–100 ng) from the individual/line is isolated and fragments suitable for analysis are prepared using the same protocol that was used for microarray preparation (Fig. 2.8). The genomic DNA is digested with the same restriction enzyme(s) and amplified using the same primer that was used to construct the genotyping array. In addition, the fragments are labeled with a fluorescent dye and used for hybridization with the genotyping array. The genotyping array is simultaneously hybridized with the fragments of the cloning vector used for genotyping array construction; these fragments are labeled with a different fluorescent dye. This is done in view of the presence of the sequences from this vector in all the spots on the microarray. Hybridization signals are detected and analyzed using specialized software, e.g., DArTsoft, which converts them into scores of 1 or 0, i.e., "present" or "absent." These scores provide the fingerprint of the individual/line, and they are also used for statistical analyses in the same way as the scores for RAPDs, SSRs, etc. The software like DArTdb, Client Interaction, DArTsoft, and DArTools required for DArT analyses have been built on the open-source software LAMP.

DArT generally detects polymorphism produced by SNPs in the restriction sites and at the sites corresponding to the selection nucleotides of the PCR primers. It also detects relatively large InDels (insertions and deletions), structural rearrangements, and copy number variations in the region between the two neighboring restriction sites. The DArT markers are distributed throughout the genome, but a majority of them tend to be located in the genetically active regions of the genome. The bias in favor of genetically active genomic regions is due to the use of methylation-sensitive restriction enzyme *Pst*I for DNA digestion. About 50 % of the DArT markers in species like barley, wheat, sugarcane, oat, sorghum, potato, etc. are highly homologous to known genes. DArT procedure has been adapted to take advantage of the Group II transposable elements MITEs, and assays have been developed for rice and some other crops.

DArT has been used for comprehensive characterization of germplasm, diversity studies, selection of parents for hybridization, seed purity/product integrity testing, and for genetic, physical (in genome sequencing), and QTL mapping. It has also been used in studies on epigenetic changes due to DNA methylation, association mapping, MAS (including gene introgression from wild germplasm), and genomic selection. A genotyping array needs to be developed for a given species only once. Genotyping arrays have been developed for several crops like wheat, rice, barley, chickpea, pigeon pea, etc. DArT automated platform genotypes for thousands of loci in a single assay and allows automated data acquisition and storage. DArT offers advantages like low costs of development and application (a few cents per data point), minimal DNA requirement, and comprehensive genome coverage. A single DArT assay takes a maximum of three working days from DNA to marker genotype data. DArT is as effective in detecting polymorphism in a polyploid species like wheat as it is in diploid species like barley. The chief limitation of DArT is the use of restriction enzymes, which are

expensive and require DNA preparations of high purity. In addition, it requires specialized equipment as well as software programs for implementation, which may not be affordable for most breeding programs/projects. In such cases, the genotyping and analysis work can be outsourced.

2.7 Variable Number of Tandem Repeats

The *variable number of tandem repeats* (*VNTRs*) are stretches of DNA composed of variable numbers of tandemly repeated sequences of, usually, 2–60 bp. VNTRs are distributed throughout the genome, and each such genomic location may be regarded as a VNTR locus. The number of tandem repeats present at a given VNTR locus varies greatly, so that each VNTR locus has several alleles. VNTRs are generally classified as minisatellites and microsatellites, which together constitute the *hypervariable DNA*. *Minisatellite* sequences are usually 0.2–2 kb long and are made up of 11–60-bp-long tandem repeat units having identical or almost identical sequences. The *microsatellite sequences*, on the other hand, are usually less than 100 bp long and consist of tandem repeats of 2–7 bp. Microsatellites are extensively used as markers in plants, and they are discussed in detail in Chap. 3 (Sect. 3.11). In the case of humans, minisatellite DNAs are concentrated in the proterminal regions of chromosomes; therefore, they are not good markers for mapping of human genome. Many different VNTR loci may share a consensus sequence. In such cases, a "polycore" probe can be constructed for Southern hybridization, which can simultaneously score alleles at up to 30 different VNTR loci. Therefore, each polycore probe generates a detailed "DNA fingerprint" of an individual. Initially, DNA fingerprinting in humans was based on polycore probes (Jeffreys et al. 1985). Fingerprinting involves digestion of the genomic DNA with a restriction enzyme that cleaves the DNA outside the regions of VNTR repeats on both the sides. Ideally, the enzyme should cut the DNA close to the ends of the VNTR sequences. The different alleles at a VNTR locus are detected by Southern hybridization using the VNTR sequence as probe. Some minisatellite probes do produce low-resolution fingerprints in plants; they can be used for variety identification (Jones et al. 1997).

2.8 Single Feature Polymorphisms

Single feature polymorphism (*SFP*) or *single position polymorphism* (*SPP*) identifies allelic variation in pairs of lines/strains/isolates of a species by using high-density oligonucleotide microarrays for hybridization with their genomic fragments/cDNAs (Winzeler et al. 1998). SFP analysis may use a ready-made gene expression microarray like Affymetrix (http://www.affymetrix.com) GeneChips or Nimblegen (http://www.nimblegen.com) arrays. Alternatively, an array may be custom made using sequence information for genes from the following sources: ESTs (expression sequence tags), mRNA sequences, known/predicted ORFs (open reading frames) from genomic sequences, unigenes listed in the NCBI database, and conserved orthologous sequences (gene sequences found in related species). For example, Winzeler et al. (1998) developed the oligonucleotide microarray as follows: for each annotated ORF in the yeast genome, at least 20 different oligonucleotides (each 25 nucleotides long) perfectly complementary to the predicted coding regions of the ORF were used as probes and arranged on a microarray (Fig. 2.9). Each 25-base-long oligonucleotide is called a *feature*, *probe*, or *oligo*, and each feature represents a unique genomic segment. In addition, for each perfectly complementary oligonucleotide probe, an oligonucleotide with a single-base mismatch in the central position was synthesized adjacent to the probe; *the probe with the mismatch served as control*. Ordinarily, only unique sequences are used, but sometimes sequences of multicopy genes have also been used. The features are generally 25 nucleotides (nt) long, but longer (45 and 55 nucleotides long) probes have also been used.

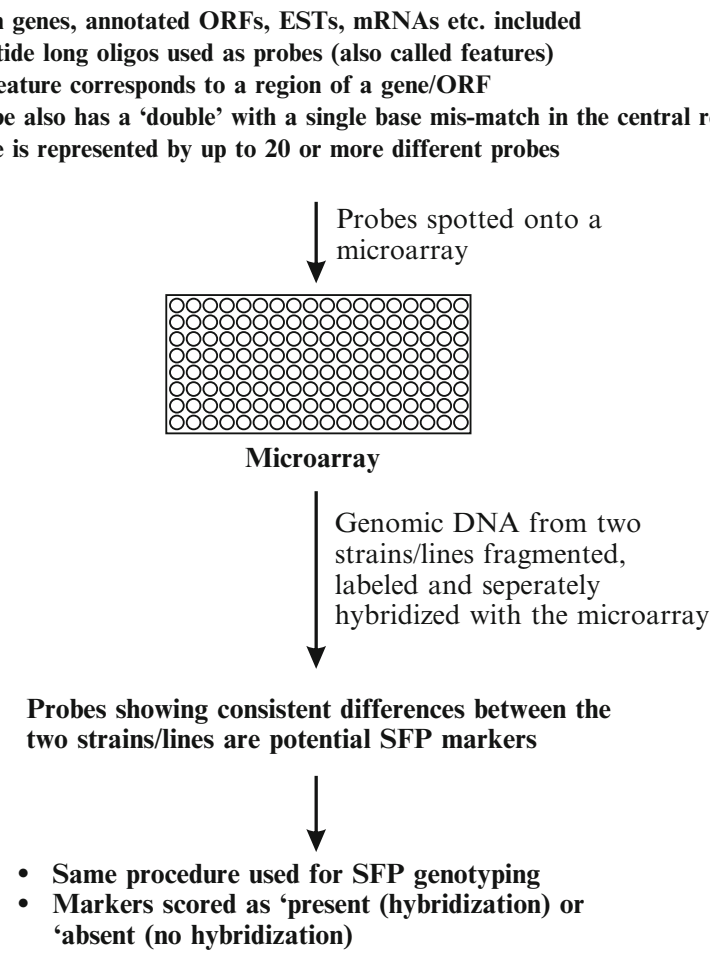


Fig. 2.9 A simplified schematic representation of single feature polymorphism (SFP). *ESTs* expressed sequence tags, *ORFs* open reading frames

Generally genomic DNAs from two distinct strains are isolated, fragmented, labeled with fluorescence, and hybridized with different sets of the microarray. The amount of fluorescence from each probe is measured, and the data are analyzed to identify those probes that show consistent differences between the two strains; these probes are potential SFP markers. Many workers, however, have used cDNA in the place of genomic fragments. But when cDNA is used, polymorphism will be detected only in those genes that are expressed in the tissue used for cDNA preparation. Therefore, cDNA would have to be prepared from multiple tissues across developmental stages and environments to capture most of the genes present in the genome. In addition, use of cDNA may interfere with SFP detection due to variation in the levels of gene expression. Generally, expression arrays designed for the same species are used for SFP discovery, but arrays developed for a related species can also be used. For example, the expression array for soybean was used for SFP discovery in pigeon pea and cowpea. But the use of expression array from a related species may lead to a higher false discovery rate for SFPs. SFPs detect sequence polymorphisms due to SNPs and InDels in or near the sequence represented by the features. SFP analysis was used for high-resolution (at distances of 11–64 kb) mapping of a locus conferring multidrug resistance as well as four other loci in yeast (Winzeler et al. 1998).

Subsequently, Borevitz et al. (2003) extended this approach to *A. thaliana*. They used an expression array with perfect match and mismatch features for genome-wide SFP analysis of an RIL population to map a morphological mutation. SFP analysis has been used in several other plant species, including barley, maize, soybean, tomato, wheat, rice, and pigeon pea.

The SFP procedure is relatively simple and can be applied to any species, for which an expression array and, preferably, a physical genome map are available. The SFP markers occur at high density and cover the whole genome, and their physical locations in the genome are known (Hill et al. 2013b). All the SFP markers are analyzed simultaneously and rapidly (in a few hours time). It is highly sensitive, and a new set of markers can be easily identified for any pair of lines/strains/isolates. SFPs segregate in a Mendelian fashion, are generally biallelic, and permit rapid mapping of genes/QTLs. SFP markers narrow down the search for candidate genes involved in the control of specific traits. The chief limitations of SFP technology are high false discovery rates (up to 40 % in some studies) in complex genomes, relatively high cost, and the need for construction of microarrays since they are commercially available for only a limited number of crop species. Further, SFPs are subject to ascertainment bias (Sect. 8.16.9) due to the use of a reference genome/strain for their discovery. Finally, a quality SFP analysis in species with complex genomes will require more DNA, a suitable complexity reduction protocol, and a higher number of replicates, which add to the cost.

2.9 Restriction-Site-Associated DNA Markers

Restriction-site-associated DNA (*RAD*) *markers* represent polymorphisms in the recognition sites for the restriction enzyme used for the preparation of the assay sample. These markers are developed by digesting genomic DNA with a selected restriction enzyme like *Eco*RI and ligating the fragments to biotinylated linkers. The fragments are now randomly sheared to produce much smaller fragments. As a result, each fragment attached to the linker contains only a short sequence located immediately on one side of a recognition site for the concerned enzyme. Streptavidin-coated immobilized beads are used to bind those fragments that are attached to the biotinylated linkers, and the rest of the fragments are removed by washing. The same restriction enzyme is then used to digest the fragments at the site where the linker is attached; this releases the fragments from the linkers and the beads. The fragments so obtained are called *RAD tags*; they comprise short genomic sequences flanking all the restriction sites for the concerned restriction enzyme present over the entire genome. Usually, two RAD tags would be recovered for each restriction site, and each of them is a potential RAD marker (Miller et al. 2007).

Polymorphic RAD markers are identified by using RAD tag samples prepared from two strains for competitive hybridization with a suitable microarray. A polymorphic RAD marker is detected when a microarray probe hybridizes with the RAD tag from only one of the two samples. The microarray used for hybridization may be a genomic tiling array, a cDNA array, or an oligonucleotide microarray. A *genomic tiling array* is a high-density microarray made up of oligonucleotide probes, which together span the entire genome of an organism. Thus, cDNA and oligonucleotide microarrays would identify a much smaller number of RAD markers than genomic tiling arrays. In fact, a microarray composed of the polymorphic RAD tags themselves would be optimal for identification of and genotyping for RAD markers. The RAD tag samples used for preparing a microarray can be enriched for informative RAD tags by subtractive DNA hybridization between the RAD tag samples derived from two different strains. The RAD tags are cloned before they are spotted onto a solid support for preparing the microarray.

The RAD tag samples to be used for RAD genotyping are ligated to linkers, amplified by PCR, and labeled with fluorescence. The RAD tag samples to be used for competitive hybridization are labeled with different fluorophores. The RAD markers are anonymous, dominant, and scored as "presence"/"absence."

The number of unique informative RAD markers detected by a given restriction enzyme depends on the frequency of SNPs in the genome and the size of genome. For a restriction enzyme with a 6-bp recognition sequence, several thousand polymorphic RAD markers would be available for most plant genomes. The RAD technique is a rapid, high-resolution, high-throughput procedure suitable for genetic analysis of most organisms. The development of RAD markers does not require prior sequence information. RAD markers have been used for a variety of studies in several organisms. This method has been adapted as RAD-Seq technique for SNP and InDel discovery discussed in Chap. 13 (Sect. 13.5).

Questions

1. Briefly explain the principle underlying restriction fragment length polymorphism.
2. Which kinds of changes in DNA generate RFLPs?
3. How can RFLPs be converted into PCR-based markers?
4. Why morphological and protein-based markers are no longer the preferred marker systems?
5. "The functional markers are the most useful markers for MAS and other applications". Examine this statement in the light of available information.
6. Explain the relevance of probes in RFLP detection.
7. "DArT is a modification of the AFLP procedure". Discuss this statement and highlight the differences between the two techniques.
8. Explain the principle of SFP (or RAD) markers.

Appendices

Appendix 2.1: Isolation and Purification of DNA from Plants

The plant DNA isolation and purification procedures can be grouped into three categories, viz., (1) CTAB method, (2) rapid DNA extraction methods, and (3) commercial DNA isolation kits. These procedures usually consist of three steps: (1) rupture and lysis of cells to obtain cell extract (tissue maceration), (2) purification of DNA, and (3) concentration of DNA. Plant tissues (fresh, freeze-dried, or frozen in liquid nitrogen) are usually ruptured by mechanical force. In general, DNAs isolated from fresh and frozen plant tissues are comparable in both quality and quantity. The particular method used for tissue grinding/maceration will mainly depend on the scale of work and the facilities available to the worker. On a small scale, mortar and pestle are widely used, but one may use a multi-pestle, a mixer mill or some other similar equipment on a moderate to large scale. The use of a mill would not only reduce the total time required for tissue maceration, but it may also improve DNA yield.

The CTAB Method

The CTAB procedure of Murray and Thompson (1980) is regarded as the standard method of DNA extraction. It is used to purify high-molecular-weight (50–100 kb) plant genomic DNA without the use of expensive equipment and time-consuming procedures. The powdered tissue is dispersed in an extraction buffer containing CTAB detergent and incubated at 50–60 °C for ~30 min. The suspension is then extracted with chloroform/octanol to remove cell wall debris, denatured proteins, etc. The extract is treated twice with chloroform/octanol, then CTAB is added, and the NaCl concentration is reduced so that CTAB–nucleic acid precipitate is formed. This precipitate is recovered through centrifugation and resuspended in 1 M CsCl, which is later removed by dialysis. In later modifications of the procedure, the precipitate is resuspended in 1 M NaCl or in TE (Tris–HCl and EDTA) buffer. The solution may be treated with RNase, and the DNA concentration can be increased by ethanol precipitation. The CTAB method has been modified by various workers to suit various needs. In one miniprep modification, CTAB is used in the homogenization buffer; the homogenate is extracted once with chloroform, followed by one ethanol precipitation and

resuspension of the pellet in water. This method is rapid so that one person can process 100–200 samples per day, and it yields adequately pure DNA for PCR. In general, this method yields ~5 times more DNA per unit weight of tissue sample than the other methods, and the DNA can be stored for long periods. However, the research workers are exposed to hazardous chemicals like CTAB, chloroform, and ß-mercaptoethanol.

Rapid DNA Extraction Methods

Several methods for rapid extraction of plant DNA have been developed (Hill-Ambroz et al. 2002; Bagge and Lübberstedt 2008). These methods have been dubbed as "quick and dirty" DNA extraction methods since the purity of DNA preparations is usually poor. In a rapid DNA extraction procedure for wheat, the tissue is placed in 0.25 M NaOH at 95 °C in a water bath for 1 min and macerated using a mortar and pestle, a 96-solid-pin replicator, or a Matrix Mill. Now 0.1 M Tris–HCl (pH 8.0) is added, the suspension is centrifuged, the supernatant is recovered, and the DNA is precipitated with 3 M sodium acetate and 100 % isopropanol. The DNA samples are then placed at −80 °C for 1 h, and the DNA is pelleted by centrifugation. The pellet is washed with ethanol, and the ethanol is removed by centrifugation. The DNA is then resuspended in TE buffer (pH 8.0) and stored at −20 °C for 30 days. Approximately 1 μg of genomic DNA was isolated from 10 mg leaf tissue at a cost of about US $ 0.10. One person can process nearly 1,000 samples per day (Hill-Ambroz et al. 2002). In a simplification of this procedure, developed for DNA isolation from rice, the leaf tissue is ground in 0.5 M NaOH, and then 0.1 M Tris (pH 8.0) solution is added to the macerate. The suspension is mixed well and centrifuged, and ultimately the supernatant containing the DNA is recovered by pouring off into a fresh tube and stored at −20 °C. The amount and the quality of DNA is enough for PCR analysis, but it cannot be stored for long periods and may not be suitable for SNP assays. Leaf tissue and endosperm tissue drilled out of dry barley seeds or excised from soaked maize seeds have been used for DNA extraction.

DNA Extraction Kits

A variety of plant DNA extraction and purification kits are commercially available. Some examples of such kits are DNeasy Mini and Maxi kits from QIAGEN, NucleoSpin Plant kits from Clontech, PureLink® Genomic Plant DNA Purification Kit from Life Technologies, PowerPlant® DNA Isolation Kit from MO BIO Laboratories, MasterPure™ Plant Leaf DNA Purification Kit from Epicentre, etc. Most of these kits are generic and can be used for DNA isolation from many plant species, but some manufacturers offer kits for specific plant species. The kits include all the buffers, reagents, plasticware, etc., required for DNA extraction and purification after the plant material has been macerated. The manufacturers provide clear-cut directions for the extraction and purification procedures, which may take 40 min to 2 h, depending on the kit and the number of samples processed. Almost all manufacturers offer Mini kits in single sample format, but some of them also provide 96-well format and/or Midi/Maxi kits in single sample format. For example, QIAGEN offers DNeasy Plant Mini Kit for isolation of up to 30 μg DNA per sample, DNeasy Plant Maxi Kit for isolation of up to 260 μg DNA per sample, and the 96-well plate format DNeasy 96 Plant Kit with typical yield of 1–15 μg of high-quality DNA per well. The NucleoSpin Plant II kit from Clontech, advertised as a next-generation kit, has improved silica membrane and affords rapid isolation of more genomic DNA of higher quality. The typical DNA yields from <100 mg of plant tissue (fresh weight) range from 1 to 30 μg DNA suitable for PCR, Southern blotting, and restriction analysis. On the other hand, the NucleoSpin Plant Midi and Maxi kits yield 20–80 μg and 60–260 μg DNA, depending on the size and source of the tissue sample.

It may be clarified that the inclusion of a manufacturer's products, procedures, services, and/or equipment for description here or elsewhere in this book is only for the purposes of illustration, and it does not in any way imply their appreciation/recommendation/endorsement. The descriptions of such products,

procedures, services, or equipment are often based on the information available from the manufacturers, but other materials have also been used.

Determination of Quantity and Quality of the Isolated DNA

The quantity and quality of the isolated DNA may be determined by a comparison of aliquots of the extracted DNA with a standard DNA of known concentration by either gel electrophoresis or spectrophotometry. The spectrophotometric method also reveals DNA purity. The absorbance or optical density (OD) for each DNA sample is recorded at 260 nm and 280 nm. If the ratio of absorbance at 260 nm to that at 280 nm for a sample is between 1.8 and 2.0, it is regarded as pure DNA. Whenever this ratio is outside the above range, the DNA sample should be subjected to further purification by ethanol precipitation. Further, an OD of 1 at 260 nm corresponds to about 50 μg/ml DNA (Sambrook et al. 1989). In the electrophoretic method, 10-μl samples of the isolated DNAs along with the gel loading dye are loaded carefully in separate wells of an agarose gel. The gel is impregnated with the intercalating dye ethidium bromide for visualization of the DNA bands containing as little as 0.05 μg DNA per band. Similarly, 1 μg of uncut lambda DNA along with the loading dye is loaded in a separate well. After 2 h of electrophoresis, the bands for the DNA samples are compared with that for lambda DNA. The quantity of DNA is determined by comparing the width of the bands and the intensity of fluorescence under UV light using the software of a gel documentation and analysis system. A high-molecular-weight DNA preparation gives rise to a single dark band close to the loading well, while a fragmented DNA sample yields a smear (Sambrook et al. 1989). Thus, both spectrophotometric and electrophoretic methods permit estimation of DNA concentration. But DNA purity is revealed by spectrophotometry and DNA quality (high-/low-molecular-weight preparation) is visualized by electrophoresis.

Appendix 2.2: Genomic and cDNA Libraries

A *genomic library* is a collection of plasmid clones or phage lysates containing recombinant DNA molecules so that the sum total of DNA inserts in this collection, ideally, represents the entire genome of the concerned organism. For the preparation of a genomic library, total genomic DNA of the organism is extracted and subjected to partial digestion with a suitable restriction enzyme (Singh 2012b). Fragments of suitable size are separated, integrated into a suitable vector, and cloned in a host like *Escherichia coli*. A genomic library may be enriched in unique sequences by using a methylation-sensitive restriction enzyme like *Pst*I. Since the repeated sequences do not contain many genes, they are far more likely to be methylated than unique sequences. As a result, the repeated sequences would be cut into much larger fragments that are not suitable for cloning. In some species like tomato, the frequency of unique sequences in *Pst*I-derived genomic library is almost comparable to that in a cDNA library and about three times as much as in a *Eco*RI-derived genomic library. In contrast, in species like rice and lentil, the frequency of unique sequences is only slightly higher in a *Pst*I-derived library than that in *Eco*RI-derived library.

Similarly, a *cDNA library* is a population of bacterial transformants or phage lysates, in which each mRNA isolated from an organism or tissue is represented as its cDNA insert in the recombinant DNAs present in this population. Construction of a cDNA library involves isolation and purification of mRNA using a suitable procedure, production of cDNA from this mRNA by reverse transcription catalyzed by the enzyme reverse transcriptase, integration of the cDNAs into a suitable vector (usually, a phage insertion vector), and cloning of the recombinant DNAs in a host like *E. coli*. cDNA library preparation is demanding, and considerable care needs to be exercised. A cDNA library would represent only those structural genes that are transcribed

Table 2.6 A comparison between cDNA and genomic libraries

Feature	Genomic library	cDNA library
Sequences present	Ideally, all genomic sequences	Only structural genes that are transcribed
Contents affected by:		
(a) Developmental stage	No	Yes
(b) Cell type	No	Yes
(c) Environment	No	Yes
Features of the DNA inserts representing a gene:		
(a) Size	As present in the genome	Ordinarily, much smaller
(b) Introns	Present	Absent
(c) 5′- and 3′-regulatory sequences	Present	Absent
(d) Sequences of a single gene present in	One or more clones	One clone
As compared to the genome:		
(a) Enrichment of sequences compared to that in the genome	In amplified genomic libraries	For abundant mRNAs
(b) Reduction in frequency	In amplified genomic libraries	For rare mRNAs
(c) Variant forms of a single gene	Not possible	Possible in cases of genes whose RNA transcripts are alternatively spliced

in the concerned tissue/organ during the given developmental stage. It is also likely to be enriched for abundant mRNA species. In addition, when RNA transcripts of a gene are alternatively spliced, two or more variant forms of such a single gene would be represented in the cDNA library. The genomic and cDNA libraries differ for several features (Table 2.6).

A genomic/cDNA library will consist of thousands of clones, and it is unlikely that all of them would be useful as probes. Therefore, the clones have to be screened for identification of those clones that are suitable for use as probes (de Vienne 2003). Some clones would fail to detect polymorphism, some may produce many bands or a complex pattern of bands, while some others may not generate any band; all such clones are rejected from the probe library. The clones forming complex band patterns would represent highly repeated DNA sequences. Many clones will yield one (in the case of homozygous individuals) or two (in the case of heterozygous individuals) bands; these clones represent unique DNA sequences and are used as probes. Some probes would give rise to more than two scorable bands; these probes most likely detect multiple RFLP loci and may be useful in some studies. The proportion of clones that detect polymorphism depends largely on the species. For example, only 5–10 % of the probes revealed polymorphism among the cultivated varieties of tomato when their DNAs were digested with three different enzymes, and the average number of alleles detected per locus was two. For this reason, it became necessary to use interspecific hybrids for preparation of RFLP maps of tomato. On the other hand, 95 % of the probes detected polymorphism among the DNAs from lines of only the dent group of maize when they were digested with three different enzymes, and the mean number of alleles per locus was more than six.

Appendix 2.3: Microarrays

An *array* is an orderly arrangement of data or items. A *microarray* is a glass slide or thin wafer of silicon glass, onto which a very large number of probes are immobilized as microdots. A *probe* is a DNA sequence representing a part or whole of a gene/cDNA single-stranded molecule. Microarrays are used for hybridization with a mixture of labeled test DNA molecules to detect the presence of sequences complementary to the probes spotted on the microarray (Singh 2012b;

Winzeler et al. 1998). Thus, microarray strategy is the exact opposite of dot blot assay, in which a series of test DNA/RNA molecules are immobilized onto a solid support and a labeled probe is hybridized with them to identify the blots having DNA/RNA molecules complementary to the probe. Each of the probes immobilized onto a microarray is a pure preparation, while the test DNA is a mixture of fluorescence-labeled DNA/cDNA fragments. The results of hybridization are visualized by confocal microscopy. A single assay using, say, a gene expression microarray permits identification of all the genes expressed in a given tissue of an organism at a given time under the given environment. Microarrays were first used in the case of yeast that has less than 7,000 genes. Every yeast gene was obtained as an individual clone, and a single-stranded sample of each gene was spotted onto a glass slide in arrays of 80 × 80 spots. In order to identify the genes expressed in yeast cells under a set of given conditions, mRNA is extracted from these cells, is converted into cDNA by reverse transcription, and is fluorescently labeled. The labeled cDNA is hybridized with the microarray, and the identity of the spots showing fluorescence, i.e., hybridization, is determined by confocal microscopy. The spots showing fluorescence represent the genes that were expressed in the cells from which the mRNA was isolated.

Microarrays are basically of two types, viz., DNA microarrays and antibody microarrays. *DNA microarrays*, in turn, are of the following two types: (1) spotted microarrays and (2) oligonucleotide microarrays. In the case of *spotted microarrays*, DNA fragments representing different genes of an organism are obtained from genomic and/or cDNA library of the concerned species or a related species and spotted onto a suitable solid support. On the other hand, *oligonucleotide microarrays* or *DNA chips* are produced by synthesizing oligonucleotides at a very high density (up to one million oligonucleotides/cm^2) directly on thin wafers of silicon glass. Each oligonucleotide has the sequence of a different gene, is located at a precise position on the microarray, and is synthesized by photolithographic solid-phase DNA synthesis. The DNA chips are inverted onto a controlled temperature hybridization chamber, into which fluorescently labeled test DNA, e.g., cDNA, preparation is injected and allowed to hybridize with the oligonucleotides. Laser excitation enters through the back of the glass support focused at the interface of the array surface and the hybridization solution. Fluorescence emission is collected by a lens and passed onto a sensitive detector, and a quantitative assay of hybridization intensity is obtained.

Microarrays are used for the following types of studies: (1) analysis of gene expression pattern in an organism as affected by the stage of development and/or environment, (2) identification of common regulatory elements by analysis of co-regulated genes, (3) analysis of already identified SNPs (these microarrays are often called *SNP chips*), (4) detection of genetic diseases, and (5) discovery and analysis of certain types of molecular markers, e.g., DArT, SFP, and RAD markers. In addition, specialized microarrays can be designed for specific purposes. For example, (6) arrays made up of probes that span across exon junctions allow detection and quantification of mRNA isoforms produced by alternative splicing, and (7) genomic tiling microarrays permit a very high-resolution mapping of the transcribed genomic regions. A *genomic tiling microarray* comprises a set of overlapping oligonucleotide probes that together represent a subset of the genome of a species at very high resolution. Analyses based on microarrays are highly sensitive and very fast, and all the genes present in the genome are analyzed in a single assay. These assays also generate quantitative data on gene expression, and the use of multiple labels of different colors may allow the use of a single microarray for assaying multiple test samples. But the construction of microarrays is expensive and requires genome sequence information. Further, there may be cross-hybridization leading to high background noise, and comparison of expression levels across experiments is often difficult.

Polymerase Chain Reaction-Based Markers

3

3.1 Introduction

The development of restriction fragment length polymorphism (RFLP) marker system amply demonstrated that DNA sequence polymorphisms could be detected and used as molecular markers. It also highlighted the great abundance and genome-wide distribution of DNA-based markers, and the novel opportunities generated by this development in various genetic and other biological investigations. But RFLP technique requires considerable preparatory work, is technically demanding, and involves expensive reagents. Therefore, efforts were made to develop simpler, less expensive, and more convenient marker systems. These efforts led to the development of several polymerase chain reaction (PCR)-based marker systems during the 1990s, which are generally called *second-generation markers*. These markers have virtually replaced the first-generation hybridization-based markers as they require much smaller quantity of DNA of relatively lower quality and are much more user-friendly and amenable to automation. Simple sequence repeat (SSR), amplified fragment length polymorphism (AFLP), and randomly amplified polymorphic DNA (RAPD) markers are some of the widely used PCR-based markers. These marker systems became possible due the development of PCR procedure by Mullis and coworkers for amplification of specific DNA sequences from DNA samples of very high complexity. At the same time, refinements in chemical synthesis of DNA ensured that PCR primers became readily available at a reasonable price. Finally, continued refinements in PCR technology enabled the PCR to become a routine laboratory technique. As a result, the PCR-based markers became greatly user-friendly and are very popular. Therefore, a brief description of chemical synthesis of oligonucleotides and PCR procedure precedes the discussion of various PCR-based marker systems.

3.2 Oligonucleotides

An *oligonucleotide*, "oligo" for short, is a short DNA fragment of few to several nucleotides (nt). Oligos are usually single-stranded, but they can also be double-stranded. Oligos are ordinarily chemically synthesized using automated oligonucleotide synthesizers. Khorana and coworkers synthesized a complete gene in 1970 using the phosphodiester method of DNA synthesis. This procedure was soon replaced by the more convenient and efficient phosphotriester approach; this method could synthesize up to 10–20 nt long oligos in a few days, and it was automated. But the present-day oligonucleotide synthesizers use the phosphite triester approach of DNA synthesis. This procedure takes 15 min for adding one nucleotide to the growing chain, and oligos as long as 50 nt can be prepared in good yields. It may be pointed out that the chemical synthesis of DNA proceeds from the 3′ to the 5′ direction

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_3, © Author(s) 2015

as compared to the progress of DNA replication from the 5′ to the 3′ direction.

Oligonucleotides have a variety of applications ranging from their use as primers to that as therapeutic agents. Oligonucleotide sequences of 12–20 nt are used as probes in nucleic acid hybridization for various purposes, including detection of DNA sequence polymorphisms. Oligonucleotides of different lengths and with specified/arbitrary sequences are used as primers for amplification of DNA fragments for the various PCR-based marker systems and for producing cDNA from RNA templates. Oligonucleotides are also used for DNA sequencing by DNA synthesis and for chemical synthesis of a complete gene that can be used for genetic transformation. In addition, oligos are used as linkers and adapters for modification of the cut ends of DNA fragments to facilitate their cloning and/or amplification.

3.3 Polymerase Chain Reaction

Karry Mullis (1990) conceived the idea of PCR in 1983 while thinking of novel approaches for DNA sequencing. Mullis and coworkers developed the PCR procedure, and Saiki et al. (1985) reported the first application of this technique. In a matter of few hours, the PCR procedure produces microgram (μg) quantities of DNA copies (up to billion copies) from even a single copy of the desired DNA or RNA segment (the *target sequence*). The DNA segment amplified by PCR is often referred to as *amplicon*. The PCR process has been completely automated and compact thermal cyclers are commercially available.

3.3.1 Generalized Procedure for PCR

PCR uses the following preparations/reagents: (1) a template DNA preparation containing the desired/target sequence, (2) a thermostable DNA polymerase, (3) a pair of ~20 nt long oligodeoxynucleotide primers that are complementary to the two 3′ ends of the target DNA fragment, and (4) the four deoxynucleotide triphosphates, viz., dATP, dCTP, dGTP, and dTTP. All these reagents are present in a suitable buffer system. The above reaction mixture is subjected to the following three steps (Fig. 3.1) for, usually, 35–40 cycles. The reaction mixture is first heated most often to 94 °C to ensure denaturation of the template DNA. The duration of the denaturation step is usually 2 min in the first PCR cycle, but it is only 1 min in the subsequent cycles. The mixture is then cooled to a temperature that would allow the primers to anneal to their complementary sequences located at the 3′ ends of the target DNA segment, i.e., the template DNA. Generally the annealing temperature is between 40 and 60 °C, and the duration of this step is 1 min. Since the primers are used at a much higher concentration than the template strands, they have a much higher chance to anneal with template strands than that for the two complementary strands of template DNA to pair with each other.

In the third and final step, the primers are extended due to the progressive addition of nucleotides to the free 3′-OH groups of the primers and, subsequently, the new strands being synthesized. These reactions lead to the extension of the two primers so that they grow toward each other; as a result, the DNA sequence located between the two primers is copied. The temperature during primer extension step is generally maintained at 72 °C, and the duration of this step is usually 2 min. Taq DNA polymerase is generally able to amplify DNA segments of up to 2 kb. However, it can amplify longer DNA segments provided it is used under certain special reaction conditions. Completion of the extension step completes the first cycle of amplification, and a new cycle begins with the initiation of the denaturation step. Thus, each PCR cycle takes merely 4–5 min.

The extension of primers continues till the strands are separated during the denaturation step of the next PCR cycle. Therefore, the products of primer extension based on the original DNA template, during the first and the subsequent cycles, are ordinarily longer than the target sequence since extension continues beyond the primer pairing sites; such PCR products are called *long product* (Fig. 3.2).

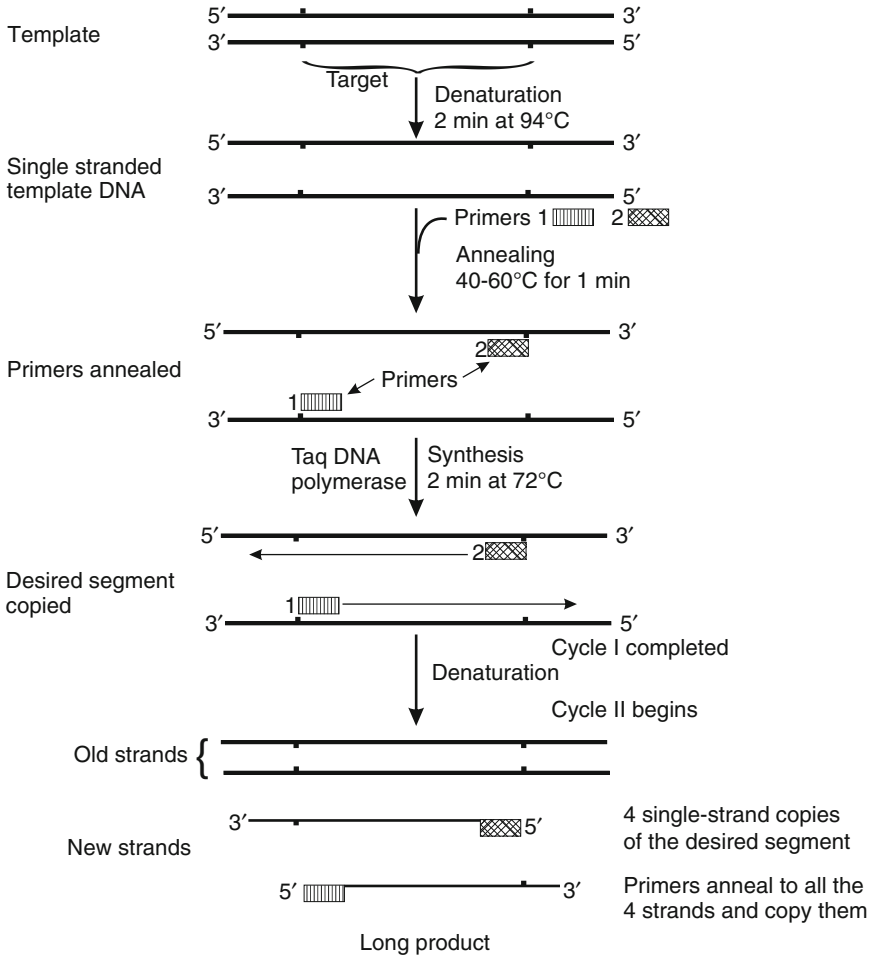


Fig. 3.1 A schematic representation of the three steps performed during the first cycle of PCR and their consequences. Note that the two primers used are complementary to the 3′ end sequences of the DNA segment to be amplified. The product of the first cycle is the "long product." During subsequent cycles, the long product accumulates linearly, i.e., only 2×40 copies will be produced after 40 cycles of PCR from a single copy of the target segment in the original DNA sample

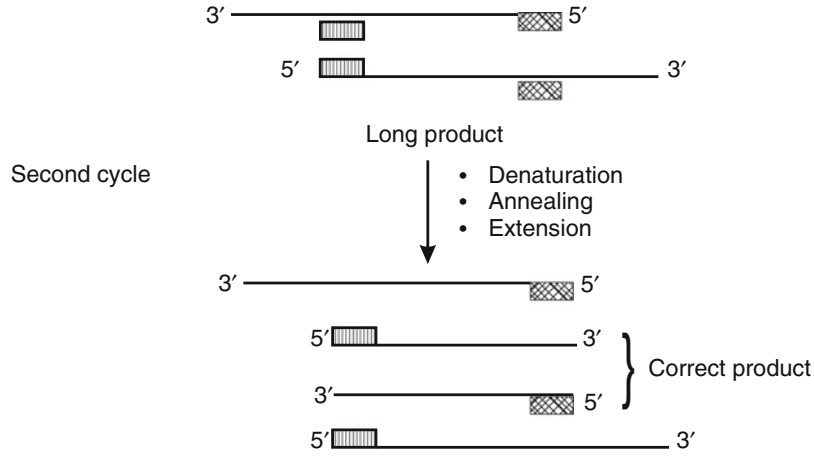


Fig. 3.2 The correct copy of the target sequence is produced in the second and the later cycles; its number increases exponentially. After 40 PCR cycles, ~2^{39} copies are expected to be produced from a single copy of the target segment

Table 3.1 Factors affecting polymerase chain reaction

Factor	Features
Template DNA	
(A) Natural features	G + C content, complexity, length of regions to be amplified, the composition (e.g., presence/absence of short repeats) of the region to be amplified
(B) Experimental features	Extraction conditions of the DNA, degree of shearing, concentration and copy number of target sequences
Characteristics of thermal block	Temperature accuracy, uniformity of temperature within the tube and between positions in the rack, ramping times
DNA polymerase	Stability, processivity, and concentration
dNTPs	Concentration
$MgCl_2$ and KCl	Ionic concentration
Organic compounds	Formamide (for SSR), glycerol, and DMSO
Temperature profiles	Annealing time and temperature, extension time, and number of cycles
Primer	Size, composition, sequence, and purity

During the subsequent cycles, primers will also anneal to the "long products" at primer binding sites located before their 3′-ends. The extension of these primers will yield the correct copy of the target sequence; these copies are known as the *correct product*. Since the "long product" is produced only from the original DNA template, it continues to increase linearly. In contrast the "correct product" is generated from both the types of PCR products so that its number doubles in every cycle, i.e., it increases exponentially. Thus, one PCR cycle increases the number of copies of the target DNA segment by a factor of two in comparison to their number at the beginning of the cycle. As a result, 2^n copies of the target DNA segment are expected to be present at the end of n cycles. But the actual number of copies generated by PCR is lower than, but quite close to, this number. The investigator has only to set the temperature and duration of each step of PCR and the number of cycles to be run in the automated thermal cyclers. After this, the machine carries out all the operations exactly as specified. After the last PCR cycle, the amplification product is separated from the template DNA by gel electrophoresis, removed from the gel, and purified; it can now be used for the specific desired purpose.

PCR is a relatively robust technique when the selectivity of primers allows for stringent annealing conditions. Purity of the template is not important provided no sequence similar to the target and of foreign origin contaminates the sample. A large number of factors can influence the success of PCR and the nature of PCR products (Table 3.1). Taq DNA polymerase at 1.25 units/25 µl of reaction mixture would give reproducible results. Taq DNA polymerase (from *Thermus aquaticus*) is perhaps the most commonly used, but Pfu (from *Pyrococcus furiosus*) and Vent® (from *Thermococcus litoralis*) polymerases are more efficient. The primer length should be at least 15–17 nt for amplification of the specific desired DNA sequence, and the melting temperature of the two PCR primers should be the same. *Melting temperature* (T_m) of a primer is the temperature at which 50 % of the template-primer duplexes would dissociate into separate strands. The annealing temperature is usually 1–2 °C lower than the melting temperature of the PCR primers, while for RAPD analyses, it is kept ~5 °C lower than the T_m of the primer. In case of RAPD analyses, ~4 µM primer should be used with ~30 ng template DNA (in 25 µl reaction mixture) to obtain sharp and reproducible bands.

3.3.2 Separation of PCR Amplification Products

DNA fragments/amplicons generated by PCR can be separated by electrophoresis in agarose or acrylamide gels. Agarose gels are easier to make and use, and the electrophoresis system

used for these gels is simpler than that for acrylamide gels. The agarose concentration in the gels depends mainly on the size range of fragments to be separated. An agarose gel of about 1 % can separate fragment of ~300–1,500 bp, and fragments differing in length by about 50 bp can be resolved. Polyacrylamide gels contain a much more uniform pore size than agarose gels and allow separation of DNA fragments with a higher resolution. A gel containing 6 % acrylamide has a fine network formed by polyacrylamide and can separate DNA fragments differing in length by even one or two base pairs. But the maximum fragment length that can be separated using this gel is usually 500 bp. Polyacrylamide gels are suitable for detection of SSR, AFLP, DNA amplification fingerprinting (DAF), and sequence-tagged sites (STS) markers, while agarose gels are well suited for RFLP and RAPD markers. The first-generation automatic DNA sequencers used capillary gel electrophoresis because it afforded automation of filling the capillaries with the polymers as well as loading of the samples. The polymer filled in capillaries of DNA sequencers is similar to polyacrylamide (de Vienne et al. 2003).

3.3.3 Multiplex PCR

Ordinarily, a single primer/pair of primers is used in one PCR reaction set up in a PCR tube to amplify a single target sequence from the given DNA sample. Often amplification of two or more different segments from the same DNA sample may be required, e.g., for analysis of some types of molecular markers. In such cases, a separate PCR reaction will have to be set up for every primer pair because of the difficulties in correct identification of their PCR products. However, if the amplification products of two or more primer pairs can be reliably distinguished from each other, these primer pairs can be used in a single PCR reaction tube; this is known as *multiplex PCR*, and the process is called *multiplexing*. The PCR products from different primers can be reliably separated by gel electrophoresis if their lengths do not overlap. Alternatively, different primers may be labeled with different fluorophores, and their PCR products can be distinguished on the basis of color differences in their fluorescence emissions. But this approach would require the fluorescence detection system of the first-generation automatic DNA sequencers. It is essential that all the primers used in a multiplex PCR have the same or almost the same melting temperature. This is essential for successful and specific amplification of all the concerned target sequences at the single annealing temperature used for the multiplex PCR. Multiplexing increases the throughput and reduces the cost and effort needed for scoring of markers.

3.3.4 Applications of PCR

PCR has many exciting and varied applications, some of which are as follows. It is used to study DNA polymorphism, including DNA fingerprinting, for which several PCR-based marker systems have been developed. PCR is used to detect the presence of transgenes introduced into organisms either by genetic transformation or hybridization. A variation of the PCR procedure, asymmetric PCR, generates copies of a single strand of the target sequence, which are used for first-generation automated DNA sequencing. PCR is also used for DNA sequencing reaction itself (thermal cycle sequencing PCR). The next-generation DNA sequencing procedures use PCR for in vitro cloning of the DNA fragments being sequenced. The enzyme reverse transcriptase is used along with DNA polymerase in RT-PCR (reverse transcription PCR) to generate DNA copies of RNA. Real-time reverse transcription PCR is used to estimate the initial quantity of the template RNA most specifically, sensitively and reproducibly. Several variations of PCR have been developed for specific applications, including inverse PCR for amplification of sequences flanking the target sequence, anchored PCR amplification of a target segment when the sequence of only one of its ends is known, overlap extension PCR for site-directed mutagenesis in the target segment, etc.

3.3.5 Advantages and Limitations of PCR

PCR is simple, relatively straightforward, very fast (requires only few hours), highly sensitive, and extremely versatile. It can amplify even a single copy of the target sequence present in a DNA sample and generate millions of copies of this sequence. PCR uses nanogram (ng) quantities of DNA, and purity and integrity of the DNA preparation are not critical. Further, even partially degraded DNA can be successfully used for PCR. It uses easy to store and relatively cheaper DNA polymerase and does not use radioactivity. However, sequence information for the two ends of the target segment must be known for designing of the primers. In general, segments of only up to 3 kb are amplified, but this length is ideally 1 kb. Taq DNA polymerase lacks proofreading activity so that it cannot remove the errors committed during replication. Further, PCR is sensitive to several inhibitors that may be present in the DNA preparation. The expected exponential amplification continues up to about 20 cycles or so, after which it enters linear phase and soon culminates in a plateau. The PCR procedure can often generate artifacts like "hybrid amplicons" and primer dimers, and it may produce erroneous results due to contaminating DNA. Primer dimers are frequently produced when the two PCR primers have partially complementary 3′ termini. They may also arise due to non-template-directed addition of some bases at the 3′ ends of the two primers, which may sometimes generate complementary 3′ ends in them.

3.4 PCR-Based Markers

PCR-based markers are considered as the *second-generation of molecular markers* and are based on DNA sequence polymorphisms detected by PCR amplification of the sample DNAs. The DNA polymorphisms are reflected in the amplification products from the target regions of the sample DNAs. The PCR procedure may use a single primer or a pair of primers, and the primers may have either arbitrary or specific nucleotide sequences. The products of amplification are separated by electrophoresis using either an agarose or a polyacrylamide gel and are visualized by staining the gel with either ethidium bromide or silver, autoradiography, or fluorescence detection. The primers used for amplification differ from one marker type to the other and form the basis of the concerned marker systems. These marker systems can be grouped into the following two categories on the basis of the primers used: (1) markers based on arbitrary sequence primers and (2) those based on specific sequence primers. More recently, (3) an intermediate group of techniques has been developed that uses either a combination of specific sequence and arbitrary sequence primers or primers composed of both fixed and arbitrary sequences. In addition, (4) some techniques combine restriction digestion of DNA with PCR amplification, and they together may be regarded as a separate group (Table 3.2). These marker systems have been extensively used for gene/QTL mapping, fingerprinting of plant genetic resources, and breeding materials including commercial varieties, analysis of genetic diversity, and studies on phylogenetic relationships.

3.5 Randomly Amplified Polymorphic DNAs

Williams et al. (1990) reported the procedure for the marker *randomly amplified polymorphic DNAs* that produces fingerprints of virtually any genomic DNA sample within a matter of hours without using radioactive reagents. A single, short (usually, 10 nt long) oligonucleotide with an arbitrary base sequence is used as primer for amplification of sequences from high molecular weight genomic DNAs of the test individuals. This primer acts as both the forward and the reverse primer for the amplification reaction (Fig. 3.3). The single primer would anneal at several sites in the template genomic DNA. Theoretically, for a 10 nt long primer, the binding sites are expected to occur, on an average, every

Table 3.2 A classification of the PCR-based marker systems in common use

Category of marker system	Marker system(s)[a]	Remarks
Arbitrary sequence PCR	RAPD, DAF, AP-PCR, ISSR	Simplest to implement; poor reproducibility
Specific sequence PCR	SCAR, STS, SSR, COS, ITP, IMP	Some to considerable developmental effort; simple to use
Combination sequence PCR	SRAP, TRAP, SCoT, CDDP, S-SAP, REMAP, RBIP, CoRAP, CBDP	Simple to implement (SRAP); database search necessary (TRAP, SCoT, CDDP, CBDP)
Restriction digestion combined PCR	AFLP, CAPSs	Technically more demanding (especially AFLP)

[a]*AFLP* amplified fragment length polymorphism, *AP-PCR* arbitrary-primed PCR, *CAPSs* cleaved amplified polymorphic sequences, *CBDP* CAAT box-derived polymorphism, *CDDP* conserved DNA-derived polymorphism, *CoRAP* conserved region amplification polymorphism, *COS* conserved orthologous sequence, *DAF* DNA amplification fingerprinting, *IMP* inter-MITE polymorphism, *ISSR* inter-simple sequence repeat, *ITP* intron-targeting polymorphism, *RAPD* randomly amplified polymorphic DNAs, *RBIP* retrotransposon-based insertion polymorphism, *REMAP* retrotransposon-microsatellite amplified polymorphism, *SCAR* sequence-characterized amplified regions, *SCoT* start codon-targeted marker, *SRAP* sequence-related amplification polymorphism, *S-SAP* sequence-specific amplification polymorphism, *SSR* short sequence repeat markers, *STS* sequence-tagged sites, *TRAP* target region amplification polymorphism

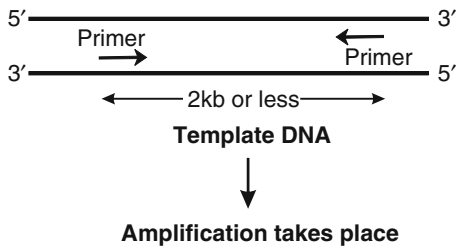


Fig. 3.3 A schematic representation of the RAPD marker system. A single arbitrary sequence primer of, generally, ten nucleotides is used for amplification. Amplification will take place if the primer binds to two sites located on the complementary strands within 2 kb of each other

4^{10} bp or 1,048,576 bp in a DNA strand, assuming a random distribution of nucleotides in the DNA strand (Appendix 3.1). However, exponential amplification can occur only when the primer anneals at two sites within ~2 kb of each other. Further, the two primer molecules should bind to the opposite strands of the template DNA so that their 3′ ends face each other; this would occur only when these two binding sites are in the opposite orientation (Fig. 3.3). The reaction conditions are normally so chosen that the number of fragments amplified is less than 20 per reaction (Fig. 3.4). Thus, a very large number of fragments can be generated by using a relatively small number of different primers. Usually, these fragments would be amplified from different regions of the genome so that several loci can be examined rapidly (Edwards 1998). Many RAPD primers may generate one to three intense bands each, which are polymorphic between the parents of a mapping population. It may be pointed out that only reproducible, intense bands should be used as markers so that the marker genotypes are scored with a degree of reliability.

RAPD method detects high level of polymorphism in plants and does not require large amounts of relatively pure DNA, and prior sequence information about the template genome is not required. It does not involve preliminary work like development of cloned DNA probes, preparation of filters for hybridization, etc., and the procedure can be automated. In addition, RAPD is safe, as it does not use radioactive components. RAPD has been used to construct high-density maps in several crop species like alfalfa, faba bean, apple, etc., in a relatively short time. This marker system has also been used to discover molecular markers linked to the desired genes in crops like tomato, lettuce, and common bean. RAPDs are dominant markers that are scored as "present" or "absent." When it

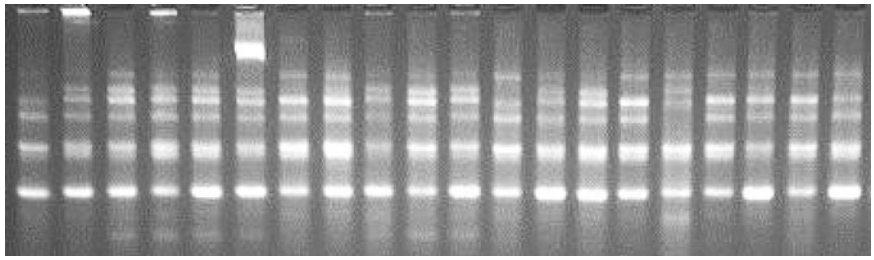


Fig. 3.4 RAPD profiles of 20 pea genotypes generated by the primer HU 12 (TGCTCAGCAG). Genotypes: 1, HUP-2; 2, Rachna; 3, DMR-42; 4, KPMR-551; 5, KPMR-615; 6, KPMR-619; 7, IPF-99-25; 8, VL-40; 9, DMR-46; 10, KPMR-660; 11, IPF-1-17; 12, IPF-1-22; 13, VL-41; 14, KPMR-662; 15, HFP-4; 16, HUDP-15; 17, KPMR-144-1; 18, DDR-49, 19, KPMR-526; 20, LFP-283 (Courtesy Kusum Yadav, Lucknow)

is important to distinguish heterozygotes from homozygotes for a locus, two RAPD markers tightly linked to this locus should be used. Further, one of the two markers should be in coupling phase, while the other marker should be associated in repulsion phase with the target locus. But this strategy will require twice the number of marker assays as that for a codominant marker. In addition, finding of two strategically located RAPD markers is not likely to be an easy task. The reproducibility of RAPD polymorphisms is low and is affected by several factors like primer to template concentration ratio, annealing temperature, and Mg^{2+} concentration (Williams et al. 1990). For example, a change of even 1 °C in annealing temperature may result in an entirely different profile of RAPD. Further, the amplification may fail due to an experimental error, but this can be scored as the "absence" allele. In many inheritance studies, RAPD markers showed significant deviation from Mendelian ratios possibly due to errors in scoring. The poor reproducibility of RAPD polymorphisms has prevented their widespread application in spite of their other highly attractive features. However, modifications of the RAPD approach have allowed the development of markers systems like SCAR, AP-PCR, RAMPO, etc., and this simple marker system still retains some relevance (Babu et al. 2014).

The information content of an individual RAPD marker is very low. RAPD markers often originate from repetitive DNA. Therefore, RAPD markers can be used as probes for locus-specific hybridization only after considerable sequence analysis of the markers. Sometimes, heteroduplex molecules may be formed between allelic RAPD products in heterozygotes, and these may give rise to false polymorphisms (Ayliffe et al. 1994). In addition, co-migrating bands may lack homology, and a single band may contain two or more different amplicons.

3.6 DNA Amplification Fingerprinting

DNA amplification fingerprinting amplifies genomic sequences using a single short oligonucleotide, typically, of 4–6 nt as primer, but primers of up to 15 bases can be used. This produces a range of up to 100 short amplified products of different lengths. The spectrum of products changes with each primer and template combination, but is characteristic for each combination. Fragments can be adequately resolved and visualized by polyacrylamide gel electrophoresis (PAGE) combined with silver staining. DAF uses less stringent conditions for annealing and primer extension reactions than PCR. Temperature variation in the thermocycler block is not as crucial in the case of DAF as it is with conventional PCR. Short extension times are sufficient for complete extension of the short products typically obtained in DAF (Caetano-Anolles et al. 1991). DAF is suitable for DNA fingerprinting of different genotypes.

3.7 Arbitrary-Primed PCR

Welsh and McClelland (1990) reported the procedure for arbitrary-primed polymerase chain reaction (AP-PCR). In *arbitrary-primed PCR*, arbitrary sequence primers of 18–32 nt are used for amplification. It is not likely that even a very large genome will have sequences complementary for an arbitrary sequence of 20 bases or more. Therefore, amplification can occur only when the annealing conditions allow primer–template pairing with mismatches at some base pairs. The first two cycles of PCR are carried out at low stringency, and during the subsequent PCR cycles, a higher stringency (achieved by increased annealing temperature) is used. In this way, up to 100 bands may be generated for each individual, which are separated by PAGE, and scored as "present"/"absent." The approach is suitable for DNA fingerprinting. Many workers consider AP-PCR to be essentially the same as RAPD, but the two procedures differ in terms of primer length, annealing conditions, number of amplified fragments, and the type of gel used for electrophoresis (de Vienne et al. 2003). This technique has now been refined to permit fragment separation by agarose gel electrophoresis. But AP-PCR is not a popular method as it involves autoradiography.

3.8 Sequence-Characterized Amplified Regions

In 1993, Paran and Michelmore developed the sequence-characterized amplified regions (SCAR) markers from selected desirable RAPD markers. However, this term is often applied for PCR-based markers derived from AFLP and other markers as well. The amplified fragment representing a desirable RAPD marker is eluted from the gel, cloned, and the nucleotide sequences of its two termini are determined. A pair of primers (usually, 20–24 not long), one forward and one reverse primer, specific for the two terminal sequences is designed. This primer pair is expected to amplify a single fragment and detect the polymorphism represented by the concerned RAPD marker in a more reliable manner. The primer pairs designed in this manner are tested for their ability to detect the concerned polymorphisms, and the successful primer pairs give rise to SCAR markers. SCAR polymorphisms are generally dominant (scored as "presence" or "absence" of a single unique band), particularly at elevated annealing temperatures (Paran and Michelmore 1993). These markers can be developed into plus/minus arrays to eliminate the need for electrophoresis. Some of the SCAR markers detect length polymorphism either directly or after digestion of the amplified fragment with a suitable restriction enzyme; the latter approach generates a marker system called cleaved amplified polymorphic sequences (CAPSs; Sect. 3.14). However, sometimes the SCAR primers fail to detect any polymorphism. In such cases, it becomes necessary to sequence both the alleles of the RAPD fragment and design the two primers based on sequence differences to ensure detection of the polymorphism (Vosman 1998). Thus, SCARs are essentially similar to STS in construction and application. They can be used for physical as well as genetic mapping, comparative mapping, and phylogenetic relationship studies.

3.9 Amplified Fragment Length Polymorphisms

Amplified fragment length polymorphism technology was developed by Zabeau and Vos (1993), and it uses restriction fragments for PCR amplification. It ingeniously combines the restriction digestion of sample DNA step of RFLP system with the PCR technique to generate a robust and highly polymorphic DNA marker system (Fig. 3.5). In the AFLP procedure, 100–500 ng genomic DNA is digested with two restriction enzymes, appropriate adapters are ligated at the ends of the resulting restriction fragments, and a much smaller set of these fragments is selectively amplified by the PCR. Strictly speaking, this marker system does not detect the fragment length polymorphism

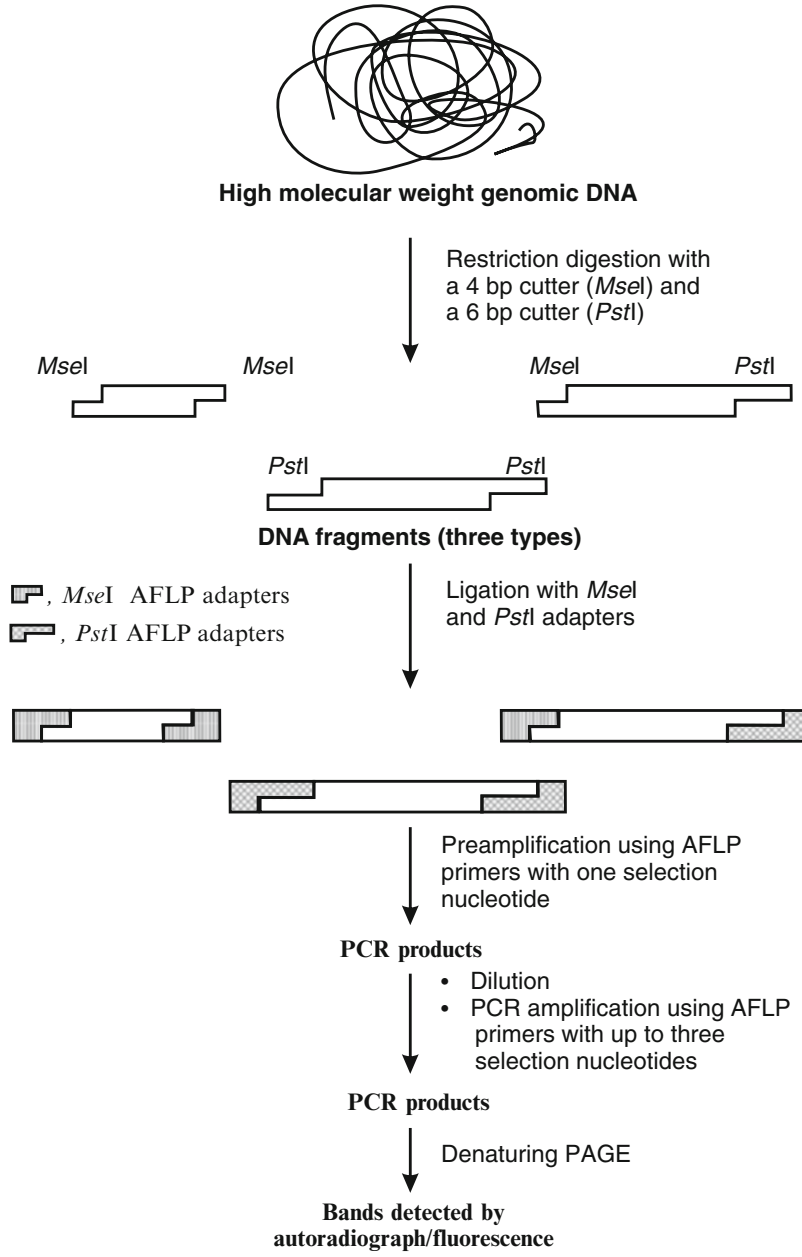


Fig. 3.5 A simplified schematic representation of the two-step AFLP method. Dilution after the preamplification step virtually removes the unamplified fragments. In the amplification step, the AFLP primer for the 6 bp cutter is labeled with radioactivity or, preferably, fluorescence (Based on Vos et al. 1995; de Vienne et al. 2003)

generated by the restriction enzymes. The restriction enzymes, in essence, only produce the set of restriction fragments from the genomic DNAs in a highly reproducible manner and also provide a dependable strategy for fragment amplification coupled with complexity reduction. The polymorphism detected by the AFLP procedure is actually generated by the selection nucleotides used in the AFLP primers. A restriction fragment will be amplified only when it has the complementary bases for the selection nucleotides in appropriate positions. On the other hand, a homologous fragment with mismatch at the selection nucleotide sites will not be amplified.

Thus, the polymorphism is generated primarily by differential amplification of the restriction fragments. Therefore, some authors prefer to call this marker system *selective restriction fragment amplification markers*, but *restriction fragment amplification polymorphism* seems to be a better term. A denaturing polyacrylamide gel is used to separate the PCR products, and up to 50–100 bands per sample are obtained. Of these, about 80 % of the bands may be polymorphic and can be used as markers. Therefore, AFLP is regarded as one of the most powerful high-density marker systems that produces ten times more informative markers per analysis than other marker systems and has high reproducibility. Further, prior sequence information is not required for this marker system.

3.9.1 The Procedure of AFLP

In the first step of AFLP procedure, sample genomic DNA is digested with two restriction enzymes (Fig. 3.5). One of these enzymes is a rare cutter, e.g., *Pst*I (6 bp recognition sequence, 5′-CTGCA/G); this enzyme does not cut methylated DNA, as result of which it creates a bias in favor of low-copy number fragments. The second enzyme is a frequent cutter, e.g., *Mse*I (4 bp recognition sequence, 5′-T/TAA); it is used to produce much smaller (256 bp $= 4^4$ bp) fragments from those generated by the first enzyme. This digestion procedure produces the following three types of fragments: (1) Type I fragments have both their ends generated by the rare cutter *Pst*I (*Pst*I-*Pst*I) and form a small fraction of the total fragments. (2) Type II fragments (*Pst*I-*Mse*I) have one end produced by the rare cutter (*Pst*I) and the other end generated by the frequent cutter (*Mse*I). (3) Type III fragments (*Mse*I–*Mse*I), on the other hand, have both their ends generated by the frequent cutter (*Mse*I) and are expected to be the most frequent; they are selectively eliminated by the following PCR procedure.

After ligation of adapters to the DNA fragments, their PCR amplification is done in two steps. In the first step, called *preamplification step*, the samples are amplified using two AFLP primers, each of which has one selection nucleotide each at its 3′ end (Fig. 3.5). An *AFLP primer* has the adapter sequence plus one to three arbitrary nucleotides at its 3′ end, and the arbitrary nucleotides are called *selection nucleotides*. The inclusion of selection nucleotides reduces the number of fragments that would be amplified by the AFLP primers. For each selection nucleotide added to an AFLP primer, the proportion of amplified fragments would be reduced to $1/16\left(= 1/4 \times 1/4\right)$ of the number of different fragments present in the mixture. In this way, 1/16th of all the three types of fragments present in the mixture will be amplified. The products of the preamplification step are suitably diluted to minimize the fragments that were not amplified in this step. The diluted mixture of the fragments is then used as template for the amplification step, in which each of the two AFLP primers has up to three selection nucleotides at its 3′ end. In addition, the AFLP primer corresponding to the ends produced by the 6 bp cutter is labeled with radioactivity or a fluorophore. The AFLP primers and the amplification conditions are so designed that they favor amplification of the type II (*Pst*I-*Mse*I) fragments. Denaturing PAGE is used to separate the PCR products, and the bands are detected by either autoradiography or, preferably, fluorescence (Vos et al. 1995). The use of fluorescence-tagged primers permits the analysis of fragments by an automated DNA sequencer, which also enables automated data collection and analysis.

3.9.2 Features of AFLP

The observed AFLP polymorphisms may result from mutations either in the recognition sequences of the two restriction enzymes used for digestion of the genomic DNA or in the sequences complementary to the selection nucleotides included in the AFLP primers. In addition, insertions within or deletions from the amplified restriction fragments will also generate polymorphism. AFLP fragments/bands are of random origin, but most of them represent unique

sequences. They are dominant markers, but it is possible to differentiate heterozygous and homozygous genotypes on the basis of intensity of the bands (Staub et al. 1996). The AFLP technique is faster and less labor intensive, and detects a large number of loci that provide far greater information than RFLP procedure. Further, AFLPs are highly reproducible, which is a great advantage over RAPDs. This marker system does not require sequence information, there is no marker development step, and it can be used in any species, including nonmodel organisms. But the AFLP marker system is laborious, time-consuming, technically demanding, and expensive to set up, and it uses restriction enzymes. It requires DNA preparations of high purity (necessary for restriction digestion), the polymorphic information content of the marker system is low (the maximum being 0.5), and in some plant species like sunflower and barley, the AFLP markers tend to cluster in the centromeric regions. AFLP markers can be used for variety/line identification, characterization of germplasm, high-resolution mapping, marker-assisted selection (MAS), and gene cloning. It is still used for genetic studies in crops species, for which little or no reference genome sequence is available. In addition, it can be used for fingerprinting of DNA clones and for identification of contigs (Vos et al. 1995).

3.9.3 Modifications of the AFLP Technique

The AFLP technique has been modified in various ways to achieve specific objectives. One modification of the AFLP procedure, called *sequence-specific amplification polymorphism* (*S-SAP*), generates a marker system that is similar to, but more polymorphic than, AFLPs. In S-SAP, the restriction fragments are generated and ligated to the AFLP adapters as usual. But in the amplification step, only one AFLP primer is used, and the other primer is based on a conserved sequence of a transposable element (TE). TEs occur in very high copy number in plant genomes, and sometimes they may be more frequent in the gene-rich regions. The use of TE-based primers amplifies only those DNA fragments that have the TE sequence. The *transposon display* procedure of van den Broeck et al. (1998) is essentially the same as S-SAP, except that it deliberately uses a hexa-cutter restriction enzyme that cuts within the chosen TE. S-SAP has been used for genetic diversity studies and linkage map construction in several species, including pea, wheat, and cashew.

In another modification, called *sequence-tagged microsatellite profiling* (*STMP*), one AFLP primer and one primer based on a SSR sequence (anchored at its 3′ end) are used for amplification of the restriction fragments after the preamplification step. This modification takes advantage of the SSR polymorphism without prior sequence knowledge and the efforts required for SSR marker development. STMP markers can later be converted to SSR markers. Another modification of the AFLP technique is called *TE-AFLP* (*three-endonuclease AFLP*) since three restriction enzymes are used to digest the sample DNA. In addition, two sets of adapters are used for amplification of the fragments. The use of third endonuclease increases the discriminatory power of the technique, and a one step amplification procedure can be used for fingerprinting of even complex genomes. The *MEGA-AFLP* (*multiplex-endonuclease genotyping approach AFLP*) is based on four or more endonucleases used for digestion of the sample DNA. However, this modification employs only a single pair of adapters for PCR amplification.

The AFLP approach has been adapted for marker genotyping by microarray hybridization as DArT (diversity array technology; Sect. 2.6) or as CRoPS (complexity reduction of polymorphic sequences; Sect. 13.4.2) for SNP (single nucleotide polymorphism) and InDel (insertion/deletion) discovery and genotyping using a new-generation DNA sequencing technology. These modifications are amenable to high-throughput marker genotyping as well as automated data acquisition and analysis.

3.9.4 Conversion of AFLP Markers

An AFLP marker of interest can be converted into a STS marker in the same way as SCAR markers are derived from RAPD markers. DNA from the AFLP band of interest is isolated, reamplified using the same AFLP primers that were used in the amplification step, and the amplification products are sequenced either directly or after cloning. Based on this sequence information, a pair of specific PCR primers is designed for amplification of the concerned DNA fragment. This strategy can generate CAPS, dCAPS, or STS markers.

3.10 Sequence-Tagged Sites

A locus that can be unambiguously defined in terms of flanking primer sequences that are used for its amplification is called *sequence-tagged site* (*STS*; Olson et al. 1989). The pair of primers for an STS locus, typically, amplifies a single band. STSs can be created in the following four ways:

1. The two ends of a RAPD fragment are sequenced, and, based on this information, a pair of PCR primers is designed for reproducible-specific amplification of the intervening segment; this strategy generates SCAR markers.
2. The two ends of an RFLP or AFLP fragment are sequenced, and specific primers are designed for amplification of the RFLP/AFLP locus.
3. STSs are often created by determining the unique sequences flanking mini- and microsatellite sites. A pair of primers specific for these unique sequences is designed for PCR amplification of each of these sites.
4. Sequences of ~400 bp long fragments of genomic DNA are determined, and primers of about 20 bp may be designed for amplification of about 200–400 bp segments. These primers are tested for PCR amplification using the genomic DNA as template. If a pair of primers amplifies a single product of the correct size, a unique STS has been identified. In human genome project, about 50 % of the primers created in this way identified unique STSs, which have been useful in creation of contigs required for physical mapping.

Thus, the creation of STS markers requires considerable amount of work, but their application requires merely the knowledge of sequences of the concerned primer pairs.

3.11 Microsatellites or Simple Sequence Repeats

Litt and Luty (1989) introduced the term *microsatellite* to describe the simple sequence fragments generated by PCR. Microsatellite sequences are also known as *short tandem repeats* (*STRs*), *simple sequence repeats* (*SSRs*), or *simple sequence length polymorphism* (*SSLP*). SSRs consist of tandemly repeated sequences of 1–6 bp, of which the dinucleotide repeats $(CA)_n$, $(GA)_n$, and $(AT)_n$ are the most frequent and highly polymorphic in eukaryotic genomes. In case of plants, $(AT)_n$ and $(GA)_n$ repeats appear to be more numerous, while $(CA)_n$ repeats constitute one of the most abundant microsatellites in mammals. (The value of n may range from 5 to 50 or even more.) Plant genomes also contain trinucleotide and tetranucleotide repeats, and the $(AAG)_n$ and $(AAT)_n$ sequences appear to be the most frequent. The average distance between two loci of a given dinucleotide SSR has been estimated as 30–100 kb. The trinucleotide and tetranucleotide SSR sequences are estimated to show similar distribution patterns. It appears that many microsatellites are uniformly distributed throughout the genome, but in some species like tomato, the SSRs may be clustered around centromeres (see Gupta and Varshney 2000).

Microsatellites differ from minisatellites (Sect. 2.7) in terms of the length of the repeating unit (11–60 bp for minisatellites) as well as the pattern of their distribution in the genome. Microsatellite sequences are almost evenly distributed in the plant genome, while minisatellites are generally confined to the

telomeres of eukaryotic chromosomes (Tautz 1989; Weber and May 1989). Microsatellite sequences are believed to have originated from unique sequences by random base substitutions and/or insertions that generated repeat motifs. Once produced, the repeat sequences expanded most likely due to slippage by DNA polymerase during replication and/or unequal crossing over. Consequently, microsatellite sequences are often highly polymorphic and SSR loci show multiple alleles. For example, in the elite germplasm of soybean, usually, only two alleles per RFLP locus are detected, while in a sample of about 100 elite soybean genotypes some microsatellite loci had up to 26 alleles. It may be reiterated that polymorphism at SSR loci is exclusively due to variation in the number of repeat units and base sequence variation is not involved. SSRs have been exploited to develop the following two types of markers: (1) sequence-tagged microsatellite site (STMS) or, simply, SSR markers, and (2) inter-simple sequence repeat (ISSR) markers.

3.12 Simple Sequence Repeat Markers

The *simple sequence repeat* (*SSR*) *markers* are a special version of STS markers, in which a microsatellite locus is amplified using a specific primer pair derived from the unique sequences flanking the SSR locus (Fig. 3.6). Sometimes, these markers are called STMS markers, simple sequence length polymorphisms (SSLPs), and even as microsatellite markers. Each SSR locus is amplified using a specific pair of primers, and the amplification products are analyzed by gel electrophoresis for the identification of different alleles of the locus. Ordinarily, a single SSR locus is amplified from a single DNA sample in each PCR reaction, and the PCR products from a single reaction are analyzed in one gel lane. The unique sequences flanking the SSR loci seem to be conserved within species and even across species within a given genus, but rarely across related genera. Therefore, SSR primers designed on the basis of genome sequence information from one species can be used in a related species as well.

3.12.1 Discovery of SSR Markers

Several innovative approaches have been used for the discovery of SSR loci. Initially, DNA inserts/restriction fragments containing microsatellite motifs may be identified from a genomic library/genomic DNA restriction digest. The genomic library used for this purpose may or may not be enriched for DNA inserts with microsatellites. The identified clones/restriction fragments are sequenced. But when genome sequence data are available, SSR loci can be identified more efficiently by analysis of the genome sequence and expressed sequence tag (EST) databases using data mining software like FASTA. The SSR markers derived from genome sequences are sometimes termed as *genomic SSRs* (*gSSRs*), while those developed from ESTs are often referred to as *expressed SSRs* (*eSSRs*). For example, one eSSR appears to be present in every 5.46 kb of wheat EST sequence. In addition, SSR markers are also derived from unigene sequences available at http://www.ncbi.nlm.nih.gov/unigene/; such markers are often called *unigene-derived microsatellites* (*UGMs*). *Unigenes* (unique gene sequences) are a set of nonredundant EST sequences from a given species so that each unigene sequence has a unique identity and genomic location. In each of the above cases, primers specific for the unique sequences flanking the SSR sequences are designed, generally, with the help of a suitable computer program. Care should be taken with respect to the following in designing of the primers: (1) GC content of the primers should be around 50 % (T_m about 60 °C), (2) their 3′-ends should be AT-rich, and (3) the frequency of primer dimer formation should be as low as possible.

3.12.2 Increasing the Throughput of SSR Markers

The cost of SSR analysis can be reduced by the following strategies: (1) pooling the PCR products from two or more separate single primer pair-based reactions and running them in a single

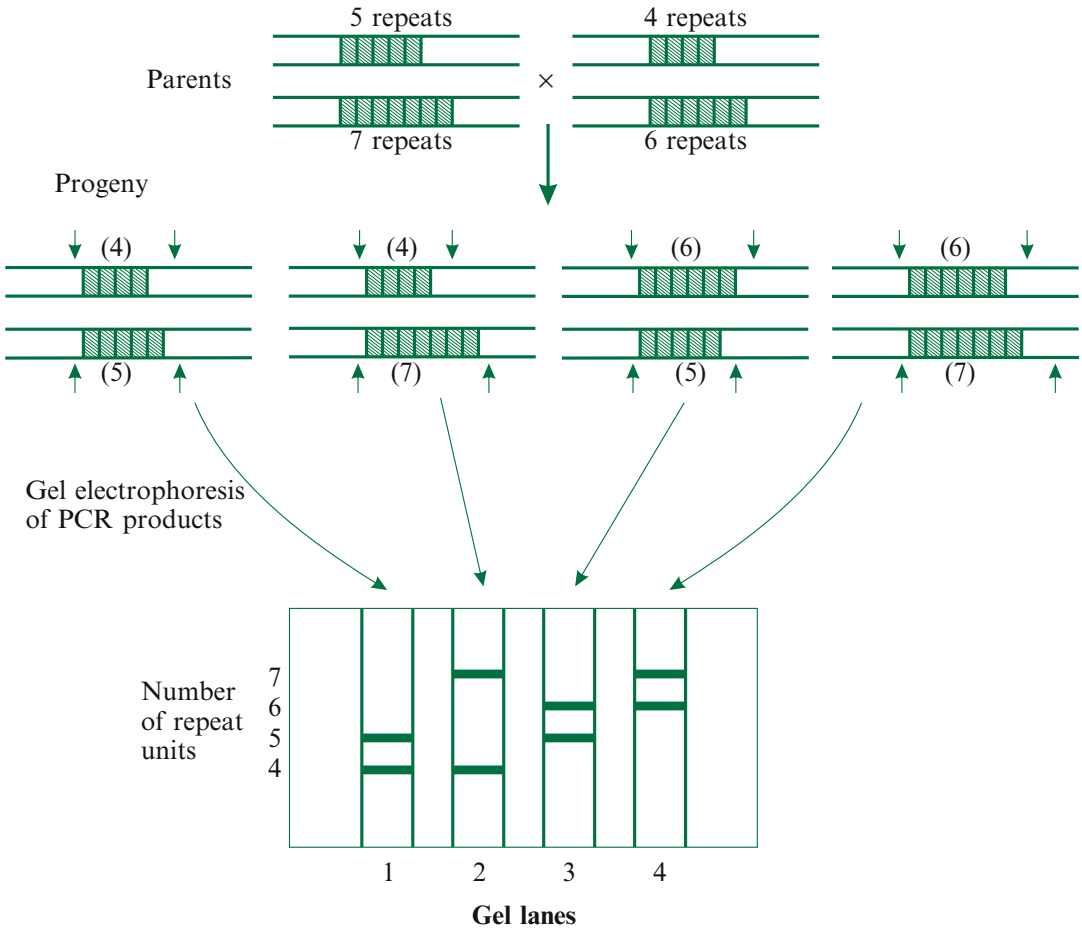


Fig. 3.6 The microsatellite (SSR) marker system: the SSR alleles result from a variation in the number of repeat units. The *arrows* indicate the sites of primer binding for PCR amplification of the SSR locus. The primers are based on unique sequences flanking the SSR locus

gel lane, (2) using a single PCR reaction tube for simultaneous amplification of two or more SSR loci, or (3) combining the above two approaches. When primer pairs for alleles at two or more SSR loci generate amplification products of different sizes to enable their unambiguous identification, their PCR products can be pooled and used for electrophoresis. If the primers for such SSR loci could be optimized for the same PCR amplification conditions, they can be used together in a single PCR reaction tube for amplification, and the PCR products analyzed in a single gel lane. This strategy, called *multiplex PCR* (Mitchell et al. 1997), leads to a significant reduction in the costs and the time needed for assays. But when the PCR products from different SSR loci have overlapping range of lengths, they can still be analyzed in a single gel lane by the following procedure. The PCR products from one reaction are loaded in the gel and allowed to run for a suitable period of time. The run is then interrupted, and products of the second PCR reaction are loaded in the gel and the run is resumed. The staggered loading of the PCR products from different reactions would allow the resolution of PCR products of similar lengths (Ribaut et al. 1997).

The PCR primers for different SSR loci can be labeled with different fluorescent labels. These primers can be used in a single PCR reaction

when they are optimized for the same PCR conditions. Otherwise, a different PCR reaction would be set up for each primer pair. In either case, the PCR products from three to five different SSR loci can be analyzed in a single capillary of an automated DNA sequencer even when the products from different loci are overlapping. It is possible to use a single capillary or gel lane for the analysis of up to 16 different SSR loci by taking advantage of both differential fluorescence labeling and differences in the lengths of PCR products (Gupta and Varshney 2000; de Vienne et al. 2003). Electronic data collection with automated DNA sequencers and data analysis using software like Genescan™ or Genotyper™ allows reliable fragment size determination and identification of SSR alleles; it also enables separation of native SSR alleles from the products of slippage during PCR amplification. But fluorescent labeling of primers is expensive and increases the cost of assays. The labeling cost can be reduced by using a universal fluorescence-labeled primer like M13(−21) in combination with the normal specific forward and reverse primers. However, the specific forward primer used in this reaction contains the M13(−21) sequence (without label) added to its 5′ end. The use of the above set of three primers for amplification labels the PCR product because the labeled M13(−21) primer will be used as primer in the second and subsequent PCR cycles. The cost of assay is reduced because the labeled universal primer is much less expensive than the labeled specific primers. Another approach for reducing the cost is the use of an array tape, in the place of microtiter plate, to drastically reduce the amounts of reagents, consumables, etc. used (Sect. 13.2.7).

3.12.3 Merits of SSR Markers

SSR markers are codominant, highly polymorphic, distributed throughout the genome in most of the cases, and exhibit simple Mendelian inheritance. SSR assay is simple, PCR-based, locus-specific, highly reproducible, amenable to automation, and has medium throughput. The amount of DNA needed per individual is small (~100 ng), the cost of assay system is low, and the assay can be handled manually. SSR markers are often transferable across different species of the same genus and even across closely related genera (Choumane et al. 2000). *Transferability of SSR markers* means that the primers for SSR markers developed for one plant species can be successfully used in some other, usually, related plant species. For example, the primers designed for *Oryza sativa* were successfully used in wild *Oryza* species and vice versa (Panaud et al. 1996). These markers are highly informative and can distinguish even closely related individuals.

SSR markers have been developed in several crop species. They are widely used for linkage mapping, cultivar identification, germplasm characterization (detection of accession duplications, seed mixtures, outcrossing, and genetic drift), analysis of gene pool variation, and MAS (Powell et al. 1996). SSRs became the "marker of choice" and dominated plant molecular research during the last decade of twentieth century and the first decade of the present century. But their pristine position is under challenge from the more abundant and ultrahigh-throughput SNP markers.

3.12.4 Limitations of SSR Marker System

One of the chief limitations of SSR markers is that their development is technically quite complicated, labor intensive, and costly. This involves construction of a genomic library, preferably, enriched for microsatellite sequences, screening the library with SSR-specific probes, sequencing the positive clones, designing of specific primers, evaluation of the primers for locus-specific amplification, characterization of copy number of the detected polymorphism, and determination of the chromosomal position of each SSR locus. But once the primers for the SSR loci are developed, marker analysis becomes easy and relatively inexpensive (McGregor et al. 2000). SSR markers permit only limited multiplexing and automations and are not

abundant enough to saturate the desired genomic regions. In addition, the cost of automation is relatively high, and often difficulties are encountered in sharing SSR marker data between laboratories due to differences in relative allele sizes detected across different genotyping platforms. Another problem arises due to the presence of null alleles at a proportion of SSR loci (~25 % of the loci in humans). When the specific primers for a SSR locus consistently fail to amplify a detectable product in some individuals, these individuals are said to have the *null allele* of the concerned locus. Null alleles are believed to be generated by mutation in the binding site for one or both of the primers, leading to a failure of amplification. The presence of a null allele at a locus will lead to an underestimation of heterozygosity at that locus (Gupta and Varshney 2000).

3.13 Inter-Simple Sequence Repeats

An *inter-simple sequence repeat* (*ISSR*) or *inter-SSR PCR* marker is based on a single primer having microsatellite sequence. The ISSR primers amplify the genomic regions flanked by the SSR sequences making up the concerned primers. The primer may consist solely of a microsatellite sequence (*non-anchored primers*) or, more often, a microsatellite sequence plus a short (usually, two nucleotides long) arbitrary sequence either at the 3′ or the 5′ end of the primer (*anchored primers*). In all these cases, amplification will occur only of such a genomic region that is flanked by the SSR sequence used as primer, and the SSR sequences flanking this region are in reverse orientation. These markers detect variation in the size of the genomic region between the two adjacent microsatellite sequences used as the primer binding sites.

The markers generated by non-anchored primers are called *single primer amplification reactions* (*SPARs*) or *microsatellite-primed PCR* (*MP-PCR*). These markers are useful only when the primers consist of tri-, tetra-, and penta-nucleotide repeats because primers containing dinucleotide repeats generally yield a smear. MP-PCR appears to offer little advantage over RAPD analysis. Further, fragments of different lengths may be obtained from the same ISSR region as a result of the primer annealing at different positions within the SSR repeats (Caldeira et al. 2002). The use of anchored primers gets around this problem, and it substantially reduces the number of ISSR fragments amplified.

The markers generated by anchored primers have been called *inter-SSR PCR*, *anchored simple sequence repeats* (*ASSRs*), *anchored microsatellite-primed PCR* (*AMP-PCR*), or *inter-SSR amplification* (*ISA*). The region amplified by such primers depends on the anchor position in the primer. If the anchor were attached to the 5′ end, the amplified fragment would include the full lengths of the two microsatellite sequences as well as the inter-SSR region. But if the anchors were linked to the 3′ end of the primer, only the region between of two SSRs, including the primer, will be amplified (Fig. 3.7).

3.13.1 Modifications of ISSR

The ISSR procedure has been modified in several different ways for achieving the desired objectives. In one modification, a 5′ anchored SSR primer can be used in combination with a RAPD primer to yield markers termed as *randomly amplified microsatellite polymorphisms*

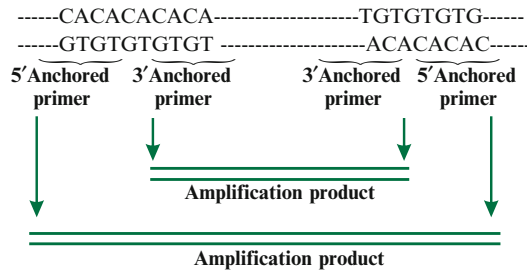


Fig. 3.7 The products generated by 3′ and 5′ anchored SSR primers. The anchored primers are about 17–32 nucleotides long and have usually two arbitrary bases at their 3′ (3′ anchored primers) or 5′ ends (5′ anchored primers) (Based on de Vienne et al. 2003)

(*RAMP* or *RAMPO*; Wu et al. 1994). These markers detect variation in the lengths of the target microsatellite as well as the region between the binding sites of the two primers. The RAPD primer binding site serves as an arbitrary endpoint for the anchored SSR primer-based amplification product. Therefore, the amplification products in RAMPs are greater in number than in the case of AMP-PCR. Since RAPD primers would have melting temperatures (T_m) ~10–15 °C lower than those of anchored SSR primers, the PCR program is so modified that the annealing temperature alternates between high and low (suited for the ISSR and RAPD primers, respectively) during the successive cycles. This approach has been used for genetic diversity studies in some plant species like barley. The amplification products may be digested with a restriction enzyme to yield *digested RAMPs* (*dRAMPs*) markers that are useful for mapping of genes/QTLs (Becker and Heun 1995).

In another modification, called *selective amplification of microsatellite polymorphic loci* (*SAMPL*), microsatellite-based primers are combined with the AFLP primers in the AFLP procedure to yield markers that are regarded as an improvement over SSRs. In case of SAMPL, one AFLP primer with three selective nucleotides and one SAMPL primer are used in combination for the amplification step (after the preamplification step) in the AFLP procedure. The best results are obtained when the SAMPL primer (18–20 nt long) is based on two different adjacent SSRs and the sequence lying between them; such sequences are known to occur in compound repeats. SAMPL primers consisting of a single SSR sequence, in contrast, generate ambiguous and less reproducible results. SAMPL bands generate dominant markers, but some of the markers may be codominant (Witsenboer et al. 1997).

Hybridization with a labeled SSR probe, e.g., $(CA)_8$, $(GA)_8$, $(GTG)_5$, $(GCGA)_4$, may be used to detect polymorphism in the amplification products obtained by using a regular RAPD primer or a 10/15 nt long non-anchored SSR primer. This method has high sensitivity at the intraspecific level, but it uses radioactivity. This marker is known as *RAMP*, *RAMPO*, *randomly amplified hybridization microsatellites* (*RAHM*), or *randomly amplified microsatellites* (*RAM*). Another marker, termed as *retroposon-microsatellite amplified polymorphism* (*REMAP*), uses a 3′ anchored microsatellite primer along with a primer based on the LTR (long terminal repeat) of a retrotransposon for PCR amplification. Consequently, REMAP can amplify three different types of DNA fragments: (1) the segments flanked by an LTR at one end and a microsatellite locus on the other, (2) sequences having a microsatellite locus at both their ends, and (3) fragments present between two neighboring insertion sites of the concerned retrotransposon.

3.13.2 Merits and Limitations of ISSR Markers

ISSR markers are more reproducible than RAPD, easy to use, cheap, have high throughput, and yield multiple polymorphic loci. Further, a prior knowledge of the template DNA sequence is not required. Generally, ISSR markers are dominant, but the use of a larger 5′ anchor can yield codominant ISSR markers. A major disadvantage of ISSR markers is that they are not highly reproducible, and some primers generate poorly reproducible band patterns.

3.14 Cleaved Amplified Polymorphic Sequences

The *cleaved amplified polymorphic sequences* (*CAPSs*) detect length polymorphism generated by restriction digestion of specifically amplified PCR products from different genotypes. They are often called *PCR-RFLP* since they were developed for easy genotyping of RFLP markers using gel electrophoresis, following PCR of the target regions (Williams et al. 1991; Fig. 3.8). Therefore, CAPSs are codominant markers. They result from alterations in the recognition sites, located within the amplification products, for the respective restriction enzymes. The restriction enzymes used for CAPS analyses should

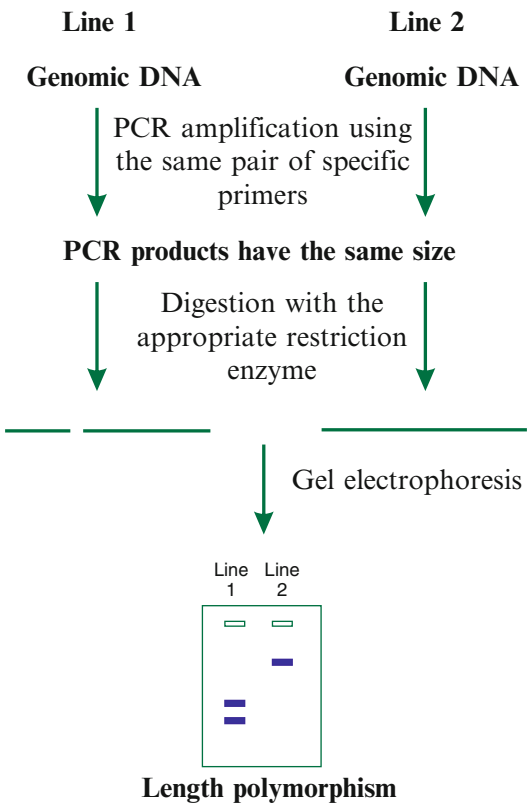


Fig. 3.8 A schematic representation of cleaved amplified polymorphic sequence (CAPS) marker system

have 4 bp recognition sequences since they are much more likely to have recognition sites within the amplification products of ~0.5–2 kb. This technique is useful when the amplified DNA fragments are large, fail to reveal polymorphism among genotypes, and contain a SNP within the recognition site for a restriction enzyme. The CAPS approach is preferable to the standard RFLP analyses. But the use of restriction enzymes for CAPS analysis adds to the assay cost and makes this marker system unsuitable for high-throughput analysis and automation.

In a variation of the CAPS method, called *derived cleaved amplified polymorphic sequence* (*dCAPS*) or *mismatch PCR-RFLP*, one of the PCR primers generates in the PCR product a recognition site for a restriction enzyme. This primer is so designed that it contains one of the SNP alleles and one or more mismatches with the target template DNA sequence. These mismatches together with the SNP allele generate a restriction site in the PCR product of this allele, but not in that of the other allele. The concerned restriction enzyme is used for digestion of the PCR products, and the SNP alleles are deduced from the restriction fragments generated from them (Michaels and Amasino 1998). The dCAPS method is simple, relatively inexpensive, and would be useful for scoring known SNP alleles and for positional cloning of new plant genes.

3.15 Single-Strand Conformation Profile/Polymorphism

Single-strand conformation profile/polymorphism (*SSCP*) is detected as differential movement of single-stranded DNA molecules, representing identical genomic regions from different individuals of a species (Orita et al. 1989). The DNA fragments used for SSCP analyses are generally obtained by PCR amplification. The differential migration of the single-stranded fragments results from differences in their secondary structures. The secondary structures of the single strands result from folding and internal complementary base pairing in short regions. The base pairing produces short double-stranded regions that stabilize the folding pattern as well as contribute to the secondary structures. The internal base pairing would depend on the base sequence. Therefore, the differences in conformations of the single-stranded molecules would reflect the differences in their base sequences.

The detection of SSCP involves heating the solution of a double-stranded DNA molecule to 95 °C so that the two strands of the DNA molecules become separated. This denatured DNA solution is now quenched, i.e., cooled very rapidly. As a result, the complementary strands do not get sufficient time to pair with each other. Instead, the single strands fold onto themselves, and internal base pairing in short regions leads to the formation of characteristic secondary structures (Fig. 3.9). The differences in secondary structures of the single strands are

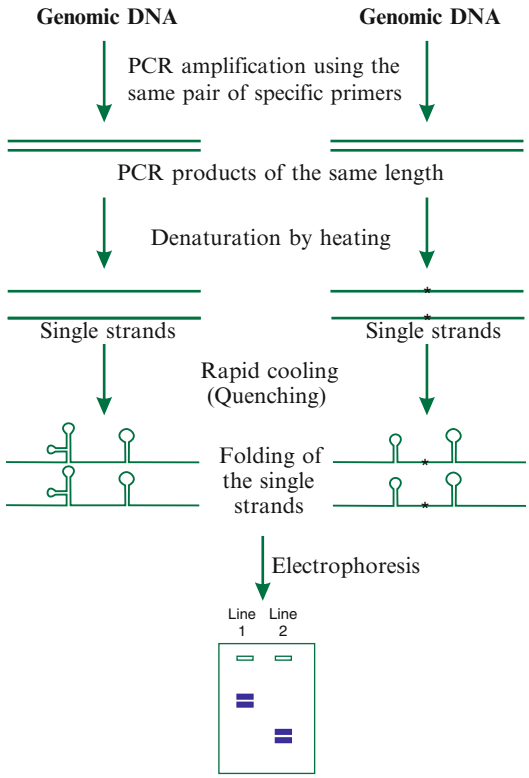


Fig. 3.9 A schematic representation of single-strand conformation polymorphism (SSCP) for discrimination between PCR products of identical lengths from the same genomic region of two lines differing for a mutation in this region (Based on de Vienne et al. 2003). * Mutation

detected by acrylamide gel electrophoresis under non-denaturing conditions. Fluorescence-labeled primers may be used for amplification of the target sequence to facilitate detection of the bands after electrophoresis. It has been estimated that in DNA molecules of up to 200 bp, 100 % of the differences in base sequence are revealed by SSCP. However, as the length of DNA duplex increases, the percentage of sequence differences detected by SSCP decreases.

The two strands of a DNA duplex usually generate slightly different secondary structures. Therefore, two bands will be observed in homozygotes (Fig. 3.8), and the heterozygotes would exhibit four bands. But each single strand of some DNA molecules can form more than one slightly different semi-stable conformation leading to the formation of multiple bands in homozygotes. SSCP procedure is useful for rapid screening of sequence differences among amplification products, when precise information about the sequence differences is not needed. This procedure is simpler and more convenient than CAPS, which requires restriction digestion of the PCR product, and D/TGGE (denaturing/temperature gradient gel electrophoresis), where a precise control of the electrophoresis conditions is necessary. SSCP has been used for mapping and genetic studies in plants only to a limited extent (de Vienne et al. 2003). The major disadvantages of SSCP are labor-intensive and costly marker development and the lack of automation.

3.16 Denaturing/Temperature Gradient Gel Electrophoresis

Denaturing/temperature gradient gel electrophoresis (*D/TGGE*) reveals differences in the movement of double-stranded DNA molecules from the same genomic regions of different individuals of a species. These DNA molecules are obtained by PCR amplification. Short stretches within a DNA duplex would differ from each other in terms of melting temperature, which depends on their base composition. For example, AT-rich stretches would have lower melting temperatures than GC-rich regions. As a result, the two strands of a DNA duplex will begin to separate earlier in AT-rich stretches than in GC-rich stretches, when the DNA molecules are subjected to increasingly denaturing conditions, e.g., during denaturing/temperature gradient gel electrophoresis. This property is exploited for the detection of sequence differences among PCR products from different individuals of a species.

The PCR products from different individuals are loaded in separate wells of an acrylamide gel. Preparing the gel with a denaturing agent, e.g., urea and formamide, can create the denaturing conditions during electrophoresis; this agent is added in a gradient of increasing concentration

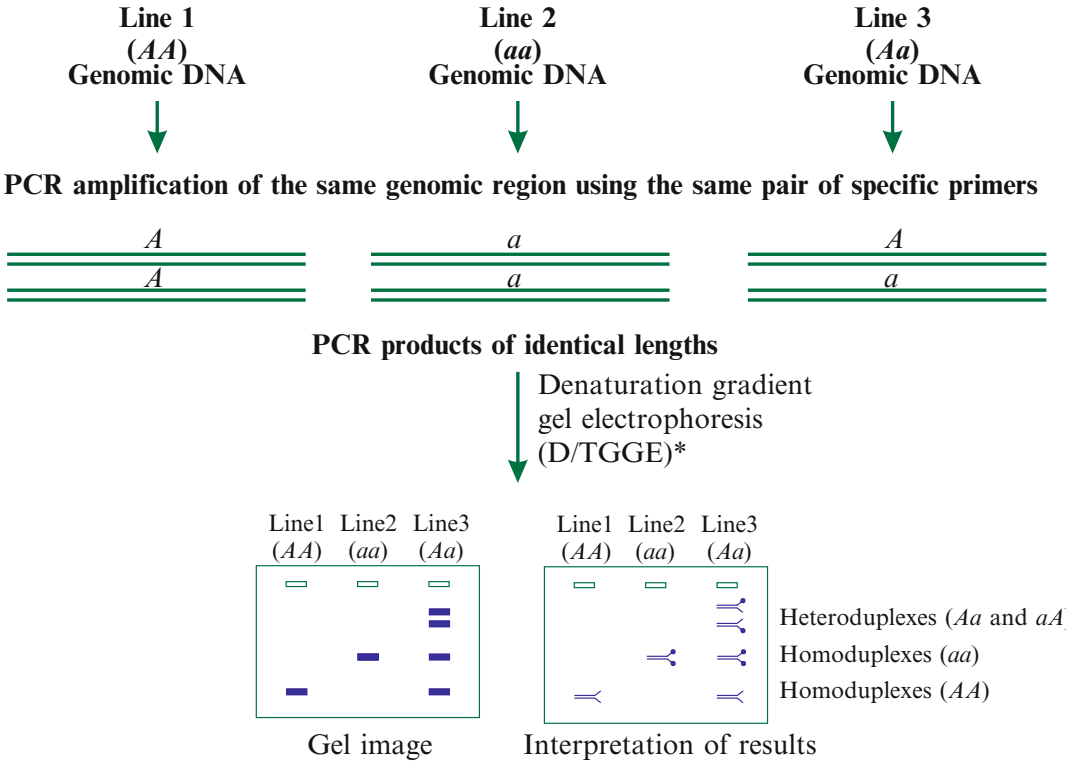


Fig. 3.10 A schematic representation of denaturing/temperature gradient gel electrophoresis (D/TGGE) to distinguish between PCR products of the same length but differing for a mutation. The *dots* at the ends of DNA strands from a line signify the presence of mutation. * the PCR products are denatured, followed by renaturation prior to D/TGGE in order to definitely heterozygotes due to the formation of heteroduplexes. PCR products migrate as duplexes in the gel till one of their ends melts to produce a branched structure and prevents further migration. Heteroduplexes form the branched structure earlier than the homoduplexes (Based on de Vienne et al. 2003)

starting from the loading wells. Alternatively, a normal acrylamide gel may be used and an increasing temperature along the gel during electrophoresis can create the denaturation gradient. The PCR products initially migrate in the gel as double-stranded molecules. As they migrate farther in the gel, they meet stronger denaturing conditions, and soon their least stable regions begin to melt. At some point in the gel, one end of the molecule would become single-stranded; this would produce a branched structure that does not migrate any further in the gel (Fig. 3.10). In most cases, a difference of even a single base pair in the least stable region of a DNA molecule of $<$300 bp would lead to a difference in the mobility of the molecule, and the variant molecule would form a different band in the gel (de Vienne et al. 2003).

D/TGGE permits identification of all heterozygous individuals by a simple step at the end of PCR. After the last PCR cycle, the denaturation step is implemented and is followed by renaturation of the PCR products; this would lead to the formation of two heteroduplexes in addition to the two homoduplexes in all the heterozygotes. The two heteroduplexes will be produced by association of each of the two strands of one allele with its complementary strand from the other allele. Heteroduplexes have considerably lower melting temperatures than the homoduplexes so that they do not migrate very far in the gel and form distinct slow moving bands. The heteroduplex bands are easily detectable even when the bands in the two homozygotes are not distinguishable (Fig. 3.9; de Vienne et al. 2003).

D/TGGE is a delicate technique and requires considerable care. The gradient must be chosen on the basis of stability features of the fragment to be analyzed, and the slope and the limits of the gradient must be carefully determined, preferably, by using suitable software. In any case, the preparation of gradient gels is time-consuming as well as prone to technical errors. In addition, the size of individual DNA fragments will determine the amount of denaturant to which they will be exposed. As a result, small DNA fragments would migrate to the bottom of the gel and might even be eluted from the gel, before they encounter sufficient amount of the denaturant for causing differences in mobility. These difficulties are overcome by temperature gradient gel electrophoresis. In case a PCR product does not have two distinct regions differing in stability, a GC clamp may be attached to the molecule. A *GC clamp* is a stretch of about 30 bp containing only GC bases. The GC clamp can be appended to the 5′ end of the PCR primer used for amplification of the fragments to be analyzed. If the base sequences of the variants of the concerned fragments were precisely known, their migration can be modeled to facilitate quick screening of the variants (de Vienne et al. 2003).

3.17 Sequence-Related Amplification Polymorphism

Sequence-related amplified polymorphism (*SRAP*) is one of several gene-targeted markers based on PCR amplification (Poczai et al. 2013); many of these markers are described in the following sections. *SRAP* is a simple marker based on open reading frame (ORF) amplification. SRAP uses two primers of 17 or 18 nt each, which have, beginning from their 3′ ends, three selective nucleotides, followed by a *core sequence* of 4 nucleotides (5′ CCGG 3′ in the forward primer and 5′ AATT 3′ in the reverse primer) and a 10 or 11 nt long arbitrary sequence (*filler sequence*) at the 5′ end (Fig. 3.11). It is important that different filler sequences are used for the forward and reverse primers. The CCGG core sequence is targeted at exons since exons

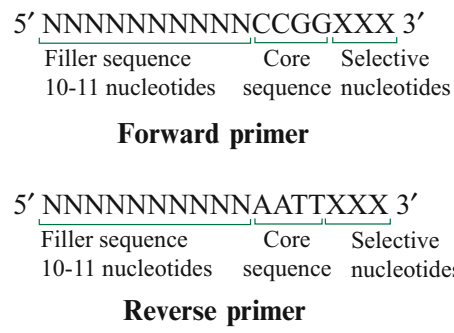


Fig. 3.11 The forward and reverse primers used for the detection of sequence-related amplification polymorphism (SRAP). Filler sequences of the two primers are arbitrary sequences, but different from each other. The sequence 5′ CCGG3′ targets exons, while the sequence 5′AATT3′ targets introns and promoter regions (Based on Li and Quiros 2001)

are more frequent in GC-rich regions. The AATT core, on the other hand, targets promoters and introns, which are normally AT-rich. The annealing temperature during the initial five PCR cycles is kept at 35 °C; it is set at 50 °C during the next 35 PCR cycles. Denaturing acrylamide gel electrophoresis is used to separate the PCR, and the bands are detected by autoradiography (Li and Quiros 2001).

In recombinant inbred line (RIL) and doubled-haploid (DH) populations of *Brassica oleracea*, SRAP markers were almost evenly distributed over the whole genome. Each primer combination generated many bands of which >10 were polymorphic. About 45 % of the bands represented already known genes that are listed in the GenBank, and 20 % of the bands showed codominance. SRAP method is simple and reliable, has moderate throughput, targets coding sequences, and generates a fair proportion of these markers behave as codominant (Li and Quiros 2001). The codominant markers will be generated by insertions and deletions since they would lead to polymorphism in the amplified fragment size. In contrast, SNPs affecting primer binding would generate dominant markers since they would either allow or prevent fragment amplification. This marker system has been used in several crops including potato, rice, lettuce, and garlic to achieve a variety of objectives, including linkage mapping, identification of

markers linked to useful genes, and genetic diversity analyses.

3.18 Target Region Amplification Polymorphism

The *target region amplification polymorphism* (*TRAP*) is a PCR-based marker system that involves in silico analysis of the EST database for designing of such primers that detect polymorphism around the desired candidate genes. TRAP uses two primers of 18 nt each; one of these primers is complementary to a sequence of the targeted EST (the *fixed primer*), while the other is an arbitrary primer (Fig. 3.12). The *arbitrary primer* has the same design as that of a SRAP primer (Sect. 3.17): the arbitrary primer may have an AT-rich core (5′ AATT 3′) and would anneal to an intron or a GC-rich core (5′ CCGG 3′) and would anneal to an exon. The fixed primer is designed as follows: EST database of the concerned species is searched, the desired EST is identified, and its sequence is used to design an 18 nt long primer with T_m of 50, 53, or 55 °C. The annealing temperature during the first five cycles of PCR is kept at 35 °C, but during the next 35 cycles, it is kept at 50 °C (Fig. 3.12).

In different plant species, the TRAP technique can generate up to 50 scorable markers of 50–900 bp from a single PCR reaction. The PCR products are resolved by electrophoresis using a 6.5 % polyacrylamide sequencing gel. These markers seem to be reproducible, and an automatic DNA sequencer in conjunction with fluorescent labels can be used for their detection (Hu and Vick 2003). TRAP system is better than SRAP as it yields markers around the target candidate genes, while the latter amplifies from all over the genome. The TRAP method has been used for germplasm characterization, fingerprinting of genotypes, and mapping of genes/QTLs (quantitative trait loci).

3.19 Transposable Element-Based Markers

Transposable elements (*TEs*) are DNA sequences that move around in the genome. They constitute >50 % of nuclear DNA and generate genetic diversity through insertion into functional genes, excision from various genomic sites, and generation of small structural rearrangements. TEs are classified into *Group I transposons* (*retrotransposons*) that transpose via RNA intermediates and *Group II transposons* that move as DNA molecules. Some

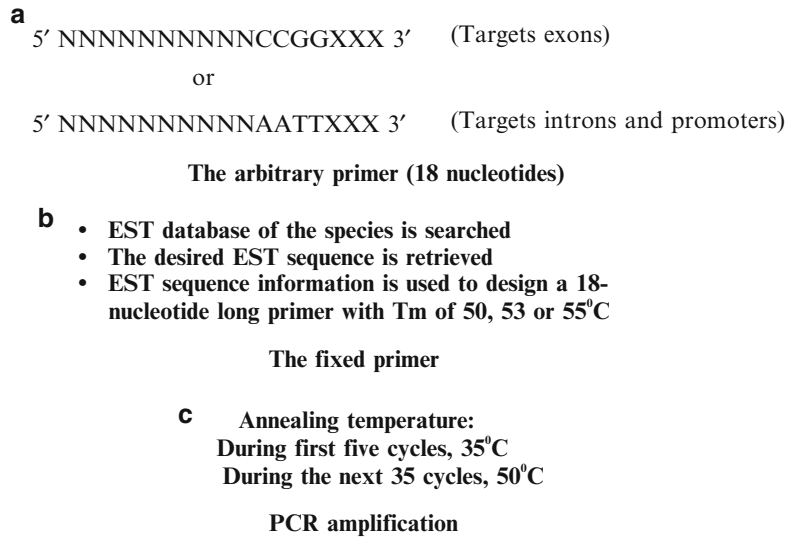


Fig. 3.12 A schematic representation of the (**a**) arbitrary and (**b**) fixed primers used for detection of the target region amplification polymorphism (TRAP) and (**c**) the significant features of the PCR amplification (Based on Hu and Vick 2003)

retrotransposons have long terminal repeats (LTRs), while others lack LTRs. Both these types of retrotransposons are present in plants usually in high copy numbers and are dispersed throughout the genome. There is a great variation in the number of copies and the sites of insertion in the genomes of even closely related species. Several marker systems are based on retrotransposons. Of these, the *sequence-specific amplification polymorphism* (*S-SAP*) seems to generate the largest number of highly polymorphic markers. S-SAP is an AFLP-like approach that displays as bands the regions between concerned retrotransposon insertion sites and the selected restriction sites (Sect. 3.9.3). In self-pollinated species like pea, S-SAP markers appear to be more informative than AFLP and RFLP markers (Ellis et al. 1998), and they have been used for phylogenetic analyses in pea.

Another approach uses primers based on LTRs of retrotransposons to amplify the region between two neighboring insertions of the element; this is called *inter-retrotransposon amplified polymorphism* (*IRAP*). The approach called *retrotransposon-microsatellite amplified polymorphism* (*REMAP*), on the other hand, uses one primer based on LTR of a retrotransposon and a second primer representing a microsatellite sequence that may be anchored. REMAP markers detect polymorphism in the genomic fragment flanked by the insertion site of a retrotransposon on one side and a SSR site on the other side. IRAP and REMAP markers are highly polymorphic, and up to 30 bands per individual may be obtained. These marker systems have been used for analysis of genetic relationships within species (Agarwal et al. 2008).

Some other transposable element-based markers are retrotransposon-based insertion polymorphism (RBIP), transposon display (TD), and inter-MITE polymorphism (IMP). The *RBIP* approach is designed to detect retrotransposon insertions at specific sites using PCR amplification (Agarwal et al. 2008). RBIP uses one primer derived from the concerned retrotransposon and a pair of primers derived from the sequences flanking this retrotransposon at the given insertion site. When the primer pair derived from the flanking sequences is used for amplification, a product would be obtained whenever there is no retrotransposon insertion in the region flanked by the primers. But when the primer based on the retrotransposon is used with a primer specific to one of the flanking regions, a PCR product would be generated only when the concerned region contains the retrotransposon. Polymorphisms can be readily detected by electrophoresis using an agarose gel. Alternatively, a simple dot blot assay using a reference PCR fragment for hybridization may be employed for analysis of the polymorphism. The dot blot assay is amenable to high-throughput automation. This method requires sequence information about the transposable element as well as the regions flanking the concerned insertion site, which involves considerable amount of work. It is perhaps the costliest and the most complicated method for detection of transposon insertions.

IMP markers are an example of markers derived from Group II transposons. The IMP technique is identical to IRAP, except for the use of primers based on MITE-like transposable elements in the place of those derived from retrotransposons. *MITEs* (*miniature inverted-repeat transposable elements*) are a family of small transposons, which are distributed widely and are plentiful in a number of plant genomes. They are often located in the terminal regions of genes and show considerable polymorphism among inbred lines. The *MITE-AFLP* method is similar to S-SAP as it uses one AFLP primer and one primer based on a MITE element for the amplification step of the AFLP procedure. The MITE-AFLP procedure has been used for studying genetic diversity and analyzing phylogenetic relationships in rice, wheat, and maize.

3.20 Conserved Orthologous Set of Markers

Conserved orthologous set (*COS*) *of genes* may be defined as a group of genes that show conservation of sequence as well copy number during the evolution of plant species. The *conserved orthologous set of markers* consists of gene-based markers derived from the conserved orthologous set of genes (Fulton et al. 2002).

The conserved set of genes is identified by computational analysis of genomic and EST sequences from a group of related species along with a well-characterized reference species like *Arabidopsis thaliana* (usually, for dicots) or rice (usually, for monocots). Each gene of the orthologous set has an orthologue in all the species of the group and often even in other distantly related species. Ordinarily, the genes included in the orthologous set are single-copy genes, but low-copy number genes may also be included. The COS gene-based markers are developed by designing a pair of specific primers for each gene set using the highly conserved sequences of exons. These primers may amplify an exonic region of the gene, but the amplified region may include at least one intron. A vast majority of the primer pairs successfully amplify genic regions, and ~90 % of the products show polymorphism. Usually, polymorphism is detected by SSCP, and the bulk (over 60 %) of polymorphisms are due to SNPs, while the rest are due to InDels.

Fulton et al. (2002) analyzed a large database of ESTs from tomato against the *A. thaliana* genome sequence. They identified 1,025 genes that are present in single- or low-copy number in the genomes of both tomato and *A. thaliana* and show high sequence conservation during evolution. They referred to this group of genes as *conserved orthologous set* or *COS markers*. In silico computational analyses and DNA gel blot hybridization were used for identification and evaluation of COS markers. A large fraction of the identified COS markers was concerned with basic metabolic processes like energy generation, biosynthesis, and degradation of cellular components. The COS markers are used for genome evolution studies, comparative mapping among even highly divergent species as well as for physical and linkage mapping of the concerned genes. COS markers have been extensively used to connect the genomes of related species belonging to the same family. Those COS markers that are conserved can be used as hybridization probes for RFLP analyses; this should allow mapping even in such species that do not have either genomic or EST sequence databases. Other COS genes can be used for the development of gene-based markers for detecting polymorphism among the PCR products using SSCP. The consensus sequences of COS markers can be used as query for homology search of genome sequence databases of other plant species to identify putative orthologous genes in these related species.

The genome sequences of three model species, viz., *A. thaliana*, *Oryza sativa*, and *Populus trichocarpa*, were subjected to comparative analysis; this resulted in the identification of 753 candidates for COS markers. Out of these, up to 359 genes were present in the EST databases of four gymnosperm species. Similarly, the Rosaceae EST databases were compared with single-copy genes of Arabidopsis to identify 1,039 RosCOS (COS set for Rosaceae) markers. Out of these, 857 genes were chosen for designing of primers flanking introns so that the PCR product included at least one putative intron. About 91 % of these primers were able to amplify *Prunus* DNA, and 90 % of the PCR products exhibited polymorphism.

3.21 Start Codon-Targeted Polymorphism

Start codon-targeted (*SCoT*) *polymorphism* markers are based on the short conserved sequence surrounding the translation initiation codon or start codon, ATG, of plant genes as reported in various studies (Collard and Mackill 2009a). The SCoT marker system uses a single 18 nt long primer to amplify the sample genomic DNAs, and the amplification products are resolved by agarose gel electrophoresis. The SCoT primer has the following invariant nucleotides: the A, T, G of the start codon (positions +1, +2, +3), G at +4, A at +7, C at +8, and C at +9. The primers also have a variable number of arbitrary nucleotides on the 5′ side of the ATG nucleotides. The GC content of the primers may range between 50 and 70 %, and they should differ from each other for at least one nucleotide at their 3′ ends. The annealing temperature during PCR is kept at 50 °C, and the primer extension time of at least 2 min is recommended. The SCoT markers are generally highly reproducible, but some primers show poor reproducibility. The amplification products are between two and

six in number, and their lengths range from 200 to 1,500 bp. This marker system is similar to RAPD and ISSR marker systems in respect of the use of a single primer, lack of sequence information requirement, and two to six amplification products in each PCR. But SCoT markers would be based on genic regions as compared to the random genomic regions in the cases of RAPD and ISSR markers. However, some of the SCoT markers may be generated by pseudogenes and even such genes that are situated within transposable elements.

Amplification of a fragment would occur when start codons of two genes are located within a reasonable distance on the complementary strands of the DNA duplex (Fig. 3.13). The SCoT markers are dominant, but few of them may be codominant due to relatively large InDels in the amplified regions; this situation is similar to that for the RAPD markers. These markers can be used for mapping of genes/QTLs, and genetic diversity analyses. A SCoT marker of interest can be converted into a STS marker to make it single band robust marker. The SCoT marker system has the potential to be used for a simplified gene expression analysis with limited resources. The cDNA-SCoT technique was developed for this purpose (Wu et al. 2013).

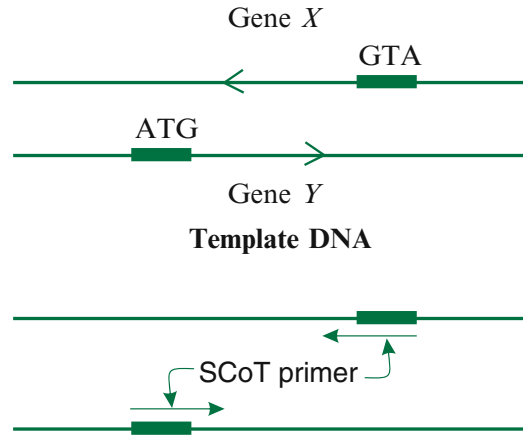


Fig. 3.13 The principle of SCoT marker system. The template DNA has different genes on the complementary strands of the DNA duplex. The start codons (ATG) of the two genes are within a distance appropriate for amplification (say, up to 1,500 bp). The ATG codons of these two genes should be located as shown in the figure for the region between them to be amplified (Based on Collard and Mackill 2009a)

3.22 CAAT Box-Derived Polymorphism

The *CAAT box-derived polymorphism* (*CBDP*) marker is a PCR-based marker similar to the SCoT marker as it uses a single primer of 18 nt that targets the CAAT box of the promoter regions of plant genes. The primer has the five-nucleotide CCAAT core flanked by 10–11 filler nucleotides on the 5′ side and 2–3 arbitrary nucleotides on the 3′ side. Singh et al. (2014a) designed a set of 25 CBDP primers and evaluated them with eight varieties of *Corchorus capsularis* and *C. olitorius*. Most of these primers generated few to several polymorphic bands in the jute varieties and in cotton and linseed as well. The CBDP marker system is similar to the SCoT marker system in many features, including the following. A band will be generated when two genes are located on the opposite strands within a distance suitable for PCR amplification. The phrase "two genes" means the CAAT boxes of the promoters of the two genes, in the case on CBDP markers, and start codons of the two genes in the case of SCoT markers. The CBDP markers would be useful for analyses of genetic diversity, DNA fingerprinting for reliable cultivar/germplasm identification, and linkage mapping of genes/QTLs and MAS.

3.23 Conserved DNA-Derived Polymorphism

The *conserved DNA-derived polymorphism* (*CDDP*) *markers* are based on conserved DNA regions of a selected set of well-characterized plant genes. For example, Collard and Mackill (2009b) analyzed the sequences of *WRKY*, *MYB*, *ERF*, *KNOX*, *MADS*, and *ABP1* genes to produce several CDDP markers. The above genes are known to participate in abiotic/biotic stress responses or developmental processes. Sequences

of the selected genes present in diverse plant species were obtained from the database and used for multiple sequence alignment analysis by ClustalW program (Sect. 14.3.9) to identify their conserved regions. These conserved sequences were used for designing primers in such a way that their GC contents were over 60 % and a single primer could have up to three degenerate nucleotides. The primers designed in either 5′–3′ or 3′–5′ direction with respect to the conserved domain sequence generated such markers that were reproducibly polymorphic. The short sequences conserved in the selected genes may be expected to be present at several locations in the plant genome and would serve as binding sites for the CDDP primers. The principle of CDDP markers is similar to that of the SCoT markers and the resistance gene analog markers. The *resistance gene analog markers* are based on primers derived from the conserved regions of genes for disease resistance of plants (Chen et al. 1998). Denaturing polyacrylamide gel electrophoresis is used to separate the PCR products. Each PCR reaction generated from 30 to 130 products, of which 27–47 % showed polymorphism in rice, barley, and wheat.

The CDDP markers are dominant and are scored as "present" or "absent." CDDP primers generate two to six fragments of 200–1,500 bp in size. This marker system is similar to the SCoT markers (Fig. 3.13) in the use of a single primer for PCR, amplification of genic regions, and the need for genes to be present at proper distance in the complementary strands. In contrast to RAPD, it uses longer primers, much higher annealing temperature (50 °C), and has high reproducibility, except in the case of some primers. CDDP differs from the conserved region amplification polymorphism (CoRAP) in the following ways. The CoRAP procedure uses two primers derived from ESTs for a specific species and requires polyacrylamide gel electrophoresis. In contrast, the CDDP markers use a single primer derived from the sequence of the selected gene present in several plant species, and they are scored by agarose gel electrophoresis. Thus, CDDP markers would target the selected plant genes, including candidate genes where known. These markers can be used for gene/QTL mapping as well as genetic diversity studies.

3.24 Conserved Region Amplification Polymorphism

The *conserved region amplification polymorphism* markers are based on pairs of primers (one fixed and one arbitrary primer) for PCR amplification (Wang et al. 2009b). The fixed primer is derived from the sequence of an EST of a given species extracted from a database like GenBank and targets the coding sequence of the gene. The arbitrary primer contains the core sequence CACGC at the 5′ end, followed by 11 arbitrary nucleotides that serve as fillers, and three bases at the 3′ end, which serve as selection nucleotides; this scheme is the same as that for the SRAP markers (Sect. 3.17). Since their core sequence is normally found in the introns of plant genes, the arbitrary primers would anneal to the majority of introns. The CoRAP primer pairs are designed for an annealing temperature of 52 °C. These markers are similar to TRAP markers (Sect. 3.18), except for the core sequences of their primers. Thus, the design of fixed primers requires sequence information of the concerned plant species. PCR amplification will occur if the two primers bind within a suitable distance from each other. The amplification products will be polymorphic if the intervening sequences had InDels, as a result of which the PCR products from different individuals/strains would differ in size. The CoRAP markers are codominant and highly reproducible. Each PCR reaction may generate 30–50 fragments of 50–1,000 bp.

3.25 Intron-Targeting Polymorphism

In case of *intron-targeting polymorphism* (*ITP*) *markers*, the primers are designed on the basis of the sequences of the conserved regions of exons flanking an intron so that the PCR product includes the intervening intron (Choi et al. 2004). Since the introns are much less conserved than exons, a high proportion of the amplified fragments may be expected to show length polymorphism due to InDels. The ITP primers are derived from the sequences of

known single-copy and low-copy number genes or from those of the ESTs available in the database. The primer pairs are designed to amplify fragments of 200–1,200 bp, which are resolved by subjecting them to agarose gel electrophoresis. The ITP markers are codominant, and the primers are transferable across the species of the same genus and, sometimes, even across genera. ITP markers are generated from genic regions, and some of them might give rise to functional markers. However, the development of ITP markers depends on prior sequence information about several target genes. The ITP markers can be used for genetic diversity analyses.

3.26 RNA-Based Molecular Markers

Several useful markers are derived by analysis of RNA (Poczai et al. 2013). For example, *SSCP analysis of cDNA* (*cDNA-SSCP*) allows estimation of relative abundance of mRNAs encoded by highly similar homologous genes of polyploid species. *RNA fingerprinting by arbitrarily primed PCR* (*RAP-PCR*) uses arbitrary sequence primers for fingerprinting of RNAs isolated from a given tissue of different individuals or RNAs obtained from different tissues of a single individual. The sequence polymorphisms detected by RAP-PCR can be used for mapping of genes. The *cDNA-AFLP* technique, as its name suggests, is an AFLP procedure that uses cDNA in the place of genomic DNA as substrate. It can discriminate between such genes that belong to the same gene family and are highly homologous and allow identification of genes related to novel processes, including stress regulation.

Questions

1. Briefly describe the procedure of PCR, and discuss its usefulness in marker development and genotyping.
2. Compare the RAPD and AP-PCR markers. Why are SCAR markers more reliable than RAPD?
3. How is complexity reduction achieved in the case of AFLP procedure? Briefly describe some of the various modifications of the AFLP procedure.
4. How are SSR markers developed? Why did they become the most widely used marker system before the SNPs became the markers of choice?
5. What are various approaches for increasing the throughput of the SSR marker system?
6. Compare the ISSR and RAPD marker systems. Discuss the applications and limitations of these markers.
7. Compare the various features of SRAP, TRAP, and CoRAP markers and discuss their usefulness in breeding programs.
8. Explain the principles of SSCP and D/TGGE markers and discuss their usefulness in breeding programs.
9. Compare the features and merits of CDDP and SCoT markers. How do they differ from RAPD markers?
10. "Transposons have been used to develop several marker systems." Discuss this statement with the help of suitable examples.
11. "The PCR technology has facilitated the development of a variety of marker systems." Discuss this statement giving suitable examples.

Appendices

Appendix 3.1: The Number of RAPD Bands Theoretically Expected from a DNA Sample

The number of RAPD bands theoretically expected from a DNA sample can be estimated on the basis of probability concept. It can be shown that the number of RAPD bands (b) of a given average size (f bp) expected from a genome of known size (N bp) amplified using primers of n nt would be given by the following formula:

$$b = 2Nf/16^n \tag{3.1}$$

The above formula is derived as follows. The probability that a specified base would occur at a given site in a DNA strand will be 1/4 since this

site could have any one of the four DNA bases. It is assumed that the distribution of nucleotides/bases is random, i.e., governed by chance, so that the four DNA bases occur in the DNA molecule in equal proportion. Surely, this assumption is unrealistic, but it is necessary for an easy estimation of the above and similar parameters. Therefore, the probability that the n bases present in a RAPD primer will be found in a DNA strand will be $1/4^n$. Exponential amplification can occur only when a second primer binding site occurs in the neighborhood of the first site; the probability of the two primer binding sites occurring together will be $1/4^{2n}$ or $1/16^n$. Since the template DNA has two complementary strands, the primer binding sites could occur on either strand at a given site. In addition, the two primer binding sites would be separated by f bp, i.e., the RAPD fragment size. Therefore, the probability of two primer binding sites occurring in a DNA duplex of f bp would be $2f/16^n$. If the size of genomic DNA were N bp, the number of expected RAPD fragments of f bp would be $2Nf/16^n$.

According to the above formula, a primer of 10 bases is expected to generate 2 bands in rice, which has the genome size of 450 Mb and 4 bands in tomato that has genome of 950 Mb. Similarly, it would produce 9 bands in maize (genome size, 2,500 Mb) and 19 bands in barley (genome size, 5,300 Mb).

Appendix 3.2: Polymerase Chain Reaction and Randomly Amplified Polymorphic DNAs

PCR was developed for amplification of a specific segment from a DNA sample of high complexity, e.g., human genomic DNA. Subsequently, this procedure was applied to achieve a variety of other objectives, for each of which the procedure was suitably modified. In a general sense, the term *PCR* signifies repeated replication of a segment of sample DNA by using suitable primer(s) and DNA polymerase. In this sense, all applications of the technique would qualify as PCR. But in a restricted sense, the term *PCR* signifies amplification of a specific sequence from the sample DNA; this PCR procedure differs in many ways from the other applications of the technique. The various features of PCR (in the restricted sense) and RAPDs are summarized in Table 3.3.

Table 3.3 A comparison between PCR and RAPD procedures

Feature	PCR	RAPD
Amplified region	Specified/known	Random
Prior sequence information of the target segment	Essential for designing the specific primers	Not required; primer sequence is arbitrary
Primer sequence	Complementary to the 3′ ends of the two strands of the target DNA segment	Arbitrary; specified by the experimenter
Number of primers	Two: one forward and one reverse	One
Primer length	15–22 nucleotides	10 nucleotides
Primer binding sites	Fixed (due to the primer sequence); usually, one	Random; usually, more than one
Annealing temperature	High (around 65 °C; ~1–2 °C less than the T_m of the primer[a])	Low (around 36 °C; ~5 °C less than the T_m of the primer[a])
Annealing conditions	High stringency	Low stringency
Number of amplified fragments/bands	One (usually) or few	Several
Reproducibility	High	Moderate to poor

[a]T_m of the primer is the melting temperature of the primer–template duplex. T_m of a DNA duplex can be estimated from the following formula

$$T_m = 4(G + C) + 2(A + T)$$

where G + C and A + T represent the numbers of purine (G + C) and pyrimidine (A + T) residues in one strand of the DNA duplex

Sequence-Based Markers

4

4.1 Introduction

The DNA markers like RFLPs, AFLPs, and SSRs were extensively used in various biological investigations and for marker-assisted selection (MAS) in both animals and plants. However, the development of many of these markers, e.g., RFLPs and SSRs, is demanding and expensive as it involves time-consuming cloning, construction of probe libraries, and/or sequencing for primer design. In addition, scoring of a number of these markers across many individuals is also expensive, labor intensive, and time-consuming. Therefore, continuous efforts were made to develop such DNA markers that are reliable, abundant, almost evenly distributed throughout the genome, and relatively cheaper, developed with minimum effort and time and are amenable to automation and high-throughput genotyping. The genome sequence data generated by the human genome-sequencing project revealed that bulk of sequence variation among different individuals was due to changes at single-base positions distributed throughout the genome. The variation in single base pairs of DNA is known as *single nucleotide polymorphism* (*SNP*). Subsequently, SNPs were found to be universal and the most abundant markers; they constitute ~90 % of the genetic variation in any organism. This marker system yields reliable and reproducible results and is amenable to automation and high-throughput genotyping (Mammadov et al. 2012).

The discovery of SNPs involves sequencing of genomic DNA or cDNA (complementary or copy DNA) from two or more individuals/lines of a given species and comparing these sequences using a suitable computer program. SNPs may also be discovered by in silico alignment and analysis of genomic/EST sequence data available in the databases of the concerned species. In either case, once SNPs are discovered, they can be genotyped using any one of more than 30 different detection methods based on one or more of the following reactions: (1) DNA hybridization, (2) primer extension, (3) oligonucleotide ligation, and (4) DNA replication. Several of these methods have been automated and scaled up for high-throughput SNP genotyping; some of these technologies are considered in some detail in Chap. 13. In this chapter, we shall discuss DNA sequencing, methods for SNP discovery, and small- to moderate-scale SNP genotyping strategies.

4.2 DNA Sequencing

The determination of base sequence of a DNA fragment is called *DNA sequencing*. DNA sequencing became feasible due to the following important developments: (1) availability of restriction enzymes, (2) development of electrophoresis techniques capable of separating DNA fragments differing by a single nucleotide, and (3) gene cloning and PCR techniques that make available very large number of copies of

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_4, © Author(s) 2015

individual DNA fragments required for sequencing. Initially, two methods, a chemical and an enzymatic method, of DNA sequencing were developed; these methods are popularly termed as *first-generation DNA sequencing procedures*. Soon the *second-* or *next-generation DNA sequencing* (*NGS*) *methods* were developed, which use PCR for in vitro cloning in the place of in vivo cloning and are much faster and cheaper (Pandey et al. 2008; Schendure and Ji 2008; Edwards 2013). At present, the *third-generation DNA sequencing* (*TGS*) *methods* are becoming commercially available; these methods sequence single DNA molecules without any cloning (Schadt et al. 2010).

4.2.1 First-Generation DNA Sequencing Methods

The *chemical method of DNA sequencing* uses specific chemical modifications of DNA bases, ultimately, leading to breaks in DNA strands at the sites occupied by the modified bases. Four separate reactions are set up for the modification of different bases, and gel electrophoresis, followed by autoradiography, allows deduction of the base sequence of the DNA strand. The *enzymatic method of DNA sequencing*, also called *Sanger–Coulson method* (Sanger et al. 1977), on the other hand, uses single-stranded DNA fragments for DNA replication catalyzed by the Klenow fragment of *E. coli* DNA polymerase I. Often the two complementary strands of a DNA fragment are sequenced in separate reactions for an enhanced reliability of the sequence data. For each strand, four separate reactions are set up. In each reaction mixture, the DNA strand, a suitable primer, the Klenow fragment, the four deoxyribonucleotides (dNTPs, viz., dATP, dGTP, dCTP, and dTTP), and the other reagents required for DNA replication are provided; at least one of the four dNTPs is radioactive to allow radioautographic imaging of the bands after gel electrophoresis. In addition, in each reaction mixture, a different 2′,3′-dideoxynucleotide (ddNTP) is also added at a concentration of about 1/100 of that of the normal deoxyribonucleotides used in the reaction.

The ddNTPs do not have a free 3′-OH group. Therefore, when a ddNTP is incorporated at a site into a growing polynucleotide chain, there is no further addition of nucleotides to the chain beyond this site. Therefore, ddNTPs are called *chain terminators* or simply *terminators*. At the concentration used here, a given ddNTP would cause chain termination at any one of all the possible sites, at which its complementary base occurs in the template DNA strand. In the end, therefore, the mixture will contain partially synthesized polynucleotide chains of different lengths produced by chain termination at every point where the base complementary to the given ddNTP is present in the template DNA strand. The DNA duplexes formed in the four reaction mixtures are denatured; the mixtures are loaded in gel lanes and subjected to electrophoresis. The bands formed in the gel lanes are visualized by radioautography, and the base sequence is read by comparing the band positions in the four lanes. This procedure enables sequencing of fragments of up to 700–800 bases.

The Sanger–Coulson method was automated to support the various genome-sequencing projects. The automated procedure uses fluorescent labels (a different label for each of the four ddNTPs) in the place of radioactivity, capillaries in the place of routine gels for electrophoresis, and computer-based sequence detection, data storage, and processing. These automated sequencers have been in use for over 30 years, and until recently most genome-sequencing projects were exclusively based on this technology. The current *read lengths*, i.e., the lengths of sequences of single fragments, are up to 1,000 bp with an error rate of 0.001 %. However, whole-genome sequencing required several sequencers located at a large center, having highly automated template preparation and other supporting facilities. In addition, the sequencing process is highly demanding in terms of both infrastructure and processing efforts, and the sequencing costs are rather high (Deschamps and Campbell 2010; Edwards 2013).

4.2.2 Next-Generation DNA Sequencing Methods

The next-generation DNA sequencing methods, also called massively parallel sequencing (MPS) technologies, are faster and cheaper and require much less template preparation than the Sanger–Coulson method. The NGS methods use PCR amplification for template preparation (in vitro *library preparation*), which takes merely 2 h, and they are amenable to very high throughput. Further, they allow simultaneous sequencing of hundreds of thousands to hundreds of millions of different DNA fragments (Schendure and Ji 2008). At present, there are five NGS methods, namely, (1) 454 sequencing, (2) Solexa method, (3) ion semiconductor sequencing, (4) Polony method, and (5) massively parallel signature sequencing (MPSS). The first three methods (454, Solexa, and ion semiconductor sequencing methods) use DNA synthesis for sequencing (*sequencing by synthesis*, *SBS*), while the Polony and MPSS methods employ oligonucleotide hybridization to the template followed by ligation to the growing chain. The MPSS is suited for quantification of gene expression; it uses multiple cycles of enzymatic cleavage and ligation to determine 17–20-bp-long "signature" sequences from the ends of cDNA molecules to distinguish and quantify the different RNA species present in the sample. The 454, Solexa, ion semiconductor, and Polony methods have already been commercialized for high-throughput sequencing and are briefly described in the following sections.

According to a survey, the use of NGS technologies in public and private sequencing laboratories of the USA and Europe had gone up to 56 % by 2010. The most frequent application of these technologies was mRNA expression profiling, followed by biomarker discovery, resequencing, diagnostics, and targeted resequencing. In 2011, Illumina HiSeq 2000 and Illumina GAIIx platforms were the market leaders in terms of sales. However, SOLiD 3 Plus was rated to have the highest accuracy as compared to the Illumina GAIIx and GS FLX systems. It is projected that NGS and TGS technologies will eventually replace the established techniques like targeting-induced local lesions in genomes (TILLING), TILLING in wild populations (Eco-TILLING), and endonucleolytic mutation analysis by internal labeling (EMAIL).

4.2.2.1 Template Preparation

The template for sequencing is single-stranded DNA (ssDNA), which can be prepared from genomic DNA, BAC clones, PCR products, and cDNA. Genomic DNA and BAC clones are randomly sheared by sonication, nebulization (mechanical shearing), or enzymatic digestion by DNase I to produce fragments of suitable size, while PCR products and cDNA may not need fragmentation. Often one may need to sequence some specific regions identified by linkage studies. In such cases, methods like "enrichment," "genome partitioning," or "genome capture" can be used for template sample preparation. These methods involve mRNA extraction, hybridization to preselected probes, or attachment of barcodes/index sequences to the fragments. After fragmentation, DNA fragments of 300–800 bp (used for shotgun sequencing) or 3–20 kb (used for paired-end sequencing) are separated. For *shotgun sequencing*, short adapters (adapters A and B) specific for both 3′ and 5′ ends are attached to each fragment; these adapters facilitate purification, amplification, and sequencing (Fig. 4.1). The fragments are now made single-stranded, and one single strand is attached to a single capture bead. These beads, along with the amplification reagents and the enzymes, are then enclosed in droplets of water-in-oil mixture. The emulsion around each bead forms a micro-reactor isolated from all other such beads. PCR amplification produces millions of copies of the single fragment attached to each bead, and all these copies become attached to the same capture bead. These beads form the in vitro library used for sequencing (Fig. 4.1).

In the case of *paired-end sequencing*, adapters are added to both the ends of the much larger fragments to facilitate their circularization. The

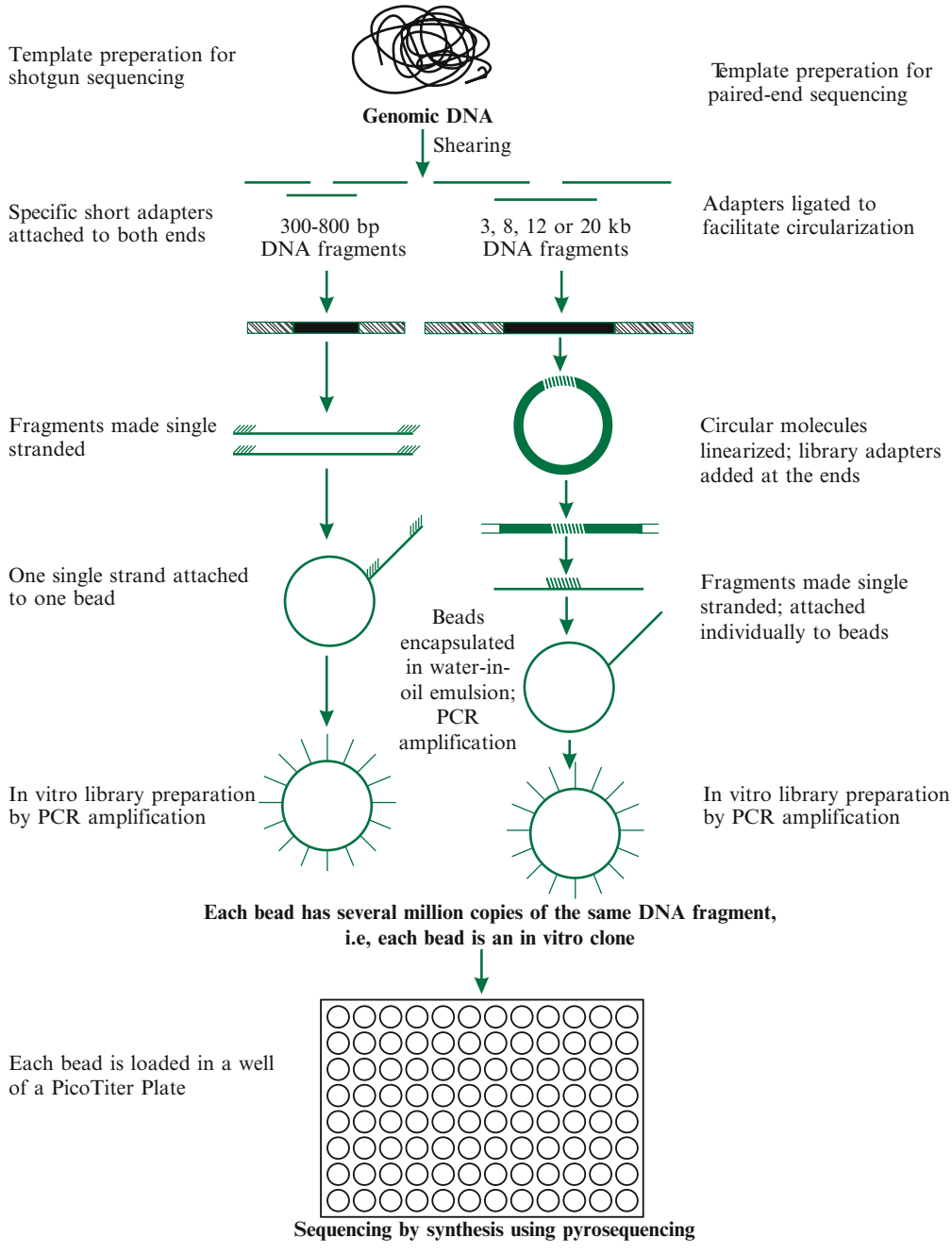


Fig. 4.1 A generalized schematic representation of the 454 sequencing method. The template preparation in other NGS methods is generally similar

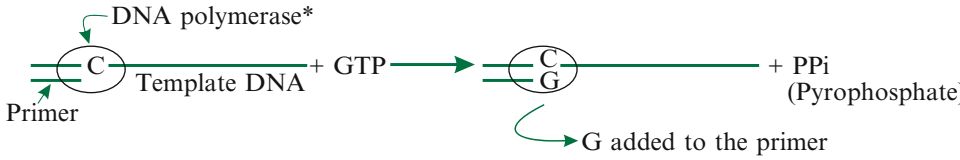
circularized DNA is fragmented, linear fragments containing the adapters are separated, their ends are polished, and library adapters (adapters A and B) are linked to both their ends. Thus each fragment has library adapters at the two ends, followed by a short segment corresponding to the ends of the genomic fragment, and finally the adapter sequence located in

the middle region of the fragment. These fragments are made single-stranded, and one single strand is attached to each capture bead. The beads are then processed in the same way as in the case of shotgun sequencing (Fig. 4.1). This description is based on template preparation for the 454 sequencing method, but the other NGS technologies also use similar strategies.

4.2.2.2 The 454 DNA Sequencing Method

This method was the first NGS technology to be commercialized in 2005 by 454 Life Sciences (now Roche Diagnostics), USA. The currently available 454 platforms are Genome Sequencer (GS) FLX System and GS FLX Titanium series. After template preparation (Sect. 4.2.2.1), the capture beads along with the attached DNA fragments are removed from the emulsion and loaded into the wells of a PicoTiter Plate. The size of wells is such that only a single bead can be loaded in each well. DNA sequencing is achieved by the pyrosequencing method. The reagents are flowed in a specific order across the plate, and the chemiluminescence signal is sensed by a sensitive CCD (charge-coupled device) camera. The computer software uses the chemiluminescence data to deduce the base sequence of the template DNA segment attached to every bead.

In *pyrosequencing*, the reaction mixture contains the template DNA, sequencing primer, APS (adenosine-5′-phosphosulfate), luciferin, the Klenow fragment, ATP sulfurylase, luciferase, and apyrase. The nucleotides dCTP, dGTP, and dTTP, and dATPαS (deoxyadenosine α-thiotriphosphate) are added to this reaction mixture sequentially one after the other. dATPαS is used in the place of dATP because it can be used by luciferase for light generation only after it has been used for DNA synthesis. In contrast, dATP will be used for producing light even when it is not used for DNA synthesis. When a dNTP, say, dGTP, is added to the 3′ end of the primer or the growing chain, one pyrophosphate (PPi) moiety is released (Fig. 4.2a). This PPi is used by ATP sulfurylase to convert APS into ATP (Fig. 4.2b), which is then used by luciferase to



(A) Release of pyrophosphate by DNA polymerase action

Adenosine 5′-phosphosulfate (APS) + PPi $\xrightarrow{\textbf{ATP sulfurylase}}$ **ATP + SO_4^-**

(B) ATP synthesis by ATP sulfurylase

Luciferin + ATP $\xrightarrow{\textbf{Luciferase}}$ **Oxyluciferin + ADP + Pi (yields light)**

(C) Luciferase uses ATP to generate visible light

dNTP (including ATP) $\xrightarrow{\textbf{Apyrase}}$ **dNDP + dNMP + Phosphate (Pi)**

(D) Pyrase action degrades dNTPs

Fig. 4.2 The various reactions catalyzed by the four enzymes used in pyrosequencing. *PPi* pyrophosphate, *Pi* inorganic phosphate, *ATP* adenosine triphosphate, * Klenow fragment (Based on de Vienne et al. 2003)

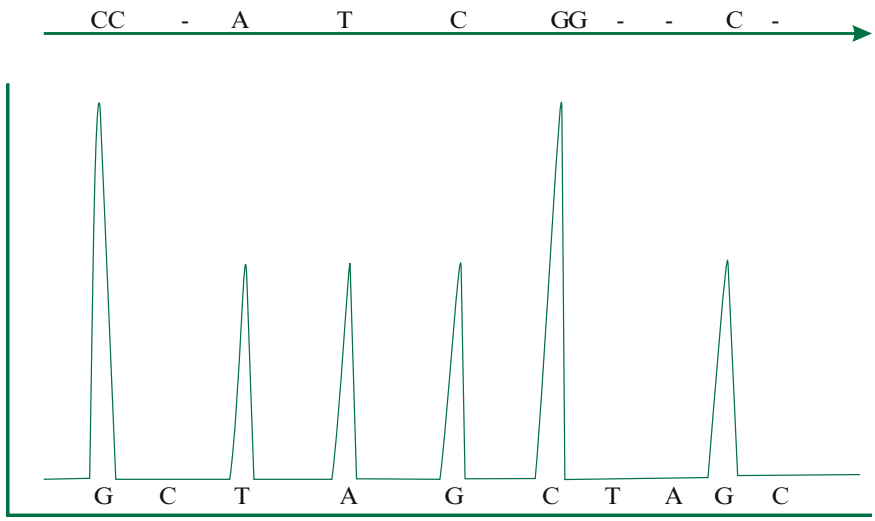


Fig. 4.3 A pyrogram pattern, and the nucleotide sequence deduced from this pattern. Production of light indicates nucleotide incorporation into the primer/growing chain. A stronger light signal, e.g., in response to the addition of the first G and the second C, reveals incorporation of the concerned nucleotide at two consecutive sites. But a lack of light signal, e.g., for the first C, shows lack of incorporation of the concerned nucleotide (Based on de Vienne et al. 2003)

generate visible light (Fig. 4.2c) (Ronaghi et al. 1996). A sensitive CCD camera detects the light, and the template nucleotide at this position is deduced (it will be C in this case). The intensity of light generated is proportional to the amount of PPi generated, i.e., the amount of nucleotide added to the primer/growing chain. Therefore, the light signal will be twice as intense if the same nucleotide occurs at two consecutive sites (Fig. 4.3). However, if the G from the dGTP was not added to the growing chain, no light will be generated. The enzyme apyrase continues to hydrolyze the unincorporated dNTPs (Fig. 4.2d) as well as the ATP produced by the ATP sulfurylase action. As a result, soon the ATP is exhausted and light production ceases; the next dNTP can now be added to the reaction mixture. Since the rate of dNTP degradation by apyrase is slower than its incorporation into the growing chain, sufficient dNTP remains available for DNA synthesis. Similarly, ATP degradation by apyrase is slower than ATP production by ATP sulfurylase so that enough ATP becomes available for light production when a dNTP is used for DNA replication.

The GS FLX system can process over one million beads at a time, and one run takes about 10 h, including template preparation. The data from paired-end sequencing can be combined with that from shotgun sequencing to readily generate a high-quality draft genome of large complex organisms. The average read length (length of individual sequences) in shotgun sequencing is ~400 bases, but bulk of the reads are of 500 bases; the GS FLX+ can now give reads of up to 1,000 bases (Edwards 2013). Read accuracy of GS FLX is over 99.6 %, while consensus accuracy is more than 99.99 %. *Read accuracy* is the accuracy of the sequence of individual reads, while *consensus accuracy* is the accuracy of the sequence of a fragment obtained as consensus of the sequences of all the reads of the fragment. In this and the Sanger–Coulson method, the error rate increases with the position of the base in the fragment due to a reduction in enzyme efficiency/concentration, leading to a reduced light signal-to-noise ratio. GS FLX can generate 400-Mb sequence in a 10-h run at a cost of US $ 5,000–7,000, while GS FLX Titanium XL+ can produce one million reads of up to 1,000 bp each (total sequence 1 Gb).

This technology can be used for de novo sequencing and assembly, genome sequencing and mapping, transcriptome analysis, analysis of epigenetic changes, etc. As the 454 method does not use chain terminators, a base will become incorporated as many times in a single cycle as its complementary base occurs consecutively in the template strand. When the same base occurs several (usually, >6) times consecutively (e.g., AAAAAAA) in the template, occasionally it is read one base less than the actual number, i.e., $n - 1$ times in the place of n times. This may lead to errors in base sequences of those stretches of template DNA, in which a base occurs more than once in tandem. Further, artifacts of single base pair deletions or insertions can be generated by signal-to-noise threshold problems.

4.2.2.3 The Illumina Sequencing Method

Illumina, USA, commercialized the Solexa NGS technology in 2007 (Bentley and Smith 2008), which is the most widely used NGS technology. The recent platforms of the series are Illumina Genome Analyzer 1 Gb and HiSeq 600 Gb. The sample DNA is fragmented, and two different adapters are ligated to their 5′ and 3′ ends. The fragments are attached to an especially prepared substrate on a flow cell, which contains a dense lawn of primers to be used in the next step of solid phase PCR. Fold-back PCR or bridge PCR produces up to 1,000 identical copies of each DNA fragment. All the copies of one fragment form an isolated cluster of molecules on the flow cell, and together they represent the in vitro clone of the fragment. All the clusters formed on a flow cell together represent the in vitro library (Fig. 4.4). The sequencing primer is now attached to the free ends of the fragments. The four dNTPs used for DNA synthesis have fluorophores linked to them; these fluorophores also serve as chain terminators. The dNTPs are added one at a time, and a CCD camera records their incorporation at the 3′ end of the sequencing primer/growing chain as fluorescence from the fluorophores attached to them. The fluorophore terminator is removed from the dNTP that has just been added to the primer/growing chain, making this nucleotide available for further DNA synthesis. A new dNTP is now added to the reaction mixture, it is incorporated at the ends of the growing chains, the fluorescence is recorded, and then the fluorophore is removed. In this way, the sequence of each DNA fragment is determined. The use of fluorophore chain terminators linked to the dNTPs eliminates the error in base sequence determination when the same base is present at two or more consecutive positions in the template strand.

Usually, read length ranges from 35 to 150 bases, and the accuracy is greater than 98.5 %. The total error-free read given by Illumina HiSeq 2000 is over 400 Gb in one run, which takes 7–8 days (Edwards 2013). MiSeq and HiSeq 2500 systems generate read lengths of up to 250 bp and have improved data capture and greater flexibility. The Illumina system can be used for de novo genome sequencing; genome resequencing for the analysis of SNPs, InDels, copy number variation (CNV), and structural variation; transcript profiling; etc. However, the PCR amplification step introduces a high error rate. The fluorescence properties of the four dyes used in this method tend to produce substitutions of A for C, G for T, and vice versa in the sequence data. In addition, the terminators of some nucleotides may not function properly so that a second nucleotide may be added to the growing chain in the same reaction cycle, generating a deletion of one base pair. However, base substitution errors are more common than insertion/deletion errors.

4.2.2.4 The ABI SOLiD Technology

The Applied Biosystems, USA, commercialized the Polony method in 2005 as SOLiD 3.0 platform (Schendure et al. 2005). SOLiD stands for "sequencing by oligonucleotide ligation detection" since this method achieves DNA sequencing by detecting oligonucleotide ligation. The DNA sample is fragmented (fragment size 600 bp to 6 kb) and processed in a manner similar to that for paired-end sequencing (Fig. 4.1). The beads along with the attached DNA molecules are immobilized in a single layer in an acrylamide matrix on a glass slide. An anchor primer is then hybridized to the adaptor sequence attached

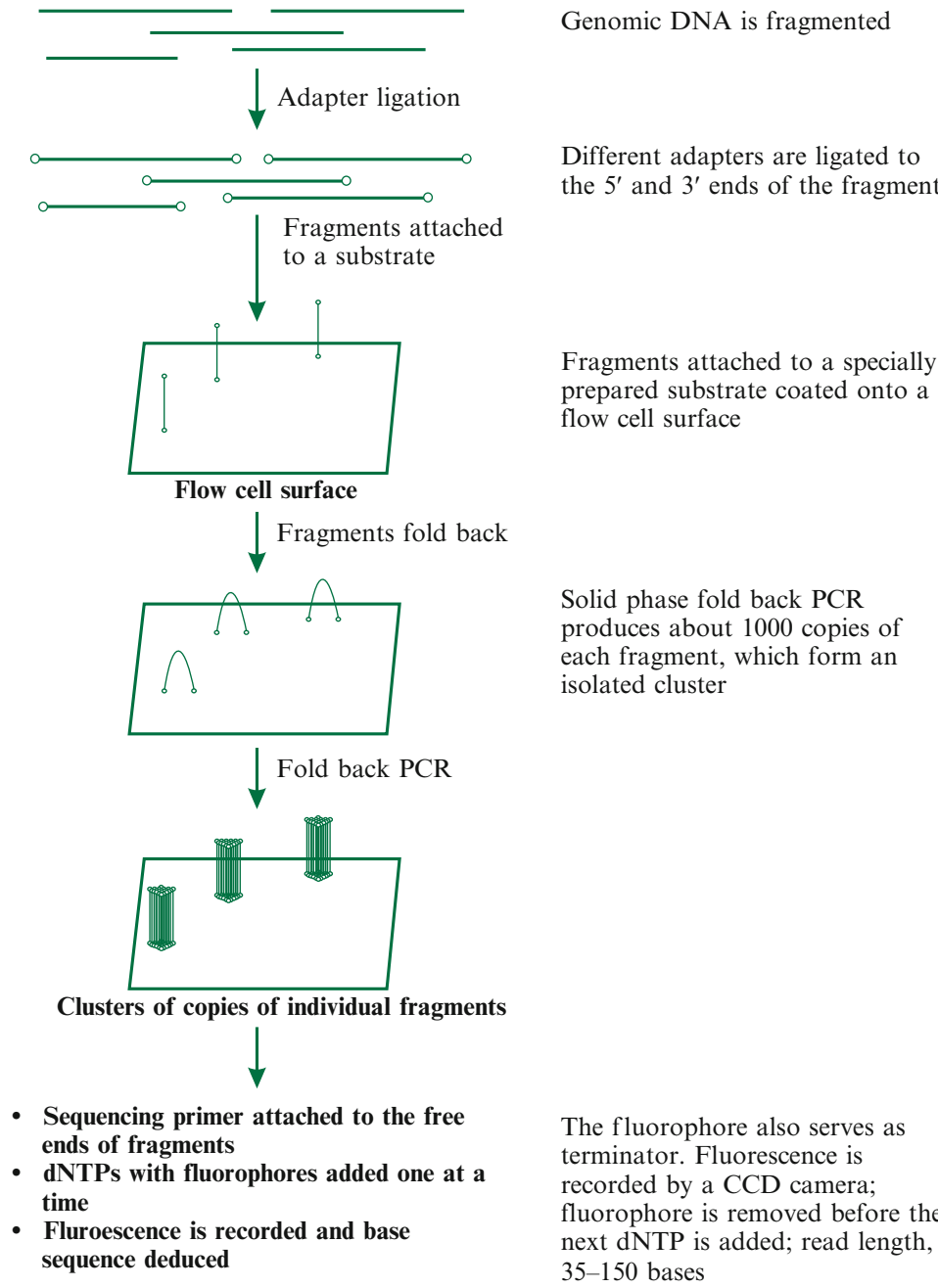


Fig. 4.4 A simplified schematic representation of the Illumina sequencing method (Based mainly on Bentley and Smith 2008)

to the template DNA. Sequencing is done by using a set of 16 oligonucleotides for hybridization with the template DNA and ligation to the 5′ end of the anchor primer/elongating chain. Each oligonucleotide is 8 bases long and is labeled with fluorophore at the 5′ end, and each member of a set of 16 oligos has a unique combination of two nucleotides at its 3′ end.

At a given time, four specific oligonucleotides of the set, each labeled with a different fluorophore, are added and allowed to pair at their 3′ ends with the template DNA. The

3′ ends of the oligonucleotides paired with the template DNA are ligated to the 5′ ends of the anchor primer molecules, the color of fluorescence is recorded, and the unpaired 5′ ends of the oligonucleotides are removed. A new set of four oligonucleotides is now added and the steps of the first cycle are repeated. After five cycles of oligonucleotide hybridization and ligation, the DNA is melted and the newly synthesized DNA strands are removed. A new anchor primer is now added that is one base shorter than the adaptor. Therefore, hybridization will begin one base upstream of the site it began in the first cycle and into the adaptor sequence. Again five cycles of hybridization and ligation are carried out, and fluorescence from each cycle is recorded. The data from the two repeats of ligation reactions are compared and analyzed to obtain the base sequence of the template strand. The repeat hybridization run using one base shorter primer allows each base to be examined twice and to fill any gaps that may remain after the first run.

The SOLiD 3.0 platform gives sequence reads of ~50 bases and generates over 20 Gb of total sequence per run, and each run takes about 6–7 days. In 2011, SOLiD 5500 and SOLiD 5500 XL systems were introduced; these systems give sequence data of up to 300 Gb per run at 99.9 % accuracy (Edwards 2013). The average error rates are lower when a good quality reference genome sequence is available and is used for error correction. In the absence of a reference genome, the error rate is higher than that for Illumina GA. Errors in base sequence arise from PCR amplification, beads carrying a mixture of fragments, incomplete dye removal, etc. The essential features of the three common NGS technologies, viz., the 454, Illumina, and ABI SOLiD technologies, are summarized in Table 4.1.

Table 4.1 A comparison among the three common NGS technologies: the 454, Illumina, and ABI SOLiD technologies

Feature	454	Illumina	ABI SOLiD
Sequencing reaction	Pyrosequencing (sequencing by synthesis)	Sequencing by synthesis	Oligonucleotide hybridization
Terminator	Not used	Used	Used
Detection based on	Luminescence generated by luciferase	Fluorescence from fluorophore	Fluorescence from fluorophore
Major error in base calling	InDels	Base substitutions	Base substitutions
Chief cause of error	Incorrect deduction of homo-polymorphic length from intensity of luminescence	Asynchronous DNA synthesis in the later cycles	Bias in fluorescence intensities in later machine cycles
Template DNA fragments attached to	Beads in microtiter plate wells	A specific substrate on a flow cell	Beads in an acrylamide matrix
Run duration[a]	10 h	7–8 days	6–7 days
Average read length (shotgun sequencing)	400 bases (GS FLX +, 1,000 bases)	35–150 bases (up to 250 bases by Hi Seq 2500)	~50 bases (SOLiD 3.0)
Total sequence data/run	400 Mb (GS FLX), ~1 Gb (GS FLX Titanium +)	400 Gb (Hi Seq 2000)	300 Gb (SOLiD 5500, SOLiD 5500 XL)
Read accuracy (%)	99.6 (99.9)[b]	98.5	–
Template preparation	Shotgun, paired end	Shotgun, paired end	Paired end
Each base examined	Once	Once	Twice
Improved base-calling algorithm[c]	Pyrobayes	Ibis and BayesCall	Rsolid
Draft genome preparation	Yes	Yes	–
Current platforms	GS FLX, GS FLX Titanium	Genome Analyzer 1 Gb, Hi Seq 600 Gb	SOLiD 5500, SOLiD 5500 XL

[a]Including template preparation
[b]The figure within parenthesis is the consensus accuracy
[c]These algorithms reduce error rates by ~5–30 % over the base-calling methods developed by the manufacturers

4.2.2.5 Ion Semiconductor Sequencing

In the case of *ion semiconductor sequencing*, a semiconductor sensing device or ion chip senses the H^+ ions produced during DNA synthesis by DNA polymerase (Schadt et al. 2010; Edwards 2013). The signals from H^+ ions are used for direct nonoptical identification of the bases present in the DNA template. The ion chip has 3.5-μm-diameter wells, each of which is located directly over each sensor. As a result, the wells confine the DNA fragments and the reagents for DNA synthesis directly over the sensors. The sequencing equipment comprises primarily an electronic detection system that is interfaced with the chip, a microprocessor to process the signals, and a fluidics system for regulating the reagent flow over the chip. The genomic DNA is fragmented, ligated to adapters, and attached to 2-μm-diameter acrylamide beads, and the DNA fragment attached to each bead is amplified by PCR (Fig. 4.5). The DNA polymerase and the sequencing primers are now attached to each template DNA molecule already attached to the beads, and the beads are then pipetted into the loading port of the chip. The well depth and the bead diameter ensure that only a single bead is loaded in each well. The four dNTPs are now added one at a time. When the DNA polymerase adds a dNTP to a primer/growing chain, there is net release of one proton (H^+), which produces a

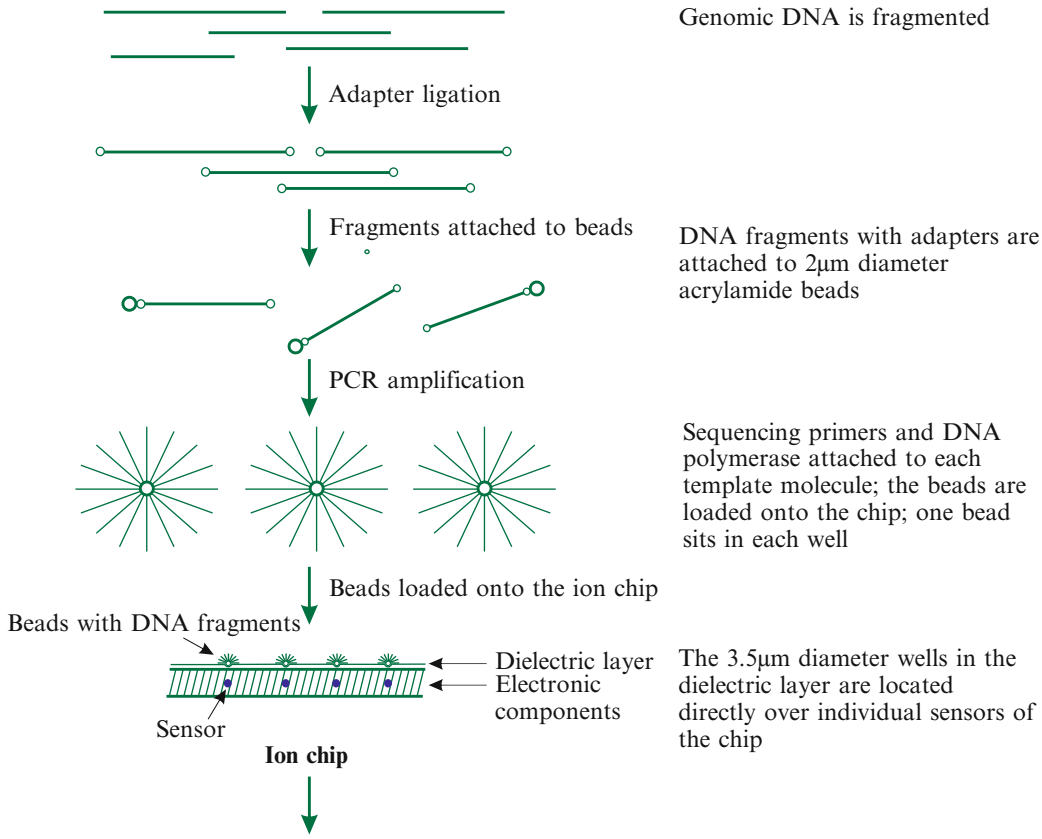


Fig. 4.5 A schematic representation of the ion semiconductor sequencing method. The Ion Torrent Proton platform generates up to 10 Gb sequence data per run; read length, 100–200 bases (Based on Schadt et al. 2010)

change in the pH of the surrounding solution. The sensor located at the bottom of each well detects this change in the pH and the signals are ultimately digitized. In case a dNTP is added to the primer/growing chain more than once due to the occurrence of its complementary base in the template DNA at more than one consecutive position, the change in pH is proportional to the number of nucleotides incorporated (0.02 pH units for each dNTP molecule added). The signal generation and detection takes ~4 s. The unused nucleotides are removed by washing before the new dNTP is added; this takes about one-tenth of a second.

Ion Torrent (acquired by Life Technologies), USA, has commercialized this technology as Ion Torrent PGM (Personal Genome Machine) and Ion Torrent Proton sequencing platforms. The Ion Torrent PGM produces 10 Mb–1 Gb sequence data per run with either 100 or 200 bases read-length protocols and sample multiplexing. But the Ion Torrent Proton platform generates up to 10 Gb of sequence data per run with 100 bp or 200 bases read-length protocols and sample multiplexing. A typical run lasts just 2 h.

4.2.2.6 Limitations of the NGS Methods

The NGS methods generate short length reads that are not easy to assemble as genome sequences because plant genomes contain extensive repeat sequences. In view of this, various sample preparation strategies like mate pair libraries/large insert libraries, paired-end reads, preparations from sorted chromosome, RNA-Seq data, optical mapping, reduced representation libraries, and information from genetic mapping are used to facilitate genome assembly. In fact, few plant genome sequences of high quality have been completed using NGS technologies. These methods use PCR for generating copies of the DNA fragments. This step inevitably introduces bias so that the quality of coverage of different genomic regions is not uniform. In addition, sequencing is based on synthesis or hybridization reaction that uses as template millions of copies of a given fragment. It is expected that reactions at all the copies of a single template fragment will occur in synchrony. This, however, may not happen and some copies may fall out of synchrony; this would introduce error in the base sequence of the given fragment. Each NGS platform provides its own software package for signal acquisition and "base calling" (deduction of bases on the basis of light color and intensity signals) with minimum error rates. In addition, several other base-calling algorithms have been developed (Table 4.2) that reduce base-calling error rates by ~5–30 % over the methods provided with the NGS platforms. But the software packages provided with the NGS platforms are the most widely used.

Sample/template preparation for NGS technologies takes several days, which often involves additional equipment costs, chemicals and other consumables, and physical space. Although NGS technologies generate sequence data at a lower cost per base sequenced, they have greatly increased the size of projects due to, among other things, the huge amounts of sequence data generated, which has created challenges for their storage, analysis and management. The third-generation sequencing methods are based on single DNA molecules, and they do not suffer from the above limitations.

4.2.3 The Third-Generation DNA Sequencing Methods

The *third-generation sequencing methods* do not use PCR amplification for template preparation because they sequence single DNA molecules (Schadt et al. 2010). For this reason, they are often called *single-molecule sequencing* (*SMS*) *methods*. The technologies being developed for TGS are quite diverse and include captured DNA polymerase, nanopores, electronic detection, fluorescence energy transfer, and transmission electron microscopy. Two of these methods emerged as feasible DNA sequencing options during 2011. Some of the TGS technologies are briefly described in the following sections.

Table 4.2 Some of the freely available NGS data analysis and SNP and genotype-calling software packages

Software	Available from	Prerequisites for application	Remarks/functions available
Single-sample calling			
SOAP2	http://soap.genomics.org.cn/index.html	High-quality variant database[a]	NGS data analysis; includes genotype caller SOAPsnp
realSFS	http://128.32.118.212/thorfinn/realSFS/	Aligned reads	SNP and genotype calling; uses allele frequencies
Multi-sample calling			
Samtools	http://samtools.sourceforge.net/	Aligned reads	NGS alignments; computation of genotype likelihoods (samtools); SNP and genotype calling (bcftools)
GATK	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit	Aligned reads	NGS data analysis; SNP and genotype calling (UnifiedGenotyper), SNP filtering (variant filtration); SNP quality recalibration (variant recalibrator)
Multi-sample and LD-based calling			
Beagle	http://faculty.washington.edu/browning/beagle/beagle.html	Candidate SNPs, genotype likelihoods	Imputation, phasing, and association, including genotype calling
IMPUTE2	http://mathgen.stats.ox.ac.uk/impute/impute_v2.html	Candidate SNPs, genotype likelihoods, fine-scale linkage map	Imputation, phasing, and association, including genotype calling
QCall	ftp://ftp.sanger.ac.uk/pub/rd/QCALL	"Feasible" genealogies at a dense set of loci, genotype likelihoods	SNP and genotype calling, generating candidate SNPs without (NLDA) and with (LDA) LD information
MaCH	http://genome.sph.umich.edu/wiki/Thunder	Genotype likelihoods	SNP and genotype calling; generating candidate SNPs without (GPT_Freq) and with (thunder_glf_freq) LD information

[a]For example, dbSNP

4.2.3.1 Helicos Genetic Analysis System

In this method, 100–200-bp-long template fragments are subjected to tailing to generate over 50-nucleotide-long poly(dA) tails at their 3′ ends, followed by blocking of the 3′ ends with a suitable treatment. These fragments are now hybridized with primers [50-nt-long poly (dT)] immobilized on a proprietary substrate within a glass microfluidics cell having 25 channels (Fig. 4.6). The dNTPs used for DNA synthesis are labeled with a bright fluorophore, e.g., Cy3 and Cy5, so that the dNTPs incorporated into single growing chains are readily detected. The four labeled dNTPs (blocked with virtual terminators) are added sequentially, one at a time. When molecules of a given dNTP are added, they will be incorporated at the 3′ ends of those primers/growing chains that are associated with the template molecules having the base complementary to the given dNTP at the proper site. The fluorescence from the incorporated nucleotide is recorded separately for each template molecule. The fluorophores of the incorporated nucleotides and the terminators are removed, and the next dNTP along with DNA polymerase is added. In this way, base sequence of each template molecule is determined.

The length of each read is ~35 bases, and up to one billion reads (and 35 Gb sequence data) can be obtained in one run. Since a virtual terminator is used, a dNTP can be incorporated only at a single site in a template during each reaction cycle even when its complementary base occurs at two or more consecutive sites in the template. The raw read error rate is generally 0.5 %, but the finished/consensus error rate tends to be much lower. Helicos BioSciences Corporation, USA, has commercialized this process as Helicos Sequencer, HeliScope™. This system generates 1 Gb usable sequence data per day (~100 times greater than the first-generation sequencers).

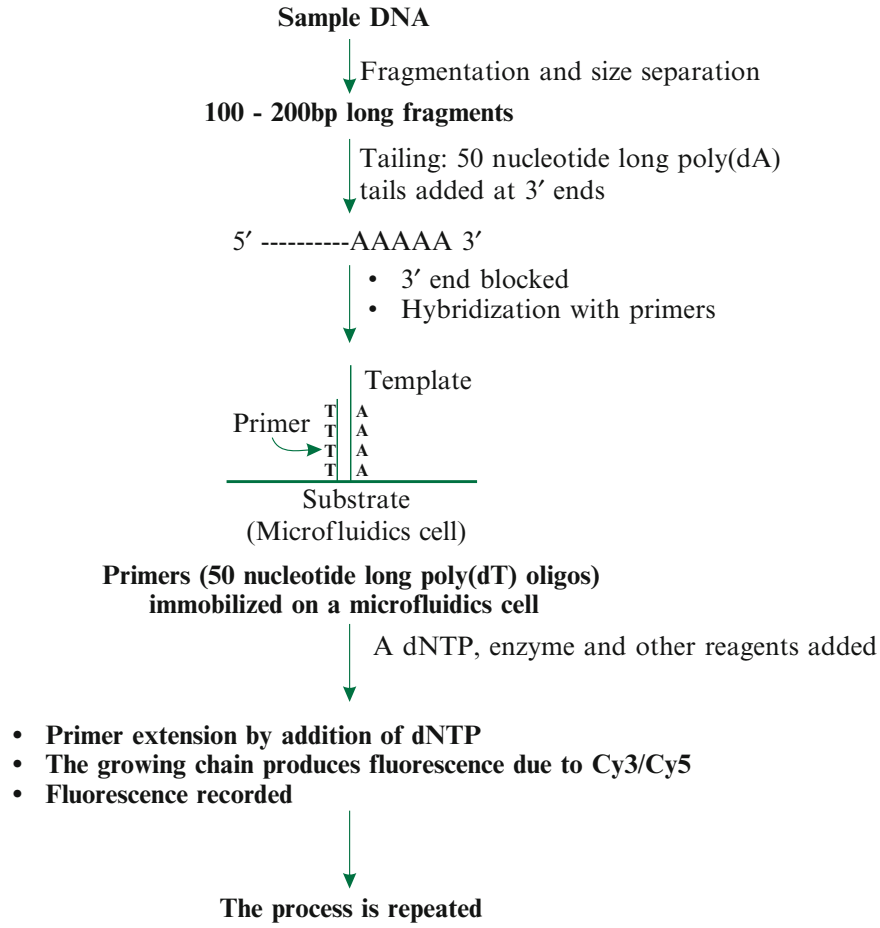


Fig. 4.6 A schematic representation of the Helicos third-generation sequencing method. Use of bright fluorophores (Cy3, Cy5) allows signal detection from replication of a single DNA molecule. This technology has been adapted for direct sequencing of RNA molecules without production of cDNA. In the case of RNA species without 3′ poly(A) tails, poly(A) tails are added to their 3′ ends. Primers (50-nucleotide-long poly(dT) oligos) immobilized on a microfluidics cell; dNTPs labeled with Cy3 or Cy5 and virtual terminators; read length ~35 bases

4.2.3.2 Single-Molecule Real-Time Technology

The *single-molecule real-time* (*SMRT*) *technology* was developed by Pacific Biosciences, USA, and was commercialized as PACBIO RS. This is the most revolutionary approach as it is based on single molecules of DNA polymerase immobilized (by biotin–streptavidin interaction) in zepoliter (10^{-21} L) wells of nanometers in diameter and depth. Each well provides a detection volume of only 20 zepoliters. High concentrations of the four dNTPs labeled with different fluorophores are used for rapid DNA replication. Each DNA polymerase molecule will use a single DNA fragment as template to add the fluorophore-labeled dNTPs to the primer/growing chain (Fig. 4.7). A highly focused detection system continuously records the fluorescence from the nucleotides added to the growing chain in each well. Since the fluorophore is attached to the phosphate moiety, it is automatically removed as the next nucleotide is added, and it diffuses out of the vicinity of DNA polymerase molecule. Since the detection system is focused onto the DNA polymerase molecule, the liberated fluorophore molecules do not interfere with the detection process. The DNA polymerase can sequence the DNA

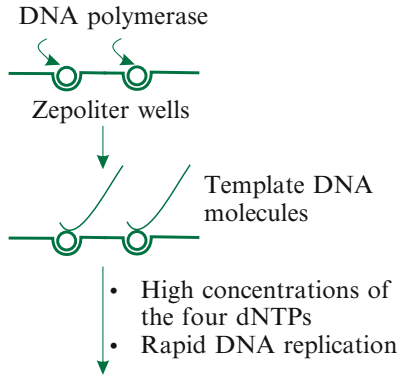


Fig. 4.7 A simplified representation of the single-molecule real-time (SMRT) method of DNA sequencing. Zepoliter $= 10^{-21}$L (Based on Schadt et al. 2010)

fragment more than once, producing multiple coverage of the same molecule (Schadt et al. 2010; Deschamps and Campbell 2010).

The sequencing platform generates 20 Gb sequence data per 30 min. The average read length is ~1,000 bp, while the maximum read length is over 10,000 bp. But an improved technology allows sequencing of up to 20 kb fragments, and efforts are being made to increase it to 40 kb. The raw read error rates may exceed 5 % mainly in the form of insertions and deletions. The use of SMRT bell sample preparation system allows sequencing of both the strands of a DNA molecule in a single cycle, which increases the consensus accuracy of sequence data. It can be used for detection of DNA methylation pattern by using suitable software and for direct RNA sequencing without the need for cDNA preparation. This method uses minimum amounts of reagents and does not require template preparation, and there are no PCR, scan, and wash steps.

4.2.3.3 The Nanopore Sequencing Technologies

In the case of most nanopore sequencing technologies, the DNA molecule and its component bases are passed through an extremely narrow hole (a nanopore), and the component bases are detected by the changes in an electrical current or optical signal caused by them (Schadt et al. 2010). Genetically engineered proteins or a suitable chemical compound may be used to construct the nanopores. The Oxford Nanopore Technologies, UK, uses BASE technology that creates the nanopore by an engineered protein (α-hemolysin). Around 2,000–8,000 nanopores are placed in a lipid bilayer built on a special application-specific integrated circuit chip. At the extracellular face of the nanopore, an exonuclease is attached, while a synthetic cyclodextrin-based sensor is linked at its inside surface; the cyclodextrin acts as the binding site for DNA bases (Fig. 4.8). The DNA sample to be analyzed is restriction digested, the digest is placed onto the chip, and one DNA fragment associates with each nanopore. An enzyme separates the two strands of the DNA duplex, and the exonuclease digests one strand, one base at a time, and passes these bases through the nanopore. Each base sequentially binds to the cyclodextrin located on the inside of the nanopore. This binding creates a disturbance in the electric current passing through the nanopore, which

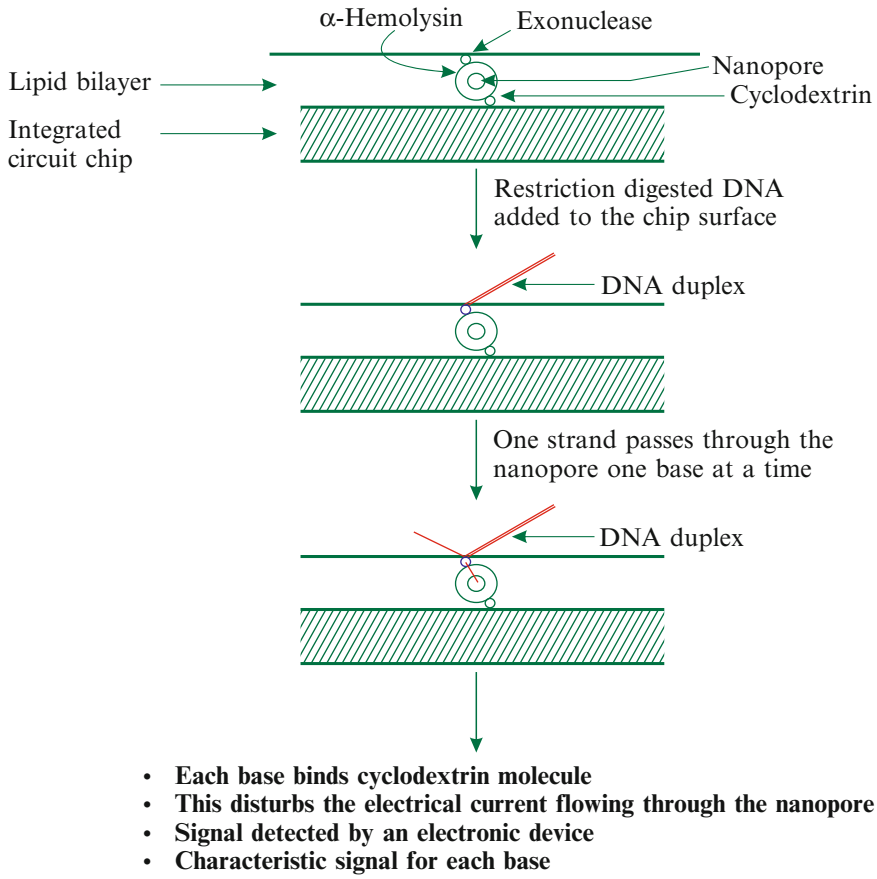


Fig. 4.8 A schematic representation of a nanopore sequencing technology (developed by Oxford Nanopore Technologies, UK). The nanopore is created by an engineered α-hemolysin; exonuclease cleaves the terminal bases one by one and passes them through the nanopore; cyclodextrin binds to the base; this creates disturbance in the electrical current flowing through the nanopore (Based on Schadt et al. 2010)

generates characteristic signal for each DNA base. This signal is sensed by an electronic device and is converted into base sequence data. This technology can detect cytosine methylation without any special chemical processing of the template.

Oxford Nanopore Technologies is preparing to launch two models, namely, MinION and GridION, for sales. MinION USB stick DNA sequencer is the size of a USB drive, is projected to cost less than US $ 1,000, works with a PC, has a lifetime of 6 h from activation, and would generate up to 150 Mb sequence data. The GridION system is designed for bigger runs, uses a standalone machine, and would be able to analyze RNA and protein as well. These systems have an error rate of 4 %.

4.2.3.4 Other Third-Generation Sequencing Technologies

Several other highly innovative third-generation sequencing technologies are in different stages of development, some of which are briefly mentioned here. IBM is developing a DNA transistor that would electronically identify individual bases in a single DNA molecule. NABsys is trying to develop the existing solid-state technologies for whole-genome sequencing based on electronic detection of bases. Genia, on the other hand, is developing a nanopore technology that relies on electrical real-time sequencing of single DNA molecules. The Starlight technology uses fluorescence resonance energy transfer (FRET) for real-time sequencing

of single DNA molecules. Another technology uses a specialized technique with a high-resolution (sub-angstrom) transmission electron microscope for identification of the DNA bases by direct imaging of the base sequence (Edwards 2013).

4.2.4 Comparison Between NGS and TGS Sequencers

The NGS sequencers are simpler to use, very fast, extremely high throughput and comparatively much cheaper, the Illumina Genome Analyzer being the cheapest. In addition, they do not require in vivo cloning and carry out the necessary template preparation in a matter of hours. Finally, they are versatile and can be used for a variety of analyses. The TGS technologies sequence single DNA molecules, are faster and cheaper, and enable a much higher throughput than the NGS sequencers. The error rate of the TGS methods is higher because the opportunity for error removal on the basis of sequencing of multiple copies of each fragment is not available. The NGS sequencers yield shorter read lengths due to the degrading effects of lasers on DNA and enzymes. Further, the washing, which must be done after each cycle, slowly reduces the amount of DNA available for sequencing. Finally, in the case of NG sequencers, asynchronous reactions may increase the error rate, which builds up through the cycles.

4.3 RNA Sequencing

Usually, RNA sequencing involves the production of cDNA by reverse transcription PCR (RT-PCR) and then sequencing of the cDNA product. If the primers for RT-PCR were correctly designed, only the desired mRNA species will be copied as DNA. Initially, Sanger–Coulson method of DNA sequencing was used for sequencing of cDNA/EST (expressed sequence tag) libraries. But this approach does not have high throughput, is expensive, and does not permit quantitative analysis of gene expression. In addition, approaches based on this strategy were generally unable to distinguish among different splicing isoforms.

4.3.1 RNA-Seq

The NGS technologies enable sequencing of complete transcriptomes in almost any population or tissue; this approach is referred to as *RNA-Seq*. RNA-Seq is used for both qualitative and quantitative analyses of genome-wide gene expression. It has also been used to discover up to hundreds of thousands of SNPs (Chepelev et al. 2009; Wang et al. 2009c) at costs similar to those from reduced representation and low-coverage methods (Sects. 13.4 and 13.6). However, RNA-Seq is more likely to discover functional SNPs than other SNP discovery methods. In general terms, the procedure for RNA-Seq consists of isolation of RNA, production of cDNA by reverse transcription, and then sequencing the cDNA population using a suitable NGS technology. The RNA preparation may comprise the total RNA or it may be a specific fraction of the total RNA, e.g., mRNA with poly(A) tails (Fig. 4.9). In the case of long RNA molecules, the RNA molecules themselves or their cDNAs are fragmented to produce, ultimately, cDNA fragments of sizes suitable for NGS sequencing. The fragments are ligated with adapters at one or both the ends, and each fragment is sequenced at one end (single-end sequencing) or both the ends (paired-end sequencing). Typically, reads of 30–400 bases long are obtained, depending on the NGS technology and the sequencing strategy (single-end/paired-end sequencing) used.

The sequence reads are aligned to a reference genome sequence to generate a *genome-wide transcription map* depicting the transcriptional status of all the genes present in the genome as well as their expression levels. But when a reference genome sequence is not available, the reads can still be assembled to produce the transcription map de novo. The available software for mapping of the reads include ELAND, SOAP,

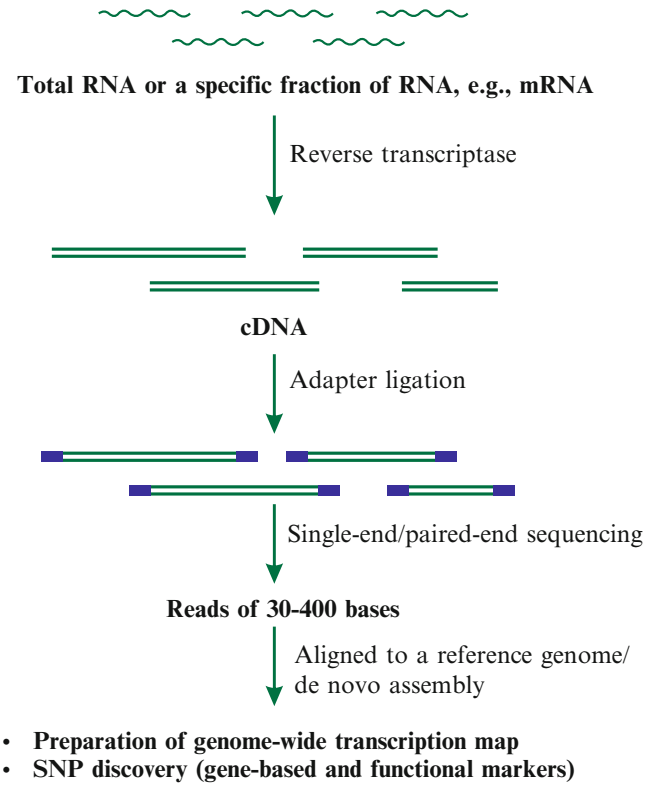


Fig. 4.9 A schematic representation of RNA-Seq technology. The RNA sample may be the total RNA or a specific fraction of RNA, e.g., mRNA. Longer RNA molecules are fragmented either as RNA or as cDNA to generate fragments of suitable size for the NGS technology to be used (Based on Chepelev et al. 2009; Wang et al. 2009c)

MAQ, and RMAP. RNA-Seq allows high-throughput qualitative as well as quantitative analyses of the entire transcriptome. It has revealed many novel features of eukaryotic genomes like overlapping in the 3′-regions of many yeast genes, novel transcribed regions in every genome studied, new splicing isoforms of known genes, 5′ and 3′ boundaries of the transcribed regions of many genes, etc.

RNA-Seq has high resolution, sensitivity, and reproducibility, generates very low background noise, and yields highly accurate quantitative data on gene expression. It requires relatively small quantities of RNA and is particularly suited for transcriptome analysis in non-model organisms. The chief limitation of this approach is the difficulty in inferring genotypes from expression data; this is complicated due to alternative splicing that produces multiple RNA molecules from a single primary RNA transcript. Further, bias may be introduced by cDNA preparation, RNA/cDNA fragmentation, and PCR amplification of the cDNA. For example, the nascent cDNA being synthesized by reverse transcriptase may dissociate from the template RNA molecule and anneal to a new RNA molecule that has a sequence similar to that of the first RNA template. This event, called *template switching*, generates a cDNA molecule made up of the 3′ region of the first RNA template and the 5′ region of the second RNA template. Reverse transcriptases can cause self-priming and, thereby, generate up to 10 % random cDNAs, which are a major source of error. Reverse transcriptases are error prone as they lack proofreading ability. Often the range of dynamic expression may need to be normalized; this becomes problematic when a reference genome is not available. Finally, efficient methods are required for storage, retrieval, and processing of large datasets and for reducing the base sequence errors.

4.3.2 Single-Molecule Direct RNA Sequencing

Helicos BioSciences, USA, has developed and commercialized the technology for direct sequencing of single RNA molecules in a massively parallel sequencing operation by the Helicos® Genetic Analysis System (Ozsolak and Milos 2011). This technology does not involve conversion of RNA to cDNA, PCR amplification, or ligation, uses only minute quantities (several femtomoles) of RNA, and provides deep sequence coverage of the transcriptome. For applications like expression profiling of poly(A)$^+$ RNA encoding genes or mapping of polyadenylation sites, the RNAs are directly used for sequencing. But in studies involving RNA species without poly(A) tails, 3′ polyadenylation of the RNA molecules is carried out. The RNA molecules are now hybridized with the poly(dT) primers immobilized onto the flow cell; the RNA molecules are "filled and locked" and sequenced by synthesis. The read lengths are up to 55 nucleotides (average, 33–34 nt). Each run may yield 800,000–8,000,000 reads per channel, and there are up to 50 channels in the 2 flow cells that can be run simultaneously. Total raw base error rate is 4–5 % (primarily deletions and insertions). A sequence aligner freely available from the Helicos BioSciences HeliSphere significantly reduces the InDel error rates.

4.4 Single-Nucleotide Polymorphisms

Single-nucleotide polymorphisms (*SNPs*, pronounced as "snips") describe variation among different individuals of a species for single base pairs at the corresponding sites of their genomes. Thus a *SNP locus* is a specific position in the genome, at which different nucleotides occur in the same DNA strand of different individuals of the species. *Therefore*, *each SNP locus has to be defined by the sequence flanking the polymorphic nucleotide*. Often insertions and deletions (InDels) are also analyzed as SNPs. The nucleotide polymorphism at a genomic position is considered as SNP only when the least frequent allele has a frequency of 1 % or more. A SNP locus can have four alleles, each allele being represented by one of the four DNA nucleotides (Fig. 4.10). However, many SNP loci have three or even two alleles; in fact, two-allele SNP loci predominate in humans. In any case, SNPs are usually scored as biallelic markers. SNPs are produced by either transition (C to T, A to G, and vice versa) or transversion (A or G to C or T and vice versa). In general, transitions seem to be more frequent than transversions. At least a proportion of C to T (and, consequently, G to A) transitions is produced due to deamination of 5-methylcytosine; this is more likely to occur in genomic regions rich in CpG dinucleotide sequence.

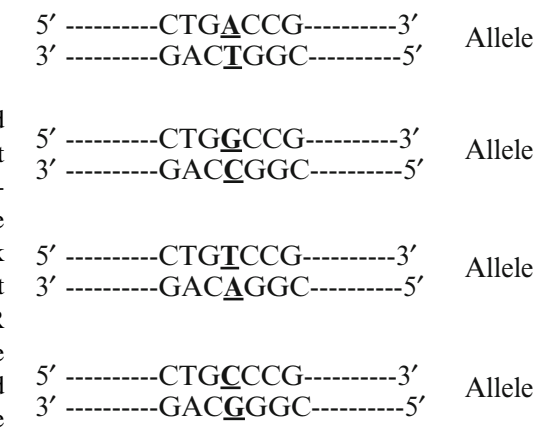


Fig. 4.10 The four possible alleles at a SNP locus. In humans, most SNP loci have only two alleles

SNPs are extremely abundant (about one SNP every 100–300 bp of plant genomes), have relatively low mutation rates, and are relatively easy to detect. SNP density varies among genomes of different species and among different genomic regions of the same species. In general, SNPs are more frequent in noncoding regions than in the coding regions due to a lack of selection pressure in the former. SNPs may generate phenotypic effects by altering either the amino acid sequence of the encoded protein or the splicing pattern of the RNA transcripts. They may also

affect promoter activity and, thereby, generate phenotypic effects. SNPs have proved ideal for automation, and high-throughput marker discovery and analysis; in addition, strategies for combining these two activities have also been developed (Chap. 13). Modern SNP genotyping platforms are supported by improved bioinformatics tools that afford robust automated allele calling and generate high-quality data. This data can be easily shared across groups and stored in common databases irrespective of the genotyping platform used. Many SNP genotyping platforms are capable of efficient, fast, and high-throughput sample processing at increasingly lower cost per data point. However, in the context of plant breeding activities, genotyping cost per sample is more relevant than per data point. Therefore, *a breeder may select an optimal number of SNP loci for each application*; *this might markedly reduce the total genotyping cost*.

The chief limitations of this marker system are the high equipment cost particularly for high-throughput genotyping. The marker development involves resequencing of even whole genomes, which is rather costly. The genotyping procedure is technically demanding and may not be a feasible proposition for many breeding programs. In such cases, it may be desirable to use commercial SNP genotyping services provided adequate funds are available.

4.4.1 Types of SNPs

SNPs are classified in a variety of ways based on different criteria, including genomic location, the effect on phenotype, etc. SNPs located in the noncoding regions of genome are called noncoding SNPs (ncSNPs), while the ncSNPs located in introns are known as intronic SNPs. Exonic SNPs or coding SNPs are found in exons and are comparable to copy SNPs (cSNPs or cDNA SNPs: SNPs discovered in cDNAs). An exonic SNP that does not lead to a change in the amino acid sequence of the concerned protein is called a synonymous SNP (synSNP), while a nonsynonymous SNP (nsSNP) alters the amino acid sequence. A human nsSNP is known as diagnostic SNP, when it is involved in a genetic disease. However, genic SNPs occur in genes and would include intronic and exonic SNPs, as well as the SNPs located in the promoter region of the concerned gene (promoter SNPs or pSNPs). Some of the genic SNPs will affect the function of the concerned gene and would give rise to phenotypic effects; these are termed as functional SNPs or candidate SNPs. But anonymous SNPs do not affect the function of a gene and do not produce a phenotypic effect; most SNPs belong to this category. A reference SNP (refSNP, rsSNP, rsID) is a SNP that serves as a reference point for defining neighboring SNPs. Each refSNP is assigned an rsID number when it is submitted to a databank, e.g., dbSNP. SNPs discovered by mining ESTs or genomic databases are generally called in silico SNPs (isSNPs) or electronic SNPs (eSNPs); these are "virtual" polymorphisms and must be validated by resequencing. Many SNPs located close to each other tend to be inherited together. The alleles of such SNPs located in the same chromosome together constitute SNP haplotype; such SNPs are referred to as haplotype-tagged SNPs. Generally, genotyping only a small number of carefully selected SNP loci from a haplotype block allows the deduction of genotypes at the remaining SNP loci of the block; these SNPs are termed as "tag" SNPs.

The SNPs in polyploid species have been classified as simple SNPs and hemi-SNPs or homoeo-SNPs. A simple or true SNP detects allelic variation between homologous loci of the same genome present in the same or different polyploid species, and it does not detect differences in their other genome(s). This group of SNPs would show typical diploid segregation in most mapping populations, is quite frequent (10–30 %), and would be the most useful for mapping. But hemi-SNPs or homoeo-SNPs, on the other hand, detect homoeologous/paralogous loci in the two or more genomes of the polyploid species and of their diploid progenitors. Therefore, these SNPs are of limited value for mapping (Deschamps and Campbell 2010).

4.5 Methods for Discovery of SNPs

It may be pointed out that all SNPs are initially discovered by sequencing, which remains the method of choice. Sequencing may involve the whole genome, a specific region of the genome, or the transcriptome. One of the major problems in SNP discovery has been the predominance of highly repetitive sequences in plant genomes. Therefore, early efforts at SNP discovery attempted to avoid repetitive sequences by resequencing unigene-derived amplicons and in silico SNP discovery by mining the EST databases, followed by their PCR-based validation. However, these approaches detected gene-based SNPs and did not discover SNPs in the noncoding regions of genes and the intergenic spaces. In addition, amplicon resequencing is expensive and labor intensive. Similarly, many of the SNPs discovered from EST databases were non-allelic in several crops because these SNPs represented paralogous sequences produced by duplication of the concerned genomic regions.

Prior to the development of NGS technologies, whole-genome sequencing was a daunting task. Therefore, it was highly desirable to minimize the sequencing effort by focusing on the genomic regions of interest; amplicon sequencing (Sect. 4.5.1) and sequence capture (Sect. 4.5.6) strategies serve this purpose. But the emergence of NGS technologies has made SNP discovery by whole-genome sequencing a feasible option. In addition, several reduced representation strategies aim at combining SNP discovery with SNP genotyping, using a suitable NGS technology, at reasonable costs (Chap. 13). Further, huge amounts of genome and EST sequence data have accumulated in various databases, which can be mined for SNP discovery. Transcriptome sequencing by RNA-Seq technology is also being used for SNP discovery.

4.5.1 Amplicon Sequencing

In this approach, a pair of specific primers is used for PCR amplification of the desired genomic region, and the PCR product (*amplicon*) is sequenced for identification of SNPs and InDels. This strategy limits the sequencing and analysis efforts to the genomic region of interest and, thereby, reduces the workload. When Sanger–Coulson sequencing is used, separate amplification of each amplicon is necessary. Further, in the case of heterozygotes or when PCR amplification is based on pooled DNA, the amplicons have to be cloned (to separate the amplicons representing the two alleles present in the heterozygotes or in different individuals) before they are sequenced. But the NGS technology has rendered these steps unnecessary since each read is generated from a single amplicon. Therefore, NGS technology permits pooling of tissues, genomic DNAs/cDNAs, or amplicons from different individuals. This approach reduces the quantum of work for template preparation and permits the discovery of all the SNPs, including the rare alleles. In case of pooling, a greater depth of sequencing should be used; in general, the depth would be greater for shorter read lengths. According to an estimate, at least 34×, 101×, and 110× sequencing depth would be needed with 454, ABI SOLiD, and Illumina GA, respectively, for separating sequencing errors from genuine SNPs. However, pooling does not permit determination of marker genotypes and haplotypes of the individuals/lines. In addition, PCR amplification of pooled DNA may lead to preferential priming of certain alleles. These difficulties can be removed by separate PCR amplification of each individual/line, using separate barcodes for each of the amplicons and then pooling the amplicons before sequencing. This approach increases the amount of work for template preparation as well as the total cost. However, it increases the usefulness of data as it combines marker discovery with genotyping of the individuals/lines.

Read length and sequencing depth are critical for detection of rare alleles, identification of InDels, and for eventual marker development. Deep sequencing minimizes false negatives, and ensures detection of genuine SNPs and discrimination of rare alleles from sequencing errors. Short reads enable discovery of InDels of 1–8 bp, while longer reads from 454 sequencing

platform permit identification of InDels of 1 to over 97 bp. Amplicon sequencing can be extended to even such species, for which sequence information is not available. For achieving this, trans-specific or universal primers are designed on the basis of conserved sequences of the target genes extracted from a related species for amplification of orthologous genes in the uncharacterized species. This approach has been successfully used in some plant species. *Orthologous genes* are those genes of different species that perform the same function. In contrast, *paralogous genes* are the genes present in the genome of the same species and have the same function; these genes are produced by duplication, polyploidization, or both.

The limitations of amplicon sequencing include size limit of amplicons (10–20 kb for long-range PCR), base substitutions due to PCR, requirement of sequence information for primer designing, amplification of paralogues by the specific primers based on sequences conserved among paralogues, overrepresentation of amplicon ends in reads, and uneven coverage of internal regions of amplicons. Many of these problems can be mitigated by suitable strategies, including rigorous quality control during sample preparation and bioinformatics tools.

4.5.2 SNP Mining

The simplest, most convenient, and highly efficient method for SNP discovery is bioinformatics analysis of the ever-increasing genomic and/or EST sequences of different individuals available in the databases of the concerned species. In addition, an investigator may sequence genome/ESTs of a genotype/line/individual of interest and analyze the sequence so obtained along with the sequences available in the database. Bioinformatics tools like PolyPhred are used for deducing the base sequence of fragments, assembly of the deduced sequences into contigs, and editing of the contigs. Suitable computer software like SNP Pipeline are then used to align the sequences and detect SNPs. Sequencing errors present in the database, particularly in the genomic regions that are not well characterized, may lead to discovery of false SNPs. Special software like POLYBAYES help minimize false discovery of SNPs due to sequencing errors. The analysis of genomic sequences will identify SNPs located in both coding and noncoding regions of the genome, while EST analysis will discover SNPs only in the coding regions. At present, most of the SNP mining activity is directed at EST databases, possibly for the above reason. Further, the search may be focused at specific regions of the genome that have been either known to be associated with the traits of interest or to contain genes with specific functions. The SNPs discovered by SNP mining are often termed as in silico SNPs (isSNPs) or electronic SNPs (eSNPs). However, *these SNPs must be validated by resequencing*.

4.5.3 Transcriptome Sequencing

Transcriptome sequencing allows rapid and inexpensive discovery of genic SNPs and avoids highly repetitive genomic regions. The NGS, RNA-Seq, and direct RNA sequencing technologies can be used for transcriptome sequencing. The sequence reads are aligned to a reference genome or to EST sequences to discover SNPs and InDels. In case a reference genome is not available, genome sequence of a related species or of the parental species may be used for sequence alignment and marker discovery. Alternatively, the sequence reads can be assembled de novo using appropriate bioinformatics tools. Analysis of EST/transcriptome sequence data also permits discovery of SSR markers. For example, an analysis of watermelon EST sequences obtained from an experiment and the EST datasets obtained from the GenBank enabled the discovery of 5,000 SSRs. Useful markers can also be found in the 3′ UTRs (untranslated regions) of mRNAs. In general, longer sequence reads are preferred for marker discovery as they facilitate sequence alignment and discrimination among paralogues in the case of polyploid species. Paired-end reads overcome

to some extent the above limitations of short reads.

Markers discovered by transcriptome sequencing will be, of necessity, gene-based markers, and a proportion of them will be functional markers. But many QTLs, regulatory sequences like enhancers, locus control regions, etc., are located in noncoding regions of the genome. As a result, it will be unable to discover markers useful for mapping of such QTLs and regulatory elements. Transcriptome sequencing coupled with appropriate experimental design would permit the determination of allele-specific differences in gene expression, estimation of the parental contributions to heterosis, and the role of genetic imprinting in development and performance. However, transcriptome analysis-based marker discovery is limited to only those genes that are transcribed in the concerned tissue/organ during the given developmental stage and under the environmental conditions prevailing at the time of sample collection. Therefore, a fair number of organs/tissues, developmental stages, and environments should be sampled to ensure the representation of most, if not all, of the genes present in the genome of the concerned species. In contrast, sequencing of hypomethylated partial restriction genomic libraries (Sect. 4.5.5) provides a more complete representation of SNPs located in genes than transcriptome sequencing and allows the detection of SNPs situated in introns, regulatory regions, and non-transcribed genes.

Assembly and analysis of NGS data requires appropriate software programs, for which a variety of options are available. A de novo assembly of RNA-Seq sequence data yields contigs, which are called *tentative ESTs* or *tentative unique sequences* (*TUSs*). Bioinformatics tools are used to filter the SNPs discovered from RNA-Seq data, and the filtered SNPs are usually validated by Sager–Coulson sequencing. For example, in one study in maize, transcriptome analysis of shoot apical meristems from two inbreds permitted the detection of 36,000 putative SNPs; these were reduced to 7,000 after stringent processing of the sequence data. Sager–Coulson sequencing was used for confirmation of a sample of 110 from these SNPs, and 85 % of them were successfully validated. Transcriptome analysis of polyploid species requires a more complex experimental design and a comparison with the diploid ancestral species for assigning the tentative ESTs to the homoeologous chromosomes. Unlike the concerned genomes, the RNA transcripts rarely contain repetitive sequences, which is a definite advantage in proper sequence alignment. The error rate of NGS sequence data is rather high, but availability of a good quality reference genome considerably reduces the error rate. It is important to use an optimum sequencing depth because a low sequencing depth would lead to higher error rate and "false negatives." Transcriptome sequencing has been successfully used in several crop species, including maize, canola, sugarcane, wheat, etc.

4.5.4 Whole-Genome Sequencing

Often SNP discovery has been based on whole-genome sequencing of a small number of selected individuals/lines; this approach remains the method of choice wherever resources and other considerations do not preclude this option. *SNP discovery is greatly facilitated by the availability of a good quality reference genome sequence*. One may reduce the sequencing effort by pooling DNAs from the selected individuals/lines and constructing a genomic library from the pooled DNA. Random clones may be picked from this library and used for sequencing. The shotgun sequences so obtained are processed, using appropriate bioinformatics tools for discovery of SNPs. It may be pointed out that the sequencing depth should be large enough (at $>20\times$ coverage) not only to yield sequence data with minimum error but also to ensure that sequence of a given genomic region is available from enough number of individuals/lines to allow SNP discovery. The term sequencing depth may refer to specific nucleotides or to the entire genome. *Sequencing depth for a specific nucleotide* represents the total number of all reads, in which a given genomic position (or base pair) from a given individual is

represented; these reads may be obtained from a single sequencing experiment or from a series of experiments. But *sequencing depth for the whole genome* is the average number of times each base of the genome (the entire genome) of an individual has been sequenced. The sequencing depth for the whole genome is generally referred to as *coverage of sequencing* and is denoted as 10×, 20×, 30×, etc., coverage or depth. The general formula for coverage (C) is $C = LN/G$, where L is the read length, N is the total number of reads, and G is the length (in bp) of the haploid genome of the concerned species. It should be kept in mind that coverage denotes average sequencing depth of the genome as a whole; therefore, some genomic regions would be sequenced at much higher depth, while some others would be sequenced at much lower depth than the coverage level. The minimum coverage level required for a study depends on many factors, including the type of study, gene expression level, the trends in published literature, etc.

The analysis of sequence data for SNP discovery proceeds in several steps. In the case of NGS data, the first step involves image analysis and base calling with the minimum error rate. This can significantly reduce false-positive SNP calls and facilitate sequence assembly, particularly when the coverage is low to moderate. The short sequence reads are then aligned onto the reference genome whenever it is available; this is known as *read alignment* or *read mapping*. The alignment algorithms should be able to handle both sequencing errors, as well as potentially real sequence differences, in the form of base substitutions and InDels, between the reference and the newly sequenced genomes. In addition, the aligners should generate well-calibrated alignment quality values, which are important for variant calling, i.e., determining the genomic positions at which at least one base differs from the reference genome. It has been recommended that Novoalign or Stampy should be used as aligners, and GATK or SOAPsnp should be used for recalibration of per base quality scores (Nielsen et al. 2011). This is followed by realignment of reads, removal of duplicate reads, and a recalibration of the quality scores for each base. Both SNP and genotype calling at a given genomic position depend on the accuracy of base calls as well as the per-base quality scores of the reads overlapping the genomic position.

SNP calling is the determination of the genomic positions at which nucleotide polymorphisms occur. It can be based on data from a single individual/line (*single-sample calling*) or it may simultaneously use data from all individuals in the sample (*multi-sample calling*). As far as possible, multi-sample calling should be used, and the calling methods should involve likelihood ratio tests or Bayesian procedures. Similarly, *genotype calling*, i.e., assigning of SNP alleles to different individuals in the sample, should be done by combining the data from all the individuals in a Bayesian framework, and information on known SNPs (e.g., those listed in dbSNP), linkage disequilibrium (LD), etc., should be included to improve the accuracy of genotype and SNP calls. A number of filtering steps based on a variety of criteria like generally low-quality scores, systematic differences in quality scores of major and minor alleles, aberrant LD patterns, strand bias, etc., may be implemented to improve the accuracy of SNP and genotype calls. Most of the software used for NGS processing carry out both SNP and genotype calls (Table 4.2). Several additional steps like local realignments, combining results from multiple SNP- and genotype-calling algorithms, etc., can be implemented to improve genotype calls. Finally, uncertainty should be incorporated in the subsequent statistical procedures for analyzing the data. It may be pointed out that analysis of NGS data is evolving rapidly, and new tools for data analysis are being continuously developed. Therefore, the choice of most suitable software package for any task will keep on changing with time.

4.5.5 Reduced Representation Approaches

Genome sequencing of a sample of individuals/lines has to be resorted to when either genome sequences are not available or it is desirable to

use genome sequences of a set of new lines/individuals. Sequencing of whole genomes is the ideal strategy, but it involves considerable time, effort, and financial and other resources. Further, sequencing of whole genomes may not be necessary for many types of studies. In view of the above, many strategies for simultaneous SNP discovery and genotyping have been developed (Sects. 13.3–13.6). In general, these methods sample a fraction of the whole genome for sequencing so that the cost and effort for marker discovery and genotyping are greatly reduced. One approach aims to enrich the sampled fragments with gene-rich regions by construction of a hypomethylated partial restriction (HMPR) library as follows. The genomic DNA of the target individual/line is digested completely with a 5-methylcytosine-sensitive restriction enzyme like *Hpa*II (5′ C/CGG 3′) with a 4 bp recognition sequence. The digest is subjected to electrophoresis; fragments of 100–600 bp are separated and used for sequencing by an NGS technology. The genomic regions having repetitive DNA are usually hypermethylated; consequently, they will be present as much larger fragments and will be excluded. This approach may eliminate ~95 % of the maize genome and enrich the selected fragments four- to five-fold for genic sequences. Sequencing of the gene-enriched fragments from two maize inbreds allowed the identification of a large number of putative SNPs. However, it restricts SNP discovery to the regions near the recognition sites of the enzyme used for digestion. Therefore, two or more 5-methylcytosine-sensitive restriction enzymes with distinct recognition sequences should be used to get a more complete representation of the genic regions.

4.5.6 Sequence Capture

Sequence capture is a targeted SNP discovery strategy applied to specific genomic regions. This strategy can be applied when the genomic region of interest is known and a closely related reference genome sequence is available. It involves designing of oligonucleotide probes or primers specific for the genomic regions of interest to permit their separation or amplification before sequencing. There are three main strategies for sequence capture, viz., SureSelect from Agilent, SeqCap from NimbleGen, and the Targeted Sequencing System from RainDance (Davey et al. 2011). All the three technologies are proprietary, require the customer to provide the target region sequences, and use in-house bioinformatics tools to design probes/primers for the target regions. The *NimbleGen SeqCap technology* uses oligonucleotide probes synthesized on microarray slides, and the lengths of the probes are adjusted to obtain a uniform melting temperature. The genomic DNA fragments are hybridized with the microarray and the captured fragments are used for sequencing. The *Agilent SureSelect* method, on the other hand, implements in-solution target sequence capture using biotinylated RNA probes of 120 nt. The genomic DNA fragments already ligated to sequencing adapters are hybridized with the probes, and the hybridized fragments are separated by exploiting the high affinity of biotin for streptavidin. In both these technologies, repeat sequences are filtered out from the probe set by using specific software programs.

The *RainDance Targeted Sequencing System*, in contrast, uses two rounds of PCR to specifically amplify fragments from the targeted genomic region. This is achieved by designing a set of PCR primer pairs using proprietary software so as to cover most of the genomic region of interest. Each primer pair of the set has at its 3′ end the sequences specific for a segment of the target genomic region, while its 5′ end comprises partial sequence of the adapter for the selected NGS technology. The target-specific primers are used for the first round of PCR amplification. In the second round of PCR, universal primers with the partial NGS adaptor sequences at their 3′ ends are used. The PCR products generated from the second round of PCR are directly used for sequencing. The SureSelect and SeqCap methods capture about 90 % of the targeted genomic region, while the Targeted Sequencing System may capture over 95 % of the region. These

technologies do not appear to introduce a substantial bias in allele representation in the sequence data. The reference genome used for designing the probes/primers should be of high quality and closely related to the population under study.

4.5.7 Validation of Discovered SNPs

Once a group of SNPs has been discovered, each locus should be evaluated to ascertain the following: (1) that it is a true SNP and not a product of sequencing error, faulty read alignment, etc., (2) that its alleles represent homologous genomic regions and not paralogous/homoeologous regions, and (3) that it segregates in a typical Mendelian fashion. The above evaluation is generally referred to as *SNP validation*. One approach for SNP validation is to resequence the concerned genomic regions of carefully selected individuals/lines to confirm that the discovered SNPs represent true polymorphisms. A more informative validation process involves designing a suitable assay for the discovered SNPs and to apply this assay to evaluate a set of diverse germplasm lines or, preferably, a segregating population. This procedure will reveal the discovered SNPs to be real or false, their ability to discriminate among the germplasm lines, and their segregation pattern in the segregating population. The choice of assay will depend mainly on the number of SNPs to be validated. The assays in common use for a large number of SNPs are Illumina's GoldenGate (Sect. 13.2.8) and Infinium assays, TaqMan OpenArray Genotyping system (Sect. 13.2.4), and Kompetitive Allele-Specific PCR (KASP) assay (Sect. 13.2.3). The length of SNP context sequence, the total number of SNPs to be genotyped, and the available funds would have to be considered while selecting one of these assays. *It may be pointed out that SNP validation still remains a cost-intensive procedure*. SNP validation in allopolyploids would be facilitated by the use of haplotype and allele frequency information, and application of bioinformatics tools like HaploSNPer. This strategy would be useful even for diploid species like barley that have highly repetitive genomes.

4.6 Methods for SNP Genotyping

The various SNP genotyping methods/platforms range from scoring of a single SNP marker to a very large number of markers assayed using high-density SNP chips, and they are suited for a wide range of applications. These methods rely on strategies that are able to distinguish between a perfect match from a single base mismatch between an oligonucleotide and the template DNA strand. These strategies are based on nucleic acid hybridization, primer extension, oligonucleotide ligation, DNA replication, or single-strand invasion coupled with cleavage of the displaced strand (Sobrino et al. 2005). The different genotyping methods include allele-specific PCR, 5′-nuclease assay, high-density oligonucleotide arrays or DNA chips, bead-based techniques, primer extension, invasive cleavage or invader technology, MALDI-TOF MS-based homogeneous MassEXTEND (hME) assay, and pyrosequencing. In addition, PCR products can be subjected to restriction enzyme digestion (cleaved amplified polymorphic sequences, CAPSs; Sect. 3.14), electrophoresis of single strands (single-strand conformation polymorphism, SSCP; Sect. 3.15), or denaturing gradient gel electrophoresis (D/TGGE; Sect. 3.16) for genotyping of known SNPs. These techniques can be broadly classified into gel-based and gel-free assays; the latter group of assays is preferred because the methods in this group are amenable to high-throughput analysis leading to economy of time and other resources.

4.6.1 Allele-Specific PCR

Allele-specific PCR is designed to amplify only one of the alleles at a SNP locus (Okayama et al. 1989). It uses a pair of primers, one of which is based on a conserved sequence present in all the alleles. The other primer of the pair is specific to the genomic region having the SNP

locus, and the base at its 3′-end corresponds to the SNP locus. When the 3′ terminal base of the second primer is complementary to the SNP allele, it pairs with the allele and supports amplification of the genomic region and yields a PCR product detectable by gel electrophoresis (Fig. 4.11a). A mismatch at the 3′ end of the primer greatly reduces the chances of amplification so that no amplification product would be detectable (Fig. 4.11b). Therefore, allele-specific PCR generates a dominant STS marker scored as "present"/"absent." But sometimes a single-base mismatch at the 3′ end of a primer is unable to prevent amplification. On the other hand, sometimes amplification may fail due to an error in the setting up of the experiment. The first difficulty is resolved by introducing a mismatch at the second base from the 3′ end of the primer (Fig. 4.11c). This mismatch will still allow amplification of the allele for which the primer has been designed but will effectively prevent amplification of the other alleles. The second problem can be overcome by using four different primers, i.e., one primer for each SNP allele, and screen every

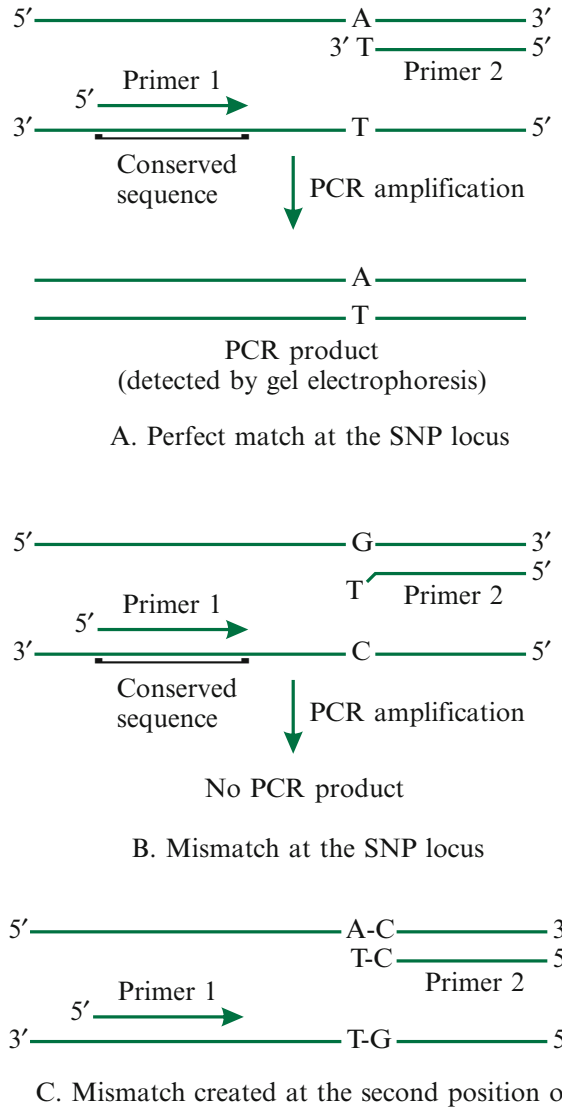


Fig. 4.11 A simplified representation of allele-specific PCR for genotyping of a SNP locus. The mismatch at the second base from the 3′ end of the primer 2 increases effectiveness of allele discrimination. This does not prevent amplification in case the 3′ terminal base is matched with the SNP allele, but it definitely prevents amplification in case of a mismatch

individual with all the four primers. If there is an experimental error, amplification will fail with all the four primers. But when there is no error, one primer is expected to generate amplification product in every individual. The four primers can be designed in such a way that each of them amplifies product of a different length, or a different fluorophore may be attached to each primer. This would allow all the four primers to be used in a single PCR tube for each individual. The allele-specific PCR is a user-friendly approach for SNP analysis by any laboratory with PCR facility. However, the overall throughput is low, and only a small number of SNPs can be analyzed by this approach. This strategy has been modified as KASP™ genotyping assay for a high-throughput SNP and InDel genotyping (Sect. 13.2.3).

4.6.2 5′-Nuclease Assay (TaqMan® Assay)

This technique gets its name from the fact that it uses the 5′-nuclease activity of Taq polymerase in real-time PCR to quantify the hybridization of allele-specific oligonucleotides with the genomic DNAs of the test individuals and deduces the SNP allele from this information. It uses two PCR primers for amplification of the target sequence, i.e., the genomic region containing the SNP locus, and a specifically designed probe, TaqMan™ probe, complementary to that region of the target sequence that has the SNP locus (Livak 1999). This probe has a fluorescent dye attached to its 5′ end and a quenching dye linked to its 3′ end. As long as the fluorescent dye molecule is located near the quenching dye molecule, there will be no fluorescence due to the quenching action of the latter. The base at the 5′ end of the probe is complementary to the SNP allele it detects. In case the 5′ end of the probe is paired properly with the SNP allele present in the target sequence, the 5′-nuclease activity of Taq polymerase will cleave the whole probe beginning at its 5′ end. This will free the fluorescent dye molecule, which will diffuse away from the quenching dye; as a result, it will now generate fluorescence (Fig. 4.12). But if the base at the 5′ end of the probe is not complementary to the SNP allele, there will be mismatch, and Taq polymerase will not be able to cleave the probe at its 5′ end. As a result, there will be no fluorescence. One may design one TaqMan probe for each allele at an SNP locus, label them with different fluorophores, and use them in a single PCR tube. In such a case, the ratios between the fluorescence of different colors will permit a highly reliable scoring of the SNP alleles.

The TaqMan® assay is homogeneous, quick (on an average, 2 h per run), and simple, PCR and data calling occur simultaneously in real-time mode, and the throughput is high. The assay generates 2,000 data points per day per person in a monoplex mode; it can also be run in a duplex mode to generate up to 3,000 data points per day per person. However, the procedure is based on a costly real-time PCR machine, and the costs of labeled probes and other consumables are high. The TaqMan® assay has been commercialized as the high-throughput TaqMan OpenArray Genotyping system (Sect. 13.2.4) by Applied Biosystems, USA. It has also been adapted for a cost-effective medium multiplexing, high-throughput SNP genotyping platform using nanofluidic dynamic arrays (Sect. 13. 2.6).

4.6.3 Molecular Beacons

Molecular beacons are specially designed oligonucleotide hybridization probes used for identification of SNP alleles. The central region of a molecular beacon is complementary to the sequences flanking the target SNP locus, including the SNP allele to be detected (Sobrino et al. 2005). The sequences on either side of the central region are universal sequences, and they are complementary to each other. A fluorophore is attached to the 5′ end of the probe, while a quenching dye is attached to its 3′ end. The probe molecules will form a hairpin structure due to pairing between their 3′ and 5′ end regions. This pairing will bring the quenching dye in close

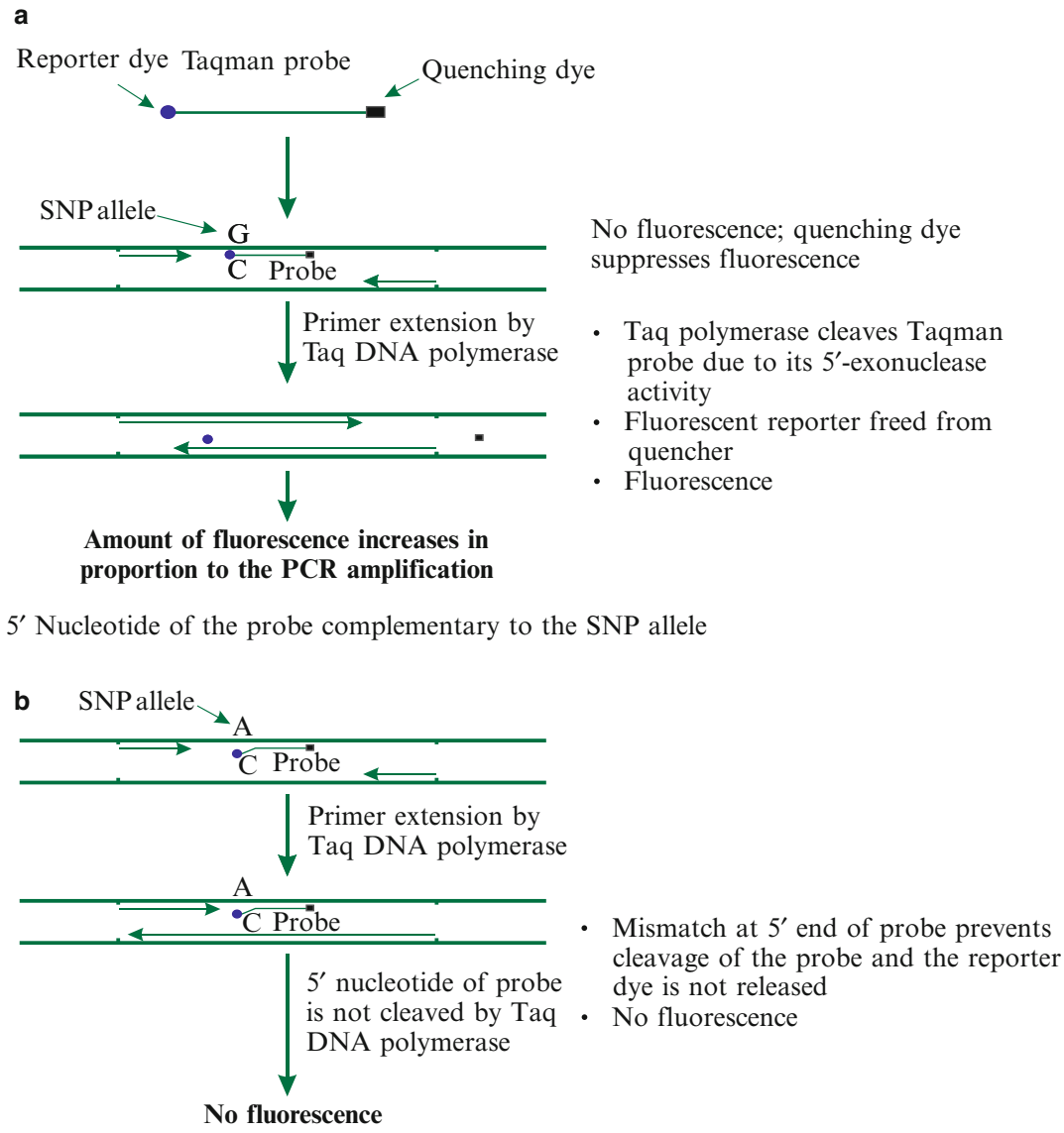


Fig. 4.12 Use of TaqMan™ probe to quantify PCR products. Primer 1 and Primer 2 are the two PCR primers; Probe is TaqMan™ probe that has a fluorescent reporter at its 5′ end and a quencher at the 3′ end. (**a**) 5′ Nucleotide of the probe is complementary to the SNP allele: fluorescence is produced as the reporter dye is released by 5′ exonuclease action of Taq DNA polymerase. (**b**) 5′ Nucleotide of the probe is not complementary to the SNP allele: mismatch at 5′ end of the probe prevents its cleavage and the release of the reporter dye. Therefore, fluorescence is not produced

proximity to the fluorophore, due to which there will be no fluorescence (Fig. 4.13a). But when the probe base pairs with the specific SNP allele, it becomes linear and the quenching dye becomes separated from the fluorophore, and fluorescence is generated (Fig. 4.13b). A molecular beacon is mixed with denatured PCR product representing the concerned genomic region of the test individual/line and allowed to anneal. If the allele at the target SNP locus is complementary to the beacon, the two will base pair and there will be fluorescence (Fig. 4.13b). But if the SNP allele were not

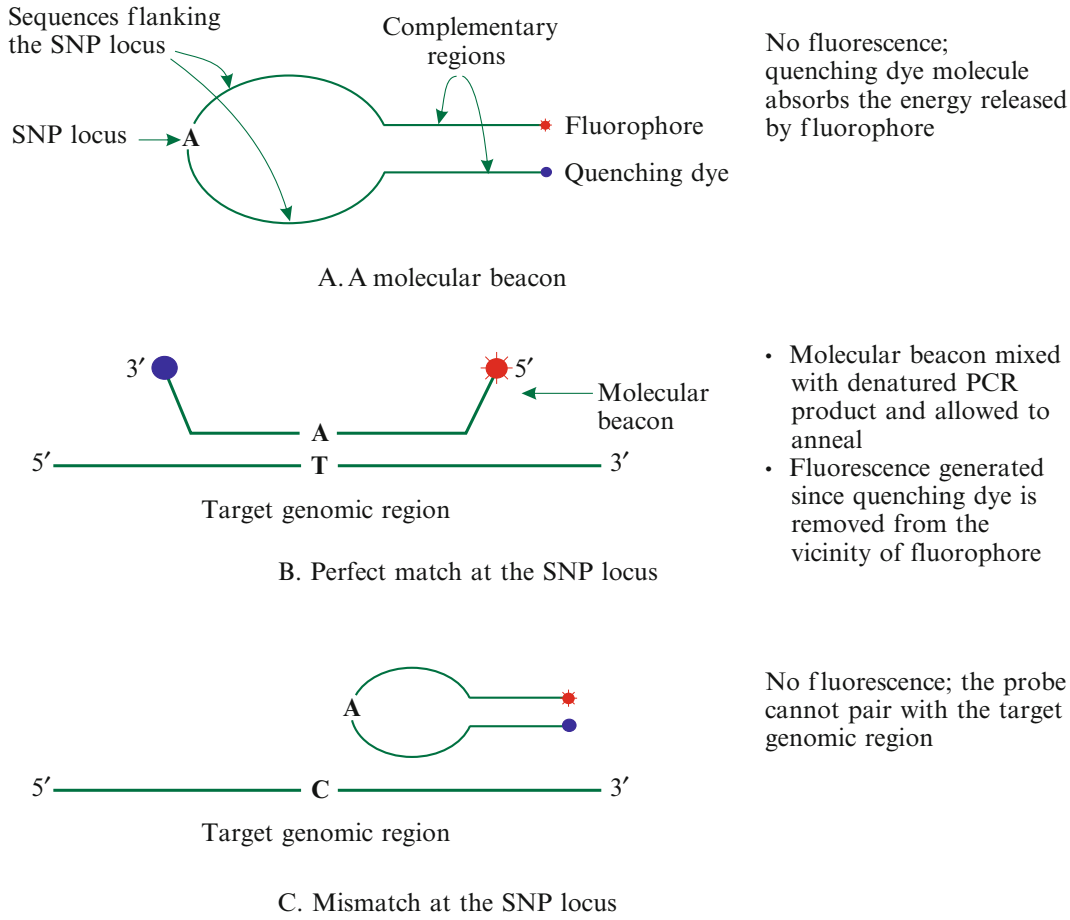


Fig. 4.13 A schematic representation of molecular beacon and its use for detection of SNP alleles. Molecular beacon is mixed with denatured PCR product representing the concerned genomic region, allowed to anneal, and then fluorescence is monitored (Based on Sobrino et al. 2005)

complementary to the beacon, there will be no pairing and fluorescence (Fig. 4.13c). A suitable sensing device detects the fluorescence signal, which is used to deduce the SNP allele.

Some degree of multiplexing can be achieved by labeling two or more molecular beacons, specific for different SNP loci, with different fluorophores and using them in a single reaction vessel. However, most detection systems use monochromatic light for excitation of fluorophores, which limits the number of different fluorophores that can be assayed together efficiently. One strategy to overcome this difficulty employs two fluorophores, one harvester fluorophore and one emitter fluorophore arranged serially, at the 5′ end of the probe in the place of single fluorophores used normally.

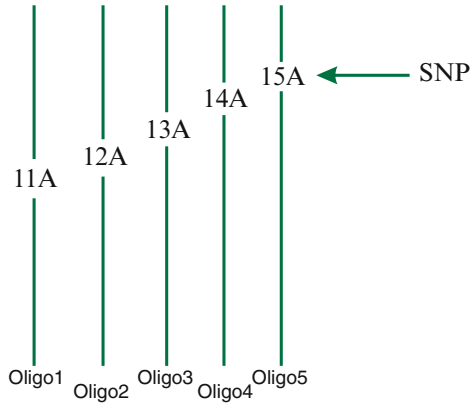
4.6.4 Microarray-Based SNP Genotyping

Microarray-based SNP genotyping requires the development of SNP microarrays/DNA chips for simultaneous genotyping at several SNP loci. A subset of the polymorphic SNP loci is selected mainly on the basis of their position in the genome, the level of polymorphism and suitability for the assay, and used to construct a microarray. A *microarray* or *DNA chip* is a

small plaque/wafer of silicon, glass, or metal, onto which one end of a large number of different single-stranded DNA molecules is covalently linked and arranged in spots (Appendix 2.3). Each spot has several copies of a single DNA molecule of 25 nt representing the SNP locus and includes the nucleotide involved in the SNP around its middle position. In order to ensure high reliability, each SNP allele is represented by five different oligonucleotides; in each of these oligonucleotides, the variable SNP is located at a different position, ranging from two bases on one side of the central base to two bases on the other side (Fig. 4.14a). At the same time, each of the oligonucleotides is spotted at two to three different locations (Fig. 4.14b), which serve as replications and help eliminate false-positive signals (possibly due to nonspecific hybridization). It may be pointed out that the SNP locus and the sequences surrounding this locus influence the hybridization efficiency. Therefore, it is very difficult to optimize the conditions for detection of a panel of SNPs using an array, and ingenious approaches are used to overcome this difficulty (Sobrino et al. 2005).

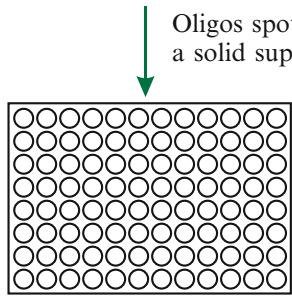
Genomic DNA from each individual to be genotyped for SNPs is used for a series of PCR reactions to amplify all the short genomic regions having the different SNPs. For this reason, each SNP locus is first converted into an STS by designing a pair of primers for its reliable amplification. The PCR products are labeled by fluorescence, and all the PCR products from a single individual are pooled and used for hybridization with the DNA chip (Fig. 4.14). The non-hybridized PCR products are removed by washing under such conditions that permit only perfectly base-paired PCR products to remain associated with the oligonucleotides spotted onto the chip. A fluorescence scanner is used to measure fluorescence at each spot on the chip, and the data are analyzed with the help of image analysis software. Since the position of each oligonucleotide on the chip is known, the alleles present at different SNP loci are readily deduced. This approach simultaneously analyzes all the SNP loci of the test individual/line. Microarrays can also be used for simultaneous genotyping of a large number of individuals/lines at a given SNP locus. This type of assay may be needed in certain situations, e.g., during MAS. In such a case, the PCR products representing the concerned SNP locus from individual plants of the relevant segregating generation are spotted onto a glass slide. This microarray is hybridized with labeled probes representing the alternative alleles of the concerned SNP locus, and plants with the desirable SNP allele are identified. This technique is referred to as *tagged microarray marker approach*; it has been successfully used in the case of humans and pea.

Wang et al. (1998) were the first to use DNA chips for SNP genotyping in humans. The microarrays have to be custom made for every species and whenever the panel of SNP loci is altered. The development of chips involves considerable amount of work like designing of STS primers for each SNP locus and construction of the oligos for every locus. Since hybridizations for all the SNPs are carried out simultaneously under the same conditions, the oligos must be designed with great care so that they all have identical requirements for perfect hybridization. This requires considerable expertise, and specialized software have been developed for this purpose. The synthesis of oligos onto the chips requires the construction of expensive "masters" for each set of oligos. Therefore, the initial development of SNP chip for a species is very costly, but the subsequent production of a large number of identical chips may be much cheaper. As a result, SNP chips are relevant only for large projects. The efficiency of discrimination between completely matched and mismatched oligos in hybridization is much lower than the ability of DNA polymerases or DNA ligases to distinguish between them. This problem is particularly aggravated in the case of microarrays since many different oligos need to be hybridized under a single set of conditions; this adversely affects the accuracy of genotype calls. Therefore, universal microarrays that can be used in any species with any set of SNPs have been developed, e.g., the Illumina's "Sentrix Array Matrix" for the GoldenGate assay (Sect. 13.8).



The numbers 11, 12, 13, etc. denote the base position corresponding to the SNP allele in the 25 nucleotide long oligos

A. Five oligos representing a single SNP locus



The five oligos are spotted in three replicates

B. Microarray

- **PCR amplification of the genomic regions corresponding to all SNP loci of an individual/line**
- **PCR products labeled with fluorescence**
- **PCR products from one individual/line pooled and hybridized with probes on the chip**
- **Washing leaves only perfectly paired PCR products hybridized with the probes**
- **Fluorescence measured at each spot of microarray**
- **Fluorescence data analyzed to deduce SNP alleles**

Fig. 4.14 A simplified schematic representation of microarray-based genotyping of SNP loci on the basis of hybridization of PCR products with probes on the microarray (Based on Sobrino et al. 2005)

4.6.5 Bead-Based Techniques

The *bead-based techniques* are similar to microarray method, but they use oligos attached to fluorescent microbeads of 3–5 μm diameter for hybridization (de Vienne 2003). The microbeads are coated with a combination of two fluorescent dyes (red and orange). Different concentrations of the two dyes are combined to generate beads of several different types. The bead types can be distinguished from each other by flow cytometry on the basis of intensity and wavelength of the fluorescent light emitted by them. One can generate 100 different types of microbeads by combining 10 different fluorescence intensities with 2 different wavelengths. To each bead type, several copies of an oligo representing a specific allele of a particular SNP locus are attached. Each SNP locus is represented by two oligos corresponding to the two alleles of the locus,

which are attached to two different bead types. Thus the set of 100 bead types will enable simultaneous analysis of 50 SNP loci, each having two alleles. All the bead types are pooled and distributed into different tubes prior to hybridization.

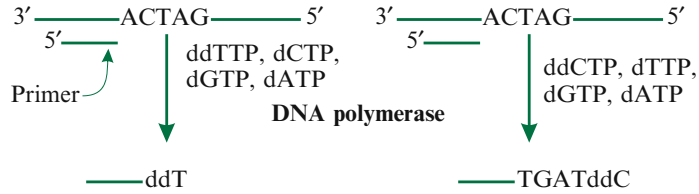
Fluorescence-labeled PCR products corresponding to the SNP loci represented on the microbeads from an individual are hybridized with the pooled microbeads. The non-hybridized PCR products are removed by washing. The beads are passed in a single row through the capillary of a flow cytometer, where they are exposed to two laser beams. The data on the levels of fluorescence in response to the two laser beams are recorded. These data enable the identification of the microbead type and, thereby, the SNP locus and its allele being examined. In addition, the fluorescence level of the PCR product, i.e., the genotyping signal, is also recorded. This signal reveals the "presence" or "absence" of the particular SNP allele. The flow cytometer can examine thousands of microbeads in a few seconds. Data from a large number of beads are collected, and the mean values of fluorescence of the PCR products for each bead type are calculated. This allows deduction of the alleles at the different SNP loci.

The technique has high-throughput potential but has the same limitations as DNA chips. The level of multiplexing is limited by the availability of only green color for the genotyping signal. At present, the use of a 96-well flow fluorometer would permit scoring of thousands of genotypes in a single 96-well format reaction. The bead-based approach has been successfully used for genotyping on the basis of allele-specific hybridization, allele-specific primer extension, single-base extension, and oligonucleotide ligation assay. The microarray- and bead-based techniques are not freely available as they are "closed" or proprietary technologies.

4.6.6 Primer Extension

The *primer extension* method involves annealing of a specially designed primer to the target PCR product, extension of the primer by one to few nucleotides using DNA polymerase (Sokolov 1990; Braun et al. 1997), and analysis of the products of extension to deduce the allele at the SNP locus. This primer is so designed that the base at its 3′ end is complementary to the base just preceding the polymorphic base of the SNP locus present in the PCR product (Fig. 4.15a). As a result, the first nucleotide added to the primer will be complementary to the polymorphic base of the SNP locus. Initially, one ddNTP and the remaining three dNTPs were used in a reaction mixture for primer extension. As a result, for each PCR product, four separate reactions, each using a different ddNTP, had to be set up. In case the ddNTP present in a reaction mixture was the first nucleotide to be added to the primer, there will be no further extension of the primer. But if one of the dNTPs was the first to be added, the primer extension will continue up to the point, at which the base complementary to the concerned ddNTP occurs in the PCR product (Fig. 4.15a). The products of primer extension are analyzed by either electrophoresis in an automated DNA sequencer or by MALDI-TOF MS (matrix-assisted laser desorption ionization time of flight mass spectrometry). The ddNTP permitting addition of only a single base to the primer is identified; the base complementary to this ddNTP will be present at the SNP locus. The primer extension approach has been developed as the homogeneous MassEXTEND (hME) assay for high-throughput SNP genotyping (Sect. 13.2.5; de Vienne 2003).

Alternatively, the four ddNTPs are used in a single reaction mixture for each PCR product so that the primer will be extended by a single nucleotide only (*single-base extension*, *SBE*). Phosphodiesterase II digestion is used to trim the 5′ ends of the products of primer extension, and the molecular weights of the shortened products are determined by MALDI-TOF MS. This permits an accurate identification of the ddNTP added to the extended primer and deduction of the SNP allele. MALDI-TOF MS analysis takes merely 4 s per sample, but the equipment is very expensive, and it requires high expertise. In addition, an extremely



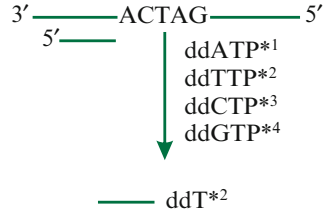
For each PCR product, four reactions are set up; in each reaction a different ddNTP and the remaining three dNTPs are included to support DNA synthesis

- Only one nucleotide (ddTTP) added to the primer
- No further extension of the primer is possible due to the addition of ddTTP
- SNP allele deduced to be A

- dTTP is the first nucleotide added
- DNA synthesis continues till ddCTP is added to the primer
- The product is much longer in this case
- SNP allele is not G

- The products from the other two reactions (with ddGTP/ddATP) will also be longer than those from the first reaction
- Products analyzed by electrophoresis/MALDI-TOF-MS

A. Primer extension (the initial scheme)



A single reaction set up for each PCR product; ddATP, ddTTP, ddCTP and ddGTP, each labelled with a different fluorophore, included in the reaction mixture; primer extended by a single nucleotide (ddTTP*2), which is identified by the fluoroscence

- **Added nucleotide identified by the fluorophore**
- **SNP allele deduced to be A**

B. Single base extension (SBE)

Fig. 4.15 A schematic representation of primer extension and its modification called single-base extension (*SBE*). The first A in the sequence ACTAG of the PCR product represents the SNP locus. *1, *2, *3, and *4, the four distinct fluorophores used to label the ddNTPs. ddNTP, dideoxynucleotide; *ddATP* 2′,3′-dideoxyadenosine triphosphate, *ddTTP* 2′,3′-dideoxythymidine triphosphate, *ddCTP* 2′,3′-dideoxycytidine triphosphate, *ddGTP*, 2′,3′-dideoxyguanosine triphosphate, *MALDI-TOF-MS*, matrix-assisted laser desorption ionization time of flight mass spectrometry (Based on de Vienne et al. 2003)

sophisticated laboratory setup is essential for an optimum use of the mass spectrometer. On a smaller scale, the four ddNTPs can be labeled by different fluorophores, each giving a different color on fluorescence. Since in a given reaction mixture, only one of the four ddNTPs will be added to the primer, the fluorescence color of the product will permit easy identification of the added ddNTP and, thereby, the deduction of the allele present at the SNP locus (Fig. 4.15b).

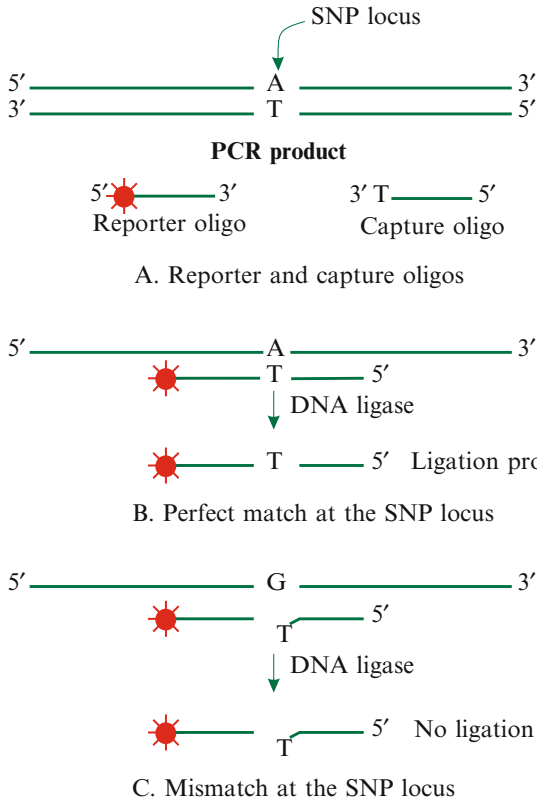
SBE approach has been used to develop diagnostic assays and microarrays for high-throughput genotyping. The SBE assay is also called *genetic bit analysis* (GBA) or *mini-sequencing*. SBE has been used to develop a diagnostic tool, in which the primer is bound to a microtiter plate well. The PCR product is denatured and allowed to anneal to the bound primer. DNA polymerase adds a single nucleotide, corresponding to the SNP site, to the primer, which allows direct determination of the SNP allele. Applied Biosystems, USA, has used this strategy for its 5–10-plex, medium-throughput genotyping system called SNaPshot®. Multiplexing is achieved by using primers of different lengths (from 23 to 60 nt). The primers for different loci differ by four to five nucleotides, and detection is based on capillary electrophoresis. The use of a 96 capillary system allows one person to generate over 10,000 data points per day.

4.6.7 Pyrosequencing

Pyrosequencing has been used for NGS technology (the 454 sequencing technology), which is suitable for both SNP discovery and SNP genotyping. The use of this technology for SNP genotyping is considered in some detail in Chap. 13 (Sect. 13.2.2).

4.6.8 Oligonucleotide Ligation Assay

Oligonucleotide ligation assay (*OLA*) utilizes a pair of oligos for hybridization with the PCR products, followed by ligation of the two oligos by DNA ligase (Landergren et al. 1988). One of the two oligonucleotides is complementary to the SNP locus, i.e., it includes the polymorphic nucleotide, and the sequence on the upstream of the SNP locus; this is known as *capture oligo*. The other oligo is complementary to the sequence on the downstream side of the SNP locus and does not include the SNP locus; this is called *reporter oligo*. The reporter oligo is labeled with a fluorophore. The pair of oligos thus represents contiguous regions including the SNP locus (Fig. 4.16a). The two oligos, the PCR product representing the target genomic region of an individual, and DNA ligase are added to a reaction mixture, heated to denature the DNA, and then cooled to permit their annealing. The two oligos would pair perfectly to the PCR product if the base at the 3′ end of the capture oligo were complementary to the SNP allele in the PCR product. DNA ligase will ligate the two oligos to generate a product having the combined lengths of the two oligos (Fig. 4.16b). However, if the base at the 3′ end of the capture oligo were



The target region is amplified by PCR. Capture oligo represents the SNP locus at its 3′ terminus, while reporter oligo is labelled and corresponds to the 5′region next to the SNP locus

PCR product, the two oligos heated to denature DNA, cooled to allow annealing of oligos to the template, DNA ligase ligates the two oligos. The quantity of ligation product can be increased by LCR

There is mismatch at the SNP locus. Therefore, DNA ligase fails to ligate the two oligos

Fig. 4.16 A simplified schematic representation of oligonucleotide ligation assay. The 3′ terminal base of capture oligo represents the base involved in the SNP. Steps of LCR are similar to those of PCR, viz., denaturation, annealing, and ligation. A thermostable DNA ligase like Taq DNA ligase is used for LCR. LCR, ligase chain reaction (Based on Sobrino et al. 2005)

not complementary to the SNP allele, there will be mismatch at this base, and the ligation reaction will be highly inefficient. Therefore, a negligible amount of the ligation product will be produced (Fig. 4.16c).

The quantity of ligation product can be greatly increased by using a thermostable ligase like Taq DNA ligase in a ligase chain reaction (LCR) procedure that is similar to PCR. The reaction mixture is repeatedly heated to denature the DNA and then cooled to allow hybridization of the two oligos with the PCR product, followed by ligation of the two oligos to generate the product. The OLA procedure can be used in combination with the DNA chip or bead-based techniques to overcome the difficulty in designing of the oligos with the same optimum hybridization conditions. However, the combined procedure is quite complex and demanding. Ligation-based assays are more amenable to multiplexing than primer extension-based assays since ligation is less prone to interference between primers. But OLA tends to be more expensive due to the use of SNP-specific fluorescent primers, while the single-base extension reaction uses a common set of fluorescent ddNTPs for all the SNP loci. The OLA assay system has been modified to develop the 96- and 192-plex assay system SNPlex™ that exploits the specificities of different DNA ligases. OLA is also used for the Illumina's highly multiplexed GoldenGate™ assay (Sect. 13.2.8; Sobrino et al. 2005).

4.6.9 Dynamic Allele-Specific Hybridization

The *dynamic allele-specific hybridization* (*DASH*) uses specific probes for hybridization with the target PCR products (de Vienne et al. 2003). It discriminates between perfect pairing and mismatch at the SNP locus of the PCR product on the basis of relative melting temperatures of the duplexes so produced and thus deduces the SNP allele. One of the two primers used to amplify the PCR product is conjugated with biotin. This PCR product is added to a microtiter plate well coated with streptavidin, to which biotin binds. Thus one strand of the PCR product remains attached with the microtiter plate well, while the other strand is washed away with alkali (Fig. 4.17). This single-stranded preparation is hybridized at a low temperature with an oligonucleotide probe specific for one allele of the SNP locus. An intercalating dye specific for double-stranded DNA (dsDNA) is added into the well. The intensity of fluorescence generated by this dye will be proportionate to the amount of dsDNA. The microtiter well is now gradually heated, and the fluorescence intensity is continuously monitored. There would be a rapid fall in fluorescence intensity as the dsDNA begins to denature. Under appropriate conditions, mismatch at a single base pair, i.e., the SNP locus, leads to an easily detectable lower melting temperature than that with perfect pairing. The sequence of the oligo used for hybridization with the PCR product together with the relative melting temperature of the duplex so formed allows deduction of the SNP allele at this locus. This assay procedure is quick and can be used for reliable scoring of all SNP types, and a suitable device for its implementation is available.

4.6.10 Denaturing High-Performance Liquid Chromatography

In the *denaturing high-performance liquid chromatography* (*dHPLC*) procedure, ion-pair reversed-phase high-performance liquid chromatography is used to separate perfectly matched DNA homoduplexes from heteroduplexes having one or more mismatched base pairs (de Vienne et al. 2003). The PCR product of a test individual is mixed with the PCR product of a reference individual that has a known allele at the SNP locus. The mixture is heated to denature the DNA and then cooled to permit renaturation (Fig. 4.18a). If the test PCR product is exactly the same as the reference PCR product, all DNA duplexes will be perfectly matched, and only one peak of elution will be detected. But if the SNP allele in the test PCR product is different from

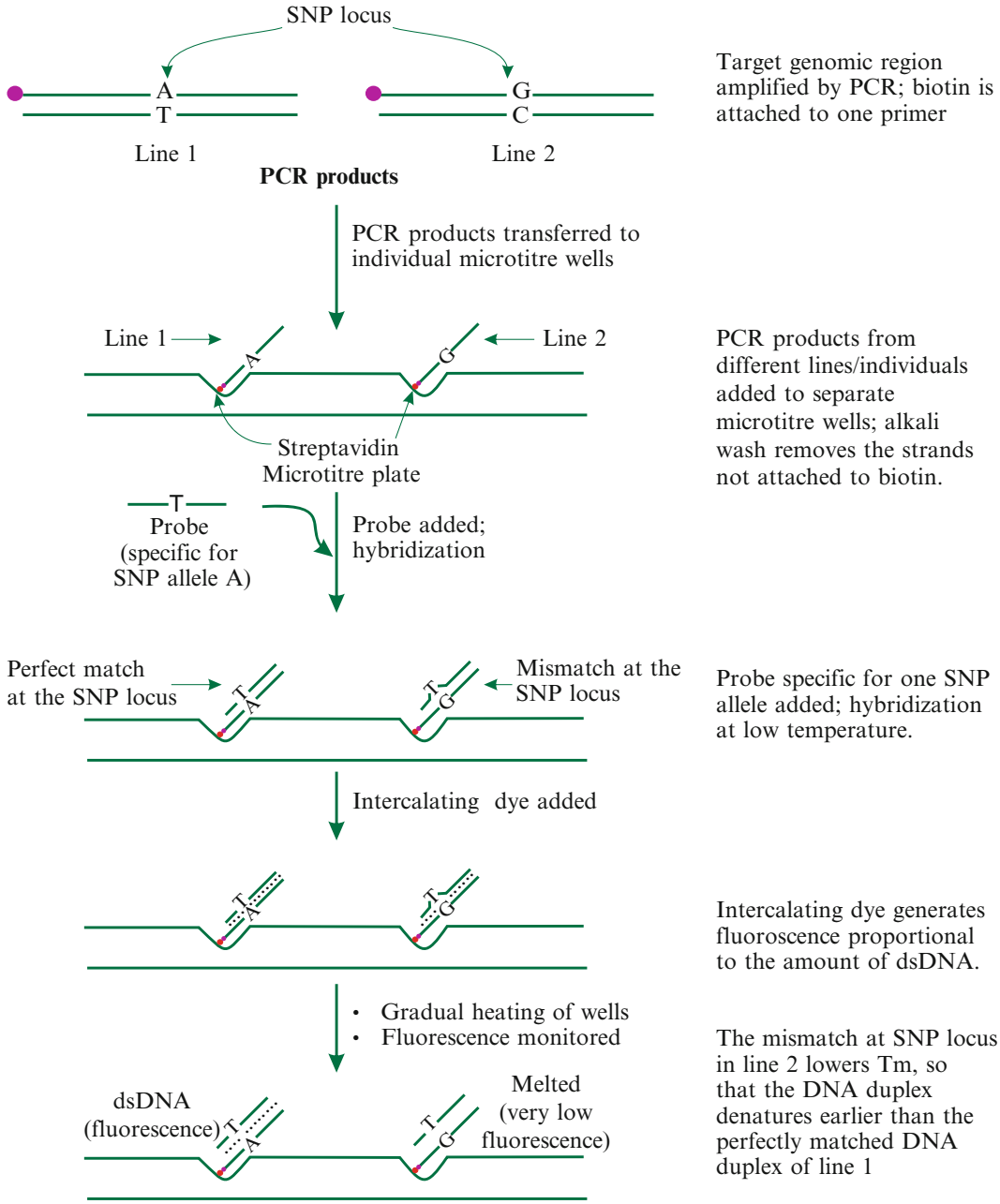


Fig. 4.17 A simplified schematic representation of dynamic allele-specific hybridization (*DASH*). Biotin specifically binds streptavidin; as a result, the biotinylated strand is retained in the microtiter wells. Each well has several copies of the concerned strand of the PCR products (only one strand is shown in each well here) (Based on de Vienne et al. 2003)

that in the reference PCR product, renaturation will produce the two homoduplexes (corresponding to the two PCR products) as well as two heteroduplexes formed by pairing between the strands of the two PCR products (Fig. 4.18b). As a result, there will be two peaks of elution in this case, one for the two homoduplexes and the other for the two heteroduplexes. The procedure requires precise control of temperature and gradient conditions.

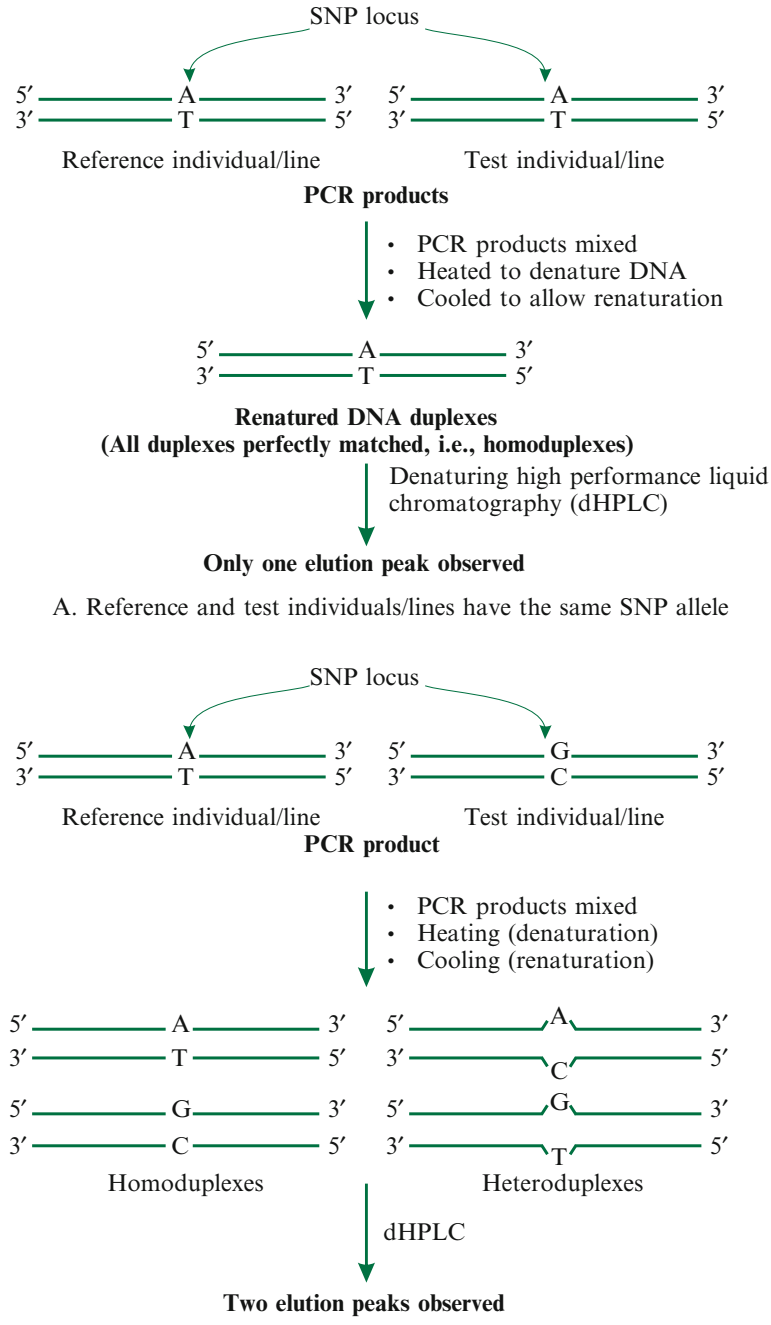


Fig. 4.18 A simplified representation of denaturing high-performance liquid chromatography (*dHPLC*). When the test and reference individuals/lines have different alleles, two elution peaks are observed: one peak corresponds to the two homoduplexes, while the other peak is due to the two heteroduplexes (Based on de Vienne et al. 2003)

Transgenomics, Inc. (San Jose, USA), has developed the fully automated dHPLC WAVE™ system for the analysis of PCR products. The dHPLC WAVE™ system has been used to develop the Masscode™ system by QIAGEN Genomics for high-throughput SNP genotyping as well as SNP discovery (http://www.qiagengenomics.com).

4.6.11 InDels as Molecular Markers

InDels are generally scored as SNPs, but Salathia et al. (2007) developed an InDel array for accurate InDel genotyping. They constructed the array using 70-nt-long oligos representing 240 unique InDel polymorphisms between two *A. thaliana* accessions. InDels of >25 bp were selected to maximize differential hybridization. For each InDel locus, 40 bp of sequence on both sides from the center of the insertion was used to derive the best 70-bp-long oligo; the GC content of the oligo was kept close to 50 %. The test DNA was sonicated, and the genomic fragments were directly labeled with Cy3 and Cy5 fluorophores. Competitive hybridization with the InDel array oligos was performed using 6 μg of the labeled genomic DNA fragments of each of the two accessions. The slides were washed to remove the free probes and the probes involved in nonspecific hybridization. The fluorescence signals were recorded with a sensitive detector, and the data were processed using appropriate bioinformatics tools to deduce InDel genotypes. The InDels were readily recognized with great precision so that there was no need for array replicates and complex statistical analyses. The InDel markers were distributed over the *A. thaliana* genome at an average distance of ~500 kb. Multi-well chips would allow groups of 16 lines to be genotyped in a single experiment. Shotgun sequencing or even partial genomic sequences should permit the application of this approach to non-model organisms for which reference genomes are not available. InDel polymorphisms have also been used for accurate mapping of recessive mutations in *A. thaliana*, identifying alternative expression isoforms of genes in *indica* and *japonica* rice and QTL mapping in salmon. Bulk of the InDels are of 1 bp, and those of 2–4 bp are the second most frequent category, while the frequency of 5 bp or longer InDels is ~10 % or less.

4.7 Epigenetic Markers

Epigenetics is the study of a change in gene function without any change in the gene base sequence. *Epigenetic changes* involve DNA methylation, RNA interference, and histone modification (acetylation, methylation, phosphorylation, and ubiquitination); these changes are also known as *epigenetic marks* (Edwards 2013). A genome-wide study of the epigenetic marks is referred to as *epigenomics*. The sites of cytosine methylation in the genome can be determined by bisulfite sequencing. In this strategy, the template DNA is treated with sodium bisulfite prior to sequencing. This treatment causes deamination of cytosine, thereby converting it to uracil. But when cytosine is methylated at 5 C, it is protected from deamination by the bisulfite treatment. Therefore, bisulfite sequencing will read normal cytosine as thymine, while methylated cytosine will be read as cytosine. The third-generation sequencing technologies are able to directly detect methylation sites. An analysis of the DNA methylation patterns in specific regions of the genome and in the genome as a whole would help understand their role in normal development and in disease. Epigenomic analyses will also elucidate the role of epigenetic changes in environmental adaptation, heritable genetic variation generated by epigenetic changes (*epimutation* and somaclonal variation), and agronomic performance of elite lines developed by breeding programs. *Somaclonal variation* is the heritable variation generated in cells and tissues grown in vitro, in the plants regenerated from them, and in the progeny of these plants.

4.8 Use of Genomics, Transcriptomics, Proteomics, and Metabolomics in Marker Development

The term *genome* denotes the complete set of nuclear and cytoplasmic genes present in an organism. *Genomics* is the field of study concerned with analysis of whole genomes in terms of their organization, including sequence, and function, including metabolic pathways and their interactions. Genomics is generally divided into the following two domains: (1) structural and (2) functional genomics. *Structural genomics* deals with determination of the complete

genome sequence and the complete set of proteins produced by an organism. *Functional genomics*, on the other hand, is the study of the gene expression patterns and the functioning of metabolic pathways. *Transcriptome* is the full complement of RNA molecules, including their quantities, produced by a cell during a specific developmental stage and exposed to a given environment. Thus *transcriptomics* aims to catalogue all the species of RNA transcripts expressed in a tissue/organ; their expression levels, splicing patterns, etc.; and the effects of developmental stages and environmental conditions on their expression.

The term *proteome* refers to the complete set of proteins produced in a cell during a specific developmental stage and under the given environmental conditions. *Proteomics*, thus, is the study of proteome using a diverse array of techniques starting with simple genetic analysis to mass spectrometry. Proteomics is usually classified into structural, functional, and expression proteomics. The discipline of s*tructural proteomics* deals with mapping of the 3-D structure of proteins and analyzing the nature of protein complexes present in a specific cell/organelle. The use of proteomics techniques for analyzing the characteristics of protein networks operating in a living cell constitutes *functional proteomics*. *Expression proteomics*, on the other hand, refers to a comparative quantitative analysis of the expression patterns of proteins between samples differing by some variable. *Metabolome* comprises all the metabolites representing the end products of cellular processes present in a cell, tissue, organ, or organism. Therefore, *metabolomics* is the systematic study of the characteristic small-molecule metabolite profiles generated by the various cellular metabolic processes.

Thus *genomic resources* of a species comprise the sum total of information about the structural and functional aspects of its genome. These resources include detailed high-density genetic maps, contig-based physical maps (including draft/completed genome sequences and their annotations), deep-coverage large-insert libraries, ESTs, gene expression levels and patterns (transcriptome), proteome, and the metabolome. The genome-sequencing projects dramatically accelerated the pace of developments in various areas of genomics, and vast amounts of data have been/are being generated for an increasingly large number of plant species. Some plant species like *A. thaliana* and rice have been investigated far more extensively and intensively than others so that the information accumulated about them is much more complete than that for other plant species. *A. thaliana*, a member of the Cruciferae family, is considered as a model dicot plant for molecular biology studies since it has a small genome size (125 Mb), low content of repetitive sequences, and short generation time and generates a large number of progeny per plant. Similarly, rice has emerged as the model monocot crop species due to its relatively small genome size and conservative genome organization. The genomes of these species have been sequenced and extensively annotated, and functions of a large number of their genes have been experimentally determined. Therefore, the genomes of these and other extensively studied species serve as reference for a variety of investigations, including identification of genes/gene families with specific functions, determination of conserved orthologous set of genes, etc.

Comparative genetic mapping of molecular markers revealed that the gene order is largely conserved (collinear or syntenic) among related plant species, e.g., among the species of grass family, and to some extent even across angiosperms. But comparisons among genome sequences revealed a much lower extent of collinearity of genes, since small-scale sequence rearrangements and InDels disturb the collinearity even between such species that are closely related. For example, comparisons of sequence-based maps reveal extensive breakdown of collinearity between wheat and rice, maize and rice, and sorghum and rice genome sequences. Knowledge of the extent of synteny and the locations of syntenic genomic regions and the patterns of chromosomal rearrangements would enable the transfer of genomic information from one species to the other. This would also

facilitate marker development for the whole genome as well as for specific genomic regions of a species based on the genomic information from a related species. Genome sequences can be analyzed with the help of suitable computer programs to identify molecular markers. For example, SSR markers can be developed by mining the end sequences of bacterial artificial chromosome (BAC) clones and screening of EST databases. All SNP markers are discovered by comparing genome and/or EST sequences of two or more lines/individuals, and SNP genotyping assays are designed on the basis of sequences flanking the SNP loci (Sects. 4.5 and 4.6). For example, comparison of genome sequences of the *indica* and *japonica* subspecies of rice has revealed several SNPs, including InDels. Similarly, the conserved orthologous sequences (COSs) are identified by comparing EST databases of a group of related species against a reference genome like that of *A. thaliana* (Sect. 3.20). Single feature polymorphisms (SFPs) are discovered by using either microarrays developed for gene expression analysis or designing microarrays based on sequences of all the annotated genes, unigenes, and ESTs of the species (Sect. 2.8). By screening the consensus EST sequences or the unigene sequences from many plant species, it is feasible to predict molecular markers like SSRs, SNPs, and COSs that could be developed as functional markers. However, all the predicted functional markers need to be confirmed and validated by appropriate genetic analyses and, ultimately, genetic transformation.

Transcriptome analysis generates a large collection of ESTs, and EST databases exist for most of the important species of plants. But the EST data have several limitations, including unidentified contaminants, chimeric sequences, paralogous and/or homoeologous sequences, and ESTs representing putatively nonfunctional transcripts. Moreover, EST databases lack the non-transcribed *cis*-acting elements and genes expressed at very low levels. However, the EST databases do serve as a rich and invaluable sequence resource for the transcribed regions of the genomes that have been exploited for a variety of purposes. Analysis of transcriptome data pertaining to segregating populations has enabled the identification of *expression QTLs* (*eQTLs*), i.e., QTLs concerned with regulation of expression levels of the genes analyzed in the study. In case a high-quality complete genome sequence is available for a plant species, annotation of the genomic regions harboring eQTLs will facilitate the identification of genes and *cis*-acting sequences involved in the regulation of gene expression relevant for various phenotypes. Efforts are being made to use metabolite levels as markers for the prediction of performance and to assess their usefulness as selection criteria.

4.9 Polymorphic Information Content of Marker Loci

The chief function of molecular markers is a clear-cut and reproducible classification of individuals/lines on the basis of DNA sequence variation. As a result, molecular markers also serve the purpose of reliable identification, based on close linkage, of the genes present in different individuals/lines. A codominant marker would also reveal the allelic states of these genes in the individuals/lines irrespective of whether they are heterozygous or homozygous for these genes. In contrast, a dominant marker will correctly identify the homozygotes but will fail to differentiate the heterozygotes from the dominant homozygotes. For this reason, codominant markers are considered to be more informative than dominant markers. Further, the usefulness of any marker locus for discrimination among different individuals/lines depends on the degree of polymorphism exhibited by the locus in the given population.

Polymorphic information content (*PIC*) of a marker locus is a measure of the degree of its polymorphism and is indicative of its usefulness in linkage and other studies. The PIC has been defined in various ways mainly depending on the biological material in which the marker locus is present and the particular use to which the marker is to be put. A simple and generalized definition of *PIC* is as follows: it is the

probability of a marker locus being polymorphic between two random individuals/lines selected from a given population. It can be readily shown that in a population homozygous for a biallelic marker locus, the PIC for the locus will equal $2pq$, where p and q are the frequencies of the marker alleles a_1 and a_2, respectively. In a homozygous population like a recombinant inbred line (RIL), there will be only two genotypes for the marker, viz., a_1a_1 and a_2a_2, and the frequencies of these genotypes will be p and q, respectively. Therefore, the probability that any two individuals randomly chosen from this population will differ at the marker locus will equal the product of the frequencies of the two genotypes multiplied by two, i.e., $2pq$. It has been shown that the same will be the situation, i.e., PIC $= 2pq$, in the case of a random mating population and in an F_2 population provided the marker locus is in Hardy–Weinberg equilibrium and the marker is codominant.

Since $2pq$ is also the frequency of heterozygotes in a random mating population, PIC is often referred to as *expected heterozygosity* (H_e) for the marker loci. The term expected heterozygosity is in use for the following reason as well. In human linkage studies, analysis of progeny from a parent heterozygous for the marker locus and affected by a dominant disease may allow one to infer the marker allele most likely linked with the disease allele. The value of PIC for a biallelic marker ranges between 0 (only one marker allele present in the population, i.e., $p = 1$ and $q = 0$ or vice versa) and 0.5 ($p = q = 0.5 = 2pq$). But as the value of p (and, consequently, that of q) deviates from 0.5, the PIC value decreases. For example, when values of p and q are 0.4 and 0.6, respectively, the PIC value declines to 0.48 ($=2 \times 0.4 \times 0.6$), while it drops down to merely 0.18 when the values of p and q are 0.1 and 0.9, respectively.

It can be readily shown that in the case of a multiallelic marker locus, the value of PIC would equal $1 - \sum {p_i}^2$, where p_i is the frequency of i^{th} allele at the marker locus. This is because the value of $\sum {p_i}^2$ would equal the sum of the frequencies of homozygotes for all the alleles at the marker locus present in the population, and that of $1 - \sum {p_i}^2$ will be the same as $\sum 2p_iq_i$. The value of PIC for a multiallelic locus ranges between zero (only one allele present in the population) and 1 (infinite number of alleles present in the population). For example, the PIC score for a marker locus with five alleles, each allele having the frequency of 0.2, will be 0.8 [$=1 - (5 \times 0.04)$]. *Thus, the PIC estimate is the property of a specific marker locus in a given population and depends on the number and frequencies of the marker alleles in the population. Therefore, PIC estimates will differ for different loci of a single marker system and for different populations for the same marker locus.*

In any study, several marker loci of a marker system are analyzed. The information from all the loci scored for the marker system may be pooled to estimate the average PIC score for the marker system. It can be shown that the average PIC score (H_{av} = average heterozygosity) for all the polymorphic markers scored for a marker system will equal $\sum H_{ei}/n_p$, where H_{ei} is the expected heterozygosity or the PIC score of the ith marker locus and n_p is the number of polymorphic loci present in the population. However, some of the marker loci may not be polymorphic in the population, but they should be taken into account while estimating the average PIC for the marker system. This can be done by multiplying the average PIC score with $ß$, i.e., the ratio of polymorphic marker loci to the total number of loci scored. Thus the PIC for the marker system would equal $ß\,H_{av}$. *However, the H_{av} estimate for a marker system is applicable to a particular population, from which it is estimated, and it may be only of limited value in other populations.*

A single assay for some marker systems permits the scoring of a single locus, while each assay for some other marker systems evaluates several marker loci. The average number of markers scored per assay of a marker system is described as its *multiplex ratio*. This ratio is different from the extent of multiplexing possible for a marker system in that it indicates the number of different markers analyzed by a single assay without application of any multiplexing strategy (Sects. 3.3.3 and 3.12.2). The multiplex ratio will be one or close to one for markers like

SSRs, SCARS, CAPSs, etc., but will be much larger for marker systems like AFLPs, RAPDs, ISSRs, etc. *Marker index* for a marker system is estimated as the product of multiplex ratio and the average PIC score for the marker system in the given population. Marker index, thus, reflects the degree of polymorphism that would be detected by each assay of the given marker system in the population. Similar to the PIC score, these indicators of the marker usefulness also will be applicable to the concerned population and would merely serve as rough indicators for other populations.

A comparison among different marker systems has been done in several crops, including soybean, barley, and wheat. In a study with soybean, SSR markers were found to have the highest expected heterozygosity, while AFLP markers had the highest multiplex ratio and the highest marker index. In comparison, RAPD markers were intermediate in terms of both expected heterozygosity and multiplex ratio, whereas RFLP markers were moderate with respect to expected heterozygosity (Powell et al. 1996). Studies with other crops have also revealed a similar picture.

4.10 Marker System Selection

RFLPs were the first DNA markers to be developed, and they were extensively used in various biological investigations, including plant breeding. But with the development of more user-friendly PCR-based markers during the 90s, the interest in RFLPs declined, and soon SSRs became the most widely used molecular markers. The dominance of SSRs began to be challenged by SNPs about a decade ago, and since then the latter have rapidly emerged as the marker of choice in view of their abundance and almost uniform distribution throughout the genome. However, the search for new marker systems continues, and so far nearly two-dozen different marker systems have been developed. The salient features of some of the common marker systems are compared in Tables 4.3, 4.4, and 4.5. It would be seen that each marker system has some desirable features that favor its plant breeding application, but some of its other features limit its usefulness. For example, RAPD technique is relatively simple and straightforward and requires much less time than RFLPs and AFLPs, but this marker system has moderate to poor reproducibility. SSR markers are highly polymorphic, PCR based, easily detectable, and codominant, but their development requires considerable time and effort. Similarly, AFLPs are highly reproducible and can be applied to any species since there is no specific marker development step, but they are dominant and anonymous, and their detection requires much more skill and instrumentation than that of RAPDs and SSRs.

The selection of a DNA marker system for a plant breeding application depends on several factors, including the objectives of the project, the financial resources available to the project, availability of the desired marker system for the concerned species, and the reproducibility of the marker system. The objective of the project would determine the scale of operations in terms of the numbers of markers and the samples to be scored during a cropping season (Table 2.4). In view of the above, the research worker has to critically evaluate each marker system for its potential utility to his/her project and select the most suitable marker system. In general, the choice will be influenced by the following features of the marker systems: degree of polymorphism, dominance/codominance of marker alleles, simplicity and speed of detection procedures, amenability for multiplexing and automation, need for prior sequence information and the amount of work required for marker development, and above all the reproducibility of the marker system. For genetic mapping, the genotyping procedure should be simple and cost-effective, and the information content of the marker should be moderate to high. In addition, the marker should be abundant and distributed across the whole genome. Cost of genotyping would depend on the amount of DNA needed for analysis, need for cloning and sequencing, the amount of potentially useful genetic information acquired per data point, the type of genetic information needed, dominance relationship of

Table 4.3 A comparison among different marker systems

Marker system	Abundance	Reproducibility	Degree of polymorphism	Locus specificity	Technical requirement	DNA quantity	Automation	Genotyping cost	Major application
RFLP	High	High	Medium	Yes	High	High	Low	High	PM[a]
RAPD	High	Low	Medium	No	Low	Low	Medium	Low	LM
SSR	Medium	Medium	Medium	No	Medium	Low	Medium/high	Low	DA
SSCP	Low	Medium	Low	Yes	Medium	Low	Low	Medium	LM
CAPS	Low	High	Low	Yes	High	Low	Low	Medium	LM
SCAR	Low	High	Medium	Yes	Medium	Low	Medium	Low	LM, PM
AFLP	High	High	Medium	No	Medium	Medium	Medium/high	Medium	LM
IRAP/REMAP	High	High	Medium	Yes	High	Low	Medium/high	Low	DA
RAMPO	Medium	Medium	Medium	Yes	High	Low	–	–	DA
SRAP/EST	Medium	High	Medium/high	–	Medium	Medium	Medium/high	Low	LM
ISSR	High	High	Medium	No	Low	Low	Medium	Low	DA
SNP	Very high	High	Medium	Yes	High	Medium	High	Low	LM

Based mainly on Meksem and Kahl (2005) and Agarwal et al. (2008)
[a]*PM* physical mapping, *LM* linkage mapping, *DA* genetic diversity analysis

Table 4.4 A rough classification of the different marker systems on the basis of their various features

	Level		
Feature	Low	Moderate	High
Detection			
Equipment cost	RAPD, SCAR, SSR, ISSR, COSs	AFLP, RFLP	SNP, SFP, DArT
Technical expertise	RAPD, SCAR, SSR, ISSR, COSs	AFLP	SNP, SFP, DArT, RFLP
Throughput	RFLP	RAPD, SCAR, SSR, ISSR, COSs, AFLP	SNP, SFP, DArT
Automation, including data acquisition and processing	RFLP, RFLP, RAPD, ISSR,	SSR, AFLP, COSs, SCAR	SNP, SFP, DArT
Assay time		RAPD, SCAR, SSR, ISSR, COSs, AFLP, SNP, SFP, DArT	RFLP
Cost per data point	RAPD, SCAR, SSR, ISSR, COSs, DArT, SNP, SFP	CAPS, AFLP	RFLP
Marker development			
Time and effort	RAPD, ISSR, SRAP	SCAR, AFLP	SNP, SFP, DArT, RFLP, SSR, COSs
Need for sequence information	RAPD, ISSR, RFLP, DArT, SRAP (not required)	SCAR	SNP, SFP, SSR, COSs
Use of bioinformatics tools	RAPD, ISSR, RFLP, DArT, SRAP (not required)	SCAR	SNP, SFP, SSR, COSs
Other features			
Reproducibility/reliability	RAPD		SNP, SFP, DArT, RFLP, SSR, COSs, AFLP
Scale of operation	RFLP	RAPD, SCAR, SSR, ISSR, COSs, AFLP	SNP, SFP, DArT
Plant material required	RAPD, SSR, ISSR, SCAR, COSs	RFLP, SNP, SFP, DArT	

Table 4.5 A summary of differences among different array-based techniques for detecting DNA polymorphisms. All the markers are scored as presence/absence and are regarded as cost-effective

	Marker system			
Parameter	SNP	SFP	DArT	RAD tag
Sequence information	Required	Required	Not required	Not required
Markers represent	Random genomic regions	Genic regions	Random genomic regions	Random genomic regions
PCR amplification	Required in some assays like MIP and GoldenGate	Not required	Required	Required
Number of markers scored per assay	High	High	Moderate	Moderate
Type of array used	Tag array on beads/glass, oligonucleotide array/ GeneChip	High-density oligonucleotide array/ GeneChip	Glass-spotted DNA microarray	Tiling microarray, oligonucleotide array/ GeneChip
Resolution	High	High	Moderate	Moderate

Based on Gupta et al. (2008)

marker alleles, amenability to automation, and the proprietary status of the technique for marker detection.

A discussion on the selection of a suitable marker system can be only in general terms, and it may not be possible to provide specific recommendations. We may begin our discussion with reference to large-scale breeding projects with adequate financial resources. In such cases, one would need a marker system capable of high to very high throughput and automated data acquisition and analysis. Four marker systems, namely, SNPs, DArT, SFPs, and RAD markers, satisfy these criteria. All these marker systems require considerable laboratory infrastructure and sophistication and moderate to large amount of marker development effort. However, SFP and SNP markers are sequence based and either good quality genome sequences should be available or de novo sequencing would be necessary for their development. In contrast, DArT and RAD markers are anonymous and their development does not require sequence information; as a result, they can be developed for any crop species irrespective of the availability of genomic resources. Therefore, the choice among them will depend mainly on the considerations of marker density requirement, cost per data point, and the availability of the marker systems for the concerned species. At present, SNPs are the preferred markers and almost all large-scale breeding programs are routinely using them. DArT markers are steadily gaining in popularity for fingerprinting, diversity studies, selection of parents, and linkage mapping, while SFPs and RAD markers have also been used.

In the case of breeding programs of small to moderate size, most of the DNA markers are available for application. However, RAPDs have limited reliability, and RFLPs are not user-friendly. Therefore, even when RFLP markers are available for achieving the desired goals, other marker systems would be preferable. When the financial resources are adequate and the desired markers are available, the choice will have to be between SSRs and SNPs. In most situations, these markers can be assayed in the laboratory, and where required SNP genotyping services are commercially available. Further, a moderate-sized breeding program can be hardly expected to de novo develop SSR and SNP markers. In case SSR and SNP markers are not available for the desired goal and genomic and/or financial resources do not support their de novo development, one has to select a marker system like AFLP, DArT, SRAP, or SCoT that does not require prior sequence information for marker development. DArT is a proprietary technology, and its development as well as detection would require substantial expenditure on equipment or the activity will have to be outsourced. AFLPs do require some expenditure on equipment, but this will be much less than that for DArT. The SRAP, SCoT and other similar markers are in experimental stages, but appear to be quite promising.

The objectives of the program also influence issues like marker density and the genomic regions to be targeted for marker genotyping. For example, a much higher marker density would be needed for association studies and genomic or genome-wide selection than those for linkage mapping and MAS. Further, even in the case of association studies, a much higher marker density would be needed in a cross-pollinated species than in a self-pollinated species. Therefore, SNPs become the preferred marker system for programs like association studies and genomic selection. It has been argued that a much higher density of SNP markers would compensate for their lower PIC as compared to that of SSRs. Similarly, when a specific region of the genome is to be targeted and/or fine mapping is to be done, an abundant marker system like SNP is preferable to the others.

Questions

1. Explain the features that make NGS technologies faster and cheaper than the first-generation technologies.
2. Briefly describe the procedure of one of the NGS technologies, and discuss the applications of the NGS technologies.

3. How do third-generation sequencing technologies differ from the NGS technologies, and what advantages do they offer in comparison to the latter?
4. Explain the meaning of PIC and discuss its significance for a marker system.
5. What are the various issues relevant to the selection of a suitable marker system for marker-assisted selection?
6. Discuss the usefulness of genomic resources in the development of molecular markers, especially single nucleotide polymorphism.
7. Briefly explain the use of primer extension for determining the SNP alleles at a given locus.
8. Discuss the use of microarrays for SNP genotyping

Part III

Linkage Maps

Mapping Populations

5

5.1 Introduction

In 1865, Mendel proposed that the development of phenotypic characters is governed by hypothetical factors, now called genes, and the alternate forms (alleles) of different genes segregate independently. Soon after the rediscovery of Mendel's findings in 1900, Sutton and Bovery proposed the chromosomal theory of inheritance in 1902, according to which genes are located on chromosomes. In 1910, Morgan provided the first experimental evidence for the chromosomal theory: he demonstrated that the inheritance pattern of white-eye gene of *Drosophila* indicated it to be located on the X chromosome. One year later, in 1911, Morgan described the essential features of linkage between genes, and in the year 1913, Sturtevant published the first linkage map of *Drosophila*. Subsequently, morphological markers were used to construct linkage maps in many species. Since the number of such polymorphisms in any species is limited, these linkage maps were sparse, i.e., the markers were spaced at considerable distances from each other. Geneticists mounted search for more abundant markers, and protein polymorphisms were the first molecular variations used to generate linkage maps. The limited number of protein polymorphisms and the environmental influence on their expression were the major drawbacks, which favored the development of DNA markers. Restriction fragment length polymorphism (RFLP) was the first DNA marker to be developed and used in mapping experiment. RFLPs have now been virtually replaced by PCR-based markers and single nucleotide polymorphisms (SNPs), which are amenable to automation and high-throughput analyses. The use of DNA markers has allowed construction of dense linkage maps in many important plant species and has enabled the mapping of the elusive quantitative trait loci (QTLs). Construction of linkage maps requires the following: (1) a suitable marker system, (2) an appropriate mapping population, and (3) software for proper analysis of the data. In this chapter, the different mapping populations are described in some detail.

5.2 Mapping Populations

A population that is suitable for linkage mapping of genetic markers is known as *mapping population*. Mapping populations are generated by crossing two or more genetically diverse lines and handling the progeny in a definite fashion. Generally, the parents used for hybridization will be from the same species. But in some cases, where intraspecific variation is limited, related species may be used as one of the parents. Mapping populations are used for determining genetic distances between pairs of loci/genes and to map them to specific locations in the genome. They also help in the identification of molecular markers that are linked to genes/loci of

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_5, © Author(s) 2015

interest; such markers can be used for marker-assisted selection (MAS) for the genes of interest. Thus, mapping populations serve as the basic tools needed for the identification of genomic regions harboring genes/QTLs and for estimating the effects of QTLs. The choice of parents, the design used for their mating to develop the mapping population, and the marker system used for mapping are determined by the objectives of the study, the cost and the accessibility of various marker systems, and the availability of a molecular map. The parents used for developing a mapping population must differ to the maximum extent at both DNA sequence and phenotypic levels. The DNA sequence level variation is essential to trace the results of recombination events. In general, the greater is the extent of DNA sequence variation, the easier it would be to find polymorphic molecular markers. When a study aims to identify and map genes governing specific traits, the selected parents must show genetic variation for the target traits. If the parents have different phenotypes for a trait, there is a reasonable likelihood that they would also exhibit genetic variation for the trait. However, sometimes environmental effects might exaggerate the phenotypic variation, which may have no genetic basis. Similarly, the absence of phenotypic variation between the selected parents may not necessarily mean that genetic variation is lacking for the trait, since different sets of genes can generate similar phenotypes.

There are basically two types of mapping populations, viz., primary and secondary mapping populations. *Primary mapping populations* are created by hybridization between two homozygous lines usually having contrasting forms for the traits of interest. *Secondary mapping populations*, on the other hand, are developed by crossing two lines/individuals selected from a mapping population; they are developed mainly for fine mapping of the genomic region of interest. The primary mapping populations are of the following different types: (1) F_2, (2) F_2-derived F_3 (F_2:$_3$), (3) backcross (BC), (4) backcross inbred lines (BILs), (5) doubled haploids (DHs), (6) recombinant inbred lines (RILs), (7) near-isogenic lines (NILs), (8) chromosomal segment substitution lines (CSSLs), (9) immortalized F_2, (10) advanced intercross lines, (11) recurrent selection backcross (RSB) populations, and (12) interconnected populations (Fig. 5.1). A summary of the characteristic features of the important mapping populations is given in Table 5.1. The specific type of mapping population to be used in a given study depends primarily on whether the concerned plant species can be subjected to self-fertilization without severe inbreeding depression, the time available for the development of the mapping population, and the trait (s) to be mapped (Schneider 2005).

5.3 Selection of Parents for Developing a Mapping Population

The selection of parents for developing a mapping population is critical to the success of map construction effort. The two lines selected as parents, designated as parent 1 (*P1*) and parent 2 (*P2*), should be completely homozygous. If necessary and where feasible, doubled haploids may be used as parents to avoid the problems due to residual heterozygosity. Since the economic significance will primarily depend upon the useful marker–trait associations depicted in the map, the genetic stocks selected as parents for generating a mapping population should differ for as many qualitative and metric traits as possible. In addition, the parents should be polymorphic for as many molecular markers as possible to afford the construction of dense linkage map. It is desirable to ascertain the polymorphism present between the two parents both at the phenotypic and genotypic, i.e., molecular marker, levels before crossing them. Another point that should be considered is whether adapted or exotic germplasm should be used for developing the mapping population. Chromosome pairing and recombination rates would be suppressed, sometimes severely, in wide crosses, and this would inevitably yield greatly reduced estimates of distances between pairs of loci (Zamir and Tadmar 1986). In general, wide crosses will

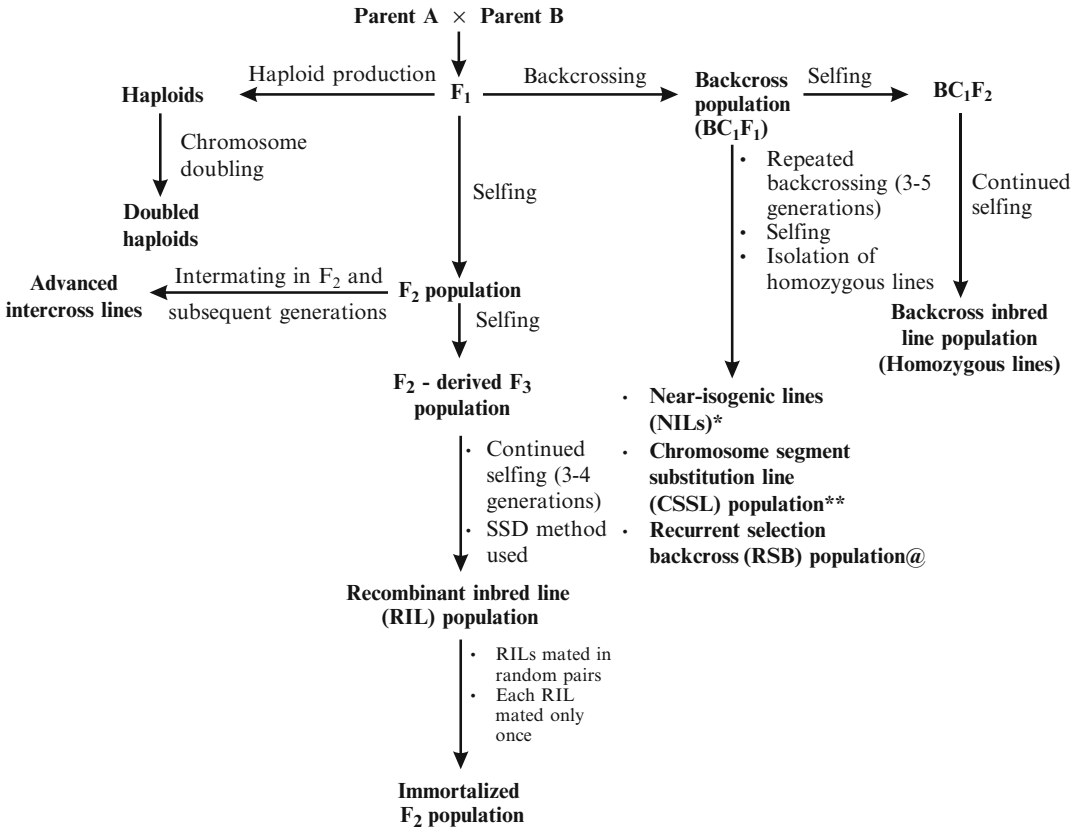


Fig. 5.1 A schematic representation of the various biparental mapping populations. * Introgression of a gene by repeated backcrossing combined with selection for the gene. ** Repeated backcrossing without selection; each line has a distinct chromosome segment from the donor parent. @ The donor parent has high value for a quantitative trait. In each backcross generation, the individual with the highest value for the trait is selected and backcrossed to the recurrent parent

generate segregating populations exhibiting a relatively large array of polymorphism as compared to that encountered in the segregating generations of narrow crosses (adapted × adapted germplasm crosses). In some crop species like chickpea, the variation in the intervarietal crosses is limited. In such cases, the use of mapping populations derived from wide crosses may be desirable. But the F_1 hybrid from such a cross should be fertile to allow the development of a mapping population. Further, the map developed from such a population should be preferably collinear, i.e., having a similar order of different loci, with the map constructed from populations derived from the adapted parents. However, some valuable inferences about the ease of introgression can be drawn even from such an interspecific map that differs substantially from that of the adapted parent due to chromosomal rearrangements.

5.4 F_2 Population

A F_2 *mapping population* comprises the progeny produced by selfing or sib-mating of the F_1 individuals from a cross between the selected parents (Fig. 5.1). The F_1 individuals would be heterozygous for all the loci for which their parents differ from each other. Each F_2 individual is expected to have a unique combination of linkage blocks from the two parents, and this

Table 5.1 A comparative summary of the important features of some of the common mapping populations

	Mapping population					
Feature	F_2	Backcross	RIL	NIL	CSSL	Immortalized F_2
Perpetuation	Ephemeral	Ephemeral	Perpetual	Perpetual	Perpetual	Perpetual[a]
Genetic composition	Homozygotes and heterozygotes	Homozygotes and heterozygotes	Homozygotes	Homozygotes	Homozygotes	Homozygotes and heterozygotes
Each genotype represented by	One plant	One plant	One line	One line	One line	One line
Generations needed to develop	Two	Two	7–8 or more	8–10	7–10	One (after RILs are developed)
Number of crosses made	One (F_1)	Two (F_1 and backcross)	One (F_1)	6 or more (F_1 and backcrosses)	6 or more (F_1 and backcrosses)	Many (hundreds per population)
Selection during population development	None	None	None	Yes (foreground and background)	Yes (foreground and background)	None
Rounds of recombination	One	One	About two	One + the number of backcrosses	One + the number of backcrosses	About two (in the RILs)
Segregation ratios for dominant and codominant markers	Different	Different	Same	Same	Same	Different
Suitable for:						
(i) Oligogene mapping	Yes	Yes	Yes	Yes	Yes	Yes
(ii) QTL mapping	No	No	Yes	Yes	Yes[b]	Yes
(iii) Fine mapping	No	No	No	Yes	Yes[b]	No
(iv) Mapping of heterosis loci	Yes	No	No	No	No	Yes
(v) Positional cloning	No	No	No	Yes	Yes[b]	No
(vi) Assessment of QTL × genotype interaction	No	No	Yes	Yes	No	No
Minimum QTL × QTL interaction	No	No	Yes	Yes	Yes[c]	No
Mapped loci belong to	Either parent	Either parent	Either parent	Donor parent	Donor parent	Either parent
Analysis covers	Whole genome	Whole genome	Whole genome	A genomic segment	A genomic segment	Whole genome

[a]A given immortalized F_2 population is, in fact, ephemeral, but the same population can be reconstructed from the component RILs; therefore, it is considered as perpetual
[b]Best mapping population for the purpose
[c]For each line

difference is the basis for detection of linkage between pairs of loci. Since F_2 generation is the product of a single meiotic cycle (in the F_1 plants), only one round of recombination can occur between any two loci. Therefore, the estimates of recombination frequencies between pairs of loci obtained from F_2 populations serve as a reference point. In a F_2 population, the ratios expected for dominant and codominant markers are 3:1 and 1:2:1, respectively. The F_2 population is grown in an un-replicated block and the target traits are scored on individual plants.

These scores would be reliable so long as the trait heritability is nearly 100 %, but they will be much less reliable for quantitative traits. In cross-pollinated crops like maize, quantitative traits can be meaningfully evaluated only with heterozygous individuals/populations such as F_2. This is because in these crops, dominance and epistatic genetic variances constitute the major proportion of the genetic variance. These variance components can only be estimated in a population consisting of heterozygous individuals. Further, in such crops, F_2 plants are crossed with suitable testers and the testcross progeny are used for evaluation in appropriate trials. Ideally, more than one tester should be used to produce the testcross progeny so that specific effects due to a particular tester genotype are excluded. In some studies, F_2 populations have been used for mapping QTLs. For example, Edwards et al. (1987) used allozyme markers segregating in two maize F_2 populations to map QTLs for 40 quantitative traits measured on individual plants; they divided the experimental area into four blocks to obtain an error term for statistical analysis.

F_2 populations are the best suited for preliminary mapping of markers and oligogenes. Creation of F_2 populations requires only two generations, which is the minimum for developing a biparental mapping population. Further, their development requires the minimum effort as compared to the other mapping populations. The F_2 populations provide estimates of additive, dominance, and epistatic components of the genetic variance. These populations capture the recombination events from both male and female parents (actually, gametes in self-pollinated crops) of the F_2 plants. They are ideal for identifying heterosis QTLs, except for the limitation of replications. Since F_2 populations are produced after one round of recombination, the markers identified to be linked with the target genes are likely to be located at a greater distance than those detected using recombinant inbred line (RIL) populations. Since each plant in a F_2 population is genetically different from the others, F_2 populations cannot be evaluated in replicated trials conducted over locations and years, except in the case of asexually propagated crops. Therefore, a precise evaluation of quantitative traits and the effects of GEI (genotype × environment interaction) on their expression cannot be done. In view of the above, F_2 populations are of limited use for fine mapping and for mapping of QTLs. F_2 populations are ephemeral, as they cannot be maintained beyond one generation, except by asexual reproduction. Further, it is not possible to construct an exact replica of a F_2 population and to increase the amount of seed of individual genotypes represented by the F_2 plants. In those crop species, where asexual reproduction is feasible, the F_2 plants can be multiplied and maintained as clones. Micropropagation can also be used for this purpose for species amenable to in vitro propagation if the effort and the expenditure were justified. But in most sexually reproducing species, a F_2 population would be available for mapping as long as the DNA extracted from F_2 plants and stored in a deep freezer is not exhausted. A F_2 population can be maintained as F_3 progeny of the F_2 plants, i.e., as F_2-derived F_3 ($F_{2:3}$) population (Schneider 2005).

5.5 F_2-Derived F_3 Population

A F_2-*derived* F_3 or $F_{2:3}$ population is obtained by selfing the F_2 individuals for a single generation and harvesting the seeds from each F_2 plant separately so that each F_2 plant is represented as an individual plant progeny (Fig. 5.1). The DNA for genotyping is obtained from individual F_2 plants or it can be reconstructed from a bulk of at least 20 plants from each F_3 family (Yu et al. 1997) since this bulked DNA may be expected to represent the genotype of the parental F_2 plant. Similar to F_2 populations, $F_{2:3}$ populations are not perpetual. $F_{2:3}$ populations are suitable for mapping of oligogenic traits controlled by recessive genes and of QTLs since data can be recorded on multiple plants in each $F_{2:3}$ family to compensate for sampling error. The mean phenotypic value from multiple plants in a $F_{2:3}$ family can be considered to represent the phenotype of its parent F_2 plant. Yu et al. (1997) analyzed a population of 250 F_3 families of rice to detect 32 QTLs governing yield and three yield component traits. The chief limitations of

the $F_{2:3}$ populations are as follows: (1) The construction of these populations requires an extra season than that of F_2 populations. (2) Most of the F_3 families are heterogeneous due to segregation of one or more genes. As a result, it is not possible to use as replicates multiple genotypically identical plants from a $F_{2:3}$ family. (3) The genotype and, particularly, phenotype of the F_3 population do not strictly correspond to that of the F_2 generation due to one more round of segregation, recombination, and inbreeding. (4) The average phenotype of the F_3 is related to but not strictly comparable to that of the parent F_2 plant for the came reasons as given above. Finally, (5) the data from F_3 are likely to underestimate dominance, overdominance, and certain epistatic components of gene action due to the increased level of inbreeding (Hua et al. 2002).

5.6 Backcross Population

Backcross populations are generated by crossing F_1 plants with either of the two parents of the concerned F_1 (Fig. 5.1). Genetic analysis can be performed only when there is detectable phenotypic segregation for the target trait in the backcross generation. Therefore, the F_1 is, as a rule, backcrossed to the recessive parent, i.e., the parent having the recessive form of the target trait. Such a backcross is called testcross, is usually denoted by B_2, and exhibits 1:1 ratio for the trait phenotype, dominant molecular markers present in coupling phase with respect to the target trait, and codominant markers in either phase. However, it would show 1:0 ratio, i.e., no segregation, for dominant markers present in repulsion phase in relation to the target trait. In contrast, progeny from backcross with the dominant parent (generally designated as B_1) would display 1:0 ratio for the trait phenotype and dominant markers present in coupling phase with respect to the target trait. However, a 1:1 ratio would be obtained in B_1 for codominant markers and dominant markers present in repulsion phase. Thus, in the case of codominant markers, the order of backcross as well as the phase of linkage is not important when only markers are to be scored. In contrast, the order of backcross is extremely important for traits showing dominance, for dominant markers, and for both dominant and codominant markers when mapping of a gene showing dominance is the objective; in these cases, only B_2 can be used.

The backcross populations offer one specific advantage as they can be further utilized for marker-assisted backcrossing (MABC) for introgression of the target traits as proposed in the advanced backcross QTL method (Tanksley and Nelson 1996). But the construction of backcross populations, like that of $F_{2:3}$ populations, requires one more generation than that of F_2 populations. Further, it requires crossing of the F_1 plants with the selected parent, which imposes additional work and may limit the population size in many crop species. The BC populations are similar to F_2 populations as they are not perpetual and cannot be evaluated in replicated trials, which makes them unsuitable for QTL mapping. In addition, they capture the recombination events of only one parent, i.e., the F_1.

5.7 Doubled Haploids

Doubled haploid (DH) plants are obtained by chromosome doubling of haploid plants usually derived by culture of anthers/pollen grains produced by F_1 plants (Fig. 5.1). In some crop species, haploids can also be produced from certain interspecific crosses. For example, when wheat or barley is crossed with *Hordeum bulbosum*, the chromosomes of *H. bulbosum* are gradually eliminated during embryo development, and embryo culture is used to rescue wheat or barley haploids. Another method for high-frequency haploid production uses "inducer" pollinator strains and is widely used in maize. The maize haploids for a DH population can be produced by pollinating the F_1 plants with an inducer strain like RWS or RWK-76. The seeds with haploid embryo have normal triploid endosperm. The haploid embryos most likely originate due to gradual elimination of the inducer strain chromosomes during embryo development.

1. The selection of haploid seeds is based on the colors of embryos and endosperms, which is

specified by the *Rnj* locus. The dominant allele *Rnj* produces violet color in both embryos and endosperms, while the recessive allele *rnj* generates colorless endosperms and embryos. The female parent (the F_1 in this case) should be homozygous *rnj rnj*, while the inducer strain should have the genotype *Rnj Rnj*. The haploid seeds produced from this cross will have colored endosperms and colorless embryos. In contrast, the diploid hybrid seeds will have colored embryos as well as endosperms, while selfed seeds will have colorless endosperms and embryos.

The frequency of haploid seeds may average 8–10 % or more depending on the inducer strain, the strain used as female, method of pollination, and the environmental factors. The available evidence shows that the ability to induce maternal haploids is under polygenic control (Geiger and Gordillo 2009). Generally, colchicine is used to double the chromosome number of haploids, since this alkaloid blocks spindle development. Seeds from individual DH plants are harvested separately and maintained as DH lines in the same way as RILs. The DH lines are completely homozygous at all the loci in the genome, and unlike RILs, they do not have any residual heterozygosity. A DH population may be expected to represent a random sample from all the homozygous lines that can be obtained from the cross provided there is no selection pressure exerted by the haploid production and/or chromosome doubling procedures. The expected ratio for the genes as well as markers in a DH population is 1:1 irrespective of the marker being dominant or codominant. DHs are similar to F_2 in that they both are products of one meiotic cycle occurring in F_1. But the frequency of recombinants would be higher in a DH population than in the corresponding F_2 population. [The frequency of recombinants in a DH population will be r, while it will be $r - (r^2/2)$ in the F_2 population, where, r is the frequency of recombination between two markers/loci.]

DH populations, like RILs, are perpetual as they can be multiplied and maintained indefinitely and can be shared among researchers/laboratories. They can be evaluated in replicated trials and are suitable for mapping both qualitative and quantitative characters. Construction of DH populations requires the same number of years as that of F_2 populations. However, their production involves tissue culture technique and greenhouse facilities. Therefore, a relatively greater technical skill is needed for their development than for other mapping populations. Further, dependable haploid production methods are not available for a number of important crops, and different genotypes of a single crop species often differ markedly in their tissue culture response. The anther culture procedure as well as colchicine treatment may induce genetic variation, which should be taken into consideration. In addition, only additive and additive × additive interaction genetic variances can be estimated from DH populations as they consist of only homozygous plants. Therefore, DH populations are not suitable for mapping heterosis QTLs. The suitability of DH populations for mapping has been demonstrated in pepper (Lefebvre et al. 1995), wheat, barley, rice, etc. (Schneider 2005).

5.8 Recombinant Inbred Lines

Recombinant inbred lines (*RILs*) are a set of homozygous lines produced by continuous inbreeding/selfing of individual F_2 plants (Fig. 5.1; Burr et al. 1988; Simpson 1989; Burr and Burr 1991). RILs are also called F_2-derived inbred lines or single seed descent (SSD) lines because they are derived from F_2 populations usually by the SSD procedure. The concept of linkage mapping using RILs was originally developed in mouse, where about 20 generations of sib-mating was conducted to achieve useful levels of homozygosity (Schneider 2005). An RIL mapping population consists of a set of random RILs derived from a suitable cross. Wherever possible, F_2 plants and their progeny should be selfed, and sib-mating should be resorted to only when selfing is not feasible for some reason. This is because the rate of decrease in heterozygosity with selfing is one-half of that in the previous generation, while that with sib-mating it is merely one-fourth. As a result, selfing requires only half as many generations as sib-mating to achieve the same level of

homozygosity. In addition, sib-mating will require twice as many F_2 plants as selfing to produce the same number of RILs. The SSD method is the best suited for developing RILs, but bulk procedure and pedigree method without selection can also be used. It is important that the generation advance is carried out under an optimal environment that affords equal survival of the various genotypes and does not impose a selection pressure against some genotypes.

The SSD procedure is followed for five or more (usually >8) generations, during which one seed is harvested from each plant of the F_2 and the later generations and seeds from all the plants are composited and planted to raise the next generation. At the end of SSD procedure, seeds from each plant are harvested separately to obtain as many RILs as there are individual plants in the SSD population. There will be some plant loss during the SSD procedure due to several reasons, viz., lack of 100 % seed germination, plant survival, and reproduction; this problem may be more acute in some crops, particularly at high plant densities (Singh 2012a). For example, Burr et al. (1988) subjected 50 F_2 plants from two maize populations to six generations of inbreeding; at the end, they had 46 lines in 1 population and only 38 lines in the second population. It is important that they had followed ear-to-row method during the inbreeding process. They harvested the whole ear from the selected F_2 plant. In the subsequent generations, individual plant progenies were raised, and in each generation, the first plant in a row was selected for selfing. Therefore, the F_2 population should be suitably larger than the desired RIL population size, and a modification of SSD procedure may be used if the plant loss is substantial (Singh 2012a). Each generation of selfing reduces heterozygosity to one-half of that in the previous generation, and there is a corresponding increase in homozygosity. As a result, in a $F_{4:5}$ RIL population, 87.5 % of the RILs will be homozygous for a given locus, while 92.25 % of the plants in the RIL population will have become homozygous for this locus (Table 5.2). It may be pointed out that the above will also be the level of homozygosity considered at the RIL and individual plant levels, respectively, at all the loci segregating in the population. A $F_{4:5}$ RIL population denotes that this population was handled as per SSD procedure up to F_4, and the seeds produced by individual F_4 plants were harvested separately to raise individual plant progenies in F_5. It is relevant to note that the level of homozygosity at the individual plant level will go on increasing as this population is advanced to $F_{4:6}$, $F_{4:7}$, etc. generations, but the homozygosity at the RIL level will remain at the $F_{4:5}$ level. The process described above yields a set of lines, each of which contains a different combination of linkage blocks from the original parents, which provides a basis for linkage analysis (Fig. 5.2).

Table 5.2 The degree of homozygosity at the levels of individual RILs and individual plants in RIL populations produced by SSD procedure for different numbers of generations

	Percent homozygosity at each locus	
RIL population	At individual plant level[a]	At RIL level[b]
$F_{3:4}$	87.50	75.00
$F_{4:5}$	93.75	87.50
$F_{5:6}$	96.875	93.75
$F_{6:7}$	98.438	96.875
$F_{7:8}$	99.219	98.438
$F_{8:9}$	99.609	99.219

[a]Homozygosity estimated as percent of homozygous plants in the RIL population
[b]Homozygosity estimated as percent of homogeneous RILs in the RIL population

The RIL population consists, almost exclusively, of the two homozygotes (e.g., *AA* and *aa*) for a locus and a rather small proportion of the heterozygote (e.g., *Aa*), depending on the number of generations, up to which SSD procedure was followed. The expected ratio of the two homozygotes in the population is 1:1. As a result, the amount of information obtained from dominant markers is the same as that from codominant markers because heterozygosity is almost negligible. In addition, RILs enable detection of markers located much closer to the target gene than is possible with F_2, DH, and BC populations. In the case of the latter populations, recombination between the marker and the target

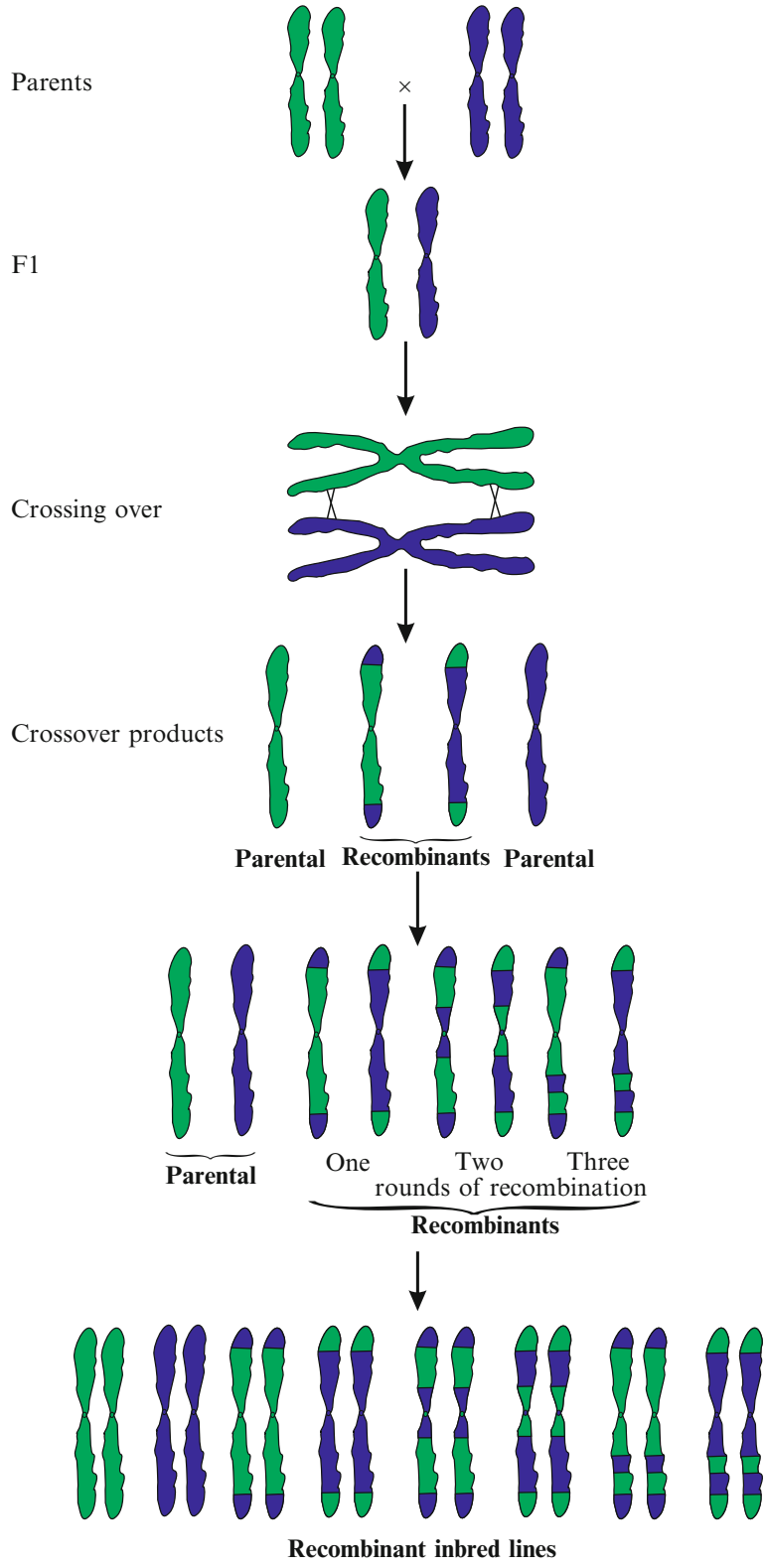


Fig. 5.2 A simplified representation of chromosome constitution of recombinant inbred lines (RILs); only one chromosome pair is shown. It is assumed that two crossing overs (one in each arm) occur in each round of recombination. In many RILs one round of recombination would take place, but in some RILs two or even three rounds of recombination would occur. The location of crossing over is assumed to be random

gene is limited to F_1 generation, i.e., there is only one round of recombination. But in the case of RIL populations, recombination would continue to occur, albeit in a progressively declining proportion of the population, for several subsequent generations. It may be pointed out that recombination will take place only in the double heterozygote *Aa Bb*. (It should be noted that crossing-over will take place almost uniformly in all the genotypes, including the homozygotes. But it will be detectable as recombination only in the *Aa Bb* genotype since only in this case the allelic combinations will be altered.) If the recombination frequency between the genes *a* and *b* were close to zero, the frequency of *Aa Bb* will be ~50 % in F_2, and it will decline in the subsequent generations by the same rate as heterozygosity. Thus, the number of generations in which recombination will involve the whole population will be about two: the F_1 will be one generation, while the F_2 and the subsequent generations together will add up to about one generation. Thus, the frequency of recombination between two linked genes in an RIL population developed by selfing would be nearly two times of that in a F_2 population, provided the two loci are <10 cM apart. *For this reason, it is often said that RILs are twice as informative as F_2 populations in terms of recombination*. But in the case of an RIL population created by sib-mating, the recombination frequency would be nearly four times as much as that in the F_2 population. This is because the rate of increase in homozygosity under sib-mating is merely one-half of that under selfing. As the distance between the marker and the target gene increases, the advantage of RILs over F_2 in terms of increased recombination frequency declines nonlinearly till the two populations become comparable for independently segregating markers and genes (Haldane and Waddington 1931). In view of the above, the chances of detection of a marker linked to the target gene are smaller in RILs than in F_2, DH, and BC populations when low marker densities are used for mapping. However, adequate marker density would not be an issue in many crop species, particularly when SNP marker system is employed. As a result of the increased recombination between closely located genes, the linkage map is expanded by a factor of about two and four, when RILs developed by selfing and sib-mating, respectively, are used for mapping, in comparison to the map based on F_2, BC, or DH population (Burr et al. 1988; Burr and Burr 1991).

Since RILs are homozygous, they can be propagated indefinitely without any further change in their genotype; this makes RILs essentially a perpetual or permanent mapping population (Burr et al. 1988). Often RILs are described as "immortal," which seems misleading since any biological entity *is* indeed mortal. Therefore, the term "perpetual" is preferable to "immortal" to emphasize the fact that these populations can be maintained/propagated indefinitely. However, RILs could become genetically variable over time, as do pure lines, due to mechanical mixture, natural outcrossing, and mutation; mutation could become important particularly for quantitative traits over long periods of time. Therefore, RIL populations should be maintained and handled with considerable care to avoid mechanical mixtures and natural outcrossing. RILs can be multiplied, shared by different researchers, and evaluated in replicated trials conducted over locations and years, which make RILs of immense value particularly for QTL mapping. The phenotypic and genotypic data and the linkage map generated from an RIL population are cumulative in that the findings from different studies using the same RIL population can be integrated, stored in a database, and shared among research workers. Finally, RIL populations yield smaller confidence limits than F_2 and BC populations when the proportion of recombination is low (Burr et al. 1988). The chief demerit of RILs is that their construction requires many (around 6–10) seasons/generations, and some parts of the genome tend to stay heterozygous for longer periods than expected from theory. In addition, the development of RILs is relatively more difficult in crops with high inbreeding depression and is problematic in obligate outcrossing species (Burr and

Burr 1991). Like DH, RIL populations can only be used for detecting additive and additive × additive components of the genetic variance.

RILs have been developed in many crops, and some RIL populations have become a public mapping tool, e.g., a population of 300 RILs obtained from the cross between Landsberg erecta and Columbia ecotypes of *A. thaliana* (Lister and Dean 1993). RILs have been widely used for the development of molecular marker linkage maps; detection of markers linked with genes governing qualitative traits like race-specific vertical disease resistance, seed or flower color, seed/fruit shape etc.; identification of markers associated with QTLs involved in the control of traits like horizontal disease resistance, yield, days to flowering/maturity etc.; mapping of genes and QTLs; and the integration of the gene/QTL maps with molecular marker maps. Several research groups have successfully developed epigenetic recombinant inbred line (epiRIL) populations. An epiRIL population of Arabidopsis was developed by crossing two parental lines that showed a little difference at the DNA sequence level, but had contrasting patterns of DNA methylation (Johannes et al. 2009). Therefore, the member lines of an epiRIL population have the same genotype, and they differ from each other only in terms of the epigenetic modifications. In contrast, the member lines of an RIL population differ from each other in their genotypes.

5.9 Immortalized F_2 Population

Gardiner et al. (1993) were the first to use the term "immortalized F_2 population" for a maize mapping population, in which the F_2 population was immortalized as follows. The F_3 progeny from a F_2 plant were intermated in two groups and the seeds from at least 20 such plants were harvested in bulk. This procedure was followed for each F_2 plant, and the resulting population was termed as "immortalized F_2 population." Later, Hua et al. (2002, 2003) developed *immortalized* F_2 (IF_2) *populations* by intercrossing a set of RILs (Fig. 5.1) and used them for genetic analysis of heterosis and detection of heterosis loci in rice. The RIL population developed from a suitable biparental cross was divided into two groups, each of which had equal number of random RILs. Each RIL of the first group was crossed with a single randomly chosen RIL of the second group with the restriction that any RIL will be involved in only one cross. This mating scheme will generate $n/2$ single crosses from a population of n RILs. Additional rounds of crosses among the n RILs can be made in the same way by fresh random pairing of the RILs of the two groups for each round of crossing. Hua et al. (2003) made three rounds of crosses among 240 F_9 RILs to generate 360 single crosses that together constituted the IF_2 population. IF_2 populations can also be developed by paired crossing of the randomly chosen RILs derived from a cross in all possible combinations, excluding the reciprocals; in this approach, the single crosses together with the parental RILs would constitute the IF_2 population. However, this approach can be used only when the number of RILs is suitably small; otherwise, the number of crosses to be made would be unmanageably large.

An IF_2 population provides a true representation of all possible genotypes, including the heterozygotes, expected in the F_2 of the cross from which the RILs were derived. Let us consider a single locus A having two alleles A and a. In the RIL population, the frequency of the alleles A and a and of the genotypes AA and aa will be 0.5 each, i.e., $p = q = 0.5$. A random mating among these RILs will produce the following three types of F_1 progeny in the frequency p^2 (= 0.25) AA, $2pq$ (= 0.5) Aa, and q^2 (= 0.25) aa, which is the same as that expected in the F_2 generation of any cross. The marker genotypes of the RILs used for creating an IF_2 population can be used to deduce the genotypes of the various F_1 progeny included in the IF_2 population. Therefore, only the RILs need to be genotyped for the markers differing between the two parents of the RIL population, and there is no need to genotype the IF_2 population itself. The IF_2 population by itself is not perpetual and is

ephemeral like any F_2 population. However, the same IF_2 population can be reconstructed from the parental RILs, which are perpetual. This feature of IF_2 is the basis for the term "immortalized" in its name. Since each F_1 progeny comprising an IF_2 population is produced from a cross between two RILs, the desired quantity of F_1 seed can be produced by fresh hybridization between the parental RILs. Therefore, IF_2 populations support replicated evaluation of F_2 genotypes over locations and permit detection and mapping of QTLs, including heterosis QTLs, and estimation of various epistatic effects. It is important that in the case of an IF_2, the plants used for measuring heterosis are hybrids themselves and not their selfed progeny as is the case with F_2-derived F_3 populations (Hua et al. 2003). The chief limitation of these populations is that their construction requires making of a large number of crosses, which could be cumbersome in at least some of the, particularly, self-pollinated crops.

Hua et al. (2003) analyzed the IF_2 population consisting of 360 single cross F_1s and identified heterotic effects at 33 loci for four traits, including yield, using modified composite interval mapping. It was observed that these heterotic effect QTLs showed little overlap with the QTLs governing the mean performance of the concerned traits. Thus, the loci involved in heterosis for a trait might be different from those that govern that trait. It was concluded that heterosis was mainly the result of single locus heterotic effects, but digenic dominance × dominance interactions also contribute to heterosis.

5.10 Near-Isogenic Lines

Near-isogenic lines (*NILs*) are pairs of homozygous lines that are identical in genotype, except for a single gene/locus. But in practice, NILs differ for the single gene and a variable length of the genomic regions flanking this locus; in addition, they may also differ for some random genomic segments located elsewhere in the genome. Thus, a pair of NILs will most likely differ for alleles at few to several loci, which justifies the use of the term "near isogenic" for such lines. NILs are generally produced by backcross procedure (Fig. 5.1), in which a donor parent (DP, a homozygous line having the trait/allele of interest) is crossed with a recurrent parent (RP, a homozygous line lacking this trait/allele), and the F_1 plants are backcrossed to the RP. The backcross (BC) generation so obtained and the subsequent BC progeny are backcrossed to the RP. In each BC generation, a strict selection is done for the trait/allele being introgressed from the DP because each backcrossing reduces the proportion of DP genome in the progeny to 50 % of that present in the previous generation. Therefore, only those individuals that have the DP allele of interest and are the most similar in phenotype to the RP are selected for backcrossing. At the end of backcross program, the progenies are selfed, and plants homozygous for the DP allele of interest and the most similar to RP in the remainder of the phenotype are selected to constitute the NIL. Thus, a NIL is essentially a segment substitution version of the RP. Repeated backcrossing eliminates the DP genomic segments unlinked to the target gene and reduces the size of DP genomic region flanking the target gene due to recombination in each BC generation (Schneider 2005). In the absence of any selection, the expected recovery of RP genome in a NIL produced by b generations of backcrossing and one terminal selfing generation will be equal to $1 - (1/2)^t$, where $t = b + 1$. Thus, an infinite number of backcrosses would be required for a complete elimination of the DP genome, but breeders generally use less than ten, most often only five to six, backcrosses to produce NILs.

Alternatively, pairs of NILs can be produced by continued selfing of the F_1 and the subsequent generations. In F_3 and later generations, progenies segregating for the target locus are identified and seeds from individual plants of such families are harvested separately to raise the next generation. Harvesting of five plants from such a family gives a 99 % probability that at least one of these plants will be heterozygous for the target locus (Pumphrey et al. 2007). In F_5 or a later generation, several plants having

the contrasting (dominant and recessive) forms of the target trait are selected from a segregating progeny, their seeds are harvested separately, and, in the next generation, individual plant progenies are grown. Harvesting of ten random plants without any reference to their phenotype would give 94 % probability of inclusion of the two homozygotes for the target locus in the sample. By this time, the plants would have become homozygous at most (~98.44 % in F_6) of their loci. A pair of homogeneous progenies, one expressing the dominant and the other showing the recessive form of the target trait, is selected from a single individual plant progeny; these progenies constitute a pair of NILs. Such pairs of NILs may be isolated from other individual plant progenies as well provided these progenies are not very closely related, i.e., they were not derived from the same F_4 or F_5 progeny. Pumphrey et al. (2007) followed this approach to develop wheat NILs for *Fhb1* QTL using three codominant SSR markers flanking the *Fhb1* locus and materials from the wheat breeding nurseries of the university of Minnesota. Bulk DNAs from five random plants from each F_4 individual plant progeny were assayed with the SSR markers to identify progenies segregating for the markers and, hence, the *Fhb1* locus. Five random plants from each segregating family were harvested separately, bulks of DNAs from five or more seeds from each of these plants were analyzed to indentify heterozygous F_4 plants, and their seeds were used to raise F_5 progenies. DNAs from at least ten plants from each F_5 family were individually analyzed to isolate the two homozygotes for the marker and the *Fhb1* loci, and the selected plants were selfed to yield pairs of $F_{4:6}$ NILs. This approach can be used for such genes/QTLs that are being used in breeding programs and for which linked markers are available. The development of NILs by this procedure does not require additional crosses, space, time, and effort, and it can be readily combined with line/variety development.

Tuinstra et al. (1997) proposed a procedure, called *heterogeneous inbred family analysis*, for rapid isolation for pairs of NILs. In this procedure, an RIL mapping population is analyzed using molecular markers associated with quantitative traits segregating in the population. This allows identification of inbred lines that are heterogeneous, i.e., segregating, for one or more of the markers. Since the RIL population would be in F_5 or more likely in F_6 generation, the plants in the heterogeneous inbred lines would be isogenic for most of the other loci. The homozygotes for the two alleles of a segregating marker locus are selected from each heterogeneous family, and the pair of homozygous lines isolated from a single inbred line forms a pair of NILs. Tuinstra et al. (1997) screened a population of 98 heterogeneous inbred families in sorghum with two unlinked RAPD markers known to be associated with seed weight. They identified three segregating inbred lines for each marker. From each segregating inbred line, a pair of NILs was isolated. Evaluation of the NILs confirmed that the two QTLs for seed weight linked to the RAPD markers were expressed in these NILs.

The NILs developed by backcrossing are identical to the concerned RPs, except for the DP genome segment having the gene of interest and, possibly, some other DP genomic segments as well. However, the two members of each NIL pair developed through selfing are identical with each other, except for the DP genome segment with the gene of interest, but they would invariably differ from some to considerable extent from the parents of concerned crosses. In either case, the pairs of NILs would differ for the alleles of the target gene and for the alleles of markers linked to the target gene; in addition, the NILs developed by backcrossing would also differ for alleles of markers located in the random DP genomic segments retained in them. The ratio of marker alleles in a group/population of NIL pairs developed by transferring the same gene from a DP into several different RPs is expected to be 1:1 irrespective of the marker being dominant or codominant. Thus, NILs developed by backcrossing differ from the respective recurrent parents for an unknown number of DP-derived molecular markers, some of which may not be linked to the gene introgressed from the DP. However, at most (but not all) of the loci

not linked to the introgressed gene, RP alleles would be restored primarily due to backcrossing per se; this will be supplemented by a successful selection for the RP phenotype. This difference between the two members of NIL pairs provides the basis for identification of markers linked with the target gene/QTL simply by comparing the allelic states of molecular markers in the RP and its various NIL derivatives. Linkage between the introgressed gene and a molecular marker would be presumed whenever a NIL and RP have different marker alleles, and the NIL allele is the same as that present in the DP. When applied to a large number of available NILs, this approach would be very useful in detecting linkage between introgressed genes and molecular markers (Muehlbauer et al. 1988). NILs may provide a convenient approach for integration of conventional genetic markers, i.e., genes, into an existing linkage map of molecular markers by identifying linkage of the introgressed genes with one or more of the already mapped molecular markers. This would allow a tentative integration of the introgressed gene into the molecular linkage map. However, multipoint linkage analysis will be required to confirm or refute the presumed linkage and to determine the specific position of the introgressed gene in the molecular marker linkage map. The prior knowledge of the putative linkage between the target gene and the molecular markers in the linkage map would greatly increase the efficiency of traditional multipoint linkage analysis.

Like DHs and RILs, NILs are homozygous and perpetual mapping resources. Many NILs are available in several crop species, e.g., soybean, tomato, rice, etc., as a result of routine breeding activities. For example, rice NILs carrying major blast resistance genes (*Pi54*, *Pita*, *Pi1*, *Pib*, *Pi2*, *Pi5*, *Pi9*) in the genetic background of Pusa Basmati 1 have been recently developed. The recovery of recurrent parent genome was hastened by marker-assisted background selection (Khanna et al. 2015). These NILs form an existing mapping resource, and can be used for the identification of markers linked to the introgressed genes/QTLs and other genetic and functional genomics investigations. The introgressed genomic region from the DP is often highly polymorphic at DNA sequence level, which is helpful in rapid identification of molecular markers located near the introgressed gene (Young et al. 1988). This approach requires analysis of only three DNA samples, viz., DP, RP, and NIL DNAs, for detection of markers having different alleles in the RP and the NIL. These markers will have the same marker allele in the DP and the NIL and will be located in the genomic region flanking the gene of interest. This is in contrast to the genetic mapping based on RILs and other mapping populations, where the whole of the population has to be tested for every marker to identify those linked with the gene of interest. Evaluation of NILs allows a more reliable assessment of QTL effects since the QTL is placed in the genetic background of the RP that is used for comparison. However, linkage drag is a potential problem in such studies, particularly when genes/QTLs are introgressed from unadapted germplasm. NILs can be used to construct high-resolution mapping populations. For example, NILs derived through selfing are intercrossed, while those derived through backcrossing are crossed with the RP to generate large F_2 mapping populations. Finally, they are quite useful in functional genomics; they can be used for gene expression profiling and for more direct hypothesis-driven experimentation (Pumphrey et al. 2007). In addition, NILs and chromosome segment substitution lines (CSSLs) are suitable for fine mapping and map-based cloning of QTLs, while RILs and DHs are not suitable for these applications (Xu et al. 2010). The development of NILs requires at least 6–8 generations of backcrossing or selfing after the F_1 generation is produced. They can be directly used for molecular tagging of only the introgressed genes, but they themselves do not support linkage mapping. Perhaps the most serious potential limitation of the NIL mapping approach concerns the extent of marker diversity between the RP and the DP genomes. Therefore, it may be advisable to first assess the marker diversity between the DPs and the RPs and use only polymorphic markers for analysis of the NILs.

5.11 Chromosomal Segment Substitution Lines

Chromosome segment substitution lines (*CSSLs*), introgression lines, or intervarietal substitution lines are a series of homozygous lines, each having a single distinct chromosome segment from a DP in the chromosome background of RP. Further, the sum total of all the DP segments present in the complete set of CCSLs equals the haploid chromosome complement of the DP. Thus, each line of the set of CCSLs has successively overlapping DP chromosome segments beginning from the top of DP chromosome 1 to the bottom of last DP chromosome; the overlapping segments ensure the representation of the entire DP genome (Eshed and Zamir 1994). The DP can be another variety of the same crop species or a related species. A line possessing a chromosome fragment from a related species has been called introgression line, while a line having a chromosome segment from a different variety of the same crop species was termed as intervarietal substitution line (Schneider 2005). However, the term introgression is used to describe gene transfer from other varieties of a crop species as well as from its related species, and in plant breeding, the term substitution line usually refers to a line, in which a whole chromosome is substituted with the same chromosome from another variety (Allard 1960). In order to avoid confusion, it is suggested that all CSSLs may be called segment substitution/segment introgression/chromosome segment introgression/chromosome segment substitution lines. However, for the CSSLs having genomic segments from related species, the term "alien" may be prefixed to the above names, yielding the terms alien segment substitution and alien segment introgression lines. A set of CSSLs can be considered similar to a genomic library with a huge genome insert in the genetic background of RP and is often referred to as introgression line library (ILL) or exotic genetic library (Eshed and Zamir 1994; Zamir 2001).

The CSSLs may be produced by backcrossing the F_1 and the subsequent progeny from a cross between the DP and the RP with the RP for six generations or so, followed by self-fertilization for two or more generations to isolate lines homozygous for the introgressed segments. Selection based on markers evenly distributed over the entire genome is used to ensure that each line of the set has a distinct but slightly overlapping DP genome segment. Eshed and Zamir (1994) developed a set of alien CSSLs through introgression of overlapping chromosome segments (average size ~33 cM) from the green-fruited *Lycopersicon pennellii* (line LA716) into the cultivated tomato (*L. esculentum*) variety M82, using a combination of backcross and pedigree programs. The F_1 from the cross M82 × LA716 was backcrossed to M82, and from 600 BC_1 plants, 99 were selected for horticultural characteristics and selfed to obtain BC_1S_1. Selfing was continued for five more generations to produce BC_1S_6 progeny, and the selfed generations were handled as per pedigree scheme. In each selfed generation, 1,500 plants were raised and 100 plants were selected on the basis of horticultural characteristics. In BC_1S_6 generation, plants were analyzed using 175 RFLP markers, and a set of such plants whose introgressed segments together represented the entire genome of *L. pennellii* was selected. These plants were backcrossed to M82; after two more backcrosses, BC_3 plants were analyzed with 350 molecular markers, and 50 such plants were selected that had a single *L. pennellii* chromosome segment, which together represented the entire 1,200 cM of *L. pennellii* genome. The selected plants were selfed and plants homozygous for the introgressed genome segments were selected to yield a library of 50 alien chromosome segment substitution lines.

CSSLs have been developed in several crop species, and in rice alone, several CSSL populations have been created (Xi et al. 2006; Xu et al. 2010). For example, Xu et al. (2010) developed a set of 128 CSSLs ($BC_5F_{2:3}$ or $BC_6F_{2:3}$) in rice in the genetic background of *indica* variety 93–11 carrying chromosome segments from the *japonica* variety Nipponbare. Analysis of the CSSLs with 254 PCR-based

markers revealed them to carry a total of 142 chromosome segments from the DP that summed up to 882.2 Mb of the DP genome that is ~2.37 times the size of rice genome. However, sequencing-based analysis of the CSSLs identified 117 new DP chromosome segments, each of <3 Mb length, that were not detected by the PCR-based markers. Multiple regression analysis of the CSSLs detected nine QTLs explaining 89.5 % of the phenotypic variance for culm length. A large effect QTL was identified in the genomic region that harbors the rice "green revolution" gene. Some of the CSSLs were superior to the RP 93–11 in some characteristics, with potential to be released as new varieties. It is important that characterization of the CSSLs with the 254 PCR-based markers took 3 years, while only 7 weeks was needed for their sequencing-based characterization.

CSSLs are a perpetual mapping resource and are suited for mapping of both oligogenes and QTLs. They can also be used for fine mapping by raising large F_2 or backcross populations following hybridization with the RP (Eshed and Zamir 1994). CSSLs do not suffer from the limitations of conventional mapping populations, such as (1) limited resolution, (2) inability to detect QTLs with small effects, and (3) interference in QTL detection due to QTL × QTL interactions. Since each CSSL is an equivalent of the recurrent parent, except for the chromosome segment introgressed from the DP, any phenotypic difference between the RP and a CSSL would be due to the DP chromosome segment. Evaluation of CSSLs in replicated trials over locations and years would allow the identification of such lines that have DP genomic segments with favorable effects on the traits of interest. CSSLs can be used for the detection of QTLs with small additive effects that are ordinarily masked by QTLs with larger effects in the usual mapping populations like F_2 and RILs. QTL identification using CSSLs does not require linkage map construction or statistical analysis. Further, each CSSL can be directly used for mapping and cloning of QTLs/genes and for development of elite breeding lines. The CSSLs developed by Eshed and Zamir (1994) were evaluated in replicated yield trials along with their hybrids with the recurrent parent and with one other tomato variety (Eshed and Zamir 1995). QTL mapping using these data revealed 23 and 18 QTLs for the total soluble solids and fruit mass, respectively; these numbers are about two-fold larger than those reported earlier using conventional mapping populations. Fine mapping of a fruit mass locus represented in two introgression lines revealed three linked QTLs. For fine mapping, the two introgression lines were crossed with the RP M82, a large F_2 population was raised and subjected to RFLP analysis to identify plants with small portions of the genomic region represented in the introgression lines, and the selected plants were selfed to isolate homozygous lines that were evaluated for QTL mapping. These findings amply demonstrate the unique mapping opportunities offered by CSSLs.

CSSLs would provide a better understanding of the number of genes governing a trait, distribution of these genes over the genome, the effects of individual genes/QTLs, and the manner in which they interact with each other and the environment (Burns et al. 2003). QTL detection using CSSLs is free from epistatic effects of the rest of the DP genome as the introgressed segments are usually small, and the QTLs are generally mapped into smaller confidence intervals. QTLs involved in heterosis may be identified by crossing individual CSS lines to a suitable tester and evaluating their F_1s. Epistatic interaction of a QTL of interest can be assayed by crossing the CSS line having the QTL with several different lines and evaluating their F_1 progeny. This can also be done by developing reciprocal CSSLs. In a pair of reciprocal CSSLs, parental line A serves as DP and line B is used as RP in one set of CSSLs, while B serves as DP and A functions as RP in the second set of CSSLs (Peleman et al. 2005).

The main disadvantage of CSSLs is that they might have undesirable traits linked to the target gene(s) because of the large introgressed chromosomal segment; this would be more likely when unadapted germplasm is used as DP. Linkage drag, where encountered, would necessitate further breeding effort, which may

be problematic in the cases of alien CSSLs due to reduced pairing and recombination between the DP and RP chromosomes.

5.12 Backcross Inbred Lines

Backcross inbred lines (*BILs*) are developed by backcrossing the F_1 from a cross between two homozygous lines to one of the parents and continued selfing of the BC_1F_1 progeny to obtain homozygous lines. Sato et al. (2003) produced a set of 98 BILs in rice by backcrossing the F_1 from the cross Nipponbare (*japonica*) × Kasalath (*indica*) to Nipponbare and continued selfing of the BC_1F_1 progeny to obtain BC_1F_5 lines. The data from BIL population were analyzed using the method for backcross F_2 population and treating the heterozygotes as missing data since a method for analysis of BIL population was not available. A possible advantage of BILs may be the increased frequency of the alleles contributed by the parent used for backcrossing. Therefore, it would be desirable to use the parent with the higher value of the target trait for backcrossing with the F_1 hybrid.

5.13 Advanced Intercross Lines

An advanced intercross line (*AIL*) population is developed by intermating the individuals of F_2 and subsequent generations from a suitable cross. Intermating in the segregating generations maintains heterozygosity in the population and allows recombination between the QTLs and the markers linked to them in every generation leading to a more precise location of the QTLs. It was estimated that the confidence interval of QTLs would be reduced by up to five-fold in AILs as compared to that in an F_2 population (Darvasi and Soller 1995). In the case of AILs, mapping resolution seems to improve for up to eight generations of intercrossing only, while it continues to improve with generation in the case of recurrent selection backcross. Further, appropriate statistical methods for modeling and analysis of the data from AILs are not available (Luo et al. 2002).

5.14 Recurrent Selection Backcross Population

Wright (1952) put forth the idea of recurrent selection backcross (RSB) procedure for isolating QTLs with large effect. In this scheme, the F_1 obtained from a cross between a homozygous line with high value for a quantitative trait (the DP) and a homozygous line with low value for the trait (the RP) and the subsequent backcross progeny are backcrossed to the RP. In each backcross generation, a predetermined number of individuals with the top phenotypic values (i.e., DP phenotype) for the trait are selected and backcrossed to the RP. RSB is proposed to be used for high-resolution QTL mapping, for which a sufficiently large number of backcrosses need to be made. Obviously, this will require considerable effort, resources, and time. Further, RSB is suited for localization of large effect QTLs, while important quantitative traits like yield are mainly governed by many QTLs with moderate to low effects.

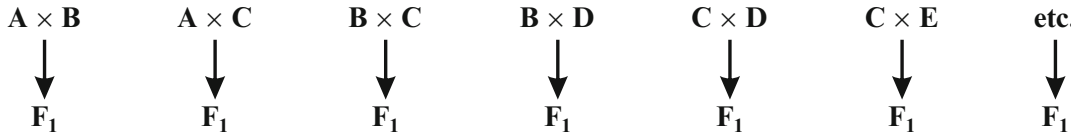
Recurrent backcrossing will lead to homozygosity at a rapid rate for RP alleles at all the loci that are not affected by the phenotypic selection for the trait. However, selection will slow down the progress to homozygosity at those loci that are involved in the control of the trait as well as those linked to these loci. Genetic drift, on the other hand, will increase the rate at which homozygosity is reached in the backcross populations. It may be expected that phenotypic selection will maintain the DP alleles of those QTLs that have large effect on the quantitative trait, while those having moderate and small effect will be lost. For example, the frequency of DP QTL allele did not change even after 50 generations of RSB if it explained 50 % of the phenotypic variance, while it disappeared completely after 30 generations if it accounted for only 15 % of the trait variance. Therefore, large effect QTL alleles from the DP and the molecular markers linked to them will be retained in a heterozygous

state during RSB. Further, recombination between the DP QTL alleles and the linked markers will take place in each generation. Therefore, the level of heterozygosity at these marker loci will go on decreasing with the increasing number of RSB generations. In addition, in a given generation, markers located farther from the QTLs will show greater reduction in heterozygosity than those located closer to the QTLs. Thus, the frequency of heterozygosity at marker loci can be used as a criterion of localizing the QTLs. Theoretical and simulation studies suggest that RSB may enable detection of markers located at or less than 1 cM from a QTL if sufficiently dense markers were available. With a fixed total number of individuals in the population, a smaller number of larger size families are more likely to reduce the effects of genetic drift than a larger number of smaller size families, but it would also reduce the resolution of mapping (Luo et al. 2002).

Recurrent selection backcross inter se intercross (RSBI) scheme is a modification of RSB scheme, in which the selected individuals are intercrossed at one or more stages during the backcrossing. Intercrossing reduces the approach to homozygosity and increases the retention of heterozygosity at DP QTL loci. However, appropriate statistical method for the analysis of data from these populations needs to be developed (Luo et al. 2002).

5.15 Interconnected Mapping Populations

Interconnected mapping populations are produced by crossing a set of homozygous parental lines in such a way that two or more crosses have one parent in common (Fig. 5.3). Interconnected populations were first used by Gilbert (1985a, b) to partition single gene effects from the overall effects of the polygenes estimated from diallel crosses. Generally, half-diallel mating design without selfs has been used, but nested, round robin (Sect. 8.5.2), factorial, or any other mating design in which two or more crosses share one parent could be used. In a diallel mating design, a set of n parental lines are mated in all possible combinations, including reciprocals and selfs to generate n^2 F_1 progeny. In a half-diallel mating design used for creating interconnected populations, the reciprocal crosses and selfs are excluded to obtain only $n(n-1)/2$ different crosses. In a factorial mating design, the n parental lines are divided into two equal groups, and each of the $n/2$ lines of the first group is mated with each line of the second group. An interconnected population may consist of F_2, backcross, RIL, or DH populations generated from each of the crosses produced as per the mating design used. As expected, F_2 and



In each cross, one of the following populations is generated:

- **F_2 population (selfing in F_1)**
- **Backcross population (BC_1F_1; F_1 backcrossed to one of the parents)**
- **DH population (haploid production, followed by chromosome doubling)**
- **RIL population (continued selfing to F_5 or F_6 using single seed descent scheme)**

[Populations from all the crosses taken together constitute an interconnected population]

Fig. 5.3 Schematic representation of an interconnected population. These populations involve several parents

backcross interconnected populations would be ephemeral, while RIL and DH populations would be perpetual.

Verhoeven et al. (2006) carried out simulation analysis of a half-diallel population for estimating variances for general combining ability (GCA) and specific combining ability (SCA), and for detection of QTLs involved in control of the trait. They concluded that the use of a larger family and a smaller number of parents is more efficient than using a small family size and a large number of parents. Further, for a fixed total population size and number of parental lines, less interconnected designs have smaller number of larger size families than more interconnected designs. Therefore, a less interconnected design like single round robin should be used in an initial study, while a more interconnected design like half diallel should be used for subsequent detailed analyses. Data from such populations have been analyzed using linear regression, general linear model, multiple QTL model, composite interval mapping (CIM), and joint inclusive CIM (JICIM; used with NAM populations) for QTL detection and mapping (see Li et al. 2011a). The software package MCQTL (Sect. 7.19.10) is designed for multi-allelic QTL mapping in multi-cross design, including diallel mating design; it uses linear regression model, and employs composite interval mapping and an interactive QTL mapping to deal with the multiple QTL models (Jourjon et al. 2005).

The usefulness of QTL findings in plant breeding depends on their general applicability and an understanding of the genetic architecture of the traits governed by the QTLs. Biparental mapping populations generate QTL information applicable to the concerned crosses, and they fail to take into account segregation of different allelic combinations of QTLs in different mapping populations and the influences of genetic background on QTL effects. Generalization of QTL findings from different biparental populations has been attempted by comparing the relative QTL positions determined from different populations by means of QTL meta-analysis (Sect. 7.12), and bioinformatics tools are being developed to facilitate this analysis (see Verhoeven et al. 2006). In contrast, joint analysis of data from interconnected populations provides more generalized information about QTL positions and effects, increases QTL detection power, enables detection and assessment of QTL × genetic background interaction, and permits identification of markers located closer to the QTLs than do biparental populations, particularly when appropriate analysis tools are used (Jannink and Jansen 2001; Verhoeven et al. 2006; Li et al. 2011a). The chief limitation of interconnected populations is that their construction requires considerable effort, time, and resources; therefore, a cooperative effort of several groups would be desirable (Verhoeven et al. 2006).

5.16 Multiparent Advanced Generation Intercross Populations

The *multiparent advanced generation intercross* (*MAGIC*) *populations* are a collection of RILs produced from a complex cross/outbred population involving several parental lines (Fig. 5.4). The parental lines may be inbred lines, clones, or individuals selected on the basis of their origin or use. MAGIC populations are an extension of the AIL proposed by Darvasi and Soller (1995; Sect. 5.13), but differ from them with respect to the involvement of multiple parents in their construction. This concept was first used in mice as "heterogeneous stocks" and later extended to plants by Mackay and Powell (2007), who also proposed the name MAGIC. Huang et al. (2015) provide an excellent discussion on various aspects of MAGIC populations, including recent achievements from, and the unique opportunities and advantages offered by the MAGIC populations. A simple approach to generate a MAGIC population is to produce a complex cross involving multiple, typically eight, parental lines and to isolate RILs from this cross. The eight parental lines are crossed in pairs to produce four different single crosses, and these single crosses are crossed in pairs to generate two

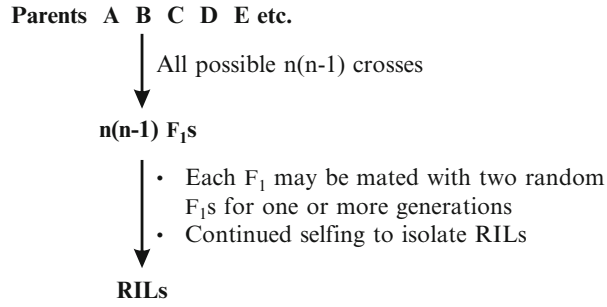


Fig. 5.4 Schematic representation of the development of a MAGIC (multiparent advanced generation intercross) population. This figure is based on the scheme used by Scarcelli et al. (2007) for developing a MAGIC population of 1,026 MAGIC lines in *A. thaliana*. MAGIC populations also involve several parents

double crosses. Finally, the two double crosses are mated together to produce an eight-parent complex cross. This complex cross is handled as per the SSD procedure to develop the required number of RILs, which together constitute the MAGIC population.

A more elaborate procedure was used to construct a MAGIC population of *A. thaliana*. This population was created by crossing 19 accessions in a diallel fashion to generate 342 F_1s, each of which constituted a family. From each family, two random plants were used as males and two random plants were treated as females. The male plants of a family were crossed with female plants of another randomly selected family, so that 342 such crosses were made; this process was repeated in the next two generations (Scarcelli et al. 2007). In the third generation, a single random plant was selected from each cross and selfed, and the 342 segregating populations were handled according to SSD procedure for six generations. In the end, up to three inbred lines, called MAGIC lines (MLs), were selected from each population, giving a total of 1,026 MLs, which are available through the Arabidopsis Stock Centre (http://www.arabidopsis.org). The MLs from a single family/cross are expected to share about 25 % of their genomes by descent. Further, each family has genomic contributions from, on an average, 9.97 accessions. More recently, Bandillo et al. (2013) used a far more elaborate procedure to develop MAGIC populations of *indica* and *japonica* ecotypes of rice. They mated 8 elite *indica* lines as per half diallel scheme, crossed the resulting 28 F_{1s} to produce 70 four-parent crosses, and then mated these F_{1s} to generate 35 eight-parent crosses. Each eight-parent F_1 and its later generations were selfed to isolate RILs. Similarly, the *japonica* MAGIC population was created. In the end, the two MAGIC populations were mated together to generate a global MAGIC population for rice.

MAGIC populations are perpetual, lack population structure, can be used for both linkage and association analyses, and can be developed at an appropriate stage during the intermating process to afford the desired mapping resolution. Since these populations are created from several parents, they are likely to show segregation for multiple traits, multiple QTLs for each trait, as well as more than two alleles for individual QTLs. They are an ideal resource for construction of high-density maps, and they allow modeling of cytoplasmic effects. In addition, the parents of a MAGIC population may be selected to represent a large part of variation present in the elite germplasm of a crop species. These populations can be used directly or indirectly for variety development when they are based on elite parental lines possessing a combination of useful traits. These populations can be used as training populations for genomic selection (Chap. 10). They are preferable as training

populations to the collections of breeding lines and cultivars because they are devoid of population structure, which is common in the latter.

A MAGIC population of over 1,000 MLs would enable assessment of two-way and three-way epistatic interactions (Cavanagh et al. 2008). A QTL mapping method based on reconstruction of haplotype mosaics of each MIL was able to map QTLs explaining 10 % of the phenotypic variation within confidence interval of 300 kb (as against 2–20 Mb for biparental populations) when 527 MILs were used for analysis. The confidence interval would decline to about 200 kb if the number of MILs used for analysis was doubled (Kover et al. 2009).

5.17 Nested Association Mapping Population

In order to combine the advantages of both linkage mapping and association mapping strategies, a structured population generated by crossing a set of diverse founder parents to one or two common parents has been suggested (Yu et al. 2008). Each selected founder is crossed to one or few common parents (nested parents) and a set of 250 RILs from each of these crosses is generated using the SSD method. For example, a population of 5,000 RILs was generated using 26 founder parents and one nested parent B73 in maize (Sect. 8.5.2; Yu et al. 2008). The nested association mapping strategy enables efficient utilization of genetic and genomic resources for genetic dissection of complex traits.

5.18 Mapping Populations for Cross-Pollinated Species

In a number of cross-pollinated species like maize, nearly homozygous inbred lines can be developed and used for creating suitable mapping populations. But in many cross-pollinated species, development of inbred lines is not feasible due to long breeding cycle, self-incompatibility, and/or severe inbreeding depression. The examples of such plant species include tree species like apple, pear, and grape and annual species like potato. The parents used for hybridization in such species are heterozygous and the mapping populations consist of the F_1 generation or backcross lines. In the case of tree species, F_1 generations from crosses between selected varieties are used for mapping. The two parents of a cross would contribute different alleles to the F_1 individuals, and linkage among molecular markers is assessed to develop a genetic map for either parent. A backcross population of potato was developed by pollinating an individual F_1 plant with one of the parents, and the F_1 generation, the parental lines, and the backcross progeny were maintained by clonal propagation (Gebhardt et al. 1989, 1991).

5.19 Linkage Mapping in Polyploid Species

Linkage mapping in polyploid species is complicated by several factors, including a complex segregation pattern, production of a larger number of genotypes for a single locus than in diploid species, multiple and co-migrating fragments in the case of markers like SSRs, and poorly characterized chromosome pairing and recombination pattern. In a polyploid species, several genotypes are produced for a single locus and the segregation ratio is rather complex. For example, in an autotetraploid species, a single locus can have five different genotypes, viz., *AAAA*, *AAAa*, *AAaa*, *Aaaa*, and *aaaa*. Segregation in an *Aaaa* individual will produce two types of gametes *Aa* and *aa* in the ratio 1:1 when there is regular bivalent formation or when there is regular quadrivalent formation, but there is no recombination between the gene and the centromere. In F_2 generation, three genotypes (*AAaa*, *Aaaa*, and *aaaa*) will be obtained in the ratio 1:2:1, and for a dominant marker, the ratio will be 3:1. But when there is quadrivalent formation and crossing-over takes place between the gene and the centromere, two sister chromatids may end up at the same pole producing the *AA* gamete; this is known as *double reduction*. The segregation distortion observed in autotetraploids is

mainly due to double reduction. The frequency of double reduction depends primarily on the genetic distance between a locus and the concerned centromere. When there is regular quadrivalent formation and the frequency of recombination between the gene and the centromere is 50 %, three types of gametes (*AA*, *Aa*, and *aa*) will be produced in the ratio 1:12:15, and the ratio in F_2 will be 1 *AAAA*:24 *AAAa*:174 *AAaa*:360 *Aaaa*:225 *aaaa* (Allard 1960). However, the above assumptions are unrealistic and the actual situation is generally not known. In any case, the complex segregation pattern makes genetic analysis in the polyploid species quite challenging. Further, since a polyploid species has more than one homologous or homoeologous genome in its haploid complement, a single SSR locus would generate more than two fragments. Usually, it is very difficult to decide which one of these fragments is allelic and which ones are paralogous, i.e., produced by loci in different chromosomes. In polyploid species like bread wheat that behave like diploids, therefore, one may deliberately select such markers that are polymorphic in only one of the homoeologous genomes. Alternatively, one may use dominant markers, e.g., AFLPs and RAPDs, that occur in a single dose in autopolyploid species so that the gametes are produced in 1:1 ratio, and a 3:1 ratio is obtained in F_2 (Sorrells 1992).

The various types of populations that can be used for mapping in polyploids are as follows: (1) mapping populations of diploid progenitors or relatives of the polyploid species, (2) a population of F_1 hybrids from a cross between the polyploid and a diploid progenitor or related species, (3) aneuploid stocks, (4) haploid populations, and (5) doubled haploid (DH) populations of the polyploid species. When gene synteny is conserved among related species, mapping in a diploid progenitor or relative of the polyploid species is distinctly advantageous. For example, this approach has been extensively used in the case of polyploid *Brassica* species. In such cases, it is important that the diploid species should exhibit high polymorphism, and the chromosomal rearrangements, if any, in the polyploid species as compared to the diploid relative should be well characterized to permit the transfer of marker information to the polyploid species. However, it is desirable that linkage analysis be done at the polyploid level because meiotic processes greatly differ between autopolyploids and diploids, the genome evolution in polyploids is extremely dynamic so that the polyploid genomes may not directly correspond to the diploid genome, and the diploid relatives of some polyploids may no longer be available (Luo et al. 2004). The polyploid species may be crossed with a diploid relative, and the F_1 population can be used for mapping. It is expected that the marker alleles contributed by the diploid relative will be easily distinguishable from those of the polyploid species. As a result, only one copy of each genome will be present in the F_1 plants, and their analysis would readily allow determination of the marker allele frequencies and reconstruction of the parental genotypes. This situation is essentially comparable to that obtaining in haploid or DH populations. In this approach, the two parents should be highly heterozygous and large mapping populations with 500 or more plants should be used. Aneuploid stocks are useful for mapping in polyploid species with moderate to low polymorphism. These stocks also allow the assignment of markers to specific chromosomes or chromosome arms; for example, such maps have been developed in wheat. The haploid and DH populations facilitate mapping by eliminating the confusion caused by heterozygosity in scoring of the marker alleles. Since the haploid/DH individuals will have only one marker allele present in each genome, any polymorphism present in an individual will be due to variation among paralogous loci (Sorrells 1992).

In the mapping strategy developed by Wu et al. (1992), the first step comprises identification of markers that segregate in a 1:1 ratio. The marker genotype data are analyzed to classify the individuals, for each marker pair, into two

groups, viz., coupling phase (++/−−) and repulsion phase (+−/−+), to identify linked markers. This is followed by grouping and ordering of the markers into a linkage map based on recombination fractions using a suitable mapping tool like MapMaker (Sect. 6.14.1). However, mapping in autotetraploids is a challenge in view of the phenomenon of double reduction. Luo et al. (2004) proposed a general theory for linkage mapping in autotetraploids and a statistical model for estimating the frequencies of double reduction and recombination between pairs of loci using both dominant and codominant marker data. Their model takes into account most of the essential features of segregation and recombination in autotetraploid species, including null alleles, alleles present in multiple dosages, segregation distortion as a result of double reduction, formation of variable numbers of bivalents and quadrivalents, and incomplete information on the relationship between phenotype and genotype. Their method involves computation of the conditional probability distribution of progeny phenotypes with the given phenotypes of their parents. Then, the expectation maximization algorithm is used for computing the maximum likelihood estimates for the model parameters. Further, the likelihood-based method enables prediction of the most likely parental genotypes at the linked loci. It may be added that efforts are being made to develop linkage maps for many important polyploid crop species like potato, sugarcane, alfalfa, etc.

5.20 Chromosome-Specific Genetic Stocks

Chromosome-specific genetic stocks facilitate localization of new mutations to specific chromosomes/chromosome arms by screening of a segregating generation derived from a cross between the new mutant and the genetic stock. The first examples of genetic stocks of this type were mutant lines with one or more visible mutations mapped to specific chromosomes or chromosome arms. For example, line W100 of *A. thaliana* is one such multiple marker line, which has specific visible mutations that identify each arm of the five chromosomes of the species (Koornneef and Vanderveen 1983). The marker line with the genotype *aa*, *bb* to *zz* is crossed with the mutant line (genotype *mm*), and frequencies of the double mutant, viz., aa mm, bb mm, and zz mm, phenotypes are scored in the F_2 generation. It is expected that the frequency of each double-mutant phenotype would be 1/16 if the two loci were segregating independently. Therefore, a significant reduction in the frequency of a double mutant from the expected 1/16 would indicate a linkage between the new mutant and the concerned mutant already mapped to a chromosome arm. In general, the closer are the two loci in the chromosome, the greater will be the reduction in frequency of the concerned double-mutant phenotype.

5.21 Natural Populations and Germplasm/Breeding Lines

The natural populations, germplasm lines, and even a collection of different breeding lines can be used as mapping populations for linkage disequilibrium-based mapping called association mapping discussed in some detail in Chap. 8.

5.22 Segregation Ratios in Mapping Populations

The genotypic segregation ratio observed in a mapping population for a marker locus depends on whether the marker is dominant or codominant and on the mapping population itself (Table 5.3). In the case of a codominant marker, the heterozygote can be clearly differentiated from the two homozygotes, while the heterozygote for a dominant marker will be identical to the homozygote showing "presence" of the marker. Similarly, some mapping populations consist of only homozygous individuals, while others have both homozygotes and heterozygotes in ratios that differ predictably with the population type. As a result, codominant markers are more informative than dominant markers in

Table 5.3 Segregation ratios at dominant and codominant marker loci in different mapping populations

	Segregation ratio					
					Backcross population	
Marker type[a]	F_2	RILs	DHs	NILs	B_1[a]	B_2[b]
Codominant	1:2:1	1:1	1:1	1:1	1:1	1:1
Dominant	3:1	1:1	1:1	1:1	1:0	1:1

[a]Dominant markers: RAPDs, AFLPs, most SCARs, ISSRs, SNPs, DArT, SFPs, RAD markers. Codominant markers, RFLPs, SSRs, CAPSs

[b]B_1 backcross with the parent having the dominant allele for the marker/trait, B_2 backcross with the parent with the recessive allele for the marker/trait

mapping populations having heterozygous individuals. Further, a clear understanding of the segregation ratios for various molecular markers in different mapping populations is critical for assessing whether the marker loci are segregating as per expectation and for deciding the statistical analyses appropriate for the marker data.

Markers like RFLPs, microsatellites, and CAPS are codominant, while AFLPs, RAPDs, ISSRs, DArT, and SNPs are ordinarily dominant markers and are scored as "presence" and "absence" alleles. Mapping populations such as RILs, DHs, and BILs equalize the two marker types because they consist of only homozygous individuals. In RIL and DH populations, the two alleles at each marker locus are present in homozygous state in 1:1 ratio in the case of both dominant and codominant marker loci. However, a BIL population will show 3:1 ratio for the RP and DP alleles of both the marker types. In contrast, mapping populations like F_2, F_2-derived F_3, backcross, and immortalized F_2 populations consist of both homozygous and heterozygous individuals. As a consequence, they would show different segregation ratios for the dominant and codominant markers. In the case of F_2, F_2-derived F_3, and immortalized F_2 populations, codominant markers segregate in 1:2:1 ratio, while 3:1 ratio is obtained for dominant markers. In a backcross population, codominant markers will show 1:1 ratio irrespective of whether the P_1 or P_2 is used for the backcross. In contrast, dominant markers will show a 1:0 ratio in B_1 population (backcross to the parent with the "presence" allele) and a 1:1 ratio in B_2 population (backcross to the parent with the "absence" allele).

5.23 Characterization of Mapping Populations

Precise characterization of the individuals/lines of a mapping population for genotypes of molecular markers (*genotyping*) and phenotypes of the traits of interest (*phenotyping*) is vital for the success of any mapping project. The molecular marker genotypes of any individual are independent of the environment. However, trait phenotypes, particularly those of quantitative characters, would be affected by the environment and are also likely to show G × E interaction. Therefore, it becomes important to precisely evaluate quantitative trait phenotypes by planting the mapping populations in replicated trials, preferably, over locations and years. This would, however, require the use of a perpetual mapping population like RILs, DHs, BILs, etc.

5.24 Problems in Mapping Studies

One of the problems encountered in mapping studies concerns limited variation at the DNA sequence level detectable as alleles of molecular markers in the elite germplasm of some important crop species. For example, crosses between cultivated varieties of tomato show exceptionally low polymorphism for RFLP markers (Miller and Tanksley 1990) and only small variation for SSR alleles (Areshchenkova and Ganal 2002). The low polymorphism is likely to result from domestication/introduction of limited germplasm and development of the modern varieties from a relatively small number of lines. This problem

may be resolved by using unadapted germplasm, including a related species, in crosses with the adapted germplasm lines to generate the mapping populations. Another problem faced in some mapping studies relates to segregation distortion for some of the molecular markers. A significant deviation of the observed segregation ratio for a marker locus from the expected ratio in a mapping population is called *segregation distortion* (Lyttle 1991). Segregation distortion may be due to meiotic drive or preferential segregation, selective abortion of male or female gametes, selective fertilization of certain gametes, selective zygotic lethality during seed development, germination and plant growth, and sampling error and/or unintended selection during mapping population development (see Xu et al. 1997). Self-incompatibility loci, hybrid sterility loci, and wide compatibility loci may also cause distorted segregation (Gebhardt et al. 1991). In addition, differential responses of pollen grains with different genotypes to the anther culture procedure could lead to segregation distortion in DH populations. Segregation distortion can occur for some specific markers in a mapping population that shows normal segregation for the rest of markers. It is, therefore, important that the "goodness of fit" of segregation ratio should be tested for each marker locus and, if necessary, data concerning markers showing a high degree of segregation distortion may be excluded from further analyses. Alternatively, one may use one of the software designed for analysis of marker data with distorted segregation.

5.25 Size of Mapping Population

It is important that the mapping population should be as large as feasible since population size is associated with several valuable aspects of mapping studies. The confidence interval for linkage estimates is smaller in a larger population than that in a smaller population (Silver 1985). When the population size is increased from 50 to 100, the 95 % confidence limit for the distance between two markers when no recombination is detected declines from 3.8 to 2.1 cM; the magnitude of the reduction in confidence interval due to increased population size decreases as the distance between the marker and the target gene increases (Burr et al. 1988). In the case of QTL mapping, population size generates Beavis effect, which signifies that the number of QTLs detected for a trait decreases as the population size decreases. In addition, the QTL effect estimates increase as the population size decreases (Sect. 7.13.4). It has been suggested that for most quantitative traits, the mapping population size should be 500 or more to minimize the Beavis effect (Bernardo 2008). Schneider (2005) has recommended a population size of ~100 F_2 individuals for producing a genome-wide overview marker map as a compromise between cost/feasibility and the resolution of linked loci, while a population of at least 200 individuals should be used for mapping of QTLs. But when the objective is positional cloning of genes, populations of several thousand plants should be used. For example, Alpert and Tanksley (1996) analyzed a population of over 3,400 tomato plants to develop a detailed marker map of the genomic region flanking a fruit weight locus. Similarly, Ashikari et al. (2005) generated a F_2 population of 13,000 individuals for fine mapping of the QTL governing grain number in rice, which was ultimately cloned and named as *Gn1*.

5.26 Choice of Mapping Population

The short-term mapping populations, such as F_2, backcross, or the conceptual near-isogenic lines developed following the bulk segregant analysis (BSA) approach (Sect. 6.6.2), can be a good starting point in molecular mapping. However, long-term mapping populations like RILs, DHs, NILs, and CSSLs, or immortalized F_2, MAGIC, or NAM should be developed for precision phenotyping of the traits of importance and for sharing of the populations among different research workers involved in global mapping projects. In fact, the development and phenotypic

characterization of mapping populations should become an integral part of the ongoing breeding programs in important crops. At this point, the roles of geneticists and plant breeders become crucial for reaping the full benefits of molecular plant breeding. Since RILs, DHs, NILs, and CSSLs are homozygous, they are not suitable for studying dominance and interaction effects, except for additive × additive interaction effects. In contrast, immortalized F_2 populations combine the benefits of perpetual mapping populations and the opportunity for studying dominance and all interaction effects estimable from F_2 populations.

Questions

1. Discuss the relevance of mapping populations in mapping of genes and quantitative trait loci.
2. Discuss the usefulness and limitations of mapping populations based on early segregating generations from biparental crosses.
3. Briefly describe the different procedures for generating near-isogenic lines and their usefulness in gene mapping studies.
4. "Recombinant inbred lines offer several advantages including the opportunity to detect markers located close to the target gene." Discuss this statement in the light of available information.
5. "Immortalized F_2 populations are the best available option for detection and mapping of heterosis quantitative trait loci." Analyze this observation critically in the light of relevant information.
6. Briefly describe the construction of chromosome segment substitution lines and discuss their usefulness in linkage mapping and plant breeding.
7. Most mapping populations are derived from biparental crosses. But some mapping populations are constructed from multiparent crosses. Briefly describe the salient features of some of the multiparent crosses and discuss their advantages and limitations.
8. Briefly discuss the populations and procedures used for linkage mapping in polyploid species.
9. Discuss the relevance of the type of mapping population and its size and the difficulties encountered in mapping studies.
10. "Doubled haploid populations are similar to recombinant inbred line populations, but they are not as informative as the latter." Critically analyze this statement in the light of relevant information.
11. Briefly describe the recurrent selection backcross procedure and discuss its usefulness and limitations.

Linkage Mapping of Molecular Markers and Oligogenes

6

6.1 Introduction

According to the second law of Mendel, the law of independent assortment, segregation of two different genes is independent of each other. However, the chromosomal theory of inheritance, proposed in 1902, led to the expectation that more than one gene would be located on a single chromosome and such genes would tend to be inherited together since each chromosome appeared to behave as a unit during mitosis and meiosis. In 1911, Morgan proposed that three sex-linked genes of *Drosophila* were linked together and described the essential features of linkage and crossing over. The tendency of two or more genes or loci being inherited together is known as *linkage*. It is now universally accepted that genes located relatively close to each other in the same chromosome show linkage. Further, new combinations of linked genes are usually recovered in the progeny; this phenomenon is known as *recombination*. Recombination between linked genes is the result of *crossing over*, i.e., a physical exchange of ordinarily strictly homologous segments between homologous chromosomes. Finally, the frequency of recombination between two linked genes is generally proportional to the distance between them. As a result, genes located close to each other in a chromosome show a lower frequency of recombination than those located farther apart. This feature of linkage was exploited by Sturtevant to generate the first ever linkage map of *Drosophila* in 1913. Since then, linkage maps have been constructed in every organism that has been the subject of genetic investigations.

6.2 Genetic Maps

A *genetic map* is a schematic representation of various genetic markers in the specific order, in which they are located in a chromosome along with the distances between them. Genetic maps have been constructed by using three diverse strategies to generate three different types of maps, viz., (1) linkage maps, (2) cytogenetic or cytological maps, and (3) physical maps.

6.2.1 Linkage Maps

A *linkage map* is a schematic representation of the relative locations of various genetic markers present in the chromosomes of an organism as determined from the frequency of recombination between pairs of markers. The recombination frequencies between marker pairs are estimated from suitable mapping populations (Chap. 5) and are converted to map or genetic distances. Based on the genetic distance, the markers are grouped into linkage groups, and their order in the linkage group is depicted as the linkage map. But the recombination frequency shows considerable variation in the different regions of the genome, and heterochromatic regions like centromeres

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_6, © Author(s) 2015

exhibit considerably reduced recombination frequencies. In such cases, cytogenetic maps depicting the physical fine structure of chromosomes can provide complementary information and enhance the usefulness of genetic maps. In addition, recombination frequency itself is a heritable trait and is affected by several factors, including sex, genetic background, and environmental conditions. A special category of linkage maps, called *functional maps*, depicts locations of different genes of the concerned species. The conventional linkage maps also depict the genes governing different phenotypic traits, but they are developed by using the concerned traits as genetic markers. The functional maps, on the other hand, are developed by using molecular markers located within genes or the gene sequences themselves are used as markers. The genes mapped in a functional map include those affecting traits of interest, genes with known function, and those comprising quantitative trait loci (QTLs). A large number of functional maps have been prepared for wheat; these maps depict genes involved in specific metabolic pathways or expressed sequence tags (ESTs) derived from mRNAs from specific organs.

6.2.2 Cytogenetic Maps

A *cytogenetic map* depicts the locations of various genes in the chromosomes of a species relative to specific microscopically observable landmarks in the chromosomes. In most cases, each chromosome has a characteristic banding pattern, which may be either naturally present, e.g., in polytene chromosomes of *Drosophila*, or is most commonly generated by specific staining protocols like Giemsa C. Even morphological landmarks like centromeres, nucleolus-organizing regions, knobs, etc., and heritable heterochromatic regions of identifiable shape have been used for mapping. Cytogenetic mapping is generally used in eukaryotes, which have relatively large microscopically observable chromosomes. Further, it is far more refined in species having polytene chromosomes. Cytogenetic mapping may use one or more of the following approaches: (1) fluorescence in situ hybridization (FISH), including multicolor FISH (McFISH), using gene sequences as probes; (2) human–mouse somatic cell hybridization, followed by genetic and cytogenetic analyses of the hybrid clones; and (3) analysis of small changes in polytene chromosomes and the genetic alterations associated with them. In addition, chromosome deletion, translocation, trisomic, monosomic, and nullisomic lines serve as valuable tools for cytogenetic mapping. Further, defined translocation breakpoints enable localization of probes to specific regions of chromosome arms. Cytogenetic maps permit the linkage groups to be associated with specific chromosomes. They also allow decision about the direction of the various linkage groups in relation to the morphology of the respective chromosomes. The information from cytogenetic maps developed on the basis of FISH facilitates the construction of physical maps and allows the BAC clones and other BAC sequences to be placed along the chromosomes.

6.2.3 Physical Maps

In a *physical map*, the genes/molecular markers are depicted in the same order as they occur in the chromosomes, but the distances between adjacent genes/markers are depicted in terms of base pairs. The distance in terms of base pairs is known as *physical distance* and is determined by either hybridization of appropriate probes or sequence alignment to a good quality reference genome. Physical mapping usually involves (1) cloning of many pieces of chromosomal DNA, (2) characterization of these fragments for size, and (3) determination of their relative locations along the chromosomes using a suitable technique like McFISH (Hass-Jacobus and Jackson 2005). The molecular markers used for linkage mapping can also be used for physical mapping (Sect. 6.17). The ultimate physical map of any genome is a good quality genome sequence that is fully annotated to depict all the functional elements of the genome. Reasonably good quality genome sequences are available for

several species, but their complete and reliable annotation remains to be accomplished.

6.3 Estimation of Recombination Rates

The two fundamentals of linkage mapping are the phenomena of linkage and crossing over. As a rule, frequencies of the parental genotypes (allelic combinations) of linked genes are much higher than their expected frequencies in the progeny. But new allelic combinations of linked genes are produced by crossing over; this is called *recombination*. The individuals having new combinations of linked genes are termed as *recombinants*. As a rule, the frequencies of recombinant genotypes are drastically lower than their expected frequencies and the frequencies of parental genotypes. In general, each event of crossing over between two linked genes produces two parental and two recombinant gametes. Further, the likelihood of crossing over taking place between two given points of a chromosome is presumed to depend on the physical distance between them. It is impractical to score the frequency of crossing over between two genes, while the recombination frequency can be readily estimated. Therefore, recombination frequency is generally taken as an approximate indicator of the distance between genes/markers and provides the basis for linkage map construction.

Let us suppose that two genes, viz., *a* (alleles *A* and *a*) and *b* (alleles *B* and *b*), are linked and two lines with the genotypes *AA BB* and *aa bb* are crossed to produce the F_1 *Aa Bb*. This F_1 will produce four types of gametes (*AB*, *Ab*, *aB*, and *ab*) and testcross progeny (*Aa Bb*, *Aa bb*, *aa Bb*, and *aa bb*). The gametes *AB* and *ab* (and the testcross progeny *Aa Bb* and *aa bb*) represent the parental allelic combinations, while the gametes *Ab* and *aB* (and the testcross progeny *Aa bb* and *aa Bb*) are the recombinant types. It should be noted that the recombinant types will be produced by one crossing over event between the genes *a* and *b* (Fig. 6.1). The frequency of recombination between the genes *a* and *b* can be estimated as follows:

$$r = \frac{\text{Number of recombinant progeny}}{\text{Total number of testcross progeny}} \quad (6.1)$$

$$= \frac{\text{Total number of } Aa\,bb \text{ and } aa\,Bb \text{ progeny}}{\text{Total number of testcross progeny}} \quad (6.2)$$

This estimation of recombination frequency (r) on the basis of phenotypic data from a testcross population is possible because this population permits visualization of the gametes produced by the F_1 hybrid. Some other mapping populations, such as backcross with the recessive parent and that with the dominant parent (in case of codominant markers/genes only) and doubled haploid (DH) populations, also allow visualization of the F_1 gametes. Therefore, r can be estimated from these populations in the same manner as described above. However, in F_2 and recombinant inbred line (RIL) and other similar populations, r cannot be directly estimated. In such populations, the maximum likelihood method has to be used to obtain the most probable estimate of r. But in the case of RILs, a simpler approach for estimation of r is to first calculate R, which is the proportion of inbred lines, in which the genes *a* and *b* have recombined.

$$R = \frac{\text{Number of inbred lines recombinant for the genes } a \text{ and } b}{\text{Total number of inbred lines in the population}} \quad (6.3)$$

Then the value of r is estimated from R following Haldane and Waddington (1931), who showed that $R = 2r/(1 + 2r)$, which leads to $r = R/[2(1 - R)]$. Therefore, when the value of r is very small, the value of R is approximately $2r$.

6.4 Genetic Distance

Since the frequency of recombination depends on distance between the two given genes, it could be used as a measure of the distance between them and as the basis for linkage mapping. However, recombination frequency cannot be directly used

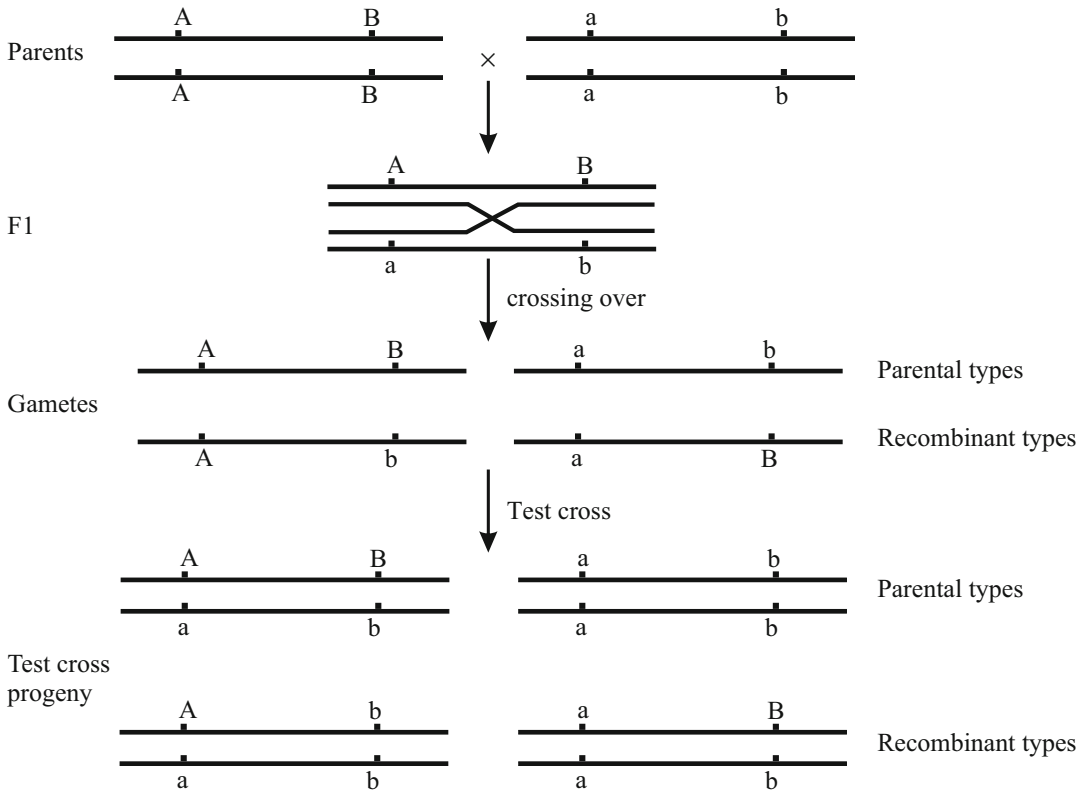


Fig. 6.1 The effect of a single crossing over between two linked genes *a* and *b*. Out of the four gametes produced by this meiotic division, two each are of the parental and the recombinant types. The four gametes can be easily monitored by examining the phenotypic expressions of the concerned traits in the testcross progeny. Since the testcross parent (*aa bb*) contributes only recessive alleles to the progeny, the phenotypes of these progenies are determined solely by the F_1 gametes. It may be pointed out that one crossing over event involves only two nonsister chromatids of a bivalent. However, the other two chromatids of the bivalent may also be involved in another crossing over event, but these possibilities are not considered here primarily to keep the discussion simple

as a measure of the genetic distance for the following reason. When two genes are located close to each other, only a single crossing over may be expected to take place between them, and each crossing over would lead to recombination (Fig. 6.1). But as the distance between these genes increases, the likelihood of two or more simultaneous crossing overs between them would also increase. The occurrence of an even number of crossing overs between two genes will yield only parental gene combinations (Fig. 6.2). As a result, the recombination frequency will no longer correspond to that of crossing over, and it will become progressively smaller than that of the latter as the distance between the two genes increases. In any case, no matter how far apart two genes are located in a chromosome, the frequency of recombination between them cannot exceed 50 %, which is the frequency of recombinants obtained with independent segregation of genes. Thus, in general, the correspondence between recombination frequency and genetic distance progressively declines with the increasing distance between the linked genes. In view of this, recombination frequencies have to be corrected for the occurrence of multiple crossovers to obtain the estimates of genetic distance from them. There are several methods, called *mapping functions*, for converting recombination frequency into genetic distance, but the two most commonly used methods are those proposed by Haldane (1919) and Kosambi (1944).

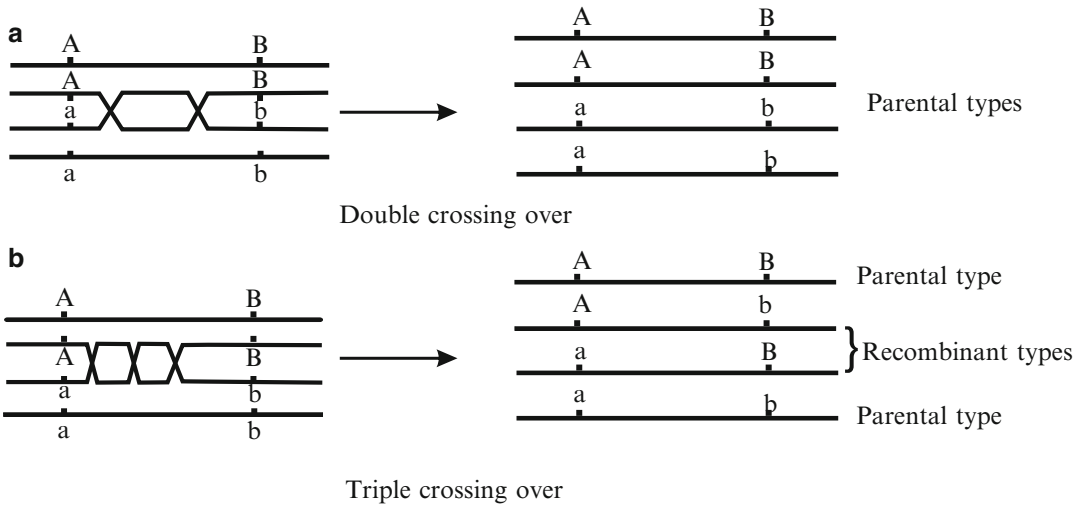


Fig. 6.2 Effect of multiple crossing over events between two genes. (**a**) When two crossing overs occur, all the four gametes are of parental type; the same will be the case for all even numbers of crossing overs. (**b**) But when three crossing overs occur, the consequences are the same as those when a single crossing over takes place (Fig. 6.1); again, all odd numbers of crossing overs will produce the same result. For simplicity, the multiple crossing overs are considered to involve the same two nonsister chromatids of the bivalent. However, they may also involve three or all the four chromatids of the bivalent, leading to different consequences

6.4.1 The Haldane Distance

The *Haldane mapping function* corrects recombination frequencies for multiple crossing over events assuming that occurrence of a crossing over does not affect the likelihood of another crossing over in the neighboring regions of the chromosome, i.e., there is no interference. Let us suppose that genes *A*, *B*, and *C* are located in the given order in a chromosome (Fig. 6.3). If there were no multiple crossing overs, a recombination between the genes *A* and *C* will be observed whenever there is a recombination between the genes *A* and *B* or the genes *B* and *C*. Therefore, the frequency of recombination between *A* and *C* (denoted by r_{AC}) will equal the total of recombination frequencies between *A* and *B* (r_{AB}) and between *B* and *C* (r_{BC}). Thus,

$$r_{AC} = r_{AB} + r_{BC} \tag{6.4}$$

But multiple crossing overs do take place, and they tend to reduce the recombination rates between the genes. In this case, two crossing overs could occur, one between genes *A* and *B* and the other between genes *B* and *C*; the frequency of this event will equal the product of the frequencies of crossing overs between the two pairs of genes ($r_{AB} \cdot r_{BC}$). Since the double crossing over event can occur in two different ways, i.e., crossing over between *A* and *B*, followed by that between *B* and *C*, and *vice-versa*, the frequency of double crossing over would equal $2r_{AB} \cdot r_{BC}$. Therefore, the observed frequency of recombination between the genes *A* and *C* will be lower (by $2r_{AB} \cdot r_{BC}$) than otherwise expected. Thus,

Fig. 6.3 A map of three genes *a*, *b*, and *c* located in the same chromosome in the order *a*–*b*–*c*

$$r_{AC} = r_{AB} + r_{BC} - 2r_{AB} \cdot r_{BC} \tag{6.5}$$

The above equation can be rewritten, simplified, transformed to make the relationships linear, generalized for any number of loci, and ultimately simplified to yield the Haldane genetic distance (*m*) in Morgans as a function of *r* as follows:

$$m = -\left(\frac{1}{2}\right)\ln(1-2r) \qquad (6.6)$$

Since map distances are generally in centimorgans (cM), and one Morgan comprises 100 cM, the above equation may be written as follows:

$$m = -50\ \ln(1-2r) \qquad (6.7)$$

6.4.2 The Kosambi Distance

The assumption of lack of interference in the Haldane function is a strategy of convenience rather than a reflection of reality. In fact, occurrence of crossing over at a chromosomal site interferes with the occurrence of another crossing over in its surrounding regions; this phenomenon is known as *interference*. As a result, the observed frequency of two simultaneous crossing overs in the neighboring regions of a chromosome is lower than expected, and the ratio of their observed to the expected frequency is termed as *coincidence* (denoted by c). Therefore, the value of $2r_{AB} \cdot r_{BC}$ (the expected frequency of double crossing over; Sect. 6.4.1) should be reduced to the fraction c. Thus, Eq. 6.5 becomes

$$r_{AC} = r_{AB} + r_{BC} - 2c\ \ r_{AB} \cdot r_{BC} \qquad (6.8)$$

In the absence of interference, i.e., with $c = 1$, this equation ultimately simplifies to $m = -(½)\ \ln\ (1 - 2r)$, which is the same as Haldane distance. But when interference is assumed, we have to assign c a value. In general, the value of c is proportional to that of r. However, this relationship is influenced by several factors, including the species, the specific chromosome of a genome, and even the particular region of a chromosome. Kosambi proposed to assign c the value of $2r$; this would yield a value of 0 for c when r takes the value of 0 and the value of 1 when r equals 0.5. With this and certain other assumptions, the equation 6.8 is simplified to give the Kosambi genetic distance (m_K) in Morgans as follows:

$$m_K = \left(\frac{1}{4}\right)\ln[(1+2r)/(1-2r)] \qquad (6.9)$$

Since the genetic distance is ordinarily expressed in cM, this formula takes the following form:

$$m_K = 25\ln[(1+2r)/(1-2r)] \qquad (6.10)$$

In general, the value of Kosambi distance for a given value of r is lower than that of Haldane distance, and this difference increases with the value of r. This is because the genetic distance estimates depend on the frequency of double crossing over, which will always be lower in the Kosambi function than in the Haldane function for obvious reason. Therefore, it is not entirely correct to say that one centimorgan is that distance between two genes, which would allow one percent recombination between them. *The correct statement would be that a distance of one centimorgan between two genes is expected to lead to one percent crossing over between them.* The Haldane and Kosambi functions of genetic distance are the most commonly used in plant genetics, but some other functions have also been proposed, which differ primarily in the relationship between the values of c and r. For example, one such function assumes the relationship between c and r to be $c = (2r)^2$. With this relationship, the value of c is 0 when $r = 0$ and $c = 1$ when $r = 0.5$. A given genetic distance function may be valid in some genomic regions, but not others, and may better fit data from some species than from other species (de Vienne 2003).

6.4.3 Variation in Genetic Distance

The estimates of genetic distance are affected by all such factors that affect recombination rates between genes. The following three factors intrinsic to biological materials are known to markedly affect recombination rates. Sex is a potent factor affecting recombination rates: in some species like tomato and barley, recombination rate is higher in the female gametes, while in some other species like maize, the opposite is the case. For example, results from reciprocal

backcrosses in tomato showed that the genetic map was about 20 % longer when the F_1 was used as female than when it was used as male. Different genotypes of the same species usually show different rates of recombination, and this difference may sometimes be more than 20 %. For this reason, genetic distances between given pairs of markers estimated from different mapping populations of the same species may not necessarily be identical. Finally, genetic distances estimated from interspecific crosses are generally considerably (in some cases, up to 65 %) smaller than those estimated from intraspecific crosses. The reduced genetic distances result from reduced recombination frequencies in interspecific crosses than in intraspecific crosses. In addition, smaller chromosomes of a species tend to show higher recombination rates than its longer chromosomes. Similarly, heterochromatic regions containing highly repetitive DNA sequences show much lower recombination rates than do euchromatic regions within a single chromosome, and the gene-rich regions corresponding to zones of transcription show higher recombination rates.

6.4.4 Relationship Between Genetic and Physical Distances

The physical distance is generally expressed as kilobase pairs (kb), Mb (megabase pairs), or Gb (gigabase pairs). In general, the genetic distance is proportional to the physical distance, but the exact relationship is quite variable and is affected by several factors (Sect. 6.4.3). A comparative study of the total genome sizes (total physical distance) and the total genetic lengths of different plant species reveals the following remarkable aspect of the relationship of genetic distance with physical distance (Table 6.1). It would be seen that the total physical distance increases from merely 0.15 Gb in *Arabidopsis thaliana* to about 16 Gb in hexaploid wheat, representing an increase of over 100-fold. In contrast, the total genetic distance increases from 630 cM in *A. thaliana* to only 3,500 cM in wheat, which is merely a 5.6-fold increase. *Thus, as the total physical distance increases, the total genetic distance also increases, but at a much lower rate than the physical distance*. This relationship is reflected in the total physical length represented by one centimorgan genetic distance: this length is merely 140 kb in *Arabidopsis*, through 750 kb in tomato to 4,600 kb in wheat.

The values presented in Table 6.1 for total genetic lengths and the lengths of DNA per cM are merely illustrative rather than definitive. This qualification is required in view of the difficulty in selecting the appropriate value for the total genetic distance for a given species. For example, more than 1,000 different genetic maps have been constructed for maize using various types of mapping populations and different parental materials. The total genetic lengths of these maps range from just 1,500 to 4,922 cM, which

Table 6.1 The relationship between genome size, physical distance, and genetic distance in certain plant species, for which saturated genetic maps have been developed

Plant species	Haploid (*n*) chromosome number	Genome size (Gb of DNA)	Total genetic length (cM)[a]	Length (kb) of DNA per cM
Arabidopsis	5	0.15	630	140
Bean	11	0.65	830	780
Maize	10	2.5	1,860[a]	1,400
Rapeseed	19	1.2	1,016	1,200
Rice	12	0.43	1,575	280
Soybean	20	1.2	2,700	440
Tomato	12	0.95	1,267	750
Wheat	21	16	3,500	4,600

[a]There are over 1,000 different linkage maps for maize and the estimates for total genetic length range from just 1,500 to 4,922 cM

represents about 3.3-fold variation. Consequently, the length of DNA/cM of genetic distance in maize would vary from merely 500 to 1,667 kb. Further, these values are the average for the entire genome assuming even distribution of the recombination frequencies over the entire genome. However, in reality, both recombination rates and the length of DNA/cM genetic distance show considerable local variation in the genome. For example, in the maize genomic region having the *bronze* locus, 1 cM corresponds to merely 14 kb as compared to 1,400 kb for the entire genome. The length of DNA/cM genetic distance is also affected by genetic as well as environmental factors.

The relationship of genetic distance to the physical distance is of considerable importance in map-based or positional cloning of genes. For example, in a species with small genome size like *A. thaliana*, a marker located at 1 cM from the gene of interest would be considered as promising for positional cloning, but the same genetic distance would be discouraging in species like maize and wheat.

6.5 General Procedure for Linkage Mapping of Molecular Markers and Oligogenes

1. The first step in mapping of markers/oligogenes is to select two genetically divergent parents expected to differ for a large number of markers and/or the trait of interest. The selected parents are crossed, and a suitable mapping population (Chap. 5) is developed.
2. The parents are tested with a large number of markers to identify polymorphic markers. The two parents would differ for the alleles of a polymorphic marker.
3. In case a molecular marker map is to be constructed, all the individuals of the mapping population are screened with the polymorphic markers; this is called *genotyping*.
4. The marker genotype data are analyzed using a suitable software package (Sect. 6.14) to estimate recombination frequencies and genetic distances between marker pairs, group the markers into linkage groups, select the most likely marker order, and finally prepare a marker linkage map.
5. In order to map an oligogene governing the trait of interest, all the individuals of the mapping population are evaluated for phenotypic expression of the trait; this is known as *phenotyping*. Since the target trait would be a qualitative trait governed by one or few oligogenes, the individuals of the population would be classifiable into a limited number of distinct phenotypic classes. For example, if resistance to a disease were governed by a single oligogene, the individual plants of the mapping population would be classified either as "resistant" or "susceptible" on the basis of their reaction to the disease.
6. The trait phenotype and marker genotype data are analyzed using a suitable computer program (Sect. 6.14) to identify the markers linked to the oligogene governing the target trait, estimate the frequency of recombination between the gene and the markers, and ultimately prepare a linkage map of the oligogene and the markers linked to it.
7. In fact, one may not genotype the entire mapping population for all the polymorphic markers to identify the markers linked to the gene for the target trait. One may use a strategy like bulked segregant analysis (Sect. 6.7.2) to identify a small set of polymorphic markers most likely to be linked to the target trait/gene.

6.6 Mapping of the Loci Present in a Chromosome

The mapping software estimate the likelihood of linkage (Sect. 6.8) as well as the genetic distance (Sect. 6.4) between all possible pairs of loci. They use the genetic distance estimates to group the loci into distinct linkage groups as well as to determine the most likely order of the loci in each linkage group. The problem of finding the "true order" of loci in a linkage

group or chromosome is known as "traveling salesman problem". The "true order" of loci in a chromosome represents the order in which the loci actually occur in the concerned chromosome. The various algorithms developed to solve the above problem are basically of two types, viz., the exhaustive search and the approximation algorithms. An exhaustive search is impractical as it requires heavy computation that is extremely time-consuming. Therefore, approximation algorithms offer the only practicable approaches; these algorithms include simulated annealing, Lander–Green, and stepwise algorithms. Many of these algorithms have been implemented in mapping programs like MapMaker/Exp, JoinMap, GMendel, etc. In an alternate approach, the linkage map construction for a chromosome begins with a small number of loci. Then more loci are added to the linkage map one at a time; this approach is called sequential map construction. Several different approaches have been proposed for sequential mapping. Of these, the unidirectional growth method has the same speed and other advantages as the other sequential mapping approaches, but it has a much higher accuracy. In this approach, the locus at one terminus of the map of a chromosome is the first to be determined. Then the other loci of the linkage group are added to the map one at a time in the direction of the other end of the chromosome. Computer simulation results showed this method to be more efficient than several other methods including simulated annealing, evolutionary strategy, and neighbor mapping methods. Thus, the unidirectional growth approach of sequential mapping is suitable for map construction with a large number of loci (Tan and Fu 2006). But the insertion algorithm, a modification of the branch and bound algorithm, begins with any pair of linked loci and then adds to this map one of the remaining loci in the appropriate position in the map. This step is repeated many times, and the loci are selected randomly for insertion into the growing linkage group. It has been found that this algorithm is more efficient than the unidirectional growth algorithm, and it was considered to be a robust and efficient algorithm for large-scale linkage mapping (Wu et al. 2011a).

6.7 Strategies for Mapping of Oligogenes

A *qualitative trait* is governed by one or few genes with large effects, and its phenotypic expression is relatively little affected by the environment. As a result, the individuals can be readily classified into two or more distinct classes on the basis of such traits, and the inheritance of such traits can be followed with confidence. Therefore, oligogenes can be treated at par with markers for mapping purposes, and the qualitative traits were the first marker type used for construction of the conventional linkage maps. Therefore, *the purpose of mapping of an oligogene with molecular markers is to identify marker(s) closely linked to the oligogene to facilitate indirect selection for the concerned trait*. As noted in Sect. 6.5, one approach for this is to screen the whole mapping population for a large number of markers. However, this would involve considerable genotyping work, and the chance of finding a marker closely linked to the gene of interest would be small. Therefore, a suitable strategy like near-isogenic lines, bulked segregant analysis, bulked segregant RNA-Seq, etc., should be used to reduce the genotyping work and to facilitate identification of those markers that are most likely to be closely linked to the gene(s) governing the target trait.

6.7.1 Use of Near-Isogenic Lines

In theory, *isogenic lines* have identical genotype, except for the alleles of a single gene. But an isogenic line (RP′) of a recurrent parent (RP) is generally produced by a backcross program (Sect. 5.10). The RP and the RP′ differ not only for the gene transferred from the donor parent (DP) but also for a variable number of loci linked to this gene and often for loci located in other chromosomes. Therefore, the RP and RP′ are

called *near-isogenic lines (NILs)* in the place of isogenic lines. When the RP, RP′, and the concerned DP are screened with a number of markers, many of the markers would be polymorphic, i.e., will differ among the RP, RP′, and DP (Kaeppler et al. 1993). Most of the polymorphic markers will differentiate the DP from the RP and the RP′; these markers will be located in the genomic regions of RP. *But some of the polymorphic markers will differentiate the RP from both the DP and the RP′; these markers will be located in those DP genomic regions that have been retained in the RP′*. A comparison between RP and RP′ would permit the identification of markers expected to be located in the transferred DP genome. However, the inclusion of DP as a control is highly desirable as it would eliminate the risk of differences between the RP and the RP′ being the product of technical and/or genetic errors. The technical errors would include generation of bands due to artifacts, while genetic errors would result from contamination due to mechanical mixture and/or cross-pollination. Two important questions, however, remain to be answered: (1) which of these markers are linked with the transferred gene, and (2) what is the genetic distance between the gene and the linked markers?

One way of resolving the first question is to evaluate several pairs of NILs developed by transferring the same gene from the same DP into several different RPs. Some of the markers differentiating the RP from the RP′ and the DP will differ among the NIL pairs; these markers will not be linked to the transferred gene and will be present in the DP genomic regions transferred randomly into the RP genome. *But some of the polymorphic markers will be common to all the pairs of NILs tested; these markers are likely to be linked to the target gene*. Another approach that provides answers to both the above questions consists of crossing an RP′ to the concerned RP to generate a mapping population like F_2. This population is screened for the target trait and the polymorphic markers, and the trait phenotype and marker genotype data are analyzed using suitable computer software (Sect. 6.14) to identify the markers linked to the target trait/gene and to obtain the estimates of the genetic distances as well. It has been suggested that markers like AFLP, ISSR, etc., that are used for fingerprinting should be used for this analysis; once a closely linked marker is identified, it may be converted into a more user-friendly SCAR marker. Several oligogenes, especially those for disease resistance, have been mapped using NILs.

6.7.2 Bulked Segregant Analysis

Bulked segregant analysis (BSA) is based on the principle of NILs. It is widely used for identifying markers putatively linked to the gene of interest with the minimum effort and expenditure. In BSA, two parents, say, a disease resistant and a susceptible line, are crossed, and a suitable mapping population, like F_2, backcross, DH, RIL, etc., population, is generated. In the case of an F_2 population, individual plants are phenotyped for their reaction to the concerned disease, and plants exhibiting extreme resistance/susceptibility to the disease are identified. Usually, equal amounts of DNA isolated from the ten most resistant and the ten most susceptible plants in each group are pooled to constitute two bulks, viz., the resistant and the susceptible bulks. DNAs from the two parents and the two bulk DNAs are screened with a large number of markers. A marker showing polymorphism between parents as well as the resistant and susceptible DNA bulks is likely to be linked to the target gene/trait, i.e., resistance to the concerned disease in this case. It may be pointed out that the polymorphism may not always be of "presence"/"absence" type, but it may be observed as a difference in band intensity between the two bulks due to the presence of one or few recombinant individuals in the bulk(s). These polymorphic markers are genotyped in all the individuals of the mapping population, and the data on trait phenotype and marker genotype are analyzed for linkage mapping of the target trait (Michelmore et al. 1991).

Conceptually, the genetic constitution of the two bulks with respect to any marker will depend on the location of this marker in the genome with respect to the gene governing the target trait that

was used for creating the two bulks. It is assumed that the target trait has 100 % heritability, i.e., there is complete correspondence between the genotype at the concerned locus and the trait phenotype. Since the bulks were constituted on the basis of the target trait phenotype, they are expected to differ for the gene governing this trait and the markers linked to this gene. In contrast, the markers not linked to this gene will segregate independently of the gene, and both the alleles of all such markers would be expected to be present in both the bulks in comparable frequencies. Therefore, the two bulks will be similar in composition for all the markers segregating independently with the gene of interest, while they would differ for those markers that are linked to this gene.

The minimum size of bulks can be determined by estimating the maximum probability of detecting linkage between an unlinked marker and the target gene, i.e., being polymorphic between the two bulks. This probability should be as low as feasible, and its magnitude depends on the type of mapping population and the dominance relationship at the marker locus. For example, the probability for an unlinked dominant marker being polymorphic between the two bulks in an F_2 population will be given by the following formula:

$$P_{fl} = 2(1/4)^n[1 - (1/4)^n] \qquad (6.11)$$

where P_{fl} is the maximum probability of detecting linkage between the gene and an unlinked dominant marker and n is the number of plants constituting the bulk. When the value of n is 10, the above equation will simplify to give the approximate estimate of the probability as 2^{-19} or 2×10^{-6}. Therefore, the probability that a marker that is polymorphic between the two bulks is linked with the target gene would be $1 - (2 \times 10^{-6})$. For this reason, usually 10 plants are used for constituting each of the two bulks. In the first study based on BSA, Michelmore et al. (1991) detected RAPD markers linked to the gene *Dm5/8* that confers resistance to downy mildew in lettuce. They created resistant and susceptible bulks of 17 F_2 plants each out of a population of 66 F_2 plants and screened them with 100 RAPD primers. They were ultimately successful in identifying three RAPD markers that were linked to the resistance gene. The BSA method can be used iteratively, i.e., new bulks can be constructed based on each new marker linked more closely to the target gene, in an effort to identify markers tightly linked to the gene (Sect. 6.9). Usually, the bulks are created from the concerned mapping population, but the BSA markers could as well be anchored in a different mapping population (Sect. 6.7.3). BSA has been extensively used for mapping of oligogenes governing qualitative traits of various crop species, and it has been extended to the mapping of quantitative trait loci (QTLs) as well.

The principle of BSA is similar to that of NILs in the following respect. (1) In both the cases, the two samples (the two bulks in BSA and the two lines of each NIL pair) are expected to be homozygous for different alleles of the target gene. Therefore, (2) the two bulks, and the two members of each NIL pair, will differ for such markers that are closely linked to the target gene. (3) Finally, the two bulks are expected to have comparable frequencies of both parental alleles of all such loci that are unlinked to the target gene. Similarly, both the members of an NIL pair are expected to be homozygous for the same parental allele of all such loci. However, the BSA and NIL strategies differ in the following respect: the dominance relationships at the target and marker loci, and the phase of linkage between them will not materially affect the results from analysis of NILs, while these factors are highly relevant in the case of BSA. This is because each member of an NIL pair will be homozygous for the target gene as well as the marker loci; as a result, the alleles of the target gene and marker loci present in the members of an NIL pair would be readily distinguished. The same will be the case when the bulks for BSA are created from DH or RIL populations, where the only confusion will arise due to those individuals that are recombinant for the target gene and the linked marker locus. But in the case of F_2 and backcross (actually, testcross) populations, the two bulks will be easily distinguishable on the basis of the dominant markers alleles linked in coupling phase with the target gene. However, the two bulks cannot be differentiated for the dominant marker alleles linked in the repulsion

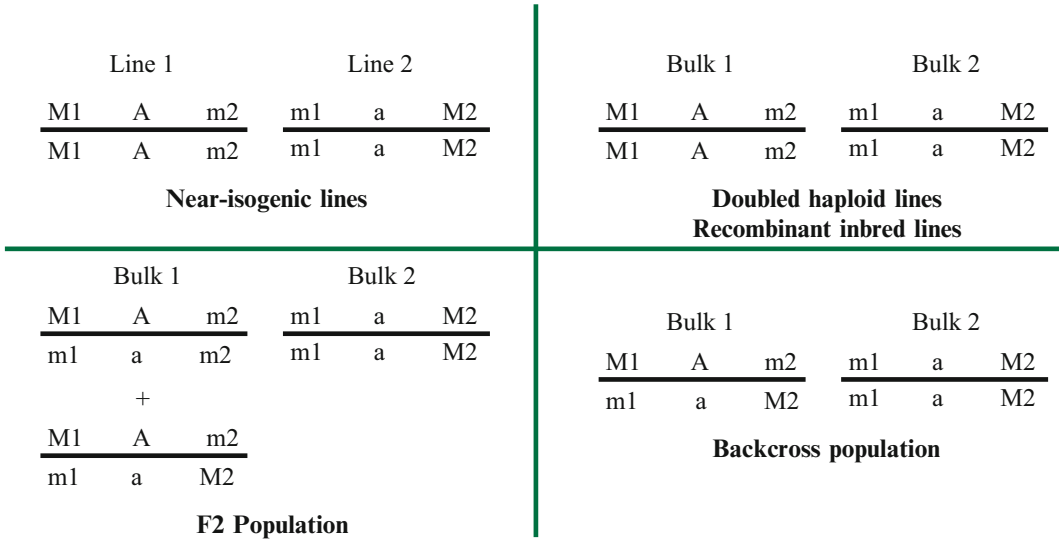


Fig. 6.4 Effect of linkage phase on the ability of dominant markers to discriminate between members of NIL pairs and the two bulks derived from doubled haploid, RIL, F_2, and backcross populations. The backcross population is produced by crossing the F_1 to the parent having the recessive phenotype for the target trait. For simplicity of presentation, it is assumed that the markers and the target gene are completely linked and there is no recombination between them. It would be seen that the marker *M1*, which is linked in coupling phase with the target gene *A*, is able to discriminate between the two F_2 and backcross bulks. In contrast, the marker *M2* is linked with the target trait in repulsion phase and is unable to differentiate between the two F_2 and backcross bulks. The two bulks (Bulk 1 and Bulk2) are constituted on the basis of phenotypes produced by the alleles *A* and *a*, respectively. It may be emphasized that the linkage phase has no effect in the cases of NILs, and DH and RIL populations

phase with the target gene (Fig. 6.4). Therefore, assuming that there is equal frequency of coupling and repulsion phase markers, about 50 % of the markers of a dominant marker system will fail to discriminate between the two bulks from these populations. Therefore, twice as many markers will have to be scored for BSA in F_2 and backcross populations than that for BSA in DH and RIL populations. One approach to overcome the above difficulty is to screen several (30–50) individual plants with the recessive phenotype of the trait, but even this approach is not very reliable (Sect. 6.7.3). This approach may, however, be useful if there were near-complete linkage and a fairly large number of plants were screened. The above difficulties will not be faced in the case of codominant markers, particularly when the target trait also shows partial dominance.

BSA offers certain advantages over NILs. (1) BSA does not require several generations of backcrossing, which is necessary for the development of NILs. (2) A proportion of marker loci polymorphic in the two members of an NIL pair are likely to map in genomic regions other than that harboring the gene of interest. But in BSA, the genomic regions unlinked to the target gene are not likely to differ between the two bulks when each bulk comprises ten or more individuals. Further, (3) all polymorphic loci detected using BSA will be segregating in the mapping population and can be mapped by analyzing the individual plants of the population. In contrast, the loci polymorphic in a pair of NILs can be mapped only after developing a mapping population from them. Finally, (4) BSA can be used for finding markers to fill the gaps remaining in genetic maps (Sect. 6.9). BSA has been extended to linkage mapping of QTLs, gene mapping using RNA-Seq, and for pooled mapping.

6.7.3 Mapping of Recessive Morphological Mutants by a Two-Step Procedure

In the *two-step procedure for mapping of recessive mutations*, the first step involves the construction of a linkage map with sufficient

number of markers using a biparental mapping population. In the second step, the mutant strain is crossed to the two parents of the mapping population to generate two F_2 populations (Castiglioni et al. 1998). From each F_2 population, 30–50 plants with the mutant phenotype are selected and screened with the markers already placed onto the linkage map. The two parents of the mapping population and about 5 wild-type F_2 plants may also be included as controls. A dominant marker present in the wild-type strains and absent from the mutant strain is expected to be present in 75 % of the mutant F_2 plants in case of independent segregation of the marker and the mutant gene, but it will be absent from the mutant F_2 plants in case of tight linkage. Thus, a marker present in the mutant F_2 plants in a low frequency would be linked to the mutant allele, and the plants having the marker would be recombinants. The recombination frequency between this marker and the mutant gene is then estimated, and based on this information, the gene is placed onto the already constructed linkage map. Dominant markers present in the mutant strain and absent from the wild-type strain are not suitable for this analysis because of the narrow window available for them. Such a marker is expected to be present in 100 % of the mutant F_2 plants in case of tight linkage between the marker and the gene, while 75 % of the plants will show the marker in case of independent segregation. Castiglioni et al. (1998) constructed a genetic map comprising 511 AFLP markers using 113 DH lines from the cross Proctor × Nudinka. The recessive mutant *branched-5*, isolated from a germplasm collection, was crossed with both Proctor and Nudinka. They selected 45 mutant and 5 wild-type plants from the F_2 population of the mutant × Nudinka cross and analyzed them for the AFLP markers. Markers linked with the mutant gene were identified, genetic distances between the gene and the markers were estimated, and the gene was placed onto the already constructed linkage map. Analysis of the mutant plants isolated from the F_2 generation of the mutant × Proctor cross supported the findings from the above analysis.

6.7.4 Bulked Segregant RNA-Seq

Bulked segregant RNA-Seq (BSR-Seq) is a modification of BSA that uses RNA sequence data from the two phenotypic extreme bulks to identify markers tightly linked to the gene responsible for the target trait (Liu et al. 2012). BSR-Seq was used to map the *gl3* (*glossy 3*) gene of maize, which affects epicuticular wax deposition on juvenile leaves to generate the glossy phenotype. RNA was isolated from leaves of the normal and glossy bulks from the F_2 population and sequenced by RNA-Seq technology. This technology also provides information about the approximate number of copies of each RNA sequence present in the sample assuming that the numbers of reads of various sequences reflect their relative concentrations in the sample. The RNA sequence data was used to discover a large number of polymorphic SNP markers. This data was analyzed by an empirical Bayesian-based BSA approach to identify the SNP markers tightly linked to the *glossy 3* gene, map it within a 2 Mb interval, and ultimately clone this gene. Theoretically, only a single allele of the marker showing complete linkage with the *glossy 3* gene should be present in the glossy bulk since this phenotype is recessive and, consequently, expected to be homozygous, while the normal bulk should have both the alleles of the marker. However, in practice, only single alleles of many unlinked SNP markers are also detected in the mutant bulk due to allele-specific expression and sampling error. The Bayesian-based BSA approach was developed to filter out such noise and identify SNP markers completely linked to the *gl3* gene.

Ideally, RNA used for sequencing should be extracted from a tissue in which the target gene expression takes place, but this is not essential. BSR-Seq combines polymorphic marker discovery with gene mapping. Therefore, it can be used even in such populations, for which marker polymorphism information is not available. The analysis of the RNA sequence data provides information on the effect of the mutant allele of target gene on global gene expression and

simplifies gene cloning efforts. BSR-Seq is efficient and cost-effective particularly in case of species with large genomes. BSR-Seq may be modified to enable mapping of dominant mutant genes, major effect QTLs, and possibly major genes influenced by modifying genes. Although replication is not essential, findings from replicated studies would be more reliable. The mapping interval size is affected by the number of individuals included in the two bulks, the depth of sequencing, and the extent of polymorphism in the mapping population. In general, the larger are the above variables, the more precise would be the mapping results. In addition, a mutant affecting allele-specific expression of genes may lead to identification of false-positive SNPs. BSR-Seq depends on a reference genome sequence. Therefore, it is affected by the quality of this sequence and the extent of structural and copy number variations between the reference genome sequence and that of the strain being analyzed. The mapping software MMAPPR (Sect. 6.14.11) has been developed for linkage mapping using pooled RNA-Seq data.

6.7.5 The MutMap Technique

The *MutMap scheme* was developed by Abe et al. (2012) for a quick, reliable, and cost-effective mapping of causal SNPs in more than 10 EMS-induced mutant lines of rice (Fig. 6.5). Mutations are induced in a homozygous line by a chemical or physical mutagen treatment. A mutant line is then selected and crossed with the parental line. In the F_2 generation, plants with the mutant phenotype are selected, and equal amounts of DNA from each selected plant are bulked. This DNA bulk is subjected to whole-genome sequencing, using a next-generation sequencing (NGS) platform with sufficient depth. The depth should be such that the set of reads for a genomic region may be expected to include sequences from almost all the mutant individuals included in the bulk. The short sequence reads are aligned with the parental reference genome and the genomic positions with SNPs are identified. These SNPs would have been induced by the mutagen treatment. Each mutant line may be expected to differ from the parent at several (up to around 2,000) SNP loci. For each SNP locus, the numbers of short reads having the parental and the mutant SNP alleles are scored, and a SNP index is calculated. *SNP index* is the ratio of the number of short reads with the mutant allele at a SNP locus to the total number of short reads covering this SNP locus. In case of a recessive mutation, all the F_2 individuals with the mutant phenotype will be homozygous for the mutant allele of the

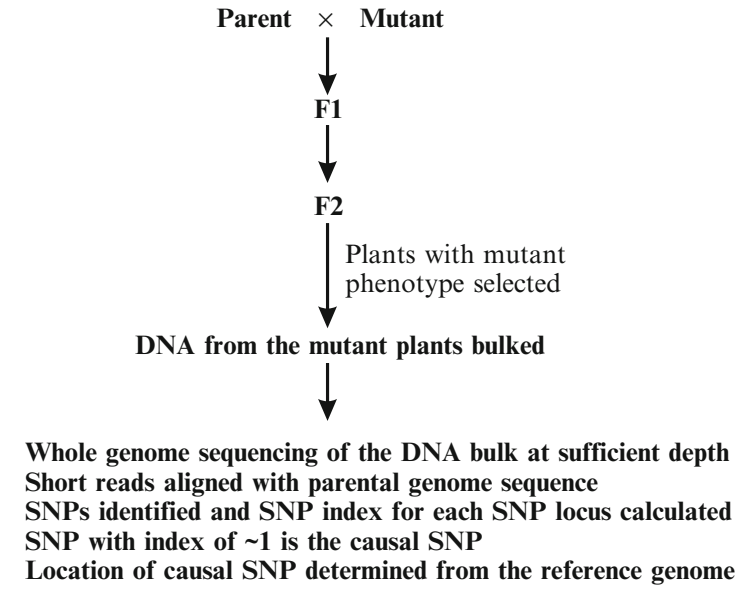


Fig. 6.5 A schematic representation of the MutMap scheme for mapping of mutant alleles of oligogenes. Sequencing is done using a NGS platform. SNP index of a SNP locus is the ratio of reads having the mutant SNP allele to the total number of reads covering the SNP locus (Based on Abe et al. 2012)

concerned gene. As a result, the mutant SNP allele involved in the gene mutation, i.e., the causal SNP allele, will also be homozygous in all these individuals. Therefore, the causal SNP locus will have SNP index of 1.0. In addition, SNP loci tightly linked with the causal SNP locus, but not involved in the gene mutation, will have SNP indices close to one. The remaining SNP loci will have SNP index of ~0.5 since the mutant and parental alleles at these loci will be in nearly 1:1 ratio in the mutant F_2 plants. The causal SNP locus can be readily mapped onto the reference parental genome. Thus, the MutMap approach involves whole-genome sequencing of a single DNA bulk and avoids marker development, genotyping of individual plants, and linkage analysis for mapping of the gene.

A variation of MutMap scheme, the *MutMap-Gap* scheme, is designed to identify mutations located in those genomic regions that are missing from the parental or reference genome. First the MutMap approach is used to identify the approximate genomic location of the causal SNP allele. Then *de novo* assembly of the missing region of the parental genome is done, the reads from the mutant bulk are aligned to this reference sequence, and the causal SNP mutation is identified and mapped. MutMap-Gap was used to isolate the gene *Pii* for blast resistance using mutant lines lacking the *Pii* function. Recently, MutMap scheme was modified on the pattern of BSA to allow mapping of mutations, including those causing seedling lethality or sterility, without the need for a cross with the parental line. In the new scheme, called *MutMap*$^+$, the seeds from mutagen-treated M1 plants are grown as individual plant progenies. Selfed seeds are harvested from wild-type plants of an M2 progeny segregating for a visible mutation, and individual plant progenies are grown. Tissues from 20 to 40 wild-type and 20 to 40 mutant seedlings are harvested from a single M3 progeny segregating for the mutant trait, and a wild-type and a mutant DNA bulk are created. The two DNA bulks are separately sequenced using NGS technology, their sequence reads are aligned with the parental reference genome, and the SNP indices are estimated. A comparison between the SNP indices for the two bulks allows the identification and mapping of the causal SNP mutations. Both MutMap and MutMap$^+$ approaches can be used to rapidly identify mutations affecting quantitative traits as well.

6.8 LOD Score and LOD Score Threshold

When two genes are segregating together, a decision has to be reached whether they are segregating independently or they are linked. This decision can be based on either a chi-square test or a LOD (logarithm of odds) score estimate. The chi-square test is simpler and far easier to carry out than estimating LOD scores, but it merely detects the presence of linkage and generates no more information. In contrast, LOD score detects linkage as well as provides an estimate of the most likely frequency of recombination between the two genes. This is because LOD score can be estimated only after assuming a value for the frequency of recombination between the two genes. Therefore, LOD scores have to be calculated for several recombination frequencies ranging from 0 to 0.5 (the maximum possible frequency of recombination with independent assortment). *The recombination frequency that yields the highest value of LOD score is taken to be the most likely value of recombination between the two genes. LOD score* (z) is the log to the base 10 of the ratio of probability of obtaining the given data assuming linkage between the two genes with a specified frequency of recombination to the probability of getting the same data with independent segregation (Morton 1955). Thus,

$$\text{LOD score}(z) = \log 10 \frac{\text{Probability of obtaining the given data assuming linkage with a specified frequency of recombination}}{\text{Probability of getting the same data assuming independent assortment}} \quad (6.12)$$

Let us suppose that two homozygous lines with the genotypes *AA BB* and *aa bb* are crossed and the resulting F_1 is testcrossed to produce the following progenies: *Aa Bb*, *Aa bb*, *aa Bb*, and *aa bb*. The genotypes *Aa Bb* and *aa bb* are the parental types, while *Aa bb* and *aa Bb* are the recombinant types. If we denote the number of recombinant types by r, the parental types would be represented by $n - r$, where n is the total number of the testcross progeny. Therefore, the frequency of recombination or recombination fraction, denoted by θ, would be r/n, and the nonrecombinant fraction would be $1 - \theta$. In case of independent assortment, the frequency of recombinant types will be equal to that of the parental types so that $\theta = (1 - \theta) = 0.5$. The probability of getting the above n testcross progeny assuming linkage between the genes a and b with recombination fraction θ will be $\theta^r \times (1 - \theta)^{n-r}$. Similarly, the probability of getting the same n testcross progeny assuming independent segregation of the genes a and b will be 0.5^n. Therefore, the value of LOD score can be estimated as follows:

$$\text{LOD score}(z) = \log_{10}[\theta^r \times (1 - \theta)^{n-r}/0.5^n] \text{ or} \quad (6.13)$$

$$= r\log_{10}(2\theta) + (n - r)\log_{10}[2(1 - \theta)] \quad (6.14)$$

A more intuitive way of calculating LOD score is as follows. Suppose the numbers of testcross progeny with the genotypes *Aa Bb*, *Aa bb*, *aa Bb*, and *aa bb* are 9, 1, 1, and 9, respectively. If the two genes were segregating independently, the probability of getting each of these genotypes will be the same, i.e., 0.25. Therefore, the probability of getting the above data with independent assortment will be

$$= 0.25^9 \times 0.25^1 \times 0.25^1 \times 0.25^9$$
$$= 0.25^{20}$$

If we assume linkage between the two genes, the recombinant fraction obtained in the testcross progeny will be 2/20 = 0.1 and the nonrecombinant fraction will be 0.9 (=1−0.1). Therefore, the expected frequency (or probability) of each of the two recombinant types will be 0.05 (=0.1/2), while that of the two parental types will be 0.45 (=0.9/2) each. Thus, the probability of getting the observed data assuming linkage with 0.1 recombination fraction will be

$$= 0.45^9 \times 0.05^1 \times 0.05^1 \times 0.45^9$$
$$= 0.45^{18} \times 0.05^2$$
$$\text{Therefore, LOD score}(z) = \log_{10}\left[\left(0.45^{18} \times 0.05^2\right)/0.25^{20}\right]$$

Thus, a LOD score of 1 signifies that linkage with the given frequency of recombination is 10 times more likely than independent segregation. Similarly, LOD values of 2 and 3 will reveal the linkage to be 100 and 1,000 times, respectively, more likely than independent assortment.

The *LOD score threshold* is the lowest value of LOD score that is accepted as evidence for linkage. *Conventionally, a LOD score of 3.0 is considered as the threshold value*. Therefore, a LOD score of 3.0 or more is accepted to indicate linkage. However, some researchers prefer a LOD threshold of 4.0. Linkage may be presumed with a LOD score lower than 3.0, but it should be stated that this is the best estimate available from the data. In some cases, the LOD score may take a negative value. It is often helpful to display the LOD scores graphically with the value of θ on the X-axis and those of z on the Y-axis. If the peak of the graph reaches z value of 3.0 or more, linkage will be accepted, and the peak will indicate the value of θ most appropriate for the data under consideration. However, the LOD score threshold will depend on the number of markers among which linkage is being tested. In case n markers are being evaluated for linkage, a total of $n(n-1)/2$ LOD score values will be estimated (one LOD score value for each marker pair). Thus, for 100 markers, a total of 4,950 LOD score values will be estimated, while for

200 markers, this number will be 19,900. In case the level of type I error is chosen as 5 %, the level generally selected for biological studies, the number of LOD score values that will equal or exceed the threshold value purely due to chance will be ~248 and ~995 for 100 and 200 markers, respectively. Therefore, the type I error of 0.1 % is chosen to keep the number of linkages detected purely by chance to a relatively low level, e.g., ~5 and ~20 for 100 and 200 markers, respectively. It may be noted that a type I error of 0.1 % will correspond to a LOD score of 3.0. More information on LOD score threshold is available in Sect. 7.8.

6.9 A Complete Linkage Map

A *complete linkage map* contains sufficiently large number of genetic markers so that every point in the genome of the species is genetically linked to at least one marker. A complete linkage map has the following features: (1) It has as many linkage groups as the haploid number of chromosomes of the concerned species, and (2) in each linkage group, the terminal positions correspond to the genomic regions immediately adjacent to the telomeric regions of the concerned chromosome. Further, (3) in theory, the total genetic length of a complete map should not increase with the inclusion of an increasingly larger number of markers in the map. Finally, (4) all the new markers included in the map should exhibit linkage with one or the other group of the markers already placed in the map.

When one or more internal regions of a chromosome are not represented in the map, due to a lack of genetic linkage with the mapped markers, the map of the chromosome will be broken into two or more parts. These parts will merge into a single linkage group as more markers located in the "gap" are mapped. Some genomic regions are poor in markers, possibly due to unusually high recombination rates in these regions, which abnormally increases the genetic distance. It may be advisable to use the BSA (Sect. 6.7.2) approach to find markers located in such regions in the place of screening the whole mapping population for a large number of random markers. In this approach, the two bulks are created for the two alleles of a marker located at one end of such a region, and the bulks are screened for a large number of markers to identify those that are polymorphic in the two bulks. The entire mapping population is then analyzed to map these markers, and some of them may actually map in the "gap" region. The same procedure may be used with the marker located at the other end of the "gap."

Mapping of a large number of marker loci with a limited number of plants in the mapping population presents insurmountable technical constraints. The bin-mapping approach overcomes these problems and allows the mapping of markers to individual bins with reasonable certainty. In the *bin-mapping approach*, the linkage map is divided into several relatively small segments called bins, and the markers are mapped within individual bins and not in the linkage map as a whole. A *bin* is a relatively small, typically 10–20 cM long, segment of a linkage group that is flanked by fixed core, anchor, or framework marker loci. (It may be noted that the concepts of "bin" and "bin mapping" in this context are slightly different from those used in Sect. 6.15.) A *core marker* is a highly polymorphic marker that is expected to be polymorphic in most, if not all, mapping populations of the given species. The anchor markers are carefully selected on the basis of their previously observed even distribution in the linkage map, high degree of polymorphism, and high reproducibility. Typically, anchor markers are SSR or RFLP markers, but some SNP markers are also used for this purpose. The marker loci already mapped within a bin do not influence the placement of new marker loci since these markers are placed with reference to the stable framework markers. Bin mapping is easily automated using ActionMap (Sect. 6.14.8) or some other suitable software. One limitation of bin mapping is that as the number of markers included in the linkage map increases, there is an increase in the number of genotyping errors; this tends to increase the estimated genetic length of the map. Therefore, the control of genotyping errors should be a priority objective in linkage mapping. MapMaker v 3.0 and other software

programs have a function that searches for genotyping errors, e.g., by detecting such recombination events that are "too close" for the given mapping population.

The representation of subtelomeric regions of all the chromosome arms in the map can be tested by evaluating the tandemly repeated telomeric sequences for linkage with the terminal markers included in the different linkage groups of the map. The recombination frequency between the telomeric regions and the terminal markers would indicate whether the marker is located near the end of the chromosome. This test has been done in tomato and maize with positive results. In case the subtelomeric regions of a chromosome are not represented in the map, the BSA approach (Sect. 6.7.2) can be used to identify markers more distal to the terminal markers included in the map. Finally, theoretical methods can be used to obtain a rough estimate of the total genetic length for a species based on the type of mapping population used and the number of markers included in the map (de Vienne 2003). The minimum number of markers needed for a complete map would depend on the total genetic length of the genome. Theoretically, one marker at every 20 cM should be sufficient, but a greater number of markers would be needed since usually the markers are unevenly distributed. Thus, theoretically, if the markers in a map were distributed at an average density to give 5 % recombination between pairs of adjacent markers, about 1 % of the marker pairs would give >25 % recombination. It may be noted that it is difficult to detect linkage between markers showing >25 % recombination.

The reference map of maize published in 1993 contained 97 markers (total genetic length 1,860 cM). The present reference map of maize available for general public use is the IBM2 map (see MaizeGDB) that has been developed from intermated recombinant inbred lines (IRILs). These IRILs were derived from the cross B73 × Mo17 by four generations of intermating among the F_2 plants, followed by continued selfing to isolate the IRILs. This reference map is divided into a number of bins and comprises thousands of marker loci.

6.10 Integration or Merger of Linkage Maps

A linkage map is specific to the particular mapping population and the marker system(s) used for its construction. As a result, multiple linkage maps have been developed for most of the crop species, and in some cases like maize, this number may be really large. When two or more linkage maps for a given species possess a minimal number of common anchor or core markers, they could be merged together to create a single more informative map called *consensus*, *merged*, or *integrated linkage map*. The process of merging different linkage maps is known as *integration* or *merger of linkage maps*. Generally, the linkage groups in a map are divided into several bins to facilitate the merger of maps. Some software packages like JoinMap (Sect. 6.14.6) and MergeMap (Sect. 6.14.7) have been specifically designed for integration of linkage maps. Integration of two or more maps generally increases marker density in the consensus map without any additional genotyping effort. Therefore, the likelihood of identifying markers tightly linked to the target genes/QTLs would be higher with a merged map than with the individual maps. Merger of maps increases marker portability, i.e., the use of polymorphic markers in more than one mapping population. Generally, the markers in a merged map are aligned with a greater precision due to the congruent anchor marker positions. Further, the inferential capabilities of consensus maps become broader since they become applicable across populations. Integrated linkage maps have been developed in several important crop species, including maize, wheat, soybean, common bean, potato, melon, etc.

Some of the main problems encountered in linkage map integration are as follows: (1) The precision of recombination frequency estimates varies greatly among the datasets for different linkage maps. (2) The type of information, e.g., the type of mapping population, the population size, and any additional information like observations on translocation and/or inversion

heterozygotes, used for the construction of individual linkage maps may also be different for the different maps. (3) An individual map might have been prepared by compilation "by hand" of data taken from the literature. (4) Often the number of common highly polymorphic markers may be limited. (5) In several species, a single marker may map at two or more loci in the genome due to genomic duplications. (6) In any case, a precise ordering of the loci placed within a single bin is challenging, and it cannot be achieved without additional data. Finally, (7) a proportion (up to ~20 % in some cases) of the markers included in the individual maps has to be excluded from the final consensus map.

6.11 Confirmation and Validation

Once linkage between a marker and the gene governing a target trait is discovered, it is necessary to ensure that this linkage is real; this is referred to as *confirmation*. Confirmation can be done by developing another mapping population from the same cross and evaluating this population for the earlier observed linkage. Alternatively, another worker may independently evaluate the same mapping population for the concerned marker-trait linkage. The next step is to determine whether the marker-trait linkage confirmed in a specific cross will hold good in unrelated germplasm; this is referred to as *validation*. Validation involves evaluation of a fairly large number of unrelated germplasm showing variation for the concerned trait for the observed marker-trait linkage. A marker that shows linkage with the target trait in diverse unrelated materials would be useful for marker-assisted selection (MAS) for the trait.

6.12 Comparative Mapping

A comparative study of linkage maps of different species is referred to as *comparative mapping*. Comparative mapping is almost as old as linkage mapping; it goes back to 1920 when Dunn is reported to have compared the linkage between genes for albinism and pink eye color in rat and mouse. Since then, numerous comparisons of linkage maps of different closely related (members of the same tribe or even family) or very distantly related (monocot and dicot) species have been undertaken. Comparative mapping in plants has been greatly facilitated by extensive linkage mapping of molecular markers. In comparative mapping, a common set of molecular markers is mapped in two different taxa, and the arrangement of these markers in their linkage maps is compared. The markers have to be orthologous and conserved across the taxa to be useful in comparative mapping. A set of *orthologous sequences* comprises those sequences from different species that originated from the same ancestral sequence. When orthologous sequences from different species are almost similar in sequence, they are said to be *conserved*. In general, single-copy DNA sequences are the most commonly used for comparative mapping, and cDNA sequences are the most likely to be conserved across related species. *RFLPs have been the most common markers used for comparative mapping, followed by CAPS markers*.

Comparative mapping reveals the similarities and differences between the genome organizations of different species. For example, a comparison of genetic maps of tomato and potato, by Tanksley and coworkers in 1992, revealed a good conservation of synteny as well as collinearity, except for five paracentric inversions. All the gene and marker loci located in the same chromosome are said to be *syntenic*, and this situation is termed as *synteny*. In contrast, the *asyntenic* loci are located in different chromosomes, and the condition is known as *asynteny*. But *collinear markers* are located in the same linear order in two different chromosomes of the same species or in the chromosomes of two different species; this phenomenon is termed as *collinearity*. Thus, two main features of a collinear run of markers are the number of markers in the run and the length (usually in cM) of the run. Both these features should be taken into account while assessing the importance of a collinear run. Generally, the

decision regarding collinearity is subjective, but a statistical test to enable objective decisions has been developed for analysis of collinearity within the maize genome.

There are basically two approaches for comparative analysis of genomes of different species, viz., comparative mapping and genome sequence-based microsynteny studies. Comparative mapping is the most accessible approach and provides a broad overview of the whole-genome organization. A variety of computer programs like MapInspect have been developed for map viewing and comparison solely on the basis of positions of different loci in the concerned maps. The microsynteny approach may cover the entire genome or it may be confined to a specific genomic region or even a specific gene. This approach yields information about the rearrangements at the DNA level, the patterns of evolution of the concerned genes, and the mechanisms involved, but it fails to yield the whole-genome picture and involves considerable effort and expenditure. But with the whole-genome sequences becoming available for an increasing number of species, microsynteny analyses are likely to become more common.

Comparative mapping has generated valuable information on (1) similarities in genome organization (synteny and collinearity) in different species, (2) patterns of genome evolution and their possible mechanisms, and (3) the possible genomic location of a gene of interest in a species on the basis of information from a related species. In general, synteny is maintained across related species and genera at the genome level. However, collinearity is usually disturbed by local chromosomal rearrangements like inversions, translocations, etc. But the order of genes appears to be conserved in smaller regions of genomes of related taxa. For example, the genomes of lentil and pea show eight well-conserved regions that together constitute 40 % of their genomes. The conservation of synteny and gene order has been extensively investigated in the grass family (Poaceae or Graminae; Bennetzen and Ma 2003). The first consensus genetic map of six species of this family, viz., rice, wheat, maize, sugarcane, sorghum, and foxtail millet, was published in 1995. This map has been periodically updated and expanded to include more grass species. The chief conclusions from this map are as follows: (1) the gross chromosome organization has largely remained conserved during the 60 million years (My) of their evolution, (2) the genomes of the present-day species are adequately represented by 30 linkage blocks of rice, and (3) these blocks would help predict the positions of genes involved in key agronomic traits of related species. It may be pointed out that there is substantial synteny even between such distantly related taxa as dicot and monocot plant species. For example, a comparative analysis of rice and *Arabidopsis* genome sequences has revealed 137 *Arabidopsis*–rice syntenic groups located at 75 sites of rice chromosomes. Further, several rice blocks mapped to more than one location in the *Arabidopsis* genome, suggesting the occurrence of genome duplication, followed by genome loss during the evolution of *Arabidopsis*. Thus, there is detectable synteny between monocot and dicot species even after their divergence for over 200 My. However, gene order conservation between monocot and dicot species is but limited. The synteny and collinearity of plant genomes have been modified by chromosomal rearrangements, which have occurred at the rate of ~1–3 rearrangements per million years.

Ancient genome duplication, followed by diploidization through genome loss, is believed to be involved in the evolution of all Poaceae crop species. This is supported by results from comparative mapping, which reveal extensive duplication in the genomes of species like maize and rice, in which 60–82 % and 53–62 %, respectively, of the genome is duplicated. Further, about 10 % of the maize genome appears to consist of multicopy sequences; this most likely is the result of duplicated genomic regions present in the diploid progenitor of maize. Other examples of plant genomes with extensive duplications are soybean, cotton, and *Brassica oleracea*. Theoretically, duplicated genes would be lost with time, but many duplicate genes are known to retain their original functions, some of them have evolved to

acquire altered expression patterns, and some others have become modified to gain new functions.

It may be expected that genes occupying homologous genomic locations in different species are likely to be homologous in function as well. The available evidence tends to favor this expectation. For example, plant height QTLs were discovered in those regions of sorghum linkage groups A, E, and H that are orthologous to the regions of maize chromosomes 1, 6, and 9, respectively, which have QTLs for plant height. In addition, the information about a desired gene from one species may be used for the isolation of an orthologous gene from a related species. For example, DNA markers from rice were used for chromosome walk in barley for isolation of the desired resistance genes. It has been suggested that species having largely syntenic genomes may be regarded to constitute a single genetic system. Therefore, markers from one species may be used in the related species to saturate specific genomic regions, and the sequence of a desired gene from one species may be used to isolate the gene from a related species. But the "unified grass genome model" has been relatively slow to develop due to the following two reasons: (1) the complete genome sequence is available for only few grass species, and (2) the collinearity observed at the linkage map level is often not seen at the genome sequence level.

A comparison of the sequences of specific genomic regions of related species provides insights into the patterns of evolution of the concerned genes and the mechanisms responsible for them. For example, a comparison among the sequences of *waxy* locus from rice, maize, wheat, and barley suggested that two introns were precisely deleted before the divergence of the ancestors of barley and wheat from those of rice and maize ~10–14 My ago. Generally, sequence conservation is the highest between the genes of most closely related species. Further, the sequence conservation is the greatest in the exons, intron–exon boundaries, and, presumably, regulatory sequences, e.g., promoters, of the genes. A comparison of the genome organization of related species would enable the identification of genomic regions that are either highly conserved or rapidly evolving. The analysis of the regions would provide insights into genome evolution, speciation, as well as domestication. The analysis of such noncoding sequences that are conserved between genomes of related species would facilitate the identification and isolation of the *cis*-acting elements needed for precise regulation of the gene expression.

6.13 Fine Mapping (High-Resolution Mapping)

For many genes, only the phenotypic effects are known, and there is no information about their protein products. Map-based cloning is one of the most promising strategies for isolation and cloning of such genes. For such cloning, high-resolution maps of the genomic regions having the target genes are a prerequisite. Linkage mapping using biparental populations usually identifies markers located at ~10 cM from the target gene, but occasionally a marker located at ~1 cM may be identified. But to be useful in positional cloning, a marker should be preferably at <0.1 cM from the target gene. Therefore, once markers linked to the target gene are identified, very large populations and a sufficiently large number of markers are used for mapping to identify markers located very close to this gene; this is referred to as *fine mapping* or *high-resolution mapping*. The following consideration would give some idea of the scale of work involved in fine mapping. In order to find a marker at a distance of 0.1 cM or less from the target gene, one has to screen a backcross population of more than 3,000 individuals for 0.95 probability of detecting at least one recombination event. Similarly, in a species with the total genetic length of 2,000 cM, a minimum of 20,000 markers have to be evaluated to achieve, on an average, a marker density of 10 markers/cM in the hope of finding a marker at 0.1 cM from the desired gene. It is assumed here that the markers are uniformly distributed throughout the genome, which is a

gross oversimplification of the real situation. It would be seen that the total number of marker assays required would be 60,000,000 (3,000 × 20,000), which is prohibitive. Therefore, a suitable strategy that allows a substantial reduction in the genotyping work must be used for high-resolution mapping.

Earlier strategies for fine mapping relied on reducing the population size to facilitate evaluation of a large number of markers in a blind search for those located in the relevant region. In one such strategy, a large mapping population is first screened with only two markers known to flank, i.e., located on either side of, the target gene. All the plants that do not show recombination between the two markers are rejected, and only plants showing recombination are retained. These plants are analyzed with a large number of new markers, and the markers located nearest to the target gene are identified. In another strategy, called *pooled-mapping technique*, plants expressing the recessive phenotype of the target trait are selected from a large segregating population of a suitable cross (Churchill et al. 1993). These plants are divided into several random pools, and each pool is analyzed with many markers. Pools containing at least one recombinant plant for a marker and the target gene are identified, genetic distances are estimated, and the most likely locus order is determined (Sect. 6.16). Churchill et al. (1993) used this approach for high-resolution mapping of a region of chromosome 5 containing the *rin* (*ripening inhibitor*) gene in tomato. Pooled mapping seems to be a highly efficient strategy, and it can be used even for QTLs. Another strategy, called *selective mapping* (Sect. 6.15), divides a large mapping population into several small samples. Each sample contains a group of individuals selected on the basis of distribution of breakpoints in specific chromosomes or chromosome regions. These samples may be used for fine mapping of the desired locus that has already been mapped to a genomic region. This approach would entail only a moderate increase in experimental effort over the effort needed for placing the target gene onto the linkage map.

The availability of genomic resources like saturated linkage maps, genome sequences, etc., has facilitated an information-guided search for markers tightly linked to a target gene. The fine mapping of rice *h2s* gene for male sterility leading to the identification of the candidate gene is a good example of such an effort. The mutant was crossed to two different lines and very large F_2 populations were generated. BSA in one F_2 population with 52 SSR markers, polymorphic in the parents, identified one SSR marker polymorphic in the bulks as well. Then 23 SSR markers flanking the above marker were assayed for polymorphism in the parents, and the polymorphic markers were used to analyze 612 male sterile F_2 plants. This enabled the identification of two SSR markers linked (at ~7.5 and 9.2 cM) to the *h2s* locus. Then 35 SSR and 74 InDel markers located within the region flanked by the above two markers were tested for polymorphism in the parents of the other cross. The polymorphic markers were then assayed with 2,400 male sterile F_2 plants, and the *h2s* locus was mapped to a 152 kb region (markers located at 0.2 and 0.3 cM from *h2s*). It would be seen that the stepwise narrowing down of the genomic region of interest and selection of markers on the basis of a comprehensive linkage map have drastically reduced the total number of markers to be evaluated. This 152 kb region was predicted to contain 22 genes. An analysis of their expression pattern indicated one candidate gene that had spatial and temporal expression patterns consistent with the h2s phenotype. This gene had a 12-base deletion in the sixth exon that is presumed to be responsible for the mutant phenotype (Qin et al. 2013).

The BSA approach has been combined with genome/transcriptome sequencing using NGS technology to achieve fine mapping. The software MultiPool (Sect. 6.14.10) is designed to analyze pooled DNA sequence data from NGS methods to identify SNP markers closely linked to the gene of interest. Similarly, the RNA-Seq approach may be combined with BSA (BSR-Seq; Sect. 6.7.4) to identify SNP markers located very close to/within the target gene. BSR-Seq was

used to map the grain protein content gene, *GPC-B1*, of tetraploid wheat to a 0.4 cM interval. Further, global, spatial, and/or temporal gene expression analysis has been used to identify the candidate genes involved in generation of this phenotype.

6.14 Software for Mapping of Oligogenes/Molecular Markers

Generally, data on several markers from relatively large mapping populations are used for construction of molecular marker maps and for mapping of oligogenes. A number of computer programs like Linkage1, GMendel, MapMaker, MapManager, etc., have been developed for this purpose. A linkage mapping software should be easy to use, have easy data preparation, provide for application of suitable statistical tools, and generate easily understandable outputs with facility of graphic visualization. Additional features like integration with other software, comparison between different analyses, evaluation of the behaviors of algorithms, etc., would be desirable for some workers, and many would prefer it to be free of cost. Researchers with an interest in software development would like the software to come with open high-quality source code so that they are able to modify and extend the program as desired. The currently available programs have been written in diverse languages and styles, with diverse user interfaces, lack interconnectivity/easy comparability, and each one of them serves a limited purpose.

6.14.1 MapMaker/Exp

MapMaker/Exp is a command-driven program designed for multipoint linkage analysis of genetic data from experimental crosses and to construct primary linkage maps (Lander et al. 1987). It simultaneously estimates all the recombination frequencies from even very large datasets for both dominant and codominant markers. The program uses a sophisticated algorithm for detecting typing errors in data, draws genetic maps as PostScript files, and can be used with a variety of mapping populations, including F_2 intercross, BC_1 backcross, F_3 (self) intercross, and sib- and self-mated RILs. The MapMaker program can work with both Haldane and Kosambi mapping functions. It uses the maximum likelihood values as the criterion for searching the best order of linked loci. As a matter of fact, it calculates recombination frequencies for adjacent intervals, assuming lack of interference among the intervals, and subsequently converts them into map distances. This is correct as long as the Haldane mapping function is used, but not when the Kosambi mapping function is applied (Stam 1993). The PC, SUN or A/UX version of the MapMaker software can be obtained on request in high density floppy discs. Alternatively, the software can be directly downloaded from the internet (genome.wi.mit.edu/distribution/mapmaker).

The raw data, including information about the type of cross, number of markers, number of progeny scored, etc., are first organized as per the file format of the MapMaker program using the "prepare data" command, and the data are saved as a text file. A pairwise analysis of data may be done to detect linkage by giving the "sequence" command and specifying the sequence of the loci. The maximum likelihood distance between each locus pair and the corresponding LOD score are calculated. If the LOD score for a pair of loci is greater than 3.0 and the distance is less than 80 cM of Haldane distance (both values represent the default setting of the program), they are considered as linked. Now the "group" command is used to divide the loci into linkage groups based on the logic that if locus *a* is linked to locus *b* and locus *b* is linked to locus *c*, then locus *a* and *c* are also linked.

The "compare" command is used to determine the most likely order of loci within a linkage group. For practical reasons, the loci within a linkage group are divided into overlapping subsets of five or so loci, most likely orders of the loci in these subsets are determined, and the subsets are overlapped to find out the order of the linkage group. The loci remaining unmapped in the linkage group are later mapped relative to the already mapped loci. After the most likely order of the linkage group is selected, the genetic distances between pairs of linked loci are updated by using an expectation maximization

algorithm. The "map" command is then given to display the map of the linkage group. The multipoint analysis facility of MapMaker takes into account such information as the genotypes of flanking markers and some amount of missing data. When a large number of loci are being mapped, the "assign" command is used in place of the "group" command. The "assign" command evaluates each new locus for linkage with the loci, called anchor loci, already located to specific chromosomes, and assigns the new loci to specific chromosomes on this basis. The results from this analysis are shown following the "list chromosomes" command.

6.14.2 RI Plant Manager

RI Plant Manager is a commercial, microcomputer-based program that requires a Macintosh Plus or later model running version 5.0 or later operating system. It is derived from RI Manager that was written for mice and is designed for genetic mapping with RIL and backcross populations. The program detects linkage between a new locus and the already known loci and determines the most likely orders for the linked loci. It displays a graphic map, including map distances, a table with all inter-locus intervals, and bibliographic reference and comments for each locus.

6.14.3 G-MENDEL

The *G-MENDEL* software has been redesigned as *G-MENDEL 3.0 PC* to operate in the Windows environment (Echt et al. 1992). It can be used for mapping in advanced backcross progeny (specifically, BC_2 and BC_3). It uses Monte Carlo analyses for locus ordering and carries out bootstrap analyses of locus orders to help select the best order. Once the locus order is obtained, it computes distances between pairs of loci from the raw distances estimated between two loci. G-MENDEL also constructs a linkage map using independent datasets that have common markers, but it does not conduct heterogeneity tests before pooling the observed and expected two-locus genotype frequencies. The map drawing function of G-MENDEL generates postscript files. These files can be imported into a graphic arts software like CorelDraw and modified as desired.

6.14.4 MultiMap

MultiMap is an expert system computer program for automated genetic linkage mapping by using heuristics for map construction; it can also be adapted for physical mapping. The order in which markers are added to the map in a nonrandom manner is based on the statistical support for order as well as the locus content. The locus content is measured by pairwise joint polymorphic information content values and genetic distances from other closely linked markers. This program has increased accuracy and speed so that the total mapping time is greatly reduced. It greatly facilitates comparison among various mapping criteria with a view to develop the most appropriate approach for linkage mapping (Matise et al. 1993). MultiMap can construct both framework and comprehensive maps, or it can expand existing framework map to a comprehensive map; it can also construct radiation hybrid maps. The user can control many of the mapping parameters that determine the types of analyses to be performed and the manner in which the maps will be constructed. MultiMap can be run automatically or interactively. In the interactive mode, the researcher is consulted at many stages for inputs concerning map construction. MultiMap is easily distributed via FTP or e-mail (tara@chimera.hgen.pitt.edu, perlin@cs.cmu.edu).

6.14.5 AntMap

AntMap is designed for construction of linkage maps using an ant colony optimization algorithm inspired by the behavior of real ant colonies

(Iwata and Ninomiya 2006). The determination of the optimum locus order becomes prohibitive when the number of loci is large. The ant colony optimization algorithm is designed to solve this problem. It can use data from F_2, backcross (BC_1), RIL (derived by selfing), and DH populations, but it cannot analyze RILs derived by sib-mating and the F_3 populations. The AntMap package carries out segregation test, classifies loci into linkage groups, and determines their optimum order. Then it constructs the linkage map, the reliability of which is indicated by bootstrap values. It performs these operations quite rapidly and nearly automatically. Source code and all the AntMap files are available from http://lbm.ab.a.u-tokyo.ac.jp/~iwata/antmap/ under GNU General Public License. AntMap can operate with Windows, Linux, Solaris, or Mac OS and requires Java 2 Platform Standard Edition (J2SE) and Java Runtime Environment (JRE) (ver. 1.4 or higher).

6.14.6 JoinMap

The program *JoinMap* was developed (Stam 1993) to use raw data from F_2, backcross, and RIL populations to prepare integrated linkage maps. A *raw data set* comprises coded genotypes for all the polymorphic markers/genes scored in the mapping population. It can also use recombination frequencies between pairs of markers/genes estimated from different experiments for developing an integrated linkage map. In addition, data from single experiments can be used for mapping. JoinMap develops the linkage map in sequential steps, and a numerical search is made at each step for the best fitting order of loci. It uses the weighted least squares method for the estimation of map distances from recombination frequencies obtained from different studies. It can also use additional information about subsets of loci for finding their best fitting order. The coding as well as the format of raw data files for JoinMap are similar to those of MapMaker, and it provides the option for Haldane or Kosambi mapping function. The current version (version 4) of JoinMap is a commercial package designed for MS-Windows version XP (Service Pack 2) platform to be run on PCs (Van Ooijen 2006).

6.14.7 MergeMap

MergeMap program is designed to integrate individual linkage maps into an accurate consensus map (Wu et al. 2011b). The investigator first assigns appropriate weights to the individual maps, which reflect the investigator's assessment of the quality/reliability of the maps. The MergeMap then converts the linkage map datasets into a suitable input data file, and the maps are merged on the basis of shared vertices into a consensus graph. The conflicts among individual maps are resolved by deleting the minimum number of markers, ordinarily from the map with the lowest weight. After this, the results are processed to generate the consensus map in the same format as the input genetic maps. The MergeMap operates in the Linux environment and depends on the boost library. It is consistently more accurate and needs less running time than JoinMap, which is currently the most popular software for this task. MergeMap can be downloaded free (for academic use only) from http://www.cs.ucr.edu/~yonghui/mgmap.html.

6.14.8 ActionMap

ActionMap automatically assigns hundreds of new loci to a fixed framework map in a single process without addition of the new markers to the framework map (Albini and Joets 2003a). This program is highly configurable, but it can be used only with inbred line and backcross populations. It has Perl and PHP scripts that automate the command steps of MapMaker. It has a set of Web forms that are used for data import. ActionMap analyzes the outputs from MapMaker to generate the file for the next step till mapping is completed. All the intermediate data, the results, as well as the raw segregation

data are stored in a database integrated into the software. It has functions that permit easy import, export, edition, update, and deletion of the data. It assigns the linkage group to a marker, computes the distances between the new marker and all adjacent loci of the linkage group, and based on these estimates determines the absolute position of the marker in the linkage group. Mapping results can be displayed either as tables or as map drawings. ActionMap can be either used online as a Web-based program or it can be freely downloaded from http://moulon.inra.fr/~bioinfo/.

6.14.9 TetraploidMap for Windows

The *TetraploidMap for Windows* is a considerably enhanced user-friendly version of the TetraploidMap program (Hackett et al. 2007). It is the only program for oligogene/marker and QTL mapping in autotetraploid species like potato. It separates molecular markers into linkage groups by cluster analysis enhanced by a graphical interface and finds the most likely order of the loci within each linkage group. It carries out interval mapping for QTLs based on a range of models and assesses thresholds by permutation tests. It has a graphical user interface, is suitable for full-sib offspring of a cross between two parents, can be used with both codominant and dominant markers, and takes into account the presence of null alleles. It has a Windows-based user interface for importing data and for displaying the results as linkage maps and QTL profile plots. TetraploidMap is free [http://www.bioss.ac.uk/ (user-friendly software)], but one needs to first register and agree to abide by its license.

6.14.10 MultiPool

MultiPool is designed for genetic mapping in experimental crosses analyzed by pooled DNA sequencing using NGS methods (Edwards and Gifford 2012). It can handle large datasets with hundreds of thousands of markers from several experimental designs with any number of replicates. It can be used for mapping of both oligogenes and QTLs. The sequence reads from a pooling experiment are aligned against a reference genome, SNPs are detected, and the allele frequencies are estimated indirectly from strain-specific read counts. These estimates are affected by several factors and are nonuniformly spaced along the genome. But the genotyping by NGS methods generates nearly saturated marker coverage for every polymorphism present in the parents of the cross. It compensates for the nonuniform noise levels by combining information about many nearby marker loci. It uses an information-sharing dynamic Bayesian network that is capable of generating robust estimates of locations of genes/QTLs and confidence intervals. The multilocus methods permit inferences to be drawn even for those genomic regions, for which strain-specific markers are not available. These methods also reduce experimental noise when many markers are available. It considers information from all linked markers for estimation of the location of a causal variant. In many cases, it was able to associate the trait phenotype with a single gene. MultiPool is freely available at http://cgs.csail.mit.edu/multipool/.

6.14.11 Mutation Mapping Analysis Pipeline for Pooled RNA-Seq

The *Mutation Mapping Analysis Pipeline for Pooled RNA-Seq (MMAPPR)* analyses BSA RNA-Seq data to identify genomic locations of recessive mutations (Hill et al. 2013a). The F_2 individuals from a cross with a mutant are divided into wild-type (both homozygotes and heterozygotes) and mutant (homozygous) phenotypic bulks. The individuals in each of the two bulks are divided into pools; tissues from individuals in each pool are bulked and used for RNA-Seq analysis. The sequence reads are processed and aligned with the reference genome sequence. MMAPPR selects polymorphic SNPs from the mapped reads, calculates SNP allele frequencies in each pool, and then estimates the locations of causative mutations. The results are

affected by bulk size and read depth. Three replications of at least 10–20 individuals each are recommended, and the default read depth is 10x. MMAPPR neither requires information about the parental strain nor a preexisting SNP map, but it does need a well-assembled reference genome. It can handle uneven recombination frequencies in different regions of the genome and manage considerable amount of noise in RNA-Seq datasets. It cannot identify genes that are missing from the reference genome or are incorrectly annotated. Further it is unable to directly identify the causative mutation if the samples are collected when the concerned gene is not expressed or the mutation lies in nontranscribed genomic region, but it still can identify the genomic region containing the mutation. It is a rapid, cost-efficient, and highly automated pipeline for mutant mapping. MMAPPR is available at http://yeast.genetics.utah.edu/software.php.

6.14.12 MapPop

MapPop is a publicly available computer program for implementing selective mapping and bin mapping (Sect. 6.15). It uses a framework map to estimate the precise positions of visible breakpoints in the map. Based on this information, it selects within minutes samples of optimum or near optimum size for bin mapping from the mapping population (having 500 individuals or less) that was used to construct the framework map. The sample selection process aims to either minimize the maximum bin length (MBL) or the expected bin length (EBL) for the sample. A cleanup routine attempts to improve the sample quality in terms of MBL or EBL. The program also implements bin mapping (Vision et al. 2000), in which new markers genotyped with the selected samples are placed at the most likely positions in the appropriate bins of the framework map. MapPop can detect and account for individual genotyping errors. It also generates a list of possible errors/filled in missing data both in the framework and the new genotype matrices. Binaries, source code, and documentation for MapPop are available at http://ars-genome.cornell.edu/software.html. MapPop ver. 1.0 can be run in Microsoft Windows 95, 98, 2000, and NT environments.

6.14.13 Next-Generation Mapping

The *Next-Generation Mapping (NGM)* is a program for quick and efficient mapping of recessive mutations using sequence data from a NGS technology (Austin et al. 2011). The NGM is a user friendly web-based tool; it is available for the analysis of NGS data at the website http://bar.utoronto.ca/NGM. The mutant to be analyzed is crossed with a mapping line to generate an F_2 population. The F_2 individuals with the mutant phenotype are pooled, their DNAs are bulked and sequenced using a NGS technology, and the reads are aligned to a reference genome. The distribution of SNPs in the genomes of the mapping line and the F_2 bulk are compared using a slightly modified version of Illumina's chastity statistic to identify the causal mutation. This statistic, termed as *discordant chastity*, measures the degree of difference between a SNP locus in the mutant bulk and the expected base in the reference genome. The SNPs are divided into "chastity threads," which are clustered into "chastity belts"; this ultimately allows the estimation of genomic position of the causal SNP. SNPs are then annotated for amino acid substitution and/or splice site disruption, filtered, and mapped. Strong resolution was obtained with average sequencing depth of 22x and genome coverage of ~70.

6.15 Selective Mapping and Selective Genotyping

Linkage mapping is based on a random sample of individuals drawn from a suitable mapping population, and there is little prior knowledge about these individuals. Therefore, a large number of different crossover sites or breakpoints can be analyzed only by genotyping a very large mapping population. The number of individuals used for analysis can be greatly reduced by

selecting them on the basis of the number and positions of breakpoints present in a mapping population. Vision et al. (2000) proposed the construction of a high-density map in two distinct phases. In the *first phase*, a high-confidence *framework map* is constructed using a large mapping population to precisely map a set of framework markers selected on the basis of even distribution throughout the genome. The framework map should be sufficiently dense, but a too dense map would be counterproductive. It has been suggested that the lowest density of framework markers should be such that the markers are evenly spaced at intervals of less than half the desired maximum bin length in the selected samples. The information from framework map is used to determine the breakpoints in each individual of the mapping population, preferably, by using a computer program (Sect. 6.14.12).

In the *second phase*, a small sample (from six plants to ~30 % of the mapping population) is drawn from the population on the basis of breakpoints. The selected individuals are such that either the MBL or the EBL (Sect. 6.14.12) for the sample is the minimum. A *bin* is an interval in a linkage group within which a breakpoint does not occur in any individual included in the sample. The ends of a bin are defined by breakpoints present in at least one individual of the sample or by the end of a linkage group (Fig. 6.6). Thus, bins are the smallest unit of resolution in the framework map, and two or more loci placed within a single bin can be ordered relative to one another only when supplementary information is generated. The individuals in the selected sample are genotyped with a large number of new markers. The positions of these markers are then inferred relative to the markers in the framework map, and the new markers are assigned to appropriate bins. This strategy, called *selective mapping* or *bin mapping*, may strike a near-optimal balance between mapping precision and the necessary genotyping effort. It may generate a high-density/saturated map with an average of ~1 marker per cM. The software MapPop (Sect. 6.14.12) carries out sample selection and selective mapping.

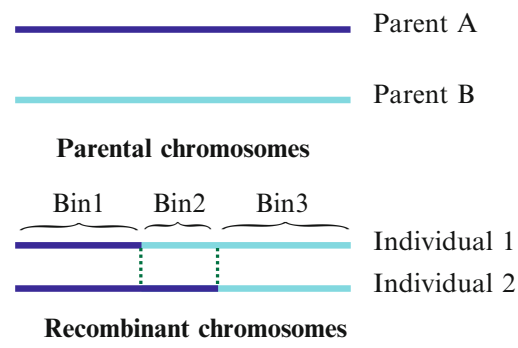


Fig. 6.6 Diagrammatic representation of the concept of "bin." Bin is the chromosomal segment within which crossing over has not taken place in the concerned mapping population. The chromosomes depicted here are homologous, and only one chromosome of a single pair is shown. The two recombinant chromosomes have only one breakpoint (representing a crossing over) each, but together they define three "bins" of the concerned chromosome. Inclusion of more individuals in the sample will divide this chromosome into more "bins" (Based on Vision et al. 2000)

The *selective genotyping*, on the other hand, is an extension of the BSA approach to facilitate linkage mapping of traits using large mapping populations of over 500 individuals with a minimum of genotyping effort. The mapping population may be F_2, backcross, RIL, DH, etc., population. The population is evaluated for the trait of interest, and 30–50 plants/lines with extreme high phenotypic values and a similar number of plants with extreme low phenotypic values for the trait are selected (Fig. 6.7). The selected plants/lines are subjected to precision phenotyping for the target trait. These plants/lines are also genotyped, either individually or by pooling their DNAs (Sect. 6.16), for a large number of markers. The data from the two groups are analyzed and based on differences in allele frequencies, the markers linked to the target trait are identified. The above approach is applicable to all such traits, the phenotypic evaluation for which does not affect survival. But when evaluation for a trait, e.g., abiotic/biotic stress tolerance, reduces survival of some of the genotypes, a random group of 30–50 plants/lines is created in the place of sensitive/susceptible

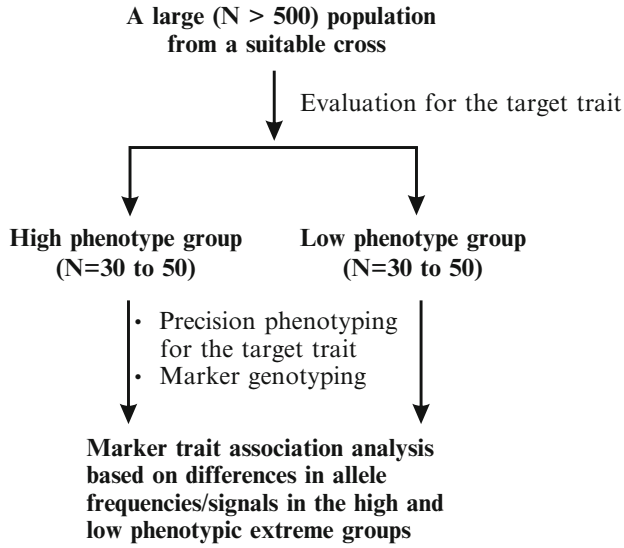


Fig. 6.7 Selective genotyping for genetic mapping of traits for which phenotypic evaluation does not affect survival; this will be the case for most of the traits. In case of traits like biotic/abiotic stress tolerance, phenotypic evaluation reduces survival of the sensitive/susceptible genotypes. In such cases, a random group ($N = 30$–50) is created in addition to the resistant/tolerant phenotypic extreme group, and marker-trait association analysis is based on comparison of random and resistant/tolerant groups (Based on Xu and Crouch 2008)

group and used for comparison with the resistant/tolerant group.

6.16 Pooled DNA Analysis

Linkage mapping using relatively large mapping populations involves considerable amount of genotyping work, which rapidly increases with the number of evaluated markers. The BSA approach was developed to reduce the genotyping effort, but individual plants of the population still have to be evaluated for the markers identified by BSA to be putatively linked to the target trait. The strategy called *pooled-mapping technique* (Sect. 6.13) is designed to reduce the genotyping work of individual plants. This strategy consists of the following steps: (1) production of a large segregating population, e.g., F_2 or testcross (for the target trait) generation, from a suitable cross; (2) selection of plants homozygous for, usually, the recessive phenotype of the target trait; (3) dividing the selected plants into several random pools of near optimum size; (4) pooling equal amounts of tissue from each plant constituting a pool; (5) DNA extraction from the pooled tissue; (6) analysis of all the pools, treating each pool as a unit, with markers putatively linked to the target gene; (7) identification of pools that contain at least one recombinant plant for a marker and the target trait; (8) estimation of recombination frequency between the marker and the gene on the basis of the proportion of pools containing recombinant plants; and (9) finding the most likely order of the markers linked with the target gene. The *recombination fractions (r)* among the loci present in the target genomic region can be estimated by the maximum likelihood method using the following formula:

$$r = (1/2k)\ln[1 - (y_A/n)] \qquad (6.15)$$

where k is the number of individuals per pool, y_A is the total number of recombinant pools, and n is the total number of pools. Further, the *optimum pool size* would be ~$1.594/2r$; it will be close to 8 for a value of 0.1 for r. Thus, the optimum pool size depends primarily on the density of markers in the target region: with low marker density,

small to moderate size pools would be optimum, while for a high marker density, larger pool sizes would be optimal. The order of loci in the target region is deduced by using a Bayesian statistical framework. Several factors like the number and the size of pools, marker density in the target region, the probability of error in phenotyping, etc., would affect selection of the correct order of markers.

The theoretical basis of above strategy is that when several markers are located within a small genomic region with the target gene, most of the individuals in a segregating population would have the parental allelic combinations of these markers and the gene. Further, only a small proportion of the individuals of the population would be recombinant for a marker and the target gene. This method has the following three requirements: (1) the target trait must be governed by a single gene, (2) it should be feasible to produce large segregating generations, and (3) it should be possible to detect in a pool the presence of a single recombinant for the target gene and a marker. In view of the last requirement, this approach can be used with dominant markers linked to the target gene in coupling phase and with codominant markers linked in either phase. Churchill et al. (1993) showed that for a marker like RFLP, one recombinant in a pool of as many as 20 individuals could be reliably detected.

The selection of homozygous recessive plants from an F_2 population will reduce the number of plants to be genotyped by a factor of four, while for a backcross population, the reduction will be two-fold. Further, creation of pools of five plants each will further reduce the genotyping work five-fold. In this way, pooled mapping would lead to a 20- and a 10-fold reduction in DNA extraction and genotyping work in the case of F_2 and backcross populations, respectively. Thus, with an F_2 population of 4,000 plants, the genotyping work will be reduced to merely 200 ($= 4,000 \times 1/4 \times 1/5$) DNA pools. Finally, BSA for a very large number of markers would enable the identification of a small number of markers putatively linked to the target trait for genotyping by the pooled-mapping technique.

The pooled-mapping approach based on BSA has been further extended and modified to facilitate mapping of both oligogenes and QTLs. The bulks used for this purpose may comprise the opposite phenotypic extremes, one phenotypic extreme and a control sample, or several pools constituted from one phenotypic extreme; all these bulks are isolated from a segregating generation of a suitable cross. The software program MultiPool (Sect. 6.14.10) can be used to analyze pooled DNA sequence data from a NGS method to identify SNPs closely linked to the target gene. Further, pooled RNA-Seq data can be used to identify SNP markers located within the target gene, and the software MMAPPR (Sect. 6.14.11) is designed for this purpose. The chief advantage of pooled DNA analysis is a dramatically reduced genotyping cost without decreasing the statistical power, particularly when large samples are used.

6.17 Physical Mapping of Molecular Markers

A *physical map* of molecular markers depicts the distances between the adjacent marker pairs in terms of base pairs. In one approach for physical mapping of markers, the genomic DNA is digested with a rare cutter restriction enzyme to generate fragments of several hundred kilobase pairs to several megabases pairs. These fragments are separated by pulse-field gel electrophoresis and transferred onto a solid support and subjected to southern hybridization using closely linked molecular markers from a dense linkage map as probes. If two molecular markers hybridize to the same DNA fragment, the length of this fragment is taken to be the maximum distance between these markers. In this way, the molecular markers of a linkage map can be localized onto the different fragments. Further, linkage relationships among the marker probes can be used to assemble the fragments into overlapping contigs spanning the entire genome

or a genomic region of interest. *Contigs* are a set of DNA fragments that represent adjoining regions in the genome, and usually pairs of these fragments have overlapping ends. This exercise generates a physical map of marker probes, which allows correlation the genetic distance to physical distance. The BAC contigs can be probed with gene-specific oligonucleotide-based probes ("*overgo*" *probes*), which are designed from expressed sequence tags (ESTs). This allows specific genes to be anchored to particular BAC clones, which facilitates linking of the genetic map with the physical map.

In another approach, the sequences of closely linked DNA markers may be aligned with a reference genome sequence to determine the locations of the markers in the genome and to ultimately develop a physical map. The reference genome can also be used for physical mapping of related species, for which neither linkage map nor genome sequence is available. For achieving this, a deep-coverage large-insert BAC library of the related species is developed. The BAC clones representing ~10 genome equivalents are fingerprinted, both the ends of each clone are sequenced, and the clones are assembled into contigs using a suitable software like FPC. The contigs are then aligned to the reference genome on the basis of end sequences of the inserts in BAC clones. This approach is being used to generate BAC-based physical maps of wild rice species.

6.18 Sources of Errors in Linkage Mapping

The genetic distances and locus orders in linkage maps are derived from the genotype and phenotype data from different mapping populations. The results from mapping studies are affected by several factors, some of which are briefly considered here.

1. *Errors in genotyping* may inflate genetic distance estimates, reduce estimates of interference, and lead to incorrect locus orders. The approaches for detecting and rectifying these errors either search for double recombinants within short distances or use appropriate computational methods.
2. *Segregation distortion* is a significant deviation of marker genotypes/trait phenotypes from the expected ratio. The expected ratio will depend on the types of mapping population and the marker used. For example, the expected ratio in an F_2 population will be 1:2:1 and 3:1 for codominant and dominant, respectively, markers/traits, but it will be 1:1 in RIL, DH, and backcross (testcross) populations. Segregation distortion is often encountered in marker data and may result from gametic/zygotic selection. There are contradictory reports on the effect of segregation distortion on genetic distance and locus order. However, it is desirable to either identify and remove, if necessary, the affected markers or use a linkage-mapping program capable of handling such data.
3. *Interference* is relevant for estimation of genetic distances from recombination values (Sect. 6.4). Haldane mapping function assumes lack of interference, while interference is common in plants. As a result, Haldane distances are, in general, longer than Kosambi distances, but they are widely used for linkage mapping in plants.
4. *Missing marker data* are another source of error in linkage mapping. They may arise due to random experimental errors leading to sporadic assay failures; in such cases, a suitable algorithm may be used to infer the missing marker scores. Alternatively, it may be nonrandom, and scoring failures may be more frequent with some markers than others; in such cases, the affected markers should be deleted.
5. In cases of visible genetic markers/traits of interest, *phenotyping errors* may occur due to misclassification of individual plants for the concerned traits. This is particularly relevant for characters having a threshold requirement, e.g., insect/disease resistance, and for quantitative traits. These errors would have effects similar to the marker genotyping errors, and every effort should be made to minimize them. The importance

of accurate phenotyping is highlighted by the development of the discipline of phenomics to address the phenotyping related issues (Chap. 15).

6. *Pooled-mapping approach* (Sect. 6.16) introduces additional sources of errors, and the *sequence data generated by NGS methods* have their own difficulties. Efforts are being made to minimize the effects of these factors by using appropriate algorithms for data analysis.
7. Finally, the accuracy of linkage-mapping results is markedly affected by the *size of mapping population*. Other things being equal, results from a larger population will be more reliable than those from smaller populations. In view of this, the framework linkage map of a species is constructed using a relatively large mapping population.

6.19 The Significance of Genetic Maps

The development of linkage maps has generated valuable insights into the genome organization of different species and enabled the use of linkage relationships between markers and genes for achieving various ends, some of which are listed below:

1. Linkage mapping provided the first substantial experimental evidence in support of the chromosomal basis of inheritance.
2. In most plant species, several probes/markers detect more than one locus in the genome, suggesting homologies within the genome. These homologies have been confirmed by genome sequencing, which revealed duplications of various sizes within the genome of a species. Homologous/duplicated regions are present even in the genomes of such diploid species as *A. thaliana* and rice that have rather small genomes. It has been postulated that during evolution, the genomes of these species had undergone partial- or whole-genome duplication, followed by genome loss leading to their diploidization. Rice is believed to have undergone at least two or three whole-genome expansion and reduction cycles.
3. A comparison of genetic maps of different related species and even distant taxa has revealed considerable conservation of the order of linked genes/markers (Sect. 6.12).
4. Linkage mapping of molecular markers provided the evidence for physical location of the elusive QTLs (Chap. 7) that contain polygenes postulated to govern the quantitative traits. This has enabled cloning of several genes located within the QTL regions, which provide an idea about the functions of polygenes.
5. Linkage mapping provides information on markers linked to genes governing traits of interest. Such markers are used for various purposes, including MAS (Chap. 9) and genomic selection (Chap. 10).
6. Close linkage between a marker and a gene of interest provides the basis for map-based or positional cloning of the gene.
7. A good high-resolution linkage map greatly facilitates genome sequencing and, particularly, genome assembly efforts.

Questions

1. Briefly describe the features of different types of genetic maps and discuss their applications and limitations. Discuss the meaning and relevance of complete linkage map.
2. "Genetic distance is related to but not the same as recombination frequency." Discuss this statement in the light of available information.
3. Briefly describe the procedure of linkage mapping of molecular markers and oligogenes.
4. Explain the concept of bulked segregant analysis and briefly describe its various modifications.
5. Briefly describe the various simplified and less demanding strategies devised for mapping of mutant alleles and even determination of the causal SNPs.

6. Explain the meaning of LOD score and its computation. Discuss the concept of LOD score threshold.
7. Discuss the various approaches for high-resolution mapping.
8. Explain the meaning of comparative mapping and discuss its relevance in plant biology and plant breeding.
9. Highlight the significance of genetic maps and the various sources of errors in linkage mapping.
10. Explain the concept of pooled DNA analysis and discuss its relevance in linkage mapping of markers.
11. Explain the concepts of selective mapping and selective genotyping and discuss their usefulness in linkage mapping of markers and genes.

Mapping of Quantitative Trait Loci 7

7.1 Introduction

Quantitative traits, by definition, show continuous variation due to polygenic inheritance and environmental influences. *Polygenes* produce small individual effects on the trait phenotype, but the effects of all the polygenes affecting a given trait are cumulative. In 1906, Yule postulated the existence of genes with cumulative action, and the experimental evidence for their existence was provided by Nilsson-Ehle in 1908. Subsequently, between 1910 and 1916, East and Emerson collected extensive data in support of polygenic inheritance in maize and tobacco. *The findings from these and subsequent studies made it clear that the continuous variation characteristic of quantitative traits resulted from the large number of polygenes involved in their control and the environmental influences on their phenotypic expression* (Singh 2009). It may be added that originally polygenes were postulated to produce only additive gene effects, but they are now known to exhibit dominance and epistatic effects as well. Since classical Mendelian methods cannot be used to follow the inheritance of polygenes, a variety of statistical tools have been developed for this purpose; these together comprise the discipline of quantitative genetics.

7.2 Quantitative Trait Loci

The development of linkage concept led to the linkage map construction and localization of various oligogenes to distinct sites in the schematic maps of specific chromosomes of different species. This gave each oligogene a physical location, called *locus*, in the genome. As a result, the oligogenes were no longer hypothetical units of inheritance. However, mapping of polygenes was not as straightforward as that of oligogenes for obvious reasons. The efforts for physical localization of polygenes began when Sax (1923) reported linkage between seed coat color and seed size, which are qualitative and quantitative traits, respectively, in common bean (*Phaseolus vulgaris*). *This work highlighted the basic principle for mapping of polygenes based on the detection of association between a quantitative trait phenotype and a genetic marker* (Sect. 2.2). But this approach does not map individual polygenes. This strategy, in fact, identifies the genomic regions associated with the expression of a quantitative trait; such a genomic region is referred to as *quantitative trait locus* (*QTL*). A QTL may contain one or more genes affecting the concerned quantitative trait. Thoday (1961) explored this concept further by combining elaborate cytogenetic techniques with genetic analysis to map QTLs for several quantitative traits in *Drosophila*. He suggested that by following the segregation of simply inherited oligogenes, the mapping and characterization of all the QTLs governing the quantitative traits should eventually become possible.

The development of DNA markers has greatly facilitated the mapping of QTLs (Tanksley 1993) leading to cloning of the genes located within some of them. The report by Paterson

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_7, © Author(s) 2015

et al. (1988) on mapping of QTLs governing fruit size, pH, and soluble solids in tomato is one of the first studies that used DNA markers for QTL mapping. They generated 237 backcross progeny from the cross between tomato (*Lycopersicon esculentum*, now *Solanum lycopersicum*) and its wild relative *L. chmielewski*. These backcross progenies were genotyped for 70 RFLP markers and phenotyped for the fruit traits. The analysis of these data uncovered six QTLs for fruit size, five for fruit pH, and four for soluble solids. In a later study, a more detailed analysis of the fruit-related QTLs was carried out in F_2 and F_2-derived F_3 family populations from the above cross, which were evaluated at three locations (Paterson et al. 1991). In this study, a total of 29 putative QTLs for fruit size, pH, and soluble solids were identified; out of these only four QTLs were detected at all the three locations. These results suggest that phenotypic evaluation of the mapping population for QTL analysis should be performed at multiple locations since evaluation at a single location may underestimate the total number of QTLs involved in the control of the concerned traits.

Over the years, not only a very large number of QTLs governing various quantitative traits but also several different types of QTLs have been discovered and mapped. QTLs have been grouped into different categories on the basis of their effect size, effect of the environment on their expression, the type of effect produced by them, and the manner of their action. *Main effect QTLs* produce direct effect on the expression of the concerned traits, while *epistatic QTLs* interact with the main effect QTLs to influence the trait phenotype. Thus, epistatic QTLs are the same as modifying genes or modifiers, and they together constitute the genetic background. A main effect QTL is described as a *major QTL* if it explains 10 % or more of the phenotypic variance for the trait, while a QTL with a smaller effect size is termed as *minor QTL*. Most quantitative traits are governed by few major QTLs and many minor QTLs. In most crop species, plant breeders would have already exploited the major QTLs. In addition, marker-assisted selection (MAS) for major QTLs is relatively easy, while that for minor QTLs is problematic (Chap. 9). The phenotypic effect of a *stable QTL* is little affected by the environment so that it is detected across environments, while an *unstable QTL* shows the opposite behavior. *Generally, most major QTLs show relatively stable expression across environments, while minor QTLs are usually sensitive to environmental variation.*

Many QTLs affect the expression level, i.e., the level of RNA transcript produced in a tissue, of various genes; such QTLs have been designated as *expression QTLs* (*eQTLs*) or *regulatory QTLs* (Sect. 8.17). The expression levels of genes can be treated as a phenotype, and variation in this trait is referred to as *expression level polymorphism*. *Metabolic QTLs* (*mQTLs*) control metabolic traits, i.e., rates of various metabolic reactions and metabolite levels. The mQTLs generally show epistatic interactions and have moderate phenotypic effects. In general, metabolic traits have much lower heritability than gene expression levels, and eQTLs and mQTLs for a specific trait are not co-localized. The quantitative variation in the cellular content of specific proteins is governed by *protein quantity QTLs* (*pQTLs*), which have been mapped in several plant species, including maize and wheat. In case of wheat, pQTLs were distributed throughout the genome, and some of them affected proteins associated with membranes. The studies aimed at the identification and mapping of eQTLs, mQTLs, and pQTL that control molecular traits constitute the field of *genetical genomics*. Finally, the QTLs involved in heterosis are called *heterosis QTLs* (*hQTLs*); these QTLs are generally different from those affecting the expression of the concerned traits.

7.3 The General Procedure for QTL Mapping

There are four salient requirements for QTL mapping: (1) a suitable mapping population, (2) a dense marker linkage map for the species, (3) reliable phenotypic evaluation for the target

trait, and (4) appropriate software packages for QTL detection and mapping. QTL mapping is generally based on biparental populations (Chap. 5). Alternatively, natural populations, germplasm collections, and breeding lines can be used for linkage disequilibrium-based association mapping of QTLs (Chap. 8). The general procedure for QTL linkage mapping is briefly summarized below.

1. As the first step, two homozygous lines having contrasting phenotypes for the trait(s) of interest are selected and crossed to develop a suitable mapping population, preferably, a doubled haploid (DH) or recombinant inbred line (RIL) population. The two homozygous lines used as parents should show a large difference for the target trait(s) and, preferably, they should have been developed by divergent selection for the trait(s).
2. The mapping population is evaluated for the target trait in replicated trials conducted, preferably, over locations and years; this is known as *phenotyping*.
3. The two parents of the mapping population are tested with a large number of markers covering the entire genome, and polymorphic markers are identified. It is important that the polymorphic markers should cover the whole genome at a sufficient density.
4. All the individuals/lines of the mapping population are now analyzed using these polymorphic markers; this is termed as *genotyping*.
5. The marker genotype data are used to construct a framework linkage map for the population, which depicts the order of the markers and the genetic distances between marker pairs in terms of centimorgans (cM).
6. Finally, the marker genotype and the trait phenotype data are analyzed to detect association between marker genotypes and the trait phenotype. In simple terms, the plants are divided into separate groups on the basis of their marker genotype. For each of these groups, mean and variance for the trait phenotype are estimated and used for comparison between the groups. In case the genotype groups for a marker differ significantly for the trait of interest, it is concluded that the concerned marker is associated with the trait, i.e., *the marker is most likely linked to a QTL controlling the trait phenotype*.

7.4 Marker and Quantitative Trait Data Structure

The data used for QTL mapping relate to marker genotype and target trait phenotype (y_i). In addition, some other relevant data, e.g., data related to age, sex, body weight, etc., may also be used as nongenetic covariates. The marker data, in fact, consist of genotype scores for markers. Suppose two homozygous lines, say, P_1 and P_2, with the marker genotypes *mm* and *MM*, respectively, are crossed to produce the F_1, *Mm*. The F_2 and backcross (BC, $F_1 \times P_1$) populations from this cross will be 1/4 *MM*, 1/2 *Mm*, 1/4 *mm*, and 1/2 *Mm*, 1/2 *mm*, respectively (Zou 2009). The scores for these marker genotypes are derived as follows: *MM*, 2; *Mm*, 1; and *mm*, 0 (the system followed by QTL Cartographer; Basten et al. 1997). The genotypes of QTLs affecting the trait of interest are denoted as *QQ*, *Qq*, and *qq*, and the phenotypes due to these genotypes are depicted as y_{QQ}, y_{Qq}, and y_{qq}, respectively. The genetic locations of the markers in the genome may be depicted in terms of their physical locations in the genome or as genetic distances (Sect. 6.4) estimated from the frequencies of recombination (r) between marker pairs (Sect. 6.3).

7.5 Methods for QTL Detection and Mapping

In simple terms, QTL mapping methods have to resolve the following three major issues: (1) the QTL genotypes of different individuals are not observed and, as a result, have to be deduced; (2) since there are potentially thousands of

possible loci in the whole genome, an appropriate genetic model for QTL analysis has to be selected from among the large number of possible models; and (3) the loci located in the same chromosome are correlated and, as a consequence, difficult to separate. QTL analysis has been and remains an area of intensive research activity as it poses a variety of challenging questions that need to be resolved for obtaining reliable and reproducible results. As a result, a large number of QTL analysis approaches have been proposed, which can be classified into the following two main groups: (1) single QTL mapping and (2) multiple QTL mapping methods (Zou 2009). Each of these groups, in turn, comprises several approaches, some of which are briefly described in this chapter. Most of these approaches use regression analysis, maximum likelihood parameter estimation, or Bayesian models for the detection of QTLs.

7.5.1 Single QTL Mapping

The *single QTL mapping methods* are able to detect a single QTL at a time. These methods do not take into account other QTLs affecting the target trait that may be present in the genome. However, quantitative traits are considered to be governed by several polygenes, which are unlikely to be located in a single QTL. Therefore, the findings from these methods tend to be less reliable than those from multiple QTL methods. But these methods are the simplest and the earliest approaches for QTL mapping and may still be relevant in certain situations. The two main methods in this category are single-marker analysis and simple interval mapping.

7.5.1.1 Single-Marker Analysis

Single-marker analysis (*SMA*), also called *single-point analysis*, is the simplest and the earliest used method of QTL detection (Soller and Brody 1976). In this method, each marker is separately tested for its association with the target trait (Table 7.1). The phenotypic means for the plants placed in the different marker genotype groups are compared to detect a QTL at or near the site of the marker. The significance of differences between the means of the marker classes can be tested by Student's *t*-test, analysis of variance, linear regression analysis, likelihood ratio test, or maximum likelihood estimation. The *t*-test can be applied when the marker genotype has only two classes. For *t*-test, individuals in the population are classified according to the genotype at a marker locus, and the significance of difference between the trait means for the two marker genotype groups is tested. A significant difference indicates the marker to be linked to a QTL affecting the trait. This procedure is repeated for every marker locus evaluated in the mapping population. The magnitude of difference between the phenotypic means of the marker genotype classes provides an estimate of the effect produced by the substitution of a single allele at the QTL locus.

Table 7.1 A tabular presentation of the findings from a hypothetical single-marker analysis for the detection of QTLs affecting a quantitative trait

Marker	Chromosome/linkage group	*P* value	R^2
A	2	<0.0001	42
B	2	0.0120	21
C	2	0.5890	2
D	4	0.0230	12
E	4	0.4312	1

The chances of detection of a QTL depend mainly on the following two factors: (1) the magnitude of the effect size of the QTL ($=y_{QQ} - y_{Qq}$) and (2) the recombination rate (r) between the QTL and the marker loci (Zou 2009). Let us suppose that the Q and q alleles of a QTL are linked with the alleles M and m of a marker, respectively, and the rate of recombination between them is r (Fig. 7.1). In the F_1 generation, the gametes MQ, Mq, mQ, and mq will be produced in the frequency 1/2 $(1 - r)$, 1/2 r, 1/2 r, and 1/2 $(1 - r)$, respectively. Therefore, in BC population, genotypes *Mm Qq*, *Mm qq*, *mm Qq*, and *mm qq* will have the frequencies of 1/2 $(1 - r)$, 1/2 r, 1/2 r, and 1/2 $(1 - r)$,

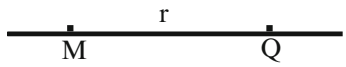


Fig. 7.1 Single-marker QTL analysis. *M*, marker locus; *Q*, QTL locus; *r*, recombination rate between *M* and *Q* loci

respectively. These individuals can be divided into two groups, *Mm* and *mm*, on the basis of their marker genotype. These groups will have the phenotypic mean of $(1-r)\ y_{Qq} + ry_{qq}$ and $ry_{Qq} + (1-r)\ y_{qq}$, respectively. Therefore, the difference in the mean phenotypic performance of the two marker groups will be

$$y_{Mm} - y_{mm} = \left[(1-r)y_{Qq} + ry_{qq}\right] - \left[ry_{Qq} + (1-r)y_{qq}\right] \quad (7.1)$$

This equation is simplified as

$$y_{Mm} - y_{mm} = (1-r)\left(y_{Qq} - y_{qq}\right) \quad (7.2)$$

Therefore, for a given magnitude of QTL effect, the larger is the value of *r*, the smaller will be the difference in phenotypic means of the two marker genotype groups and, at the same time, the smaller will be the likelihood of this difference being significant.

When the number of marker classes is more than two, the data can be subjected to one-way analysis of variance with fixed effects, which amounts to linear regression analysis. In this analysis, the individual markers constitute the single factor, and the different genotypes for this marker correspond to the levels of this factor. A significant *F* value indicates real difference between the marker genotypes for mean phenotypic values for the target trait and a linkage of the marker with a QTL affecting the trait. The fraction of phenotypic variation explained by the concerned marker (R^2, the coefficient of determination) is obtained as the ratio of the marker sum of squares to the total of the marker sum of squares and the error sum of squares. But regression analysis is the most frequently used because the estimate of R^2 provides an estimate of the QTL effect size (Table 7.1). Computer programs QGene and MapManager QTX are the normally used packages that implement SMA. Further, solved examples of SMA are available in Liu B-H (1998). The results obtained from SMA are generally presented in tabular form depicting the marker, the chromosome, or the linkage group in which the marker is located (if known), *P* (probability) value denoting the probability of linkage between the marker and a QTL governing the trait, and the fraction (as percent) of the phenotypic variance accounted for by this QTL (Table 7.1).

SMA is (1) computationally the simplest, (2) can be performed using common statistical software, and (3) does not require marker linkage maps. (4) It is generally used prior to the application of other methods of QTL mapping primarily to detect missing data. This analysis suffers from some serious limitations as follows. (1) As the magnitude of *r* increases, the likelihood of detection of a QTL with a given effect size decreases. (2) This analysis does not provide an estimate of recombination rate between the QTL and the marker, as a result of which (3) the position of QTL in the genome remains unknown. Further, (4) a high value of *r* would lead to the same result as a small effect size of the QTL. This leads to a downward bias in the estimate of the QTL effect size since the value of *r* will rarely be zero. (5) This method cannot determine whether one or more QTLs are associated with a marker. (6) The method has low QTL detection power. Finally, (7) it gives many "false-positive" signals because when the rate of false positives (Type I error) is fixed at 0.05 for one test, the actual error rate for the study will be much higher because several markers are tested in any given study.

7.5.1.2 Simple Interval Mapping

Lander and Botstein (1989) developed the *interval mapping (IM)* procedure, which is generally known as *simple interval mapping (SIM)*. This method is regarded as the second level method of QTL mapping. SIM has become a standard QTL mapping procedure and has been further

extended as composite interval mapping (CIM) and multiple interval mapping (MIM) procedures. SIM requires a marker linkage map for QTL search as it uses neighboring marker pairs to define marker intervals and searches QTLs within these intervals. SIM makes a systematic linear or one-dimensional search for a QTL at several locations, say, at every 1 or 2 cM, within each marker interval. The SIM model considers at a time a single QTL affecting the concerned trait, and each marker interval is analyzed independent of the other marker intervals. The SIM genetic model is as follows:

$$y_i = \mu + ax_i + e_i \tag{7.3}$$

where, y_i is the trait phenotype of ith individual (i takes the value of 1 to n, where n is the number of individuals in the mapping population), μ is the grand phenotypic mean, a is the QTL effect, x_i is the indicator of QTL genotype, and e_i is a random error term assumed to have mean of 0 and variance as σ^2. The term x_i represents the number of positive alleles at the QTL locus; for example, it is 1 and 0 for QTL genotypes Qq and qq, respectively. It is assumed that there is no QTL in the marker interval being examined, i.e., the null hypothesis (H_0) is that $a = 0$.

The values of μ, a, x_i, and σ^2 are seldom known. The conditional probabilities of different genotypes (x_i) of the presumed QTL within each marker interval can be estimated on the basis of genotypes of the marker pair defining the concerned interval. In a given marker interval (Fig. 7.2), the QTL location is assumed to range from that at marker M_1 to that at marker M_2 and at every possible location between the markers

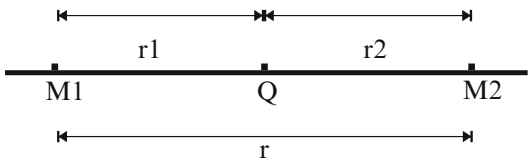


Fig. 7.2 Interval mapping of QTLs. *Q*, QTL locus; *M1* and *M2*, marker loci flanking the QTL locus; *r1*, *r2*, and *r*, recombination rates between *M1* and *Q*, *M2* and *Q*, and *M1* and *M2*, respectively

M_1 and M_2. The values of r_1 (the rate of recombination between the marker M_1 and the QTL Q) and r_2 (the rate of recombination between the marker M_2 and the QTL Q) will change with the QTL location within the interval. The values of r_1 and r_2 can be used to estimate the probabilities of different genotypes of the QTL at any given location within an interval. For example, the conditional probabilities of the different genotypes of the QTL in a testcross population will be as those given in Table 7.2. These frequencies are readily estimated on the basis of the following considerations. Let us suppose that the two markers, M_1 (alleles, M_1 and m_1) and M_2 (alleles, M_2 and m_2), are linked with recombination fraction r between them. Further, the two genotypes crossed to produce the F_1 are M_1M_1 M_2M_2 and m_1m_1 m_2m_2. Among the gametes produced by the F_1 and scored in the testcross progeny, the frequency of parental marker genotype combinations will be $1 - r$ and that of the recombinant types will be r. Now we consider a QTL Q (alleles Q and q) located between the two markers with recombination fraction r_1 between M_1 and Q and r_2 between M_2 and Q so that $r_1 + r_2 = r$. The nonrecombinant gamete M_1 Q M_2 will be produced only when there is no recombination between M_1 and Q as well as between M_2 and Q. The frequency of this event will be $(1 - r_1)$ $(1 - r_2)$. The frequency of the other nonrecombinant gamete, m_1 q m_2, also will be $(1 - r_1)$ $(1 - r_2)$. These frequencies will be divided by $(1 - r)$ to give the fraction of the testcross population represented by them. The genotype M_1 q m_2 will result when there is recombination only between M_1 and Q; therefore, it will have the frequency r_1 $(1 - r_2)$. This frequency will be divided by r to obtain the fraction of the population constituted by the given genotype. Similarly, the frequencies of the other five recombinant genotypes can be estimated. This estimation generates conditional values for x_i for the given QTL location in the marker interval under consideration.

The values for μ, a, and σ^2, however, are still unknown, and they are treated as missing values.

Table 7.2 Probabilities of different QTL genotypes with the given marker genotypes in a backcross (=testcross) population. See Fig. 7.2 for explanation of r_1 and r_2 (based on Zou 2009)

Marker genotype		Probability of QTL genotype[a]	
Marker M_1	Marker M_2	Qq	qq
M_1m_1	M_2m_2	$[(1 - r_1)\ (1 - r_2)]/(1 - r)$	$(r_1\ r_2)/(1 - r)$
M_1m_1	m_2m_2	$[(1 - r_1)\ r_2]/r$	$[r_1\ (1 - r_2)]/r$
m_1m_1	M_2m_2	$[r_1\ (1 - r_2)]/r$	$[(1 - r_1)\ r_2]/r$
m_1m_1	m_2m_2	$(r_1\ r_2)/(1 - r)$	$[(1 - r_1)\ (1 - r_2)]/(1 - r)$

[a]For QTL genotype Qq, $x_i = 1$, while for genotype qq, $x_i = 0$

A linear regression program that maximizes the likelihood function is used to obtain the maximum likelihood estimates (MLEs) for μ, a, and σ^2 with the above estimates of x_i. Lander and Botstein (1989) found the maximum likelihood estimation with missing data, specifically, the expectation maximization algorithm to be the most convenient for this purpose. The *MLEs* are the values of these parameters (μ, a, and σ^2) that maximize the likelihood of obtaining the observed phenotype data with the given marker genotype data. Now MLEs for these parameters are estimated under the assumption that there is no QTL in the marker interval so that the value for a becomes zero. The above two MLEs are used to estimate the LOD score, which indicates the likelihood of a QTL being present in the marker interval. LOD scores are estimated at various positions in the entire genome and are ordinarily presented as a graph (Fig. 7.3). In this graph, the marker positions in the linkage map are depicted on the X-axis, and the LOD scores are plotted on the Y-axis. The point where LOD score peaks and exceeds the threshold value (Sect. 7.8) is considered to harbor a QTL for the target trait. In fact, the QTL position is described by an interval called confidence or support interval (Sect. 7.9).

A regression interval mapping method was proposed by Haley and Knott (1992) to save computation time. In this method, the QTL genotype x_i is replaced by c_i, which is the conditional expectation of x_i, estimated from the flanking marker genotypes, to give the following formula:

$$y_i = \mu + ac_i + e_i \tag{7.4}$$

The LOD score is computed as $n/2 \log_{10} (RRS_0/RRS_1)$, where RRS_0 and RRS_1 are residual sum of squares under the null (=there is no QTL in the interval) and alternative (=there is a QTL in the interval) hypotheses, respectively. The regression analysis is straightforward and approximates the maximum likelihood method of SIM. A weighted least-squares method has been proposed to improve the efficiency of regression analysis. Further, Zeng (1993, 1994) has proposed a method that combines the features of both the likelihood and the regression approaches.

SIM is considered to be (1) statistically more powerful in QTL detection than single-marker analysis. (2) It provides a LOD score curve that allows localization of the QTL onto the linkage map. (3) The QTL position is represented by a support interval. (4) The QTL effect estimates are more reliable as they are not confounded with the rate of recombination between the QTL and the marker. Finally, (5) SIM takes into account missing marker genotype data, which enhances reliability of the findings. The chief limitation of interval mapping is that (1) the estimate of QTL position in the genome and that of QTL effect are biased when two or more QTLs affecting the trait of interest are linked. For example, (2) it tends to detect a single "ghost" QTL when there are actually two QTLs located close to each other. (3) Implementation of SIM requires more computation time than single-marker analysis. Finally, (4) it tends to detect only large effect QTLs. As a result, the effect size estimates for the detected QTLs tend to be biased upward; this effect is often termed as *selection bias*.

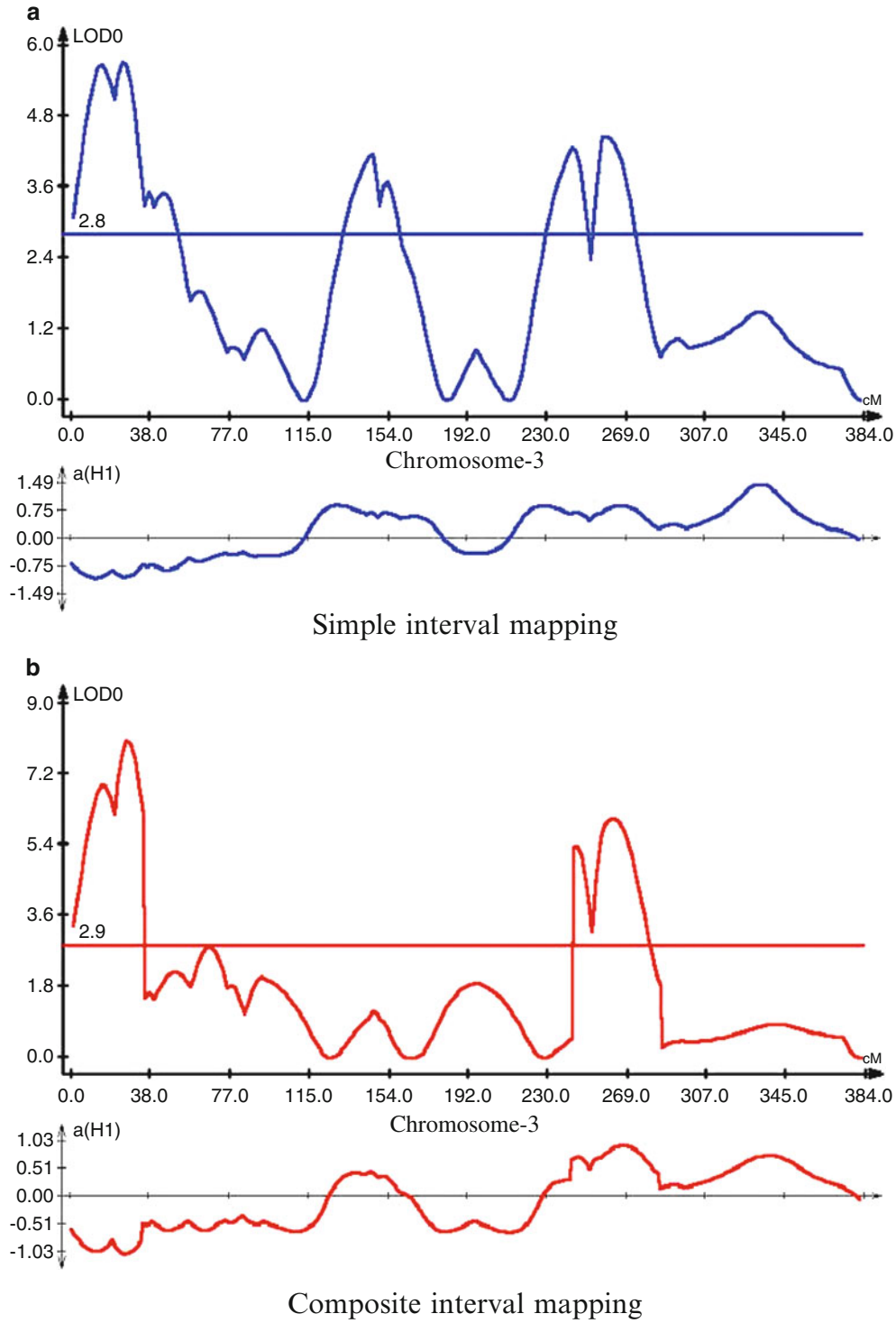


Fig. 7.3 LOD score graphs obtained by (**a**) simple interval mapping (SIM) and (**b**) composite interval mapping (CIM) for 1,000-grain weight in rice (courtesy Balram Marathi, Hyderabad)

7.5.2 Multiple QTL Mapping

The single QTL mapping methods were developed to detect a single QTL at a time. But quantitative traits are ordinarily governed by multiple QTLs. As a result, single QTL mapping methods are likely to yield biased results. *Multiple QTL mapping (MQM)* combines multiple regression analysis with SIM to include all the significant QTLs in the genetic model used for mapping (Jansen 1994). MQM offers the following advantages: (1) consideration of other QTLs affecting the trait tends to reduce residual variation and (2) increase the QTL detection power, (3) linked QTLs can be detected as separate QTLs, (4) the estimates of QTL effects are more reliable than those with single QTL methods, and (5) QTL × QTL interaction can be detected. But when too many markers are included as cofactors in the model, the QTL detection power tends to decline in comparison to SIM. The main multiple QTL mapping methods include (1) composite interval mapping, (2) multiple interval mapping, and (3) Bayesian multiple QTL mapping.

7.5.2.1 Composite Interval Mapping

Composite interval mapping (CIM) combines interval mapping with multiple regression analysis (Jansen 1994; Zeng 1994). CIM controls the effects of QTLs present in other marker intervals of the same chromosome in which the QTL is being tested and in other chromosomes as well; this increases the precision of QTL detection. CIM first carries out single-marker analysis; it then typically builds up the model as multiple QTL model using stepwise or forward regression method. In this approach, the marker with the highest LOD score is selected first; then the marker with the second highest LOD score is added, and the two markers are reevaluated for significance. If both the markers remain significant, the marker with the next highest LOD score is added to the model, and the significance of the three markers is reevaluated. In this manner, all the markers that remain significant when brought together are fitted into the model as cofactors, and the entire genome is scanned for QTL detection and mapping. The cofactors serve as proxies for other QTLs since these markers are detected to have significant association with the target trait. The inclusion of markers as cofactors in the model improves the analysis in the following two ways. (1) If the cofactor were not linked to the interval under examination, i.e., the *target interval*, the QTL detection power is increased. (2) But if the cofactor were linked to the target interval, it may help separate the QTL present in the target interval from the QTL for which the cofactor serves as the proxy.

In case of a backcross population of size n genotyped with $m + 1$ ordered markers, the additive effect QTLs are detected by the following linear regression model:

$$yi = \mu + axi + \sum_{j=1}^{m+1} bjmij + ei \qquad (7.5)$$

where yi represents the trait phenotype value in the ith individual, μ is the overall mean of the model, a denotes the QTL effect, xi is the genotype of the supposed QTL, mij is the genotype of the individual i at the marker locus j that is selected as cofactor to remove the confusing effect of the other QTL (mij equals 1 for homozygotes and -1 for heterozygote), bj is the regression coefficient of the trait phenotype on the marker locus j conditional on all other markers, and ei is the remaining random error term, which is assumed to have normal distribution. In the multiple regression analysis based on this model, the estimate of partial regression coefficient for the phenotypic values of the trait on the marker genotype is dependent on only such QTLs that are situated within the marker interval being tested for QTL. The regression coefficient is not affected by QTLs located in other marker intervals. Further, when unlinked markers are included in the multiple regression analysis, the error variance is reduced and, as a result, the power of QTL detection is increased. It is assumed that (1) the residual errors are normally distributed, (2) gene action is additive, and (3) the linked QTLs are separated by at least one blank marker interval and they do not occur

in contiguous intervals. The last assumption is more likely to be satisfied when marker intervals are short and the QTLs are loosely linked.

In the algorithm of Zeng (1994), the effect of QTL being tested and the regression coefficients of the marker variables located in other intervals of the genome are estimated simultaneously, and the regression coefficients of background markers are estimated afresh for every marker interval tested. As a result, the regression coefficient for the same marker may differ due to a change in the genomic position of the QTL being tested. This algorithm of CIM is unable to completely prevent absorption of the effect of QTL being tested by the background marker variables; this may lead to a bias in QTL effect estimates. CIM is a relatively simple procedure. It has been implemented in the freely available software QTL Cartographer (Sect. 7.19.3). As a result, it has become the most widely used method for QTL mapping in biparental populations. When CIM is implemented properly, it is the best interval mapping method based on linear regression model and maximum likelihood principles. The chief limitation of CIM algorithm is (1) the arbitrariness in selection of the cofactors for QTL analysis. In fact, different methods of cofactor selection, e.g., unlinked marker control, all marker control, and the standard model using stepwise regression (window size 10 cM), may produce different and sometimes contradictory results. (2) CIM is unable to detect interacting QTLs; as a result, it is inefficient when epistasis is present.

7.5.2.2 Inclusive Composite Interval Mapping

The *inclusive composite interval mapping (ICIM)* uses a modification of the CIM algorithm, which uses all the marker information to build the linear regression model of CIM (Li et al. 2007a). The modified algorithm aims to ensure the complete fulfillment of the two properties of the algorithm of the Zeng (1994) CIM model. Since the number of QTLs would be lower than the number of markers used for QTL analysis, standard stepwise regression analysis is used to discover the markers that are the most important for the QTL analysis; this in turn identifies the significant QTLs affecting the trait. The markers having significant regression coefficient estimates are selected as background markers or cofactors; this is done only once during the entire analysis, and the regression coefficients for the remaining markers are set at zero. ICIM is not much affected by the choice of probability levels for the inclusion/exclusion of cofactors, but a lower probability level would reduce the chances of detecting false-positive QTLs. Stepwise regression analysis is used to estimate the effects of significant markers before the interval mapping of Lander and Botstein (1989) is applied to the adjusted phenotypic data. In the case of a QTL that is located in the center of a marker interval, the effect of QTL is equally distributed between the two markers defining the interval, and the QTL maps in the middle of the interval. But when a marker is located close to one of the two markers, this marker will absorb most of the effect of this QTL, and it will map onto this marker. ICIM is capable of detecting dominance and two-gene epistasis.

In general, (1) ICIM detects a greater number of true-positive QTLs and a smaller number of false-positive QTLs than CIM. (2) The selection of cofactors does not suffer from arbitrariness. (3) The form of ICIM is simpler and the speed of convergence is faster than those of CIM, while it retains the optimal properties of CIM. (4) ICIM shows visibly high LOD scores in such genomic locations where QTLs are detected. This improves the QTL mapping power and the precision of ICIM over CIM; these conclusions are supported by extensive simulation studies. Further, (5) even in the presence of epistasis, both CIM and ICIM can effectively locate QTLs and estimate the additive effects of these QTLs, provided the narrow sense heritability estimate of the trait is not too low. Finally, (6) the results obtained from ICIM were more or less comparable to those from more complex and time-consuming Bayesian models of QTL mapping.

7.5.2.3 Joint Inclusive Composite Interval Mapping

The ICIM algorithm was extended as *joint inclusive composite interval mapping (JICIM)* for the analysis of data from multiple cross populations that have one common parent, e.g., nested association mapping (NAM) populations (Li et al. 2011a). JICIM, like ICIM, uses a two-step statistical method. The first step consists of stepwise regression analysis for identifying markers with significant regression coefficients. Following this, the coefficients of the remaining markers are set as zero. The use of stepwise regression analysis to estimate the parameters of the model avoids over-fitting of the model. In the second step, one-dimensional scanning of the marker intervals is done in a manner similar to that of ICIM. The influence of QTLs located in intervals other than the one being scanned is excluded by adjusting the phenotypic values using the regression coefficients. The existence of QTL in the interval being scanned is tested using the null (H_0) and alternative hypotheses (H_1). Expectation maximization algorithm is used to ultimately estimate the additive effect of each supposed QTL in every family.

The use of JICIM for the analysis of NAM population data allows simultaneous testing for the segregation of multiple (>10) alleles of QTLs. The JCIM consistently shows higher QTL detection power when the QTL position overlaps a marker than when the QTL is located in the center of marker interval. In fact, QTL detection is the most difficult when the QTL is situated in the middle of a marker interval. The likelihood of locating a QTL inside 1 cM of one-LOD score support interval was 85 % when the QTL position overlapped a marker. In contrast, the analysis of biparental populations of the size in hundreds allows mapping of QTLs within one-LOD score support interval of at least 10 cM. But for rare QTLs (a QTL segregating in one family only), analysis of single biparental populations is preferable to JICIM. JICIM can be extended to other multiple cross populations with common parents, e.g., eight-way cross, diallel mating design, etc. Some earlier approaches developed for the analysis of interconnected populations include the method proposed by Jannink and Jansen (2001) for multiple QTL analysis in a population created from a simple diallel between three inbred lines. This method was reparametrized and further extended by Jansen et al. (2003) for multiple QTL analysis in interconnected populations generated by other mating designs involving a single generation of mating.

7.5.2.4 Multiple Interval Mapping

The *multiple interval mapping (MIM)* approach is devised for simultaneous QTL mapping in multiple marker intervals (Kao et al. 1999). MIM avoids the complicated procedure used in CIM for the selection of background markers, but it uses several selection methods like forward search method and forward and backward selection methods to search for the best genetic model. Both of the above selection methods are implemented by QTL Cartographer, which guides the user through the model selection procedure. However, the different model selection methods usually lead to different results. The MIM genetic model includes the number, location, and interaction (epistasis) between the QTLs as follows:

$$yi = \mu + \sum_{j=1}^{k} a_j x_{ij} + \sum_{1 \le j < r < k} b_{jr} x_{ij} x_{ir} + e_i \qquad (7.6)$$

where k is the number of putative QTLs; x_{ij} is 0 and 1 if the QTL genotype of the ith individual at the jth QTL is qq and Qq, respectively (similarly, x_{ir} is either 0 or 1 for the rth QTL of the ith individual); a_j is the main effect of the jth QTL; and b_{jr} is the epistatic interaction effect between jth and rth QTLs. Since all QTL positions are unknown, all x_{ij}s and x_{ir}s are missing; hence the conditional probabilities of QTL genotypes are estimated from the concerned flanking marker genotypes. The maximum likelihood estimates of the various parameters of the above genetic model are then obtained by using the expectation maximization algorithm.

MIM is able to take into account epistatic interactions, if present, among the multiple QTLs included in the model. The chief limitations of MIM are as follows. (1) As the number of QTLs included in the model is increased, there is an exponential increase in the number of parameters. As a result, the MIM implementation is computationally intensive. Further, (2) there is a problem saturation when the model has a large number of QTLs, and the number of covariates is larger than the number of samples. Finally, (3) the selection of appropriate model from among the innumerable models that are possible is a challenge since an appropriate and reliable criterion of model selection is difficult to develop.

7.5.2.5 Bayesian Multiple QTL Mapping

Bayesian QTL mapping has been designed for the detection of multiple QTLs. It treats the number of QTLs as a random variable and uses reversible-jump Markov Chain Monte Carlo (MCMC) procedure for specific modeling (Satgopan et al. 1996; Banerjee et al. 2008). In a Bayesian model, a prior distribution is selected, from which the posterior distribution is derived, and inferences are drawn from the posterior distribution. Both CIM and Bayesian methods use maximum likelihood functions. The advantages of prior distribution decline with the increase in sample size. Therefore, for most biparental mapping populations (population size in hundreds), the Bayesian method may offer little advantage over the conventional mapping, particularly when high-density maps are available, and the genotype data are nearly complete. The Bayesian mapping methods are flexible in handling the ambiguity related to the QTL number, locations of the QTLs, and missing genotypes of QTLs. Bayesian models estimate the probability that a QTL exists in a given marker interval; this feature is regarded as the major advantage of these methods. The Bayesian mapping has not been used widely partly because of the following reasons: (1) difficulties in choosing a prior distribution; (2) complexities of computation, including that of the posterior distribution; (3) challenges in deciding the acceptance probability for each change in the dimension; and (4) lack of user-friendly software. Some models have been implemented in QTL Cartographer, FlexQTL, INTERQTL, R/QTLBIM, etc. (Sect. 7.19).

7.5.2.6 Some Other Approaches for QTL Mapping

The analysis of huge SNP genotype data requires a method that is fast, efficient, and capable of fine-scale mapping of QTLs. The various QTL linkage mapping methods use one of the following procedures: full likelihood, nonparametric analysis of linkage, and variance component estimation. The SMA is the simplest QTL mapping method, but it suffers from high rate of "false-positive" signals as well as low power of QTL detection. The full likelihood methods are computationally intensive and the nonparametric methods have their own limitations. In the variance component methods, marker genotypes are used to identify the QTL alleles that are identical by descent. The *haplotype-based variance component method* proposed by Meuwissen and Goddard (2001) assumes QTL effects to be random. It estimates the probability of QTL alleles being identical by descent on the basis of marker haplotype similarity. The QTL variance components are estimated by restricted maximum likelihood method from the inferred identity-by-descent (IBD) probability matrices of the QTL alleles. These matrices are constructed from IBD probabilities calculated from the two ancestral haplotypes. The effects of environmental factors and those of polygenic background can be easily incorporated into this QTL mapping method. This method is more powerful than SMA and tends to have a continuous profile of QTL detection, which reduces the false-positive signal rate. But the computations of variance components are very time-consuming and they are difficult to converge. The software GridQTL implements this algorithm and uses a large public grid of computers in parallel for analysis.

The HAPim method proposed by Boitard et al. (2006) uses LD information for interval mapping by a likelihood maximization

procedure. This method does not require the values of t (the time when LD was created) and N (the effective population size). In contrast, the IBD method of Meuwissen and Goddard (2001) uses $t = 100$ and $N = 100$ as default, which may not be appropriate for at least some studies. The QTL positions obtained by the HAPim method were comparable to those estimated by the IBD method. A variance component method based on the Bayesian approach allows detection and mapping of interacting QTLs. This method uses information on both linkage and linkage disequilibrium for its covariance structure and accommodates additive, dominance, as well as epistatic effects of the multiple QTLs. Therefore, this method allows the detection and fine mapping of both main effect QTLs as well as those generating epistatic effects (Lee and van der Werf 2007). Bink et al. (2012) developed the approach for the incorporation of parental IBD matrices in linkage analysis by two Bayesian QTL mapping methods called Threshold IBD model and the Latent Ancestral Allele Model. They carried out simulation analyses and showed that the incorporation of parental IBD information considerably improved the power and, above all, the accuracy of QTL linkage mapping, including the QTL position and the QTL effect size.

Fang (2012) proposed a fast *expectation maximization algorithm under fixed effect model* (*EMF*). EMF assumes the QTL to be biallelic and solves model effects by using an expectation maximization algorithm. EMF avoids construction of IBD matrices and restricted maximum likelihood method for variance component estimation, which saves considerable computation time. The results from simulation studies show that this method is computationally much faster than the variance component method; it is comparable to the latter in terms of power of QTL detection as well as estimation of parameters, and both of them outperform single-marker analysis and interval mapping. But EMF has lower QTL detection power than the variance component method when the QTL is multiallelic; however, it can be modified to handle multiallelic QTLs.

7.5.3 Some Remarks on QTL Mapping

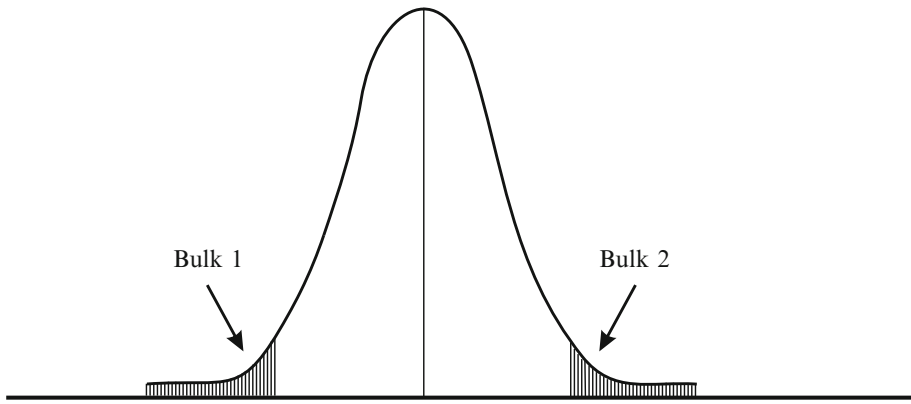
Most of the methods described above are designed for use in mapping populations derived from a single cross between two inbred/homozygous parents. However, plant breeding programs generally use complex crosses involving two or more inbred/homozygous parents. The data from such crosses may be analyzed separately for each population, and the results so obtained may be compared and combined in some fashion. However, this may reduce the QTL detection power. Some methods for QTL mapping were developed for concurrent analysis of data from all such populations. For example, CIM has been extended to permit analysis of data from multiple cross populations sharing one common parent (Sect. 7.5.2.3). Methods for analysis of complex crosses have been implemented in some software packages like QTL Express and MCQTL.

Most QTL mapping methods assume that the data on trait phenotype are normally distributed with reference to the genotype for each QTL since this supposition leads to a great simplification in the form of the likelihood function. But this assumption may not be always fulfilled. When it is suspected that the distribution of data for a trait phenotype is not normal, one option is to transform the original data so that the distribution becomes normal. Alternatively, the significance of the estimates of genetic effects can be tested by applying the model-free Wilcoxon rank-sum statistics. In addition, some other semi-parametric and nonparametric methods have been developed for QTL mapping. For example, a semi-parametric QTL mapping approach uses the exponential tilt; many of the parametric models have been derived from this semi-parametric model.

7.6 Bulked Segregant Analysis for QTL Mapping

The idea of *selective genotyping* of the 5 % most extreme phenotype progeny for QTL mapping was first explored by Lander and Botstein

(1989). This idea is similar to the bulked segregant analysis (BSA) scheme, also called tail analysis or selective DNA pooling (SDP), that was originally developed for the mapping of oligogenes. The general features of the BSA scheme used for QTL mapping are essentially the same as those for oligogene mapping (Sect. 6.7.2). For example, an oilseed rape F_2 population consisting of ~2,500 plants was segregating for glucosinolate content. Bulks with high and low extremes of glucosinolate content were created and scored for 2,000 AFLP loci to identify those loci that were polymorphic between the two bulks. These polymorphic AFLP loci were scored in ~200 random plants from the F_2 population. This procedure identified three QTLs governing glucosinolate content; these findings were confirmed by independent studies (Peleman et al. 2005). *QTL-Seq* is an extension of BSA; it has been designed for QTL mapping by subjecting the "high" and "low" bulks of DNA to whole-genome resequencing. The "high" and "low" bulks of DNA contain DNAs from 20 to 50 plants showing the extreme high phenotype and the extreme low phenotype, respectively, for the target quantitative trait selected from a suitable mapping population (Fig. 7.4). The short sequence reads obtained from an NGS method are aligned against the reference genome sequence of one of the parents of the mapping population, and SNP-index plots of the "high" and "low" bulks are compared. The genomic region having a QTL affecting the target trait would display contrasting patterns of SNP indices in the plots for the two bulks. This approach was successfully applied to rice RIL and F_2 populations to identify QTLs for partial resistance to the blast disease of rice and those for seedling vigor. The results from a simulation study suggest that QTL-Seq would be able to identify QTLs over a range of experimental variables (Takagi et al. 2013). The QTL-Seq and MutMap+ (Sect. 6.7.5) schemes are similar in their general approach, but they differ in the following respect: MutMap+ is designed for the detection of the causative SNP responsible for a mutant phenotype, while QTL-Seq is used to map the approximate locations of QTLs governing the target trait.



A. Distribution of the mapping population for phenotype of the target trait

1. Creation of Bulk 1 (extreme low phenotype) and Bulk 2 (extreme high phenotype) for the target trait
2. Equal amounts of DNA from all the plants in each 'bulk' are taken to generate the two DNA bulks
3. The two DNA bulks are sequenced separately using a NGS method
4. Short sequence reads are aligned against a reference genome of one of the parents
5. 'SNP' index for the high and low bulks computed and compared
6. The SNPs showing contrasting patterns in the two bulks denote the QTLs governing the target traits

Fig. 7.4 A simple schematic representation of the QTL-Seq approach of QTL mapping using short sequence reads generated by a NGS method (based on Takagi et al. 2013)

In the SDP approach, like in BSA, two DNA bulks are created from the individuals constituting the two tails of the target trait distribution in the mapping population. Several different approaches have been proposed for the estimation of QTL position and the confidence interval from marker genotype data obtained from the above two DNA pools. In the *fractioned-pool approach*, the individuals present in each of the two tails are randomly divided into several independent sub-pools. Each sub-pool is genotyped separately, and the data are analyzed to obtain marker allele frequencies in the two tails, which allow QTL detection and mapping. This concept has been extended as *fractioned-pool design (FPD)* to provide a complete and reliable analysis system for QTL mapping (Korol et al. 2007). The FPD analysis does not require normal distribution of the trait. But sufficient number of individuals should be present in the two "tails" for a reasonably high QTL detection power. FPD can use several such statistical tools that could earlier be used only with the individual genotyping procedure. For example, it uses permutation tests for QTL detection, and estimation of confidence intervals for QTL positions and QTL effects are based on jackknife or bootstrap resampling techniques. FDP findings are more reliable than those from the analysis of a single pool per "tail" of the trait distribution. It is a highly cost-efficient method for genome-wide QTL analysis in F_2 and *BC* populations. FPD can be extended to SNP microarray data analysis leading to dramatic reduction in genotyping costs.

7.7 Multiple Trait QTL Mapping

In QTL analysis experiments, usually data on more than one quantitative trait are collected, and often these traits are correlated. There are two ways in which the data from multiple traits may be processed: (1) the data for each individual trait may be analyzed separately or (2) the data for all the traits may be analyzed together by taking into account the trait correlations. A joint analysis of correlated traits is preferable in view of the following three advantages. (1) Inclusion of information from the correlated traits can increase the power of QTL detection. In general, as the number of traits in an analysis increases, the number of relevant significant QTLs involved in their control also increases. (2) A joint analysis can enhance the precision of QTL effect estimates. Finally, (3) it provides appropriate formal procedures to test a number of biologically interesting hypotheses, including whether the observed trait correlations are due to QTL pleiotropy or a close linkage among QTLs affecting individual traits. *Pleiotropy* refers to a single gene influencing the phenotypic expression of more than one trait. Testing these hypotheses is key to understanding the biochemical pathways underlying complex traits, which is the ultimate goal of QTL mapping. Dense marker coverage would be helpful in the test for and separation of multiple linked QTLs and, thereby, help resolve the issue of pleiotropy. It may be pointed out that the real basis of trait correlations has important implications for plant breeding: *undesirable correlations due to close linkage offer a chance for breaking them, while those due to pleiotropy offer little, if any, such chance*.

Jiang and Zeng (1995) were the first to propose a version of CIM (based on maximum likelihood approach) for joint analysis of multiple traits. Since then, several different methods have been developed, including (1) methods based on maximum likelihood approach; (2) those based on least-squares approach; (3) a dimension reduction technique like principal component analysis, discriminant analysis, or use of canonical variables associated with the traits; (4) MCMC algorithm; (5) a Bayesian approach using reversible-jump MCMC; (6) a Bayesian shrinkage analysis with a fixed-interval approach; and (7) a seemingly unrelated regression model implemented by the Bayesian approach. Each of these approaches has one or more weaknesses. For example, all the multivariate methods use the traditional multivariate regression model, which assumes the same genetic model for all the correlated traits. However, this assumption is unrealistic since most correlated traits are likely to show different

modes of genetic control. The seemingly unrelated regression model allows different genetic models to be used for the different traits, but the currently available model can accommodate only strictly additive effect QTLs. But this model can be extended to include QTL × QTL and QTL × environment interactions as well. Some software packages that carry out joint QTL analysis of multiple traits are QTL Cartographer, QGene 4.0, and FlexQTLM.

The QTLs involved in the control of different correlated traits usually map in the same genomic region; such a genomic region is often referred to as *QTL hotspot*. A QTL hotspot may contain hundreds of different genes, e.g., an estimated 600 genes in a QTL hotspot containing QTLs for different biomass-related traits in poplar (*Populus* sp.). Another type of QTL hotspot is observed for eQTLs, called *eQTL hotspots*; these hotspots are genomic regions containing eQTLs, which affect the expression of several different genes located in the same genomic region. As a result, a single polymorphism in such a hotspot would lead to widespread changes in gene expression. According to a hypothesis, an eQTL hotspot represents a common master regulator gene linked to the eQTL. For example, cyclin H has been shown to be a transcriptional regulator of those genes that constitute a known hotspot.

7.8 LOD Score and LOD Score Threshold

In case of SMA, the significance of association between the markers and the traits is assessed by the particular test that is appropriate for the statistical analysis that was used for detecting the given association. For example, in case of analysis of variance, the significance of the F value is used as an indication of the significance of the detected QTL. In case of the likelihood ratio test (LRT)-based methods like SIM, CIM, and ICIM, the QTL position is indicated by a peak of the LOD score (Sect. 6.7) profile that either equals or exceeds a predecided value. The value of LOD score that must be either equaled or exceeded by the observed values of LOD score for being considered as significant is referred to as *LOD score threshold*. LOD score is readily derived from LRT since LRT = 2 ln 10 LOD, where ln is the natural log. Since 2 ln 10 is approximately 4.61, LRT = ~4.61 × LOD. The LOD score threshold is affected by many factors, including the size of the genome, the density of markers, and the amount of missing data. The value of LOD score threshold for each marker interval may be obtained from a *chi-square table* (with one degrees of freedom) since LRT statistic approximates this distribution. This is particularly true when a relatively small number of markers are added to the model and the sample size is large. However, this significance level is inadequate because in any QTL analysis study, one evaluates the entire genome, and not a single-marker interval, for the presence of a QTL.

An empirical threshold value for LOD score may be obtained from a *permutation test* (Churchill and Deorge 1994). In this method, the marker genotypes for the individuals of the sample are kept unchanged, while their trait phenotype values are randomly shuffled. Thus, every individual retains its original marker genotype, but it is assigned a random value for the trait phenotype from among the observed values with the restriction that each observed value is used only once. As a result, the original association, if any, between the trait phenotype and the marker genotype is totally disrupted. QTL analysis is now done using the marker genotype and the "shuffled" phenotype data, and LOD score is determined for a given position in the genome. This process is repeated, usually, 1,000 times for a given genomic position, and the LOD scores so obtained are examined to obtain the LOD score threshold value. *Since all the LOD scores obtained by the permutation process represent "false" marker-trait associations, the proportion of LOD scores exceeding a given value would give the probability (P) of Type I error if this value were used as the LOD score threshold*. The *level of Type I error* denotes the probability of detection of a "false-positive" QTL when, in fact, there does not exist a QTL. One may select a threshold value at which P equals 0.05, 0.01 or, sometimes, even 0.001.

Table 7.3 LOD score thresholds estimated from 1,000 permutations of the original data on root thickness from a population of 203 RILs of rice (based on Churchill and Deorge 1994)

	LOD score threshold	
Type I error level*	For the whole experiment	For a single comparison[a]
0.05	2.51	1.34
0.01	3.24	2.13

*The probability of "false-positive" signals for the existence of a QTL affecting the target trait

[a]Average of LOD threshold values across all the points in the genome at which QTL search is made

The above process will generate a LOD score threshold value for comparison at that specific genomic position, for which the 1,000 LOD scores were obtained by the permutation test. The above procedure may be repeated for every genomic position at which the presence of a QTL is to be tested. All the LOD score values computed in this way are considered together to obtain the LOD score threshold value. This procedure gives the LOD score threshold value for the entire experiment; this value, as a rule, will always be considerably higher than that for the single-point comparison (Table 7.3). Thus, the permutation test generates LOD thresholds specifically for the experiment, for which it is computed. Further, it is time-consuming, and its naive application may inflate the Type I error rates. However, this method does not depend on the distribution pattern of phenotype data. A new resampling procedure has been proposed for genome-wide testing of significance, using a score statistic that is comparable to LOD score. However, LOD score remains the most widely used test and is implemented by almost all QTL mapping software, which also generate the estimates of threshold LOD scores, usually, based on permutation test. But before the introduction of the permutation test, a LOD score of 3.0 (sometimes, even 2.0) was widely employed as the threshold LOD score value. If required, the probability of a QTL being present at a testing position can be calculated. In fact, the LOD score indicates the probability of a QTL being present at the genomic position being tested in the case of mapping methods based on likelihood ratio test.

7.9 QTL Confidence/Support Interval

The position of a QTL in the linkage map is often depicted as a bar beside the map. If QTLs for other traits are located in the same region, they are denoted by additional bars placed side by side (Fig. 7.5). The length of this bar represents an interval, called *confidence interval* or *support interval*, in which the QTL is likely to be located. The confidence interval extends on either side of the point at which the LOD score peak is located. The confidence interval provides directions for future experiments and indicates the genomic region to be probed by fine-mapping strategies. The QTL mapping methods do not yield direct estimates of the QTL position and the support interval. Lander and Botstein (1989) proposed the widely used *LOD score drop-off method* for determining the confidence interval. In this method, the confidence interval comprises the interval demarcated by the map positions on either side of the LOD score peak, at which the LOD score drops to one less than the peak LOD score value. This confidence interval is often known as *one-LOD support interval*. This interval, however, depends on the effect size of the QTL, and it fails to behave as true confidence interval.

The bootstrap method of resampling can be used to construct *empirical confidence intervals* with reasonable coverage. *Bootstrap* is a method for resampling, in which the trait phenotype values for the different individuals in the sample are replaced by random phenotype values drawn from the sample. In this procedure, the same trait

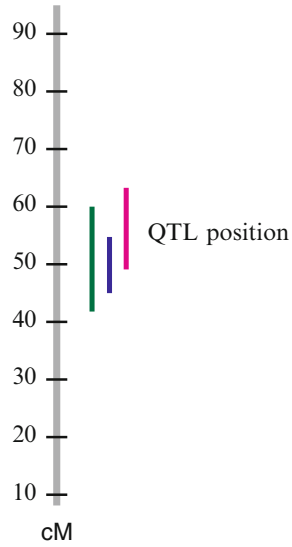


Fig. 7.5 The location of QTLs for traits 1, 2, and 3 depicted in the linkage map as green, blue and red bars. Since the QTLs are co-localized, they are depicted side by side. The lengths of bars correspond to the respective QTL confidence intervals

phenotype value may get assigned to more than one individual in the sample, while some observed values might not be included. In general, bootstrap confidence interval is slightly conservative unless significant replicates are used. Further, the coverage of these intervals for QTL location is critically affected by the QTL position within the concerned marker interval, and this method is computationally demanding. An *approximate Bayes credible interval* for QTL position is claimed to give dependable coverage independent of the effect size of QTLs, density of markers, and the size of samples.

7.10 Confirmation and Validation of QTL Mapping Results

A study that leads to detection and mapping of a QTL governing a trait of interest is called *primary study*. The results from primary studies should be confirmed and validated by later studies termed as *replication studies*. *Replication studies ensure that the detected QTL is real and verify the QTL position and effect reported by the primary study*; this is the *confirmation of QTL mapping results*. A replication study may be conducted by the same workers by constructing a new mapping population from the same parents or closely related parents. It may also be based on the same population that was used for the primary study, provided some other worker(s) perform the replication study. The confirmation of QTL mapping results is necessary before the markers linked to a QTL can be used for MAS for the QTL. This is needed because the population size in most QTL mapping studies is small (<500), which leads to low QTL detection power and introduces a large bias in the estimates of QTL effect size. However, confirmation of most QTL analysis results is not done primarily due to resource, including time, constraints.

QTL validation consists of confirmation of the marker QTL association and the QTL position in unrelated germplasm and assessment of the effect of genetic background on QTL expression. It would also be very useful to know if a QTL has some undesirable effect on the performance of lines expressing this QTL. The most common approach for QTL validation is the analysis of additional mapping populations developed from parents other than those used in the primary study; this would allow validation of marker-trait association, the position of QTL, and the effect size of QTL. But the development of additional mapping populations and their analysis requires considerable effort, resources, and time. The analysis of a range of cultivars and elite germplasm lines may also be used for the validation of marker-trait association. However, a comparison among a set of near-isogenic lines (NILs), developed by using a single donor parent and different recurrent parents, provides a more dependable means of QTL effect validation. This is because in each NIL pair, the QTL is placed in the genetic background of the recurrent parent, which is used for comparison. NILs are usually developed by 5–6 backcrosses, which requires considerable effort and time. Therefore, it has been proposed to combine QTL discovery, QTL effect assessment, and the use of these QTLs for line development, e.g., by the advanced backcross QTL analysis (Sect. 9.11.2) and the inbred

enhancement and QTL mapping procedures (Sect. 9.11.1).

A method called heterogeneous inbred family analysis was proposed for quick development of NILs from RIL mapping populations (Sect. 5.11). In this procedure, the RILs are screened with molecular markers linked to the concerned QTLs to identify those RILs that are heterogeneous or segregating for these markers. NILs for the concerned QTLs are then isolated from these heterogeneous inbred lines: the homozygotes for the marker alleles linked to the two QTL alleles isolated from a single line comprise a pair of NILs. The evaluation of these NILs allows determination of the positions of concerned QTLs and provides estimates of the QTL effects in somewhat different genetic backgrounds. However, these estimates are applicable only to the concerned mapping population. Further, if the population size is not large and the RIL population has been carried to F_6 or a later generation, the number of different NILs obtainable for a given QTL would be limited. However, the analysis of several different NIL pairs would be desirable, as it would provide a more reliable estimate of QTL effects in different genetic backgrounds. In view of the above, Pumphrey et al. (2007) proposed a scheme for rapid isolation of NILs for a QTL (or a gene) from the breeding materials developed using a parent with the concerned QTL (Sect. 5.10). This approach permits rapid isolation of pairs of NILs with the concerned QTL in several different genetic backgrounds, which affords a more reliable assessment of the QTL effects over genetic backgrounds. This strategy will be useful for such QTLs that are present in donor lines being used in several breeding programs. But this procedure would not be useful for a novel QTL discovered in a line that is not being commonly used as a parent in hybridization programs.

7.11 QTL Fine Mapping

QTL fine mapping consists of identification of markers located very close to, preferably at <1 cM from, the concerned QTL. It involves facilitating the occurrence of crossing overs as close to the target QTL as possible and scoring the crossover products to identify markers located very close to the QTL. Some of the common strategies for QTL fine mapping are as follows.

7.11.1 Homozygous Lines Derived from Near-Isogenic Lines

A backcross program is used to produce a pair of near-isogenic lines (NILs), one of which carries the target QTL. The two members of an NIL pair are crossed to produce a large F_2 or backcross population, and plants heterozygous for the donor genome segment carrying the QTL are identified. These plants are selfed to isolate plants homozygous for this segment. Progenies of these plants are evaluated and genotyped for markers located in the target genomic region to identify markers, which are more closely linked to the QTL of interest than the earlier markers.

7.11.2 Intercross Recombinant Inbred Lines

Intercross recombinant inbred lines (IRILs) are produced by few to several generations of random mating or inter-mating among the individual plants beginning in the F_2 generation. This step increases the likelihood of recombination between the QTL and the marker loci closely linked to it. At the end of inter-mating, either RILs or doubled haploid (DH) lines are isolated. After n generations of inter-mating, the recombination rate (R_n) observed between any two loci in the RIL population will be $(n + 2)$ times the normal recombination rate (r), i.e., $R_n \approx r(n + 2)$. Thus, the genetic distance between two loci will become expanded by the same factor. Therefore, a distance of 1 cM would be detected as 10 cM after eight generations of inter-mating (de Vienne and Causse 2003). The RIL/DH lines so obtained are analyzed by a suitable technique, including pooled mapping (Sects. 6.15 and 7.9), QTL-Seq (Sect. 7.9), or BSR-Seq (Sect. 6.7.4) to identify markers close to the target QTL.

7.11.3 Recurrent Selection Backcross QTL Mapping

Recurrent selection backcross (RSB) QTL mapping uses an RSB population (Sect. 5.15) that is derived from a cross between two homozygous lines, which differ for a quantitative trait and for a large number of dense (say, at every 1 cM), evenly distributed markers (Luo et al. 2002). The selection during RSB for the quantitative trait will maintain the large effect QTLs and the markers linked to it in heterozygous state, while other markers will become fixed for the recurrent parent (RP) marker allele. In every generation, recombination would occur between the QTL and the linked markers, and the marker locus involved in recombination will become homozygous for the RP allele. After a sufficient number of backcrosses, the selected population is genotyped for markers, and mean and variance for marker heterozygosity are estimated and used for QTL mapping. The QTL will be located at or close to the marker showing the highest heterozygosity. Theoretical and simulation analyses showed that this method could reduce the QTL confidence interval to 1 cM or less; the mapping resolution increases with the number of RSB generations. In addition, practical application of RSB method does not require complicated statistical modeling of the experimental data.

7.11.4 Genetically Heterogeneous Stocks

In out-crossing species with a short generation time, a *heterogeneous stock* can be generated by crossing several inbred strains/lines and maintaining the population by random mating or mating in pairs a suitably large number of individuals (Sect. 5.18). The chromosomes of an advanced generation of a heterogeneous stock would have undergone many rounds of recombination since the initiation of the stock. These chromosomes will contain small blocks contributed by the different founder inbred parents, and these blocks will become increasingly smaller with the advancing generation. Therefore, these populations would permit a fine mapping of QTLs/genes. For example, the 60th generation of a mice heterogeneous stock was estimated to allow mapping of QTLs within 0.5 cM interval with the analysis of <2,000 animals.

7.11.5 Multiparent Advanced Generation Intercross Population

The *multiparent advanced generation intercross (MAGIC)* populations are a collection of RILs produced from a complex cross/outcross population involving several parental lines. The parental lines may be inbred lines, clones, or individual plants selected on the basis of their origin or use (Sect. 5.17). The progeny from the complex cross or outcross may be inter-mated for one or more generations before the isolation of RILs. A QTL mapping method based on the reconstruction of haplotype mosaics for each of 527 RILs could map QTLs that explained 10 % or more of the phenotypic variation within 300-kb interval in *A. thaliana* as against 2–20 Mb for mapping in biparental populations (Kover et al. 2009).

7.11.6 Reverse QTL Mapping

The above strategies aim to keep the relevant genomic regions in heterozygous state for extended periods of time so that crossing over may occur very close to the QTL. In addition to the long periods required, some of the approaches like MAGIC require considerable effort and investment; therefore, they can be developed and maintained only as community resources. In contrast, the *reverse QTL mapping (RQM)* method uses a two-step screening of a very large, e.g., of 2,000 plants, segregating population like F_2 to achieve the same end. First of all, a random sample of ~200 plants is used for QTL analysis to identify major effect QTLs for the target trait. Then markers flanking these

QTLs are used to screen the entire population, and plants recombinant at a given QTL locus, and homozygous nonrecombinant at the other QTL loci, are identified. F_3 or clonal progeny of these plants are phenotyped and genotyped with sufficient number of additional markers located in the concerned QTL region. These data are analyzed to map the QTL more precisely often within a sub-centimorgan interval. Similarly, the other major QTLs for the trait can be fine mapped. RQM was used to map a QTL involved in the control of erucic acid content in oilseed rape within an interval of <1 cM (see, Peleman et al. 2005).

7.11.7 Combination of QTL Mapping and Transcriptome Profiling

The schemes described so far combine increased opportunity for recombination with linkage mapping to map QTLs often within <1 cM intervals. But QTL mapping can be combined with transcriptome profiling to enable identification of even the candidate genes for the target trait. For example, a set of 161 rice RILs was developed from the cross Pusa 1266 (high grain number) × Pusa Basmati 1 (low grain number). These RILs were genotyped with 166 SSR markers that were almost evenly distributed over the entire genome; this data was used to construct a framework linkage map for the cross. QTL mapping based on 3 years of phenotype data identified one consistent major effect QTL, *qGN4-1*, for grain number located in a 6.16 Mb region of the long arm of chromosome 4. Then six more markers were used to analyze this genomic region, and the QTL interval was reduced to 11.1 cM or 0.78 Mb. This region contains at least 117 expressed genes. Microarray-based transcriptome profiling of the two parents using tissues in early panicle development stage revealed differential expression of eight genes located in this genomic region. These genes are strong candidate genes for the QTL *qGN4-1* governing grain number (Deshmukh et al. 2010). A similar approach was used to identify 30 differentially expressed genes related to the QTLs associated with salt tolerance in rice. In this case, transcriptome profiling of the parents was not helpful. But transcriptome profiling of RNA pools from 10 extremely salt-tolerant and 10 extremely salt-sensitive RILs subjected to salt stress identified 30 differentially expressed genes; two of these genes were located in the QTL interval for salt tolerance (Pandit et al. 2010).

7.12 QTL Meta-Analysis

A *meta-analysis* attempts to combine results from many different studies concerning a single research issue with a view to identify common patterns, sources of disagreements, and any other relationships among the findings of these studies. Meta-analysis has been used mainly in the fields of medical, social, and behavioral sciences. The first application of meta-analysis to QTL mapping was related to QTLs for yield located on chromosome 3 of maize (Goffinet and Gerber 2000). When we consider several QTL mapping experiments, they are likely to be based on mapping populations having different parents, which may be segregating for different QTLs affecting the target trait. As a result, different studies may identify different QTLs for the same trait. In addition, since only a limited number of recombinations can occur between a marker and a QTL in most mapping populations, the position of a QTL determined in an experiment is only an approximation of the "actual" position of the QTL. Therefore, the positions of a single QTL detected in different studies may differ from each other due to the sampling of different recombination events and other experimental factors like sample size, accuracy of phenotyping and/or genotyping, etc. Further, the average confidence interval reported for a QTL position in most QTL mapping studies is ~10 cM or more, which would include several hundreds of genes. Finally, the QTL detection methods tend to detect QTLs with large effects more often than those with small effects. As a result, the number of QTLs affecting a trait is usually

underestimated, while the effect size of the detected QTLs is overestimated.

QTL meta-analysis attempts to integrate the findings from different QTL studies to determine the "actual" number of QTLs affecting a trait, estimate their "actual" positions in the genome, and reduce their confidence intervals. The meta-analysis generally uses information on linkage maps, the QTL positions, and the confidence intervals of the QTLs as reported in the different studies. In case confidence interval is not reported in a study, it can be approximated by the following formula: $CI(95) = 530/(N\lambda)$, where CI(95) is the confidence interval at 95 % probability level, N is the size of mapping population, and λ is the proportion of phenotypic variance accounted for by the concerned QTL. QTL meta-analysis assumes that (1) the different QTL mapping studies are independent, i.e., the individuals of the different mapping populations have been derived independently, (2) the number of QTLs controlling a quantitative trait is finite, and (3) these QTLs co-segregate in the different mapping populations. Further, it is assumed that (4) the QTLs detected in a single study are independent of each other; this assumption would fail unless QTL mapping was done by a procedure like multiple QTL model of CIM.

The first step in QTL meta-analysis is to carry out a library search for different studies on QTL mapping for the target trait of the species (Fig. 7.6). A consensus linkage map is then constructed (Sect. 6.9), and the QTLs detected in different studies are projected onto the consensus map. This projection uses a simple scaling rule based on the relationship between the original map distance separating two markers flanking a given QTL and the distance between the corresponding markers in the consensus map. The QTL confidence interval in the consensus map is estimated on the basis of the average correspondence between the lengths of the original and the consensus linkage groups. Finally, the different QTL positions are subjected to clustering using a suitable approach to distribute them into distinct clusters representing the "true" QTLs or meta-QTLs detected by meta-analysis. The results from QTL meta-analysis can be visualized graphically for their quick and easy appreciation. QTL meta-analysis reduces the large number of QTL positions detected in different mapping studies into a relatively small number of meta-QTLs. For example, 18 different mapping studies detected a total of 34 different QTLs on chromosome 8 of maize for three flowering related traits. Meta-analysis of these

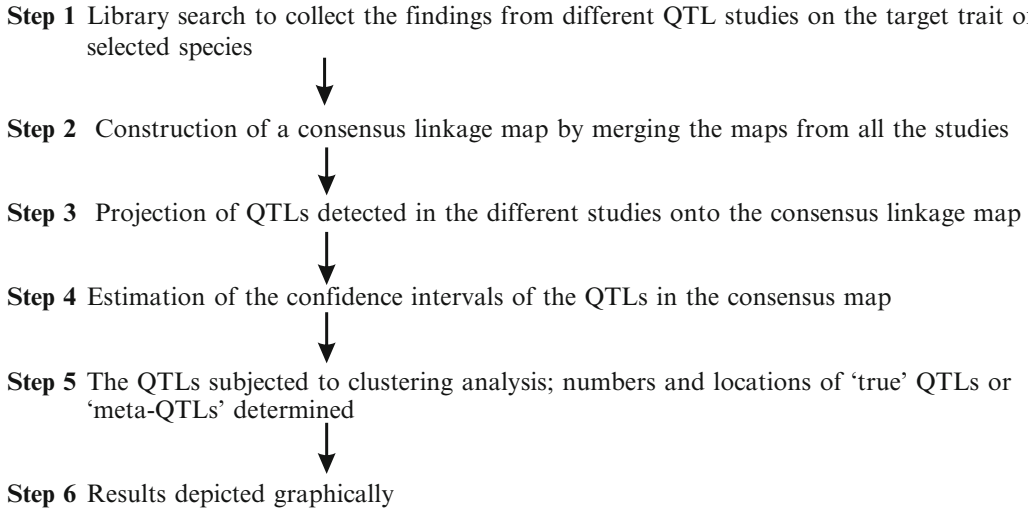


Fig. 7.6 A simplified representation of the steps involved in QTL meta-analysis

studies identified only five meta-QTLs. QTL meta-analysis permits positioning of the meta-QTLs onto a consensus linkage map that also depicts the locations of various genes mapped in the species. In addition, the confidence interval of the meta-QTLs is generally much shorter than those of the originally detected QTLs. These two features would greatly facilitate positional cloning of candidate genes involved in the control of the concerned traits.

The methods for whole-genome QTL meta-analysis have been developed and implemented as MetaQTL and BioMercator ver. 3 software packages. BioMercator was the first software package developed for QTL meta-analysis and has been extensively used. BioMercator ver. 3 is a much improved version of this package; it achieves linkage map compilation in a single step, imposes no limit on the number of meta-QTLs per chromosome, and supports high-density maps with no limitation on the number of loci (Sosnowski et al. 2012). MetaQTL is a free package for whole-genome QTL meta-analysis (http://bioinformatics.org/mqtl). It uses weighted least-squares procedure for constructing a consensus linkage map, uses a Gaussian mixture model for clustering the QTLs from different studies into meta-QTLs, and offers graphical visualization of the results (Veyrieras et al. 2007).

7.13 Inconsistent Estimates of QTL Effects

The estimates of QTL effects for a single trait of a given species vary from one study to the other due to one or more of the following reasons: (1) segregation of different QTLs in different mapping populations, (2) QTL × genetic background interactions, (3) QTL × environment interaction (QTL × E interaction), and (4) the Beavis effect.

7.13.1 Segregation of Different QTLs in Different Populations

Ordinarily, QTL analyses are based on biparental mapping populations, and the parents of different populations differ in their pedigree and selection history. As a result, different mapping populations would segregate for only some of the QTLs affecting a trait, which are likely to differ from one population to the other. Therefore, the QTL effects estimated from the different mapping populations may not be consistent. The above would be particularly true for traits like yield that are governed by several QTLs each with small effect (Bernardo 2008).

7.13.2 QTL × Genetic Background Interaction

When a trait is governed by one or few major effect QTLs, a unique QTL would be detected in a germplasm line and in all such lines that are derived from this line. But this unique QTL may show variable expression in different genetic backgrounds, i.e., QTL × genetic background interaction. For example, a unique major effect QTL, *Fhb1*, responsible for resistance to *Fusarium* head blight was identified in Sumai-3 line of wheat. This QTL was transferred into 13 different genetic backgrounds: in 12 cases, it had positive effect on *Fusarium* head blight resistance, but in one case it produced negative effect (Pumphrey et al. 2007). The QTL × genetic background interaction is a general form of epistatic interaction and can be detected by the analysis of interconnected mapping populations (Sect. 5.16). *Epistasis* describes an interaction between two or more different genes so that the expression of a gene is modified due to the influence of other interacting genes. Epistasis would be observed when the interacting genes are involved in the same metabolic pathway or a gene participates in the regulation of expression of other genes governing the target trait. According to one view, epistasis is present in breeding populations, but its contribution to the total genetic variance is relatively small. In most cases, the epistatic QTL effect size is much smaller than the effect size of the main effect QTLs. But some epistatic QTLs may have effect size comparable to that of the main effect QTLs. Further, there is evidence that epistasis would be much more important for traits governed by

several QTLs with small effects than for those governed by few large effect QTLs (Bernardo 2008).

7.13.3 QTL × Environment Interaction

When a single mapping population is evaluated in several environments, effect size estimates for the same QTL may vary from one environment to the other, and some of the QTLs may not even be detected in some of the environments; this is called *QTL × environment (QTL × E) interaction*. For example, in a large ($N = 344$) mapping population of maize, significant QTL × E interaction was observed for about one-third of the 107 QTLs detected for plant height, grain yield, and three yield-related traits (Melchinger et al. 1998). However, the failure to detect a QTL in some of the environments may not necessarily be due to QTL × E interaction, but it may be the result of an unusually high error variance in the concerned environments. In any case, both significant QTL× E interaction and high error variance in some environments reduce the transferability of results from QTL analyses across environments (Bernardo 2008). In addition, QTL× E interaction reduces heritability and the effectiveness of selection for the trait. However, the estimation of QTL × E interaction requires phenotypic evaluation in replicated trials conducted under different environments, which is an expensive and demanding task.

7.13.4 The Beavis Effect

The number of QTLs detected and the estimates of QTL effect sizes are markedly affected by the size of mapping population. Beavis (1994) used both simulation analysis and empirical data from 400 maize F_3 families from the cross B73 × Mo17 to demonstrate the following: *the smaller is the size of mapping population, the smaller is the number of detected QTLs for a trait and the larger are the estimates of their effects*; this phenomenon is known as *Beavis effect*. When data from the 400 F_3 families were analyzed, four QTLs were detected for plant height, and their R^2 estimates ranged from 3 to 8 %. But when data from four sets of 100 random F_3 families were analyzed separately, only 1–3 QTLs were detected in each set, and the R^2 estimates ranged from 8 to 23 %. Similar results have been reported from subsequent studies as well (see, Bernardo 2008). Simulation studies by Beavis (1994, 1998) showed that the QTL effects are greatly, slightly, and negligibly overestimated with population sizes of 100, 500, and 1,000, respectively.

The *inconsistent estimates of QTL effects, particularly for traits controlled by many QTLs with small effects, limit the transferability of QTL findings across populations and environments*. Therefore, (1) QTL effects should be estimated for each population to take care of the problems due to the segregation of different QTLs in different mapping populations and QTL × genetic background interactions. (2) These estimations should preferably be done in the target environment to avoid QTL × E interaction. In addition, (3) the mapping population used for QTL analysis should be sufficiently large ($N = 500$ or more) to minimize the Beavis effect. However, these requirements impose considerable burden on plant breeding programs, and they may become prohibitive for many programs (Bernardo 2008).

7.14 QTL Detection Power and Precision of QTL Mapping

QTL analysis consists of QTL detection, QTL mapping, and QTL fine mapping. Although QTL detection and mapping are done simultaneously, they are distinct activities both in logical and statistical terms. The probability of detecting a QTL with a given effect size and the stated level of Type I error is known as *power of QTL detection*. In general, the QTL detection power of an experiment at the given level of Type I error depends on the effect size or the strength of QTL and the mapping population size. The *strength of a QTL* denotes the proportion of total phenotypic variance for the target

trait explained by the QTL. A QTL that explains 20 % of the phenotypic variance is considered as strong QTL, and its detection is almost as easy as that of oligogenes. On the other extreme are weak QTLs that account for merely ~1 % of the phenotypic variance; the detection of such QTLs will require a mapping population of over 1,000 individuals/lines. The biological relevance of minor effect QTLs depends primarily on the LOD score threshold chosen for the study. In addition, when a QTL has two alleles, the power of detection is proportional to $p(1\text{-}p)$, where p is the frequency of the less frequent allele. Obviously, the QTL detection power would be the greatest when $p = 0.5$. In case of biparental mapping populations, $p = 0.5$. Several factors affect the power of QTL detection, e.g., the number of QTLs controlling the trait, the presence of epistasis, trait heritability, the type of mapping population and its size, marker density in the linkage map, the method used for QTL analysis, and the LOD score threshold used for QTL detection (Liu 1998). In fact, insufficient marker density severely limits both the power as well as precision of QTL mapping.

Another aspect of QTL analysis relates to the precision of mapping. In simple terms, the *precision of QTL mapping* is inversely proportional to the size of the confidence or support interval that defines the genomic position of the QTL and to the standard error for the genetic effects of QTL alleles. Thus, precision denotes the dispersion of the repeated independent estimates of the genomic location of the QTL or those of the effect sizes of the QTL alleles. The precision of a QTL mapping study depends on mapping population size, the density of the molecular markers employed, and the genetic variation for the target trait present in the mapping population. Generally, the size of confidence interval is inversely proportional to mapping population size and to the square of the QTL effect size. Another term, *accuracy of mapping*, reflects the closeness of the estimates of QTL location and the size of QTL effect obtained from a study to their "true" values. However, it is impractical to estimate the accuracy of mapping since the "true" genomic position and the "true" effect size of any QTL remain unknown.

7.15 Factors Affecting Results from QTL Mapping

The findings from QTL mapping studies are affected by several factors (Table 7.4), the most important of which are (1) genetic properties of the QTLs, (2) the genetic background, (3) the size of the mapping population, (4) environmental effects, and (5) experimental error.

7.15.1 Genetic Properties of QTLs

The genetic properties of QTLs include QTL effect size, the presence of and the strength of linkage with one or more other QTLs affecting the same trait, the interaction with other QTLs or the genetic background, and the sensitivity of QTL expression to environmental effects. QTLs with large effects (accounting for 10 % or more of phenotypic variance) are far more likely to be detected than those having smaller effect size. When two QTLs are closely linked (located at $<$20 cM), they will tend to be identified as a single QTL unless the size of the mapping population is larger than 500 individuals/lines. QTLs sensitive to environmental effects may be detected in some environments, but not in others, particularly when their effect size is relatively small.

7.15.2 Genetic Background

The genetic background in which a QTL is placed has considerable impact on the results of QTL mapping. This will become an issue, especially when findings from different mapping populations are compared. QTL mapping allows the analysis of QTL $\times$ QTL interactions, but this requires the same QTL to be placed in more than one genetic background, which is both costly and time-consuming.

Table 7.4 Some common factors that affect the results from QTL analysis

Factor	Effect on QTL analysis
Type of mapping population:	
(1) F_2	QTLs having additive effects detected and their degree of dominance estimated; QTL × QTL interaction cannot be estimated
(2) Backcross	Negative alleles from the recurrent parent not detected; biased estimate of gene effects when dominance is present; QTL × QTL interaction cannot be estimated
(3) RIL, DH	Estimation of QTL position more accurate; only additive effects detected; QTL × QTL interaction can be estimated
Size of mapping population	The larger is the population size, the greater is the chance of detecting QTLs with smaller effects, and more precise is the QTL position. Normally, the population size should be 200 or more, but it should not be less than 50
Marker density	*A QTL is best detected by a marker that occupies the same site as the QTL*. As the distance between the marker and the QTL increases, the chances of QTL detection and the magnitudes of QTL effects decrease. Therefore, markers should be spaced preferably at <10 cM
Magnitude of QTL effect	QTLs with large effects are almost always detected; smaller effect QTLs are difficult to detect unless large populations are used. For example, a population of 1,700 may allow detection of QTLs contributing merely 1 % to the phenotypic variance
Linkage between QTLs	Closely linked QTLs, i.e., QTLs located within a 20 cM region, are usually detected as a single QTL with a population size of <500
Experimental error:	
(1) Marker genotyping	Genotyping errors and missing genotype data can affect the marker order in the linkage map and the distances between marker pairs
(2) Trait phenotyping	*A reliable trait phenotyping is of utmost importance for a reliable QTL analysis*. Reliability of phenotyping is increased by replication and by trials conducted over locations and years/seasons
Trait heritability	Lower trait heritabilities reduce the chances of QTL detection and increase the error in QTL location.
	Increasing the number of replications in the trial used for phenotyping, conducting the trials across locations and years/seasons, and removing the residual variation due to other QTLs minimize these problems
Method of QTL mapping:	
(1) Single-marker analysis	Probability of QTL detection declines as the distance between a QTL and the marker increases; two or more QTLs linked to a single marker are detected as a single QTL; QTL location cannot be determined
(2) SIM[a]	More powerful than single-marker analysis, cannot separate two or more QTLs present in the same marker interval, QTL location is not accurate
(3) CIM	Minimizes the background effects of other linked QTLs
(4) MIM	Maps multiple QTLs and detects QTL × QTL interaction (the results from both CIM and MIM are highly dependent on the genetic model used for QTL analysis, as well as marker cofactor selection for CIM)

[a]*SIM* single interval mapping, *CIM* composite interval mapping, *MIM* multiple interval mapping

7.15.3 Type and Size of Mapping Population

The type of mapping population would determine the types of gene effects detected by QTL analysis. For example, the analysis of RIL and DH populations would permit the estimation of additive and additive × additive effects only. On the other hand, the additive and dominance effects are totally confounded in a backcross population. Further, mapping population size is perhaps the most important factor of a QTL analysis experimental design. In general, the larger is the population size, the greater would be the probability of detecting smaller effect size QTLs. An increase in population size enhances the QTL detection power, the precision of QTL location, and the precision of QTL effect size estimation.

7.15.4 Environmental Effects on QTL Expression

Phenotypic expression of quantitative traits is markedly affected by the environmental factors, and the extent of influence primarily depends on the trait concerned. In general, QTLs with relatively large effects are more stable over environments than those with small (<10 % of the phenotypic variance) effects. Therefore, QTL analysis results for traits having high heritability are much more reliable than those for traits with low heritability. It is, therefore, desirable to conduct phenotypic evaluation over environments, usually represented by locations (including field and glasshouse/greenhouse) and years, to be able to assess the stability of QTL expression, and to identify QTLs whose expression relatively stable across environments.

7.15.5 Experimental Error

Phenotypic evaluation of the mapping population and its marker genotyping are the two chief sources of experimental error. In view of the environmental effects on quantitative trait expression, measurements on individual plants are not very reliable. Further, phenotypic evaluation should be based on carefully conducted replicated trials. Errors in genotyping and missing marker data would affect the order of markers in the linkage map and the distances between the mapped markers; this, in turn, may affect the estimated QTL locations.

7.16 Advantages of QTL Linkage Mapping

1. Linkage mapping detects and maps each of the QTLs governing the target trait within relatively short confidence intervals.
2. QTL mapping identifies markers flanking the QTL regions; these markers can be used for MAS, including recombinant selection, for the concerned QTL.
3. It provides an estimate of the QTL effect size on the trait phenotype. Thus, breeders get a rough idea of the usefulness of incorporating a given QTL in their breeding programs.
4. Joint QTL analysis of multiple correlated traits can distinguish between close linkage and pleiotropy as the basis of the trait correlations. This would indicate whether negative trait correlations may be broken or not in breeding programs.
5. High-resolution QTL mapping can locate a QTL in a very small (<1 cM) confidence interval, which greatly facilitates cloning of the genes located in the QTL region.
6. Selective DNA pooling can be combined with transcriptome analysis to identify a limited number of candidate genes located in the genomic region harboring the QTL for the target trait.
7. Appropriate experimental designs and QTL analysis methods are available for the detection and estimation of QTL × QTL and QTL × environment interactions.
8. QTL analysis based on biparental populations presents some unique advantages over association mapping (Chap. 8). For example, association mapping cannot identify and map rare functional alleles of genes/QTLs, but this can be easily achieved by linkage mapping. This is because the rare allele will be present in one of the two lines crossed to generate the concerned mapping population. This will ensure the frequency of rare allele to be 50 % in the mapping population, which will facilitate its mapping by increasing QTL detection power.

7.17 Limitations of QTL Mapping

1. Since the mapping population is initiated by crossing two parents selected for the purpose, genetic variation in the quantitative traits of the population is limited to the differences between the two parents.
2. The effects of only two alleles of the genes/QTLs can be studied in most mapping

population. In real situation, many, if not all, genes/QTLs may have more than two alleles each. However, multiple alleles of genes/QTLs can be analyzed in interconnected populations like MAGIC and NAM.

3. QTL mapping has low-resolution power because only few meiotic divisions occur during the period between the hybridization of the parents and the use of the resulting populations for mapping. As a result, a QTL position may span from few to tens of centimorgans (typically, 5–20 cM). This region often corresponds to several megabases (on an average, 1.2–4.8 Mb), which may typically contain hundreds of genes.
4. In view of the above, often high-resolution mapping has to be undertaken to map QTLs within sub-centimorgan intervals or even identify candidate genes (Sect. 7.11.3). This requires additional effort and adds to the cost.
5. Even when a QTL with large effect is identified, it is very difficult to identify the gene responsible for this QTL effect. In any case, QTL mapping does not indicate the number, the nature, and the function of the genes present in the detected QTLs.
6. In fact, a major effect QTL may often consist of many closely linked QTLs with small, sometimes even opposite, effects on the target trait.
7. The creation of biparental mapping populations requires time and effort, and in some species like tree species, this may not be feasible.
8. A QTL detected in a biparental population may not be equally effective in other genetic backgrounds. This necessitates validation of the detected QTLs in unrelated germplasm.
9. Similarly, the markers linked to QTLs/genes identified in a biparental population need to be tested for their applicability to other unrelated genotypes/populations. This involves additional investment of resources, including time and effort.
10. Often, by the time a new QTL is discovered, it may already have been transferred into the breeding populations using conventional breeding approaches. This limits the usefulness of MAS for this QTL in breeding programs.
11. Different QTLs would be detected for the same trait when different populations are used for mapping. Further, interactions between the QTLs detected in different populations cannot be studied.
12. Only large effect QTLs located close to a marker locus will be reliably detected. Further, it may be difficult to even detect QTLs with strong epistatic effects and QTLs sensitive to environmental influences unless suitable experimental and analytical designs are used.

7.18 Nature and Function of Polygenes

Molecular markers have greatly facilitated the identification of QTLs that harbor polygenes, i.e., the genes that were proposed to explain the inheritance of quantitative traits. Results from various QTL mapping studies reveal that a small number of QTLs produce relatively large effects, while most of the QTLs have small effects. A QTL generally represents a large genomic region that may contain several, even hundreds, of different genes. Thus, QTL mapping studies have clearly established the physical locations of polygenes, but they have not been able to reveal the nature and function of polygenes. Fine mapping has facilitated positional cloning of several QTLs in plants (Table 12.4). In addition, QTLs from several other organisms, including humans, have also been cloned. In only few cases, the genomic region identified to contain a QTL had a single gene, and a vast majority of QTLs had more than one gene; in few cases, up to 38 genes were identified in one QTL (Salvi and Tuberosa 2005, 2007). In addition, QTLs are known to occur in clusters in plants, and in some cases one of the QTLs in the cluster may exert the major influence on the trait phenotype. For example, extensive association studies for flowering time in maize have revealed that many variants clustered in a few common loci affect this trait (Buckler et al. 2009). This kind of

genetic architecture has been described as the common gene hypothesis, and it would lead to a high heritability for the trait, but a low association between QTLs and markers. Several of the cloned plant QTLs encode transcription factors, some others code for known enzyme activities like invertase, while others encode other kinds of proteins, including various response regulators.

7.19 Software for QTL Mapping

Some methods of QTL analysis like SMA and regression interval mapping can be performed using standard statistical software. But other QTL analysis methods require special software packages for their implementation. Generally, the available software packages implement SMA, interval mapping, regression interval mapping, and CIM. But software for Bayesian interval mapping, MIM, and multiple trait analysis are also available. Most of these software packages are listed at http://www.linkage.rockefeller.edu.soft. Most of these packages would give similar, if not the same, results for the same datasets, but they differ with respect to the required data format, computer platform used, user interface, graphic output, etc.

7.19.1 MapMaker/QTL

MapMaker/QTL is one of those QTL mapping software that were the first to become available to the scientific community (Lincoln et al. 1993). It implements interval mapping and nonparametric mapping methods for non-normal phenotype data. It is a companion program of MapMaker and uses the linkage maps prepared by MapMaker for QTL analysis. It operates on most computer platforms, has command-driven user interface, and lacks a graphic user interface. But one can save the output graphs as postscript files. It can be downloaded free from http://hpcio.cit.nih.gov/lserver/MAPMAKER_QTL.html.

7.19.2 PLABQTL

PLABQTL (Utz and Melchinger 1996) is the most suited for the analysis of topcross progenies, but can be used for the analysis of data from F_2 and later segregating generations (including RILs), backcross, and DH populations. Additive and dominance gene effects can be fitted in its genetic model. It performs SIM and CIM using a fast multiple regression procedure. Cofactors for CIM can be selected by the user with the help of stepwise regression and Akaike's information criterion. PLABQTL can handle missing marker or phenotype data, detect outliers in these data, rapidly calculate and compare LOD curves for different model assumptions, and analyze QTL $\times$ E interactions. All input and output files are in ASCII format and largely compatible with other programs.

7.19.3 QTL Cartographer

The original version of QTL Cartographer was command driven and not very user-friendly. But its Windows-compatible version, *WinQTLCart* (Wang et al. 2005, 2012), is very user-friendly and provides a powerful graphic interface. It is compatible with Windows 2000, XP, or Windows 7, and its latest version is 2.5_011 (Wang et al. 2012). It imports and exports data in a variety of formats, calculates empirical threshold LOD scores by permutation, and estimates confidence intervals for QTL positions by the bootstrap method. It implements single-marker analysis, SIM, CIM, MIM with epistasis, Bayesian interval mapping, multiple trait analysis, and multiple trait MIM analysis and maps categorical traits as well. However, it lacks program for marker linkage map construction. Therefore, it imports linkage map results from MapMaker and uses it for QTL analysis. It is available free at http://statgen.ncsu.edu/qtlcart/WQTLCart.htm.

7.19.4 MapManager QT/QTX

MapManager QT/QTX is user-friendly, has a fine graphic interface, and is available at http://mapmanager.org/mmQTX.html (Manly and Olson 1999; Manly et al. 2001). These programs detect and map QTLs by fast regression-based SMA, SIM, CIM, and interactive QTL search. Both QT and QTX programs compute empirical threshold LOD scores by permutation. They also generate the QTL confidence intervals by the bootstrap method (MapManager QT) or the quick Piepho method (MapManager QTX). They support the analysis of data from advanced backcross, advanced intercross, and recombinant inbred intercross (RIX; F_1s from a diallel mating among a set of RILs; a form of immortalized F_2, Sect. 5.10) populations. Marker linkage maps for the programs are constructed by a companion program (Manly et al. 2001). MapManager QTX creates data files that are compatible with MS Windows as well as Mac OS versions of this program. MapManager QTX is not being further developed.

7.19.5 R/QTL

R/QTL has been designed as an add-on program to the statistical language/software R, which is freely available from http://www.r-project.org/. It is a powerful and open source that can operate on several platforms. The R/QTL software is available for free at http://www.rqtl.org/ (Broman et al. 2003). Its updated version (Arends et al. 2010) performs SMA, SIM, regression interval mapping, CIM, and MQM. The R/QTL MQM procedure has a higher statistical power than many other methods for the detection and separation of the effects of multiple linked and unlinked QTLs. R/QTL is capable of superior handling of missing data and determining the significance thresholds for the detection of QTLs and QTL hot spots. Further, *cis–trans* and QTL interaction effects can be visualized with the help of this program. R/QTL can be scaled up for the analysis of large genetic genomics datasets and used for automated procedures. It calculates empirical threshold LOD score by permutation and estimates confidence intervals for QTL positions by bootstrap. It can implement non-parametric mapping methods for non-normal phenotype data and has a companion program for constructing marker linkage maps. R/QTL emphasizes model improvement, and its application requires some basic R programming skill. J/QTL is a Java graphical user interface (GUI) for R/QTL to help users deficient in R programming skill.

7.19.6 R/QTLBIM

R/QTLBIM carries out Bayesian interval mapping (Yandell et al. 2007) and is available at http://qtlbim.org/. This program allows Bayesian model selection for the mapping of multiple interacting QTLs and can handle both continuous and binary or ordinal traits. It uses data from experimentally inbred lines, allows epistasis and interacting covariates, and performs a genome-wide search for potential QTLs. R/QTLBIM is built on R/QTL (Broman et al. 2003) and requires R/QTL version 1.03 or a later version for support; R/QTLBIM 1.7.7 is the latest stable version of this package.

7.19.7 QTL Express

QTL Express is the first software package with web-based user interface for QTL analysis in outbred populations. It is suitable for half-sib outbred populations and F_2 populations from crosses between inbred and outbred parents (Seaton et al. 2002). It is user-friendly and performs single or multiple QTL mapping by regression approach. It requires marker linkage map, trait phenotype, and marker genotype as input files. The analysis is performed in two steps, viz., estimation of IBD probabilities for specific chromosomal positions using multiple marker data and fitting a statistical model to the observations and the IBD coefficients. A general linear model is fitted to the phenotype data since this model allows inclusion of additional fixed

effects and covariates that explain the phenotypic variation in the trait. Either one or two QTLs are fitted in the linear genetic model, while additional (known) QTLs can be fitted as covariates. In case of populations derived from crosses, QTL effects can be specified as additive and dominance effects with the option for QTL × QTL interaction as well as a fixed effect, viz., sex or population. For outbred line crosses, the QTL model may have a parent of origin, i.e., imprinting effect as well. In case of F_2 populations, the founder parents are treated as fixed for the alternative alleles of the QTLs. This package computes genome-wide or chromosome-wide threshold LOD scores by permutation and estimates confidence intervals for QTL positions by bootstrap.

7.19.8 FlexQTL

FlexQTL (www.flexqtl.nl) is based on Bayesian theory and is implemented via Markov Chain Monte Carlo simulation. It can be used for QTL mapping in any pedigreed population, including F_2 and backcross populations. It uses a marker linkage map for mapping of the QTLs. It can also estimate probabilities of genes being identical by descent provided the pedigree and marker data are known. FlexQTL model allows inclusion of nongenetic variables like treatments, years, locations, etc., but it is unable to directly estimate G × E interactions. It can handle missing phenotype and/or marker genotype data and can analyze multiple traits simultaneously to reveal pleiotropic behavior of QTLs and correlations of polygenic and residual components of the quantitative traits. It treats each QTL as biallelic and can estimate additive, dominance, and imprinting effects for each QTL. The Bayesian approach provides clear visual presentations of the statistical aspects of all parameters relevant to QTL analysis (Bink et al. 2008, 2014).

7.19.9 INTERQTL

INTERQTL implements Bayesian QTL mapping (Jannink and Wu 2003) in multiple interconnected populations (Sect. 5.16). It can be used with backcross (BC_1 and BC_2), F_2, DH, and RIL populations and can map multiple QTLs on the same or multiple chromosomes. This package consists of Bayesian analysis, simulation, and interface modules. The first two modules can be used independently in either DOS or Windows, while the third module organizes and runs the first two modules in Windows desktops (Windows 3.X, 98, NT). It permits the users to prepare input files and to tailor a specific analysis to their needs and visualize posteriors. The INTERQTL genetic model includes only additive effects of QTLs. It allows users to model either random or fixed QTL effects and to preset allele number and configuration based on prior information, but the same can also be inferred by analysis. INTERQTL accommodates missing marker data by multi-marker calculation of conditional QTL genotype probabilities.

7.19.10 MCQTL

MCQTL operates in the UNIX environment. It is designed for the mapping of QTLs having multiple alleles in multi-cross designs as well as the biparental populations. In case of multiple related families, it may treat the within family QTL effects to be fixed or allow diallel modeling of the QTL effects. It uses only additive genetic model, is based on linear regression method, and scans the whole genome by CIM. It allows an interactive QTL mapping to allow the handling of multiple QTL models. The markers used as cofactors in the model are selected by forward (whole-genome scan) or backward (chromosome by chromosome scan) stepwise methods. The LOD threshold is estimated by permutation. The MCQTL procedure is flexible and robust, and a family can be added or dropped without the need for recomputation of QTL genotype probabilities. The output is in the form of XML-formatted files and graphic files. MCQTL is available for free to academic institutions and nonprofit organizations at http://www.genoplante.com (bioinformatics products; Jourjon et al. 2005).

7.19.11 QGene

QGene 4.0 package is written in Java, has a rich GUI, and can operate on any computer that supports Java (Joehanes and Nelson 2008). It implements several QTL mapping methods, including SIM, CIM, MIM, and multitrait methods, many of which are not available elsewhere. It can display superimposed QTL profiles for any number of chromosomes, many methods of analysis, and a large number of traits. It can generate simulated genotype data and maps for all mating designs. It can also perform these operations for multiple correlated traits. It carries out segregation test, assesses normal distribution of traits, estimates correlation, and transforms data where required. But it cannot estimate QTL × QTL interaction, is able to handle only genetic covariates, and cannot subject data from trials conducted in multiple environments to mixed model analysis. In addition, it cannot perform nonparametric QTL mapping. QGene is amenable to third-party addition of new features. It can handle data from several types of populations, including those developed by unorthodox mating designs. But it cannot use data from outcross, half-sib, and multi-cross populations. It is available at http://coding.plantpath.ksu.edu/qgene, and the source code can be obtained on request.

7.19.12 Some Other Software Programs

MQTL runs on DOS or Sun OS and implements a simplified version of CIM using large datasets from multiple environments. It estimates environmental effects as well as QTL × E interaction. *Multimapper* operates in the UNIX environment and works as a companion program of QTL Cartographer. It builds and implements Bayesian multi-QTL models automatically and generates plots of QTL probabilities at different positions. It is best suited for mapping QTLs within a linkage group that has been indicated by some other program to contain multiple QTLs. *Epistat* is a DOS-based interactive program designed primarily for the detection and analysis of epistatic interactions between QTLs. But it does not carry out interval mapping. *QTL Cafe* is a program written in Java and runs in a Java-enabled World Wide Web browser. The program *IciMapping* is a very user-friendly integrated software that prepares marker linkage maps as well as carries out QTL mapping.

Questions

1. Explain the meaning of and procedure for QTL confirmation and QTL validation, and discuss their relevance.
2. How is the LOD score threshold for QTL mapping determined? Explain the meaning and relevance of QTL support interval.
3. Different studies often identify different QTLs for the same trait in the same species. Discuss the reasons for this situation and the approach that may be used to identify true QTLs.
4. List the various software programs for QTL analysis, and briefly describe the important features of any two of these packages.
5. What are the various approaches for QTL analysis? Which of these approaches would you use for QTL analysis and why?
6. What is QTL analysis? Briefly describe the procedure for QTL linkage mapping and discuss its advantages and limitations.
7. "The results of QTL analyses are affected by a variety of factors." Comment on this statement in the light of available relevant information.
8. Briefly describe the salient features of composite interval mapping and the modifications thereof.
9. Discuss the relevance of bulked segregant analysis approach to QTL analysis.
10. Explain the meaning of quantitative trait locus, the distribution and organization of QTLs in the genome, and the various types of functions performed by them.
11. Discuss the various approaches for QTL fine mapping.

Association Mapping

8

8.1 Introduction

The mapping approaches are basically of two types, viz., family mapping and population mapping. In *family mapping*, populations constructed by crossing generally two homozygous lines (Chap. 5) are used for linkage mapping of markers and genes/QTLs (Chaps. 6 and 7). Thus, these populations comprise closely related families derived from common parents using a specific mating scheme (Myles et al. 2009). In *population mapping*, generally referred to as *association mapping* (*AM*), the mapping population consists of a diverse set of individuals/lines drawn from natural populations, e.g., random mating populations of wild species; wild relatives of crops like wheat, barley, maize, rice, etc.; as well as breeding populations. These populations can also be regarded as groups of many families of rather small (one individual per family in extreme cases) size. In addition, AM can use populations designed for family mapping. In such cases, it exploits the linkage disequilibrium (LD) resulting from hybridization between the lines used as parents of these populations as well as the historical LD present between them. AM uses LD between markers and the concerned genes/QTLs for identifying marker-trait associations. AM is also known as association analysis, LD mapping, and structured association mapping. The AM approach was originally developed by human geneticists for measuring genetic proximity of loci to each other and to map oligogenes. Subsequently, AM approach was extended to mapping of QTLs and still later to mapping in plants, including crop plants and perennial tree species. The AM approach is expected to identify markers located much closer to the genes of interest than is feasible with conventional linkage mapping. This is expected because LD analysis utilizes all the recombination events that would have occurred between the gene and the marker in the past in the population being used for AM (Fig. 8.1). In contrast, linkage mapping uses only those recombination events that occur between the gene and the marker after the two selected parents are crossed. The AM approach offers some other advantage over linkage mapping, but it suffers from some limitations as well (Table 8.1).

8.2 The General Procedure for Association Mapping

The general procedure for genome-wide association mapping in plants is briefly outlined here based on Abdurakhmonov and Abdukarimov (2008). But the exact details of the procedure will depend on the chosen study design and whether or not the population shows structure.

1. *Association mapping population*. A large random sample from a natural population, a germplasm core collection, a collection of breeding lines including cultivars, or a

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_8, © Author(s) 2015

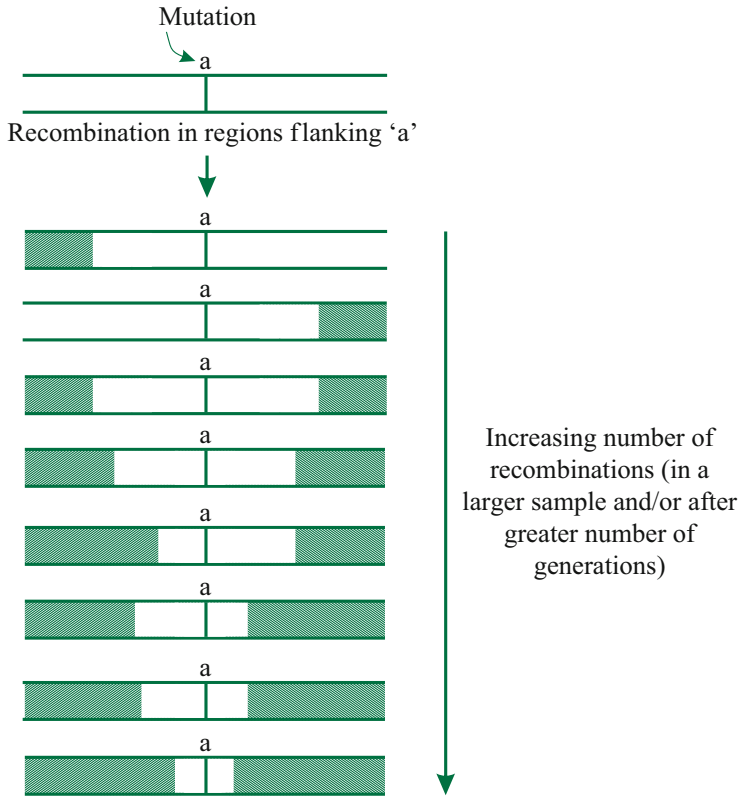


Fig. 8.1 Recombination reduces the extent of linkage disequilibrium (LD) around a locus, say, locus *a*. The frequency of recombination increases as the distance of a given locus from the mutant allele increases. Therefore, as the number of meiotic events included in a sample increases, the chances of recovering chromosomes with recombinations closer to the mutant allele also increases. The number of meioses included in a sample can be increased by increasing the sample size, but this will not be practical beyond a point. In LD-based association mapping, this is achieved by including recombinations that have occurred in the past generations in the population, from which the sample for AM study is drawn. Therefore, the older is a mutation, the smaller will be the region of high LD around the mutant allele *a* (Based on Ardlie et al. 2002)

population derived from multiparent crosses of the concerned species is used for AM. The sample should include as much genetic diversity present in the population/germplasm collection as is practically feasible. This sample constitutes the *association mapping population*, *association mapping panel*, or, simply, *association panel*.

2. *Phenotyping*. The selected sample is evaluated for the various traits of interest; this is called *phenotyping*. Phenotyping should be preferably based on replicated trials conducted over locations and years to minimize environmental effects. The trials should be conducted using a suitable experimental design like randomized block design, augmented design, nested design, etc. *A precise and reliable phenotyping is critical to any mapping effort* (Sect. 8.3).
3. *Genotyping for population structure analysis*. The sample is then *genotyped*, i.e., tested with a set of molecular markers (preferably SSR markers) that are evenly distributed over the entire genome of the species. These markers should be unlinked, i.e., should be located more than 40 cM apart in the genome (Pritchard et al. 2000a, b).
4. *Structure and kinship analysis*. The marker data are analyzed to detect and estimate the population structure of the sample using the

Table 8.1 A comparison between linkage and association mapping approaches

Feature	Linkage mapping	Association mapping
QTL effect size	Effective for moderate to large effect QTLs; ineffective for QTLs with small effect size	Effective for QTLs with much smaller effect size than in linkage mapping
Effectiveness with low allele frequencies	Effective[a]	Ineffective
Number of alleles detected per locus	Only two alleles can be detected	All the alleles present in the sample are detected
Type of information on marker alleles used for mapping	Information on identity by descent	Current approaches use information on identity by state
Need for QTL result confirmation/validation	Confirmation as well as validation required	Often confirmation is done by replication studies
Populations used for mapping	Produced by crossing selected parents	Natural populations, breeding materials, germplasm lines, lines produced from multiple crosses
Recombination events exploited	Those occurring after the crosses are made	All the recombination events that occurred since the LD was created
Identified markers linked to QTL/gene	Few to several centimorgans (cM) away from gene/QTL	Much closer than those by linkage mapping
Mapping based on	Recombination frequency between the loci	Linkage disequilibrium (LD) between the loci
Familial relatedness	Minimized by controlled crossing	Minimized by kinship coefficient estimation and its use in association mapping
Population structure	Minimized by controlled crossing	Minimized by estimation of Q or P and its use in AM
Feasibility in different species	Feasible in annual and biennial species, *not feasible in perennial species*	Feasible in annual, biennial, and perennial species
Integration of QTL discovery with breeding	Novel breeding schemes proposed for the purpose (Sect. 9.11)	Integration feasible when breeding materials are used for AM
Number of markers needed to cover the whole genome	Low (10^2) to moderate (10^3)	High (10^5 for small genomes) to very high (10^9 for large genomes)
Conclusions applicable to	The concerned populations unless validated in other materials	The concerned species or subspecies

[a]Allele frequencies increase as a result of biparental crosses

STRUCTURE program and the extent of kinship among the individuals of the sample using the TASSEL program.

5. *Genotyping for LD analysis*. The sample is also genotyped with a sufficiently large number of molecular markers that cover the entire genome as densely as is feasible (Table 8.1) so that LD between markers and the loci of interest can be detected. The pattern of LD in the concerned genomic regions of the species and the extent of LD observed among different populations of the species would determine the number of markers required for adequate coverage of the whole genome. *SSR and SNP marker systems are the most widely used for this purpose*.
6. *AM and LD analyses*. A model-based analysis of relatedness between the phenotype and the genotype data is done to detect and quantify LD between the markers and the genes/QTLs governing the traits of interest. The estimates of population structure and kinship are used as covariates in the model to minimize false associations between the markers and the genes/QTLs of interest. Since these analyses are computationally intensive, suitable computer programs are used for their implementation.

8.3 Phenotyping

Accurate phenotyping of the AM panel is a prerequisite for arriving at valid conclusions. An increase in the number of individuals/lines included for phenotyping enhances the power of AM much more than an increase in the number of markers used for genotyping (see Ingvarsson and Street 2011). The precision of phenotyping can be improved by replicated tests/trials conducted over locations and years, whenever inbred/homozygous lines or clones are used for AM. The data from different replicates of a line can be averaged to minimize the environmental effects and measurement errors, and the mean values are then used for association analysis; this is referred to as *two-stage association mapping*. Alternatively, one may resort to *one-stage association mapping*, in which data from all the replicates are directly used for association analysis (Stich et al. 2008). Available evidence suggests that the results from these two approaches may be either comparable or they may differ considerably in terms of the power of mapping. Further, conducting phenotypic evaluation over locations and years will allow estimation of genotype × environment interaction effects, which are of considerable importance for almost all quantitative traits. Further, replicated phenotyping increases the power of QTL detection (Kang et al. 2008; Stich et al. 2008; Zhu et al. 2008).

Precise phenotyping of large samples in replicated tests will require efficient field designs and appropriate statistical methods, particularly when the fields are heterogeneous. Experimental evidence for the efficiency of different field designs in tackling field heterogeneity can be obtained only through trials conducted in fields having different levels of heterogeneity, which may be a great challenge. Such studies will require a strong collaboration between geneticists and statisticians in addition to considerable resources and effort. Further, certain aspects of phenotype development and correlation among different traits need to be considered during phenotyping. For example, when diversity panels include genotypes adapted to different growing conditions, their phenotypic evaluation under uniform conditions may not generate reliable data. A *diversity panel* is a sample that includes as much genetic diversity of the parent population as is practically feasible. Therefore, care should be taken to create diversity panels having accessions with similar adaptation, photoperiod requirements, etc. Some traits like flowering time influence the expression of other correlated traits. As a result, lines differing in traits like flowering time may differ for the correlated traits as well. Finally, evaluation of traits like resistance to biotic and/or abiotic stresses in a trial will interfere with the phenotypic expression of other traits like yield in the same trial. Therefore, these and similar factors must be taken into account during phenotyping (Zhu et al. 2008). It may be mentioned that the discipline of phenomics is devoted to the development and refinement of high-throughput precision phenotyping techniques (Chap. 15).

8.4 Genome-Wide and Candidate Gene Approaches for Association Mapping

There are two general approaches for association mapping, viz., genome-wide and candidate gene approaches. In *genome-wide association studies* (*GWAS*), the markers used for genotyping are distributed, preferably evenly and densely, over the whole genome. In this approach, all the loci involved in the control of all the traits showing variation in the sample can be evaluated in one go. The number of markers used for genotyping would be much larger in cross-pollinated than in self-pollinated species because the LD decays much faster in the former than in the latter (Table 8.2). It is important that a genome-wide linkage map of markers of the concerned species must be available to permit the selection of an appropriate set of markers. In addition, considerable resources and effort will be required for reliable phenotyping of the variable traits. Finally, when a large number of markers are used, thousands of independent comparisons

Table 8.2 Some studies on the extent of LD and AM in plants

Plant species	Extent of LD	Traits mapped	Mapping approach[a]
Cross-pollinated species			
Maize	4–41 cM; 200 bp–500 kb	Endosperm color and several metric traits	GLM, SA, MLM, WGA
Sorghum	50 cM	Not available	Not available
Sugarcane	10 cM	Not available	Not available
Silage maize and ryegrass	200 bp–2 kb	Cold tolerance, flowering time, forage quality	ANOVA, multiple linear regression
Several forest trees	100 bp–2 kb	Some metric traits	ANOVA, LD, and QTL mapping
Self-pollinated species			
Arabidopsis	50–100 cM; 10–250 kb	Several quantitative traits, including flowering time	One-way ANOVA, simple regression, SA, MLM
Barley	10–50 cM; 300 bp–500 kb	Disease resistance and many metric traits	Pearson correlation, regression, ANOVA
Rice (*Oryza sativa* var. *indica*, *japonica*; *O. rufipogon*)	20–225 cM; 5–500 kb	Many metric traits (yield and quality traits)	DA, MLM, mixed model with multiple QTL effect
Bread wheat (*T. aestivum*)	<1–10 cM	Several seed quality traits and blotch resistance	GLM-Q, LMM
Potato	0.3–3 cM	Several metric traits and disease resistance	Standard two sample *t*-test, GMM
Soybean	10–50 cM	Seed protein content	WGA

Based on Abdurakhmonov and Abdukarimov (2008)
[a]*ANOVA* analysis of variance test, *DA* discriminant analysis, *GLM* general linear model without population structure, *GLM-Q* general linear model using population structure matrix *Q* or the least square solution to the fixed effects GLM, *GMM* general mixed model, *LMM* linear mixed model, *MLM* mixed linear model, *SA* structured association, *WGA* whole-genome association

among marker loci will have to be made. This would necessitate a large sample size (one thousand or more individuals) to permit the detection of QTLs with moderate effect size. F_1-derived mapping populations like RILs are highly suited for genome-wide scanning for QTLs since only a few hundred markers need to be evaluated, and they provide greater statistical power to evaluate the effect of a genomic region than AM. These difficulties can be resolved by using a population that has experienced bottleneck in the recent past, e.g., elite germplasm of maize, since bottleneck leads to a substantial increase in LD in the entire genome (Sect. 8.16.7). The increased LD would reduce the number of markers to be evaluated as well as the size of sample to be studied.

Another way around the above problems is to restrict the analysis to the genomic regions having the candidate genes/QTLs for the trait(s) of interest; this is known as *candidate gene approach*. A *candidate gene* is a gene that is expected, on the basis of previous knowledge, to be involved in the control of a trait of interest. Generally, information from several different sources, e.g., comparative genomics, genome sequence annotation, transcript profiling, QTL analysis, etc., is used to identify the candidate genes. After this, the genotyping effort is focused in the genomic regions with the candidate genes. This greatly reduces the target genomic region, which can be analyzed with a high density of markers. Further, the total number of markers used as well as the sample size will also be considerably reduced. A limitation of this approach is that the involvement of genes not included in the list of candidate genes in the development of the trait phenotype cannot be assessed. In addition, usually candidate genes are discovered from loss of function mutations in laboratory strains. Therefore, it is difficult to determine as to how well these mutations relate to the variation present in the trait in natural populations. In spite of these difficulties, the candidate gene approach has been used to

identify genes involved in the control of many traits, including morphological, phenological, and stress resistance traits (Ingvarsson and Street 2011). This approach may be able to identify a QTL where genome-wide AM fails to detect a significant marker-trait association after false discovery rate (FDR) correction is applied. In addition, the use of this approach along with GWAS tends to increase the power and precision of QTL detection (see Gupta et al. 2014).

In many species, complete genome sequences are not available, and development of RILs may not be possible. As a result, it would not be feasible to carry out genome-wide association mapping in such species, except for using whole-genome sequencing for SNP genotyping. In such cases, *gene space-based association mapping* or *gene space study* may be carried out using SNP data generated from sequences of transcribed genes. The term *gene space* refers to that fraction of the genome, which corresponds to protein coding genes. This term, in addition, describes the distribution pattern of the genes as well (Jackson et al. 2004). In the cases of eukaryotic genomes that contain large amounts of repetitive DNA, this term also relates to the notion of gene-rich regions located within gene-poor regions comprising mainly of repeat sequences. Thus, gene space would include untranslated regions and conserved noncoding sequences associated with genes. Gene space studies are expected to generate useful information, although they would not cover the relatively much larger nontranscribed regions of the genome (Ingvarsson and Street 2011). Further, in silico *association mapping* or *haplotype association mapping* uses phenotype and genotype data on inbreds and breeding lines routinely developed in the breeding programs. The trait phenotype and marker data are generally collected during regular breeding programs, but the marker data may be generated de novo if they were not available. These lines are tested against a dense genome-wide consensus marker linkage map to determine association between haplotypes and the traits of interest. This strategy saves considerable resources as it avoids the creation of AM panels and collection of genotype and phenotype data specifically for AM. However, it should be viewed as complementary to, and not a replacement of, regular AM (Zhu et al. 2008).

8.5 Populations Used for Association Mapping in Plants

The population used for AM is one of the main factors affecting the success of AM. The population may be based on a natural/breeding population or it may be a family-based population. AM can also be performed in biparental and multiparent populations, but single biparental populations are generally not used for AM. Generally, doubled haploid, F_3, etc., families derived from several biparental crosses generated by mating a group of inbreds in diallel scheme or in a random manner are used for AM. In case of multiparent populations, two populations, namely, multiparent advanced generation intercrosses (MAGIC) and nested association mapping (NAM) populations, have become very popular since they allow both AM and linkage mapping and can even be used for variety development.

8.5.1 Population-Based Association Panels

AM can be performed in all panmictic populations that harbor considerable LD in genomic regions involved in the control of the target phenotypic traits. In addition, it uses samples drawn from natural/breeding or synthetic populations that are not amenable to linkage mapping. The AM populations relevant to breeding programs are derived from germplasm collections, inbred lines/cultivars developed by breeding programs, and synthetic populations derived from a group of inbred lines. The AM panel from a germplasm collection may either be a random sample or a "core" set of germplasm accessions. The various types of populations differ for a variety of features, including the level of LD, the mapping resolution, and the power of

Table 8.3 The relevant features of various mapping populations available for association analysis in plant breeding programs

Feature	Germplasm bank	Elite breeding material	Synthetic population
Sample	Core collection accessions	Lines and cultivars developed in breeding programs	Individuals or lines drawn from the population
The composition of sample	Does not change	Changes with time as new materials are developed	Changes with time as the generation advances
Traits analyzed	Highly heritable and domestication traits	Low heritability traits like yield	Depends on the evaluation scheme
Level of LD	Low	High	Intermediate
Population structure	Medium	High	Low
Allelic diversity in the sample	High	Low	Intermediate
Resolution of AM	High	Low	Intermediate; increases with generation
Power of association analysis	Low	High	Intermediate; decreases with generation
The use of markers associated with the target traits	Marker-aided selection (MAS)	MAS	Incorporated in selection index

Based on Breseghello and Sorrels (2006)

QTL detection (Table 8.3; Breseghello and Sorrels 2006). In addition, more QTLs would be detected in populations based on exotic germplasm, but these QTLs will usually be relevant for introgression from exotic into the elite germplasm. However, they would not be useful for marker-assisted selection (MAS) in breeding programs since these programs are ordinarily based on elite germplasm. But the use of elite germplasm for AM would identify superior QTL alleles present in superior lines that are used in breeding programs; therefore, these QTLs would be directly useful for MAS (Wurschum 2012).

Inbreds can be maintained perpetually, evaluated in replicated trials, and shared among researchers for repeated and varied investigations. A panel of diverse inbred lines can be carefully created to represent the maximum possible diversity of the species. For example, a panel of 300 diverse maize inbred lines has been developed. Similarly, a panel of 377 sorghum inbred lines representing all major cultivated races, i.e., tropical lines from diverse geographical and climatic regions, and important breeding lines developed in the USA and their progenitors has been created. In case of barley, an assemblage of 3,840 lines that includes progenies derived by pedigree programs and diverse germplasm lines is being used by Barley CAP initiative. In addition to the existing soft winter wheat panel, four regional association panels are being developed to represent both winter and spring wheat types and grain hardness. An association panel comprising diverse germplasm lines may identify a new QTL or new superior alleles of an already known QTL. Such a discovery would make available to the breeders markers linked to the novel QTL, which can be directly used for MAS.

The various populations used for association mapping may be grouped into the following five categories on the basis of kinship and population structure: (1) ideal populations with little population structure and familial relationship (kinship); (2) populations with little population structure, but moderate familial relationship; (3) populations with moderate population structure and moderate familial relationship; (4) populations with moderate population structure, but little familial relationship; and (5) populations with strong population structure and variable familial relationship (Table 8.4). Since most plant materials will be adapted to the conditions of various localities in which they have been growing, exposed to natural and/or artificial selection, and are likely to be subjected to breeding, they would belong to the category four listed above (Zhu et al. 2008).

Table 8.4 The types of populations and study designs suitable for them

Population type	Example	Population structure	Kinship	Appropriate design
I	–	Little	Little	Regression, GC[a]
II	–	Little	Moderate	Mixed model, GC
III	Maize association panel	Moderate	Moderate	Mixed model, SA, GC
IV	–	Moderate	Little	SA, GC
V	Self-pollinated species	High	Variable	EMMA

Based on Yu et al. (2006), Kang et al. (2008), and Zhu et al. (2008)
[a]*GC* Genomic control, *SA* structured association, *EMMA* efficient mixed model association

8.5.2 Family-Based Association Panels: NAM Population

The *nested association mapping* (*NAM*) *population*, proposed by Yu et al. (2008), can be used for both linkage mapping of QTLs and AM. The NAM scheme was developed for maize: it uses RILs developed from a diverse set of parents, requires a smaller number of markers than GWAS in population-based association panels, and has higher resolution than QTL linkage mapping. The maize NAM panel has 5,000 RILs developed by crossing the inbred B73, as reference inbred, with each of the 25 diverse inbred lines selected to represent a substantial portion of the global maize inbred line genetic diversity. From each of the above 25 crosses, 200 RILs were developed by six generations of selfing without selection. This NAM population is estimated to represent 135,000 recombination events and has been genotyped for 1,106 SNP markers (Kump et al. 2011). Similar NAM populations may be developed in other crop species following appropriate mating schemes like diallel mating, North Carolina design II, eight-way cross, single/double round robin, etc., designs to generate sets of RILs. In a *round robin mating scheme*, each member of a set of inbred lines is mated as male to a defined number of inbred lines and as female to an equal number of other inbred lines so that each inbred is involved in equal and defined number of crosses. In *single round robin scheme*, each inbred is mated as male to one inbred and as female to one other inbred of the set. For example, if 5 inbreds (inbreds 1, 2, 3, 4, and 5) are mated as per single round robin scheme, the following five crosses will be made: 1×2, 2×3, 3×4, 4×5, and 5×1.

The NAM strategy has higher power than AM because the controlled crosses made for generating NAM populations minimize population structure and familial relatedness (Fig. 8.2). Further, the frequencies of otherwise rare alleles are increased in the biparental families making up the NAM population. In addition, the RILs comprising a NAM population can be used for linkage mapping. The large number of the RILs substantially enhances the power of linkage mapping. The NAM strategy facilitates cost-effective genome-wide scans and allows sharing of the NAM panel with researchers. The main statistical challenge with NAM and similar methods relates to the estimation of probability that alleles of various loci that are identical in state are also identical by descent. This problem is likely to become a minor issue since near-complete genome sequences of most species of interest would soon become available. But some questions related to NAM scheme need to be answered, e.g., the optimum number of parental lines to be used for generating NAM populations, the basis of selection of parental lines, the number of reference lines to be used, modifications needed to adequately address the issues of population structure, and the genetic architecture of the traits of interest (Myles et al. 2009).

8.5.3 Family-Based Association Panels: MAGIC Population

The *multiparent advanced generation intercross* (*MAGIC*) *populations* comprise a set of RILs

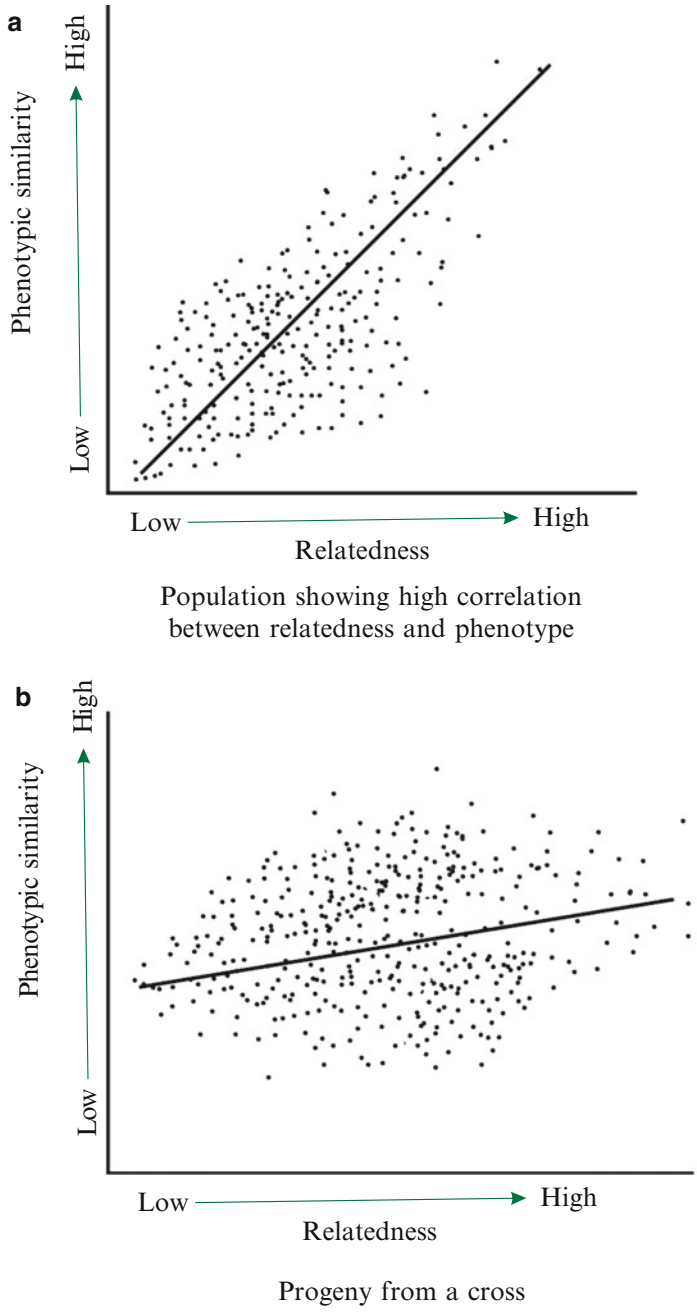


Fig. 8.2 The relevance of relationship between relatedness (kinship) and phenotype in AM studies. (**a**) In a population showing high correlation between relatedness and phenotype, closely related individuals have more similar phenotypes, while distantly related individuals have more dissimilar phenotypes. In such cases, random genetic markers distributed over the entire genome will show strong association with phenotype, and association mapping will have very low power to detect QTLs. (**b**). But when crosses are made between individuals/lines from such populations and biparental populations are produced, the correlation between phenotype and relatedness is greatly reduced. Therefore, the power to detect QTLs is greatly increased. The NAM (nested association mapping) populations, as a result, increase the power of QTL detection (Adapted from Myles et al. 2009)

produced from a complex cross or a set of crosses involving multiple parents (Sect. 5.16). These populations can be used for both linkage and association mapping of multiple traits for which the parents differ, and multiple alleles at the target loci can also be detected. The development of these populations is accompanied with several rounds of recombination, which increases the precision and resolution of QTL mapping. MAGIC populations can be derived from breeding lines and germplasm lines of interest to breeders. In such cases, these populations can also be used, either directly or indirectly, for variety development. In addition, they present opportunities for studying the interactions of genome segment introgressions and chromosomal recombinations. MAGIC populations have been developed in wheat and rice (using *indica* and *japonica* lines) and used for QTL mapping and variety development in rice.

8.6 Linkage Disequilibrium for Biallelic Loci

In a random mating population, the gene and genotype frequencies remain constant generation after generation. Changes in gene and genotype frequencies are produced by mutation, migration, selection, and random drift, which are often called evolutionary factors. If a gene has two alleles *A* and *a* with frequencies *p* and *q*, respectively, the genotype frequencies at this locus will be p^2 *AA*, $2pq$ *Aa*, and q^2*aa*. These genotype frequencies are known as Hardy–Weinberg equilibrium frequencies or simply equilibrium frequencies. In case the equilibrium is disturbed by one or more of the above factors, it is restored in the next generation after the causal factor is removed. When we consider two independently segregating genes (alleles *A*, *a* and *B*, *b*), the frequencies of their allelic combinations *AB*, *Ab*, *aB*, and *ab* would equal the products of frequencies of the respective alleles of the two genes. Therefore, the observed frequency of allelic combination *AB* (=*pAB*) will be the product of observed frequencies of alleles *A* (=*pA*) and *B* (=*pB*), that of *Ab* will be the product of frequencies of alleles *A* and *b* (=*pAb*), and so on. Thus, the four gametic combinations *AB*, *Ab*, *aB*, and *ab* will have the frequencies *pAB*, *pAb*, *paB*, and *pab*, respectively. At equilibrium, *pAB*.*pab* equals *pAb*.*paB* (Fig. 8.3). In case the equilibrium is disturbed by some factor, *pAB*.*pab* does not equal *pAb*.*paB*. The difference between *pAB*.*pab* and *pAb*.*paB* is referred to as *disequilibrium* (*d*), i.e., $d = (pAB.pab) - (pAb.paB)$. After each generation of random mating following the removal of the disturbing factors, the value of *d* declines to ½ of that in the previous generation. Thus, it would take several generations for the population to approach equilibrium for the concerned genes. For example, 1.6 % of *d* will remain even after six generations from the generation the disturbing factors were removed. In this case, disequilibrium is the result of disturbing effects of one or more of the evolutionary factors on gene and genotype frequencies in the population.

Allele	A (0.7)	a (0.3)	Total
B (0.4)	AB (0.7 × 0.4 = 0.28)	aB (0.12)	0.4
b (0.6)	Ab (0.42)	ab (0.18)	0.6
Total	0.7	0.3	

Independent segregation of genes A and B

Here, $pAB.pab = pAb.paB$

$(0.28 \times 0.18) = (0.12 \times 0.42)$

$0.0504 = 0.0504$

$D = (pAB.pab) - (pAb.paB)$

$= 0.0504 - 0.0504$

$= 0$

Fig. 8.3 The frequencies of different allelic combinations produced by independent segregation of alleles of two genes (*A*/*a* and *B*/*b*); this produces an estimate of "zero" for *D* (a measure of LD)

Disequilibrium can also result from linkage between the genes *a* and *b*. The term *linkage disequilibrium* (*LD*) signifies that a specific allele at one locus occurs with a specific allele at the second locus more often than expected on the basis of random assortment of the two loci. The two loci may represent two markers, two genes/QTLs, or one gene/QTL and one marker. Thus, in simple terms, *LD describes a nonrandom association between alleles of two or more loci*. As a result, the allelic combinations of the concerned loci observed in the population deviate significantly from their frequencies expected on the basis of independent assortment. Thus, the value of *pAB* does not equal that of *pA·pB* and so on, and that of *pAB·pab* differs from that of *pAb·paB* (Fig. 8.3). It may be added that in self-pollinated populations also, the value of *pAB·pab* will be equal to that of *pAb·paB* when the two genes are segregating independently. However, the two values will differ from each other if the two genes were linked, and the magnitude of this difference will increase with the strength of linkage. In this case, the difference (*d*) between *pAB·pab* and *pAb·paB* is termed as LD. In each generation of random mating, the magnitude of *d* will decline by the value *rd*, where *r* is the frequency of recombination between the two loci. This decline in LD is known as *LD decay*. Since *r* will be much smaller than 0.5, the LD decay will be much slower than the decline in disequilibrium between genes segregating independently. Further, the magnitude of LD will decrease with the genetic distance between the two loci since it is inversely related to the frequency of recombination between them. In each generation, there will be recombination between the two loci during meiosis, which will lead to a decline in the magnitude of LD. In simple terms, LD between two loci decays both temporally (as the generation advances) and spatially (with the increasing distance between the two loci).

In historical terms, when a new mutation arises, it will exhibit complete LD with the alleles at flanking loci because the mutant allele will always be present with them. For example, when gene *B* having only a single allele (allele *B*) is located close to a marker having two alleles (alleles *A* and *a*), there will be only two haplotypes for these two loci, viz., *AB* and *aB*. But when the gene *B* mutates to produce the allele *b*, this mutation can occur either in a chromosome with the marker allele *A* (as shown in Fig. 8.4) or in a chromosome having the marker allele *a*. After this mutation, there will be three haplotypes, viz., *AB*, *Ab*, and *aB* (Fig. 8.4), and the new mutant allele will always be present with marker allele *A*. A *haplotype* is the combination of alleles of two or more loci present in the same chromosome that tend to be inherited together. The fourth haplotype, i.e., *ab*, will be produced only when a recombination takes place between the two loci involving the chromosomes with *Ab* and *aB* haplotypes (Fig. 8.4). Therefore, the recombination will lead to a decline in the level of LD between the two loci. Obviously, the magnitude of LD will be greater in the cases of more recently produced mutant alleles than in the cases of those produced in relatively distant past. This is because a smaller number of recombination events are likely to occur in the case of former than in the case of latter. Thus, LD analysis can be used to deduce the historical aspects of genetic variation in terms of the contributions of mutation and recombination to the level of LD observed between pairs of loci in a given population; this aspect has been briefly explained in Fig. 8.5.

The above consideration assumes that LD arises due to linkage only and that its magnitude is directly related to the strength of linkage. *It may be clarified that the phenomenon of LD is quite different from linkage*. In *linkage*, alleles of two genes are inherited together because they are located close to each other in the same chromosome. On the other hand, *LD* is the occurrence of nonrandom associations between alleles of two loci in a population irrespective of their physical location in the genome. It should be noted that linkage between loci would generate LD, but significant LD can be observed between even unlinked genes due to epistatic selection (Sect. 8.16.2). In association mapping, efforts are made to filter out all other influences on LD estimates to, ideally, retain the effects of only linkage and use this information for identification

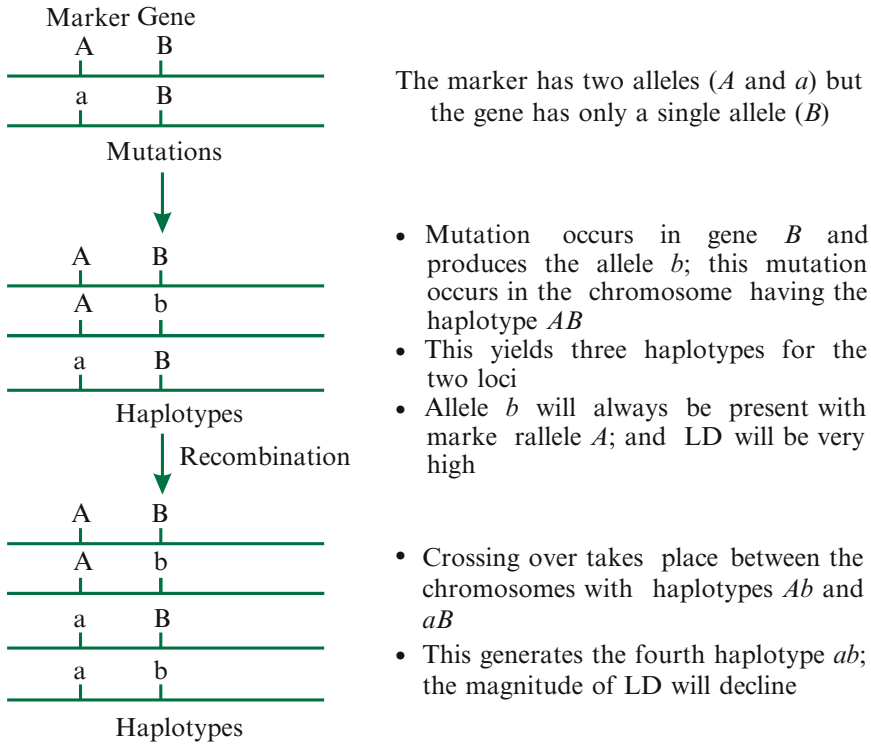


Fig. 8.4 A schematic representation of the contributions of mutation and recombination to the magnitude of LD. Only mutation can produce the third haplotype, *Ab* in this case. Recombination would, however, produce the fourth haplotype, *ab*. Mutation can also produce the fourth haplotype, but mutation rates are several orders of magnitude lower than those for recombination. As more recombination events occur between the two loci, the frequencies of the four haplotypes will move closer to their equilibrium frequencies. As a result, there will be progressive decline in the magnitude of LD between the two loci (Based on Ardlie et al. 2002)

of markers closely linked to the genes/QTLs governing the trait(s) of interest.

LD is also called *gametic phase disequilibrium* (*GPD*) or *gametic linkage disequilibrium* (*GLD*). But *zygotic linkage disequilibrium* (*ZLD*) is defined as a deviation of joint zygotic frequencies from the expected values assuming zero zygotic associations. *Most of the statistical properties of ZLD are similar to those of GLD and the results from them are generally comparable, but ZLD detects LD more extensively than GLD. Further, GLD is ideally applied to random mating populations that exhibit Hardy–Weinberg equilibrium.* But many natural populations diverge from Hardy–Weinberg equilibrium due to a variety of genetic events, including mutation, migration, selection, bottleneck, and population structure. In such populations, ZLD is the most appropriate measure of LD. LD has been extensively used to map and ultimately clone many genes.

8.7 Measures of Linkage Disequilibrium

The concept of LD was first proposed by Jennings in 1917, but Lewontin developed its estimation in 1964. Several different measures of LD have been proposed primarily to estimate LD between two loci each having two alleles. But some of these measures have been modified for application to other situations like two loci with more than two alleles and more than two loci. The statistical significance of LD estimates is determined by Fisher's exact test when the two

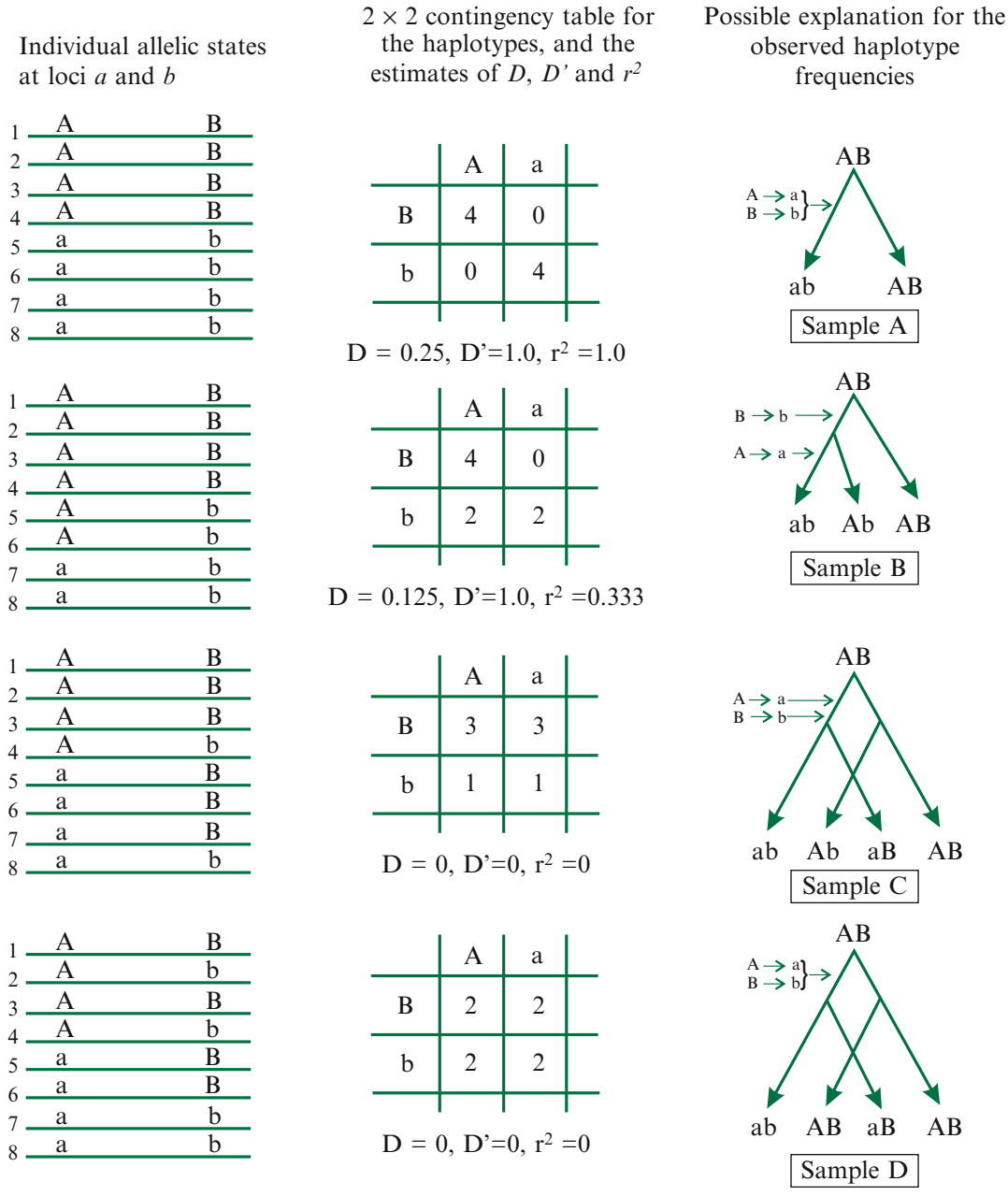


Fig. 8.5 Some hypothetical examples of observed haplotype frequencies and estimates of LD (D, D' and r^2) from them. The possible interpretations of the data in terms of mutational and recombination histories at the concerned loci are shown in the accompanying figures. In *sample A*, only two of the four haplotypes are observed in equal frequency. D' and r^2 estimates are 1. The A and B alleles appear to have mutated in the same chromosome around the same time. Further, there has been no recombination between the two loci after the mutational event. In *sample B*, three of the four possible haplotypes are present. The haplotype *AB* is more frequent than the haplotypes *Ab* and *ab*. The estimate of D' is 1.0, while that of r^2 is only −0.333. It appears that mutation of *B* to *b* occurred first, followed by that of *A* to *a* in the same lineage of chromosomes. There has been no recombination between the two loci. In *sample C*, all the four haplotypes were observed. The allele *a* is more frequent than allele *b*, suggesting that the mutation of *A* to *a* is older than that of *B* to *b*. Both the mutations have occurred in the same chromosome lineage. Recombination between the two loci has generated the fourth haplotype *Ab*. Further, the recombination event appears to be rather recent. In *sample D* also, all the four haplotypes were recovered, but in equal frequency. The two mutations (*A* to *a* and *B* to *b*) seem to have occurred around the same time in the same chromosome lineage, and there has been recombination between the two loci (Based on Flint-Garcia 2003; Gaut and Long 2003. Note: Data in samples C and D indicate random association, i.e., independent assortment, between the alleles of the two loci)

Table 8.5 Formulas for computation of the various estimates of LD

LD estimate	Formula	Remarks
D	$(pA.\ pab) - (pA.\ paB)$ or $pAB - (pA\ .\ pB)$	Depends on allele frequency; not in common use
D'^{a}	D/min $(pA.pb, pa.pB)$ when $D>0$	Most often used in plants
	D/min $(pA.pB, pa.pb)$ when $D<0$	
r^2 or Δ^2	$D^2/(pA{\cdot}pa{\cdot}pB{\cdot}pb)$	Most often used in plants
δ^{a}	$D/pB.pbb$	Approximation of δ*; also known as λ and P_{excess}
δ^{*a}	$\frac{pA[\{(pAB/pA)/(paB/pa)\}-1]}{1+pA[\{(pAB/pA)/(paB/pa)\}-1]}$	Frequently used in epidemiology
D	$D/(pB.pb)$	Specifically recommended for case-control studies
Q	$D/(pAB.pab + pAb.paB)$	Used in case-control studies; range −1 to +1
λ	$(pAB.pab)/(pAb.aB)$	Used in population genetics

Based on Devlin and Risch (1995)

[a]When disease causing allele is rare and the sampling of haplotypes is random, $\delta = \delta^{*} = D' = [D/(pa.pB)]$, where B is the allele producing disease and A and a are the marker alleles

loci have two alleles each and by multifactorial permutation analysis when more than two alleles occur at one or both the loci (Flint-Garcia 2003).

8.7.1 Two Biallelic Loci

At present, several different measures of LD, including D, D', d, r^2, Q, δ, δ*, and λ, are available for estimating LD between two loci, each having two alleles (Table 8.5). D is the basic estimate of LD, but it is not in common use. The estimates r^2, D', d, δ, and Q are different standardized versions of the D estimate. D is estimated as the difference between the observed frequency of an allelic combination in a sample and the product of the observed frequencies of the concerned alleles (Devlin and Risch 1995; Fig. 8.6). Thus,

$$D = pAB - (pA \cdot pB) \tag{8.1}$$

$$D = pab - (pa \cdot pb) \tag{8.2}$$

$$D = (pAB \cdot\ pab)\ -\ (pAb \cdot\ paB) \tag{8.3}$$

where D is LD between loci a and b; pAB, pAb, etc. are the observed frequencies of the allelic combinations AB, Ab, etc. in the sample; and pA, pB, etc. are the observed frequencies of the alleles A, B, etc. in the same sample. The value of D depends on allele frequencies. As a result, it may vary widely among different pairs of loci even when all the pairs are in complete LD. Therefore, the numerical value of D has little usefulness in determining the strength of LD, and it is not suitable for comparing the levels of LD among different loci and among various studies. In view of this, alleles with frequencies lower than 5 % or even 10 % are generally not included in estimation of D (Gaut and Long 2003).

D' and r^2 are the most relevant estimates of LD for plant species. D' is a standardized version of D calculated by dividing D with its maximum possible value obtainable from the given allele frequencies at the two loci (Devlin and Risch 1995; Fig. 8.6):

$$\left|D^{'}\right| = Dab/\min(pA \cdot\ pb,\ pa \cdot\ pB) \quad \text{when the value of } Dab \text{ is } > 0 \tag{8.4}$$

$$\left|D^{'}\right| = Dab/\min(pA \cdot\ pB,\ pa \cdot\ pb) \quad \text{when the value of } Dab \text{ is } < 0 \tag{8.5}$$

where Dab is the estimate of D between loci a and b and pA, pa, pB, and pb are observed frequencies of alleles A, a, B, and b, respectively (Lewontin 1964). Further, the terms min $(pA \cdot\ pb, pa \cdot pB)$ and min $(pA \cdot pB, pa \cdot pb)$ signify

Allele	A	a	Total
B	0.58	0.02	0.60
b	0.12	0.28	0.40
Total	0.70	0.30	

pAB = 0.58; pA=0.7; pB=0.60; pa=0.3; pb=0.4

A. Frequencies of gametic/gene combinations

D = pAB - (pAb. paB)
= 0.58 - (0.7 × 0.6)
= 0.58 - 0.42
= 0.16

D' = DAB / min (pA.pb, pa.pB)
= 0.16 / min (0.7 × 0.4, 0.3 × 0.6)
= 0.16 / min (0.28, 0.18)
= 0.89

r^2 = DAB^2 / (pA. pa. pB. pb)
= 0.16^2 / (0.7 × 0.3 × 0.6 × 0.4)
= 0.0256 / 0.0504
= 0.51

B. Estimation of D, D' and r^2

Fig. 8.6 Estimation of D, D', and r^2 from frequencies of the four types of allelic combinations (gametic combinations of the two alleles of the genes *A*/*a* and *B*/*b*). *pAB* = frequency of the allelic combination *AB*; *pA*, *pa*, *pB*, and *pb* = frequencies of the alleles *A*, *a*, *B*, and *b*, respectively. Min(*pA.pb*, *pa.pB*) signifies that the smaller of the two values, viz., *pA.pb* and *pa.pB*, will be used as the denominator

that the lower of the two estimates listed within each parenthesis will be used as the denominator. This measure of LD minimizes the effects of low allele frequencies on LD estimates. The estimates of D' will range between 0 and 1 even when the allele frequencies at the two loci are not identical (Flint-Garcia 2003). A D' value of 1 indicates complete LD, and it will be obtained only when no recombination would have taken place between the alleles at the two loci after the origin of these alleles. In addition, the concerned alleles should also not be separated by either gene conversion or recurrent mutation. In such cases, only three of the four possible haplotypes of the two loci will be observed in the sample (Fig. 8.4). Further, estimate of D' will be <1 only when all the four allelic combinations, i.e., haplotypes, are observed in the sample. D' takes into account primarily the recombination history since mutation and gene conversion are relatively much less frequent events. However, D' estimates of <1 do not have a clear interpretation. Therefore, statistically significant values of D' that are close to 1 can be safely considered to indicate genomic regions of low historical recombination, but intermediate values of D' cannot be relied upon as measures of LD. D' estimates are strongly influenced by small sample size particularly for loci with rare alleles. In such cases, high values of D' can be obtained even when the concerned loci are in equilibrium. This property makes D' estimates unreliable for comparing LD across loci and from different studies (Ardlie et al. 2002).

The value of square of the estimate of correlation coefficient between the alleles of the two genes gives the estimate of r^2 or Δ, and its magnitude ranges from 0 to 1. The estimate of r^2 is zero when alleles of the two genes segregate independently. This value will be one when, and only when, the two loci have identical allele frequencies in addition to lack of recombination between them. Unlike D', the intermediate values of r^2 can be easily interpreted. Further, the value of r^2 is related to the amount of information provided by one locus about the other. The r^2 estimates take into account differences in allele frequencies and show much less inflation than D' estimates (Ardlie et al. 2002). The following formula is used to estimate r^2 (Devlin and Risch 1995; Fig. 8.6):

$$r^2 = \frac{(Dab)^2}{(pA \cdot pa \cdot pB \cdot pb)} \qquad (8.6)$$

where pA, pa, pB, and pb are the observed frequencies of alleles A, a, B, and b, respectively, in the population and (Dab) is the estimate of D between the two loci.

The D' estimates provide a more reliable estimate of physical distance between loci than the estimates of D and r^2, since the latter are dependent on allele frequencies. Further, estimations of D and r^2 implicitly assume constant population size over generations, but this assumption is often violated. However, D' and r^2 are the most widely used measures of LD. Estimates of D' are strongly influenced by small sample size and yield unreliable results, particularly when comparing loci with low allele frequencies. In comparison, estimates of r^2 are more reliable under low allele frequencies. Further, estimates of D' measure only recombination history, while those of r^2 reflect both mutation and recombination histories (Flint-Garcia 2003). Therefore, it is desirable to verify the magnitude of LD detected by D' by estimating r^2, particularly when the allele frequencies are low, before a comparison is made across loci. *In general, r^2 seems to be the most appropriate measure of LD for AM*. Usually, r^2 values above 1/3 are considered to be useful for LD mapping (Ardlie et al. 2002; Gupta et al. 2005).

8.7.2 Two Loci with Multiple Alleles

Some markers like SSRs have multiple alleles, and many traits themselves may be governed by genes/QTLs having multiple alleles. In case of multiple alleles, first of all D' estimate for each pair of alleles (denoted by D_{ij}') of the two loci is obtained. Then weighted average (depicted as D') of these D_{ij}' estimates is computed to obtain an overall estimate of LD between all the alleles at the two loci using the following formula (Hedrick 1987):

$$D' = \sum_{i=1}^{k} \sum_{j=1}^{l} p_i q_j \mid D'_{ij} \mid \qquad (8.7)$$

where p_i and q_j are the frequencies of ith and jth alleles at the two loci having k and l alleles, respectively. D' estimates appear to be much less affected by allele frequencies and standardization of D' seems unnecessary. Methods for computing D' for maximum likelihood estimates using an expectation maximization (EM) algorithm and its use for mapping of multiallelic markers and QTLs have been developed. The chief problem in LD estimation in such situations arises due to difficulties in inferring the haplotype phase when more than one locus is heterozygous and multiple loci are considered. Several approaches like pedigree analysis, gamete characterization, etc., were evaluated for inferring the haplotype phase. One of these approaches uses EM algorithm to obtain the maximum likelihood estimates of gene frequencies for LD estimation. This approach has been applied to both animal and plant systems (see Gupta et al. 2005).

8.7.3 Multiple Locus Methods

Estimates of LD based on several loci will be required for preparing whole-genome LD maps. There are two approaches for obtaining multilocus LD estimates: (1) bottom-up approach and (2) top-down approach. In the *bottom-up approach*, one begins with individual loci and then measures multilocus LD. The estimation of multilocus LD is mostly based on the bottom-up approach (Geiringer 1944; Lewontin 1974). One method for handling multilocus data uses the LD estimate λ; the λ estimate for each marker and gene is used to obtain log-likelihoods that are summed up to yield a multiplying test for LD. The other approaches for treating multilocus data may be grouped as follows: (1) composite likelihood methods, (2) least square methods, (3) haplotype segment sharing methods, and (4) entropy-based method. These methods either use information from one marker at a time (*single point methods*) or from multiple loci at the same time (*multipoint methods*). The latter methods may be based on frequencies of haplotypes or of individual alleles at several marker loci. Multipoint methods are suitable for fine mapping of QTLs and are still being refined. In the *top-down approach*, higher-order LD coefficients are first determined; they are then broken down into lower-order LD estimates (see Gupta et al. 2005).

Most of the algorithms for the computation of higher-order LD were provided by Geiringer (1944). In 2004, Gorelick and Laubichler defined the LD at a single locus as the gene frequency at that locus; this definition greatly simplifies the calculations. They extended and simplified the approach of Geiringer (1944) and developed the algorithm for top-down approach for estimation of higher-order LD. In this approach, multilocus LD is first computed using an explicit formula. This LD estimate can be easily decomposed into its lower-order LD components. The estimates of LD for two, three, four, and six loci following this method are consistent with those obtained by the bottom-up approach. The highest-order LD estimate from this approach has to be interpreted along with all the lower-order LD estimates derived from it. Multilocus LD estimates are seldom used in experimental studies as their calculation requires a large number of input data. However, they are of theoretical importance and can be useful in the analysis of multilocus epistasis (Gorelick and Laubichler 2004).

8.8 Graphic Representation of LD

LD estimates are mostly obtained for pairs of loci. Pairwise LD estimates for a large number of markers may be depicted graphically to get an idea of the pattern of LD blocks, i.e., the genomic segments exhibiting persistence of LD, in the species. There are two methods of graphic display of LD values, viz., LD decay plot and color-code triangle plot, disequilibrium matrix plot, or LD heatmap. *LD decay* is the decline in the magnitude of LD between two loci due to recombination between them. In *LD decay plot*, pairwise values of LD (estimated as r^2 or D' and depicted on the X-axis) are plotted against the genetic distance (in cM) or physical distance (in base pairs, bp) between pairs of the markers (depicted on the Y-axis, Fig. 8.7). A nonlinear logarithmic regression curve of r^2 values on the genetic/physical distance is drawn to depict the generalized relationship between them. The LD decay plot may represent pairwise LD values between markers covering a specific region of the genome, or it may summarize the LD values between pairs of markers distributed throughout the genome. The distance at which r^2 value equals 0.1 or D' value equals 0.5 on the regression curve is generally considered as the average distance up to which LD persists in the species. But in AM, a higher threshold value of LD ($r^2 \geq 0.2$) is used as a cutoff point.

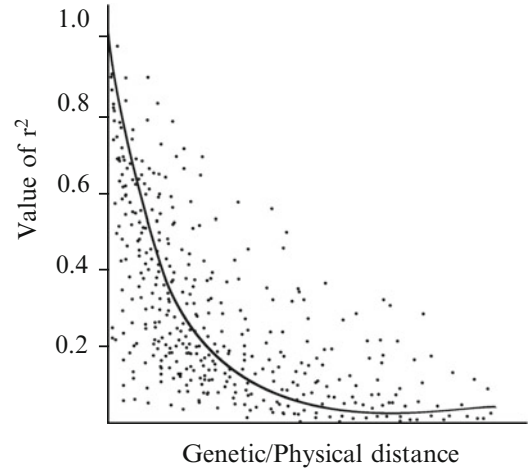


Fig. 8.7 LD decay plot for a hypothetical locus. The value of LD between pairs of loci tends to decrease with genetic/physical distance between them, but many values of LD between closely located loci may be much higher or lower than expected (Flint-Garcia 2003)

The *color-code triangle plot* can be generated by using programs like *TASSEL*, *GOLD* (Graphical Overview of Linkage Disequilibrium), a R-program like LDheatmap, PowerMarker, etc. The triangle plot represents a specific region of the genome or a single gene, and the significant LD values between pairs of several markers covering the region along with their *p*-values are depicted as colored cells above and below, respectively, of the diagonal. The markers covering the region are depicted on the X-axis, and the genetic or physical distances among them are shown on the Y-axis. The significant values of LD between marker pairs are depicted above the diagonal as colored blocks. The color of a cell depends on the value of LD as indicated in the color-code bar on the right side of the quadrangle. On the lower side of the diagonal, the *p*-values of

the LD estimates obtained by rapid 1,000 shuffle permutation test are depicted. Again the color of a cell depends on the *p*-value as shown in the color-code bar. Large red blocks along the diagonal of the triangle plot indicate high levels of LD between the loci located in the blocks and suggest that there has been limited or no recombination between these loci since the formation of LD blocks. Presence of large LD blocks facilitates association mapping of complex quantitative traits and considerably reduces the number of markers required to cover the given genomic region. When reliable estimates of the average size of LD blocks existing in a genome are available, one may safely determine the minimum number of markers required to efficiently cover the entire genome for genome-wide AM (see Abdurakhmonov and Abdukarimov 2008).

The LD decay plot provides the average pattern of LD blocks present in the gene/genomic region represented in the plot. In contrast, the color-code triangle plot identifies the specific segments in the gene/genomic region where LD blocks exist as well as the extent of individual LD blocks. It will be seen, particularly from the LD decay plot (Fig. 8.7), that the magnitude of LD between marker pairs declines as the distance between them increases. However, the magnitude of LD shows considerable random variation among different pairs of markers located in different regions of the gene/genome at comparable distances from each other (Flint-Garcia 2003). Further, this variation is much more pronounced at shorter distances, suggesting that LD decay rate in different genomic regions shows marked differences. The variation in the magnitude of LD also arises from several other factors as discussed in Sect. 8.16.

8.9 Useful LD

The concept of *useful LD* relates to the level of LD that is useful for association mapping. In general, much higher values of D' are required for useful levels of LD than those of r^2 because D' tends to overestimate the level of LD. The *half-length of* D' is the physical/genetic distance at which the value of D' between two loci declines to 0.5; this value greatly overestimates the distance over which LD would be useful for mapping. In contrast, a value of r^2 greater than 1/3 is generally taken as the minimum useful value of LD for mapping. The description of LD values in terms of *p*-values is likely to create confusion because *p*-values are largely dependent on sample size. Therefore, *p*-values cannot be used for comparing the levels of LD between different studies based on different sample sizes. In many situations, a significant *p*-value of LD may reveal little about the biological significance of the LD estimate. For example, a r^2 value of even 0.01 can be statistically significant if the sample size were 1,000 or larger, but such estimates of LD are unlikely to be useful for mapping (Ardlie et al. 2002).

8.10 The Extent of LD in Plant Species

The pattern of LD has been extensively investigated in maize, barley, rice, wheat, and *Arabidopsis* (Table 8.2). LD pattern varies from one species to another. For example, LD extends to >500 kb in *Oryza sativa* ssp. *japonica*, to ~75 kb in *O. sativa* subsp. *indica*, and to merely ~40 kb or lower in *O. rufipogon* (Mather et al. 2007). Further, different groups of materials of a single plant species may show considerably different extents of LD. For example, in maize, studies with several populations using different marker systems have revealed that LD patterns vary substantially from one population to the other and also with the marker type used. The maize populations used in these studies were sets of diverse inbred lines, synthetic populations subjected to generations of random mating, and diverse germplasm accessions. In most studies, a rapid decay in LD (r^2 declining to <0.25 within 200 bp) was observed for most of the genes. But in some groups of materials, the LD extended to up to 41 cM. Further, some genomic regions showed little LD decay over up to 105 cM (Table 8.6). The differences in LD patterns in

Table 8.6 Effect of germplasm and genomic regions on the extent of LD in maize

Germplasm/genomic region	Extent of LD	Reference
Material studied		
Flint group inbreds (whole genome)[a]	41 cM	Stich et al. (2005)
Dent group inbreds (whole genome)[a]	26 cM	Stich et al. (2005)
Genomic region		
Chromosome 2 (flint group)[a]	105 cM	Stich et al. (2005)
Chromosome 7 (dent group)[a]	103 cM	Stich et al. (2005)
Gene *su1* (*sugary 1*)	>12 kb	Remington et al. (2001)
Gene *adh1* (*alcohol dehydrogenase 1*)	500 kb	Remington et al. (2001)

[a]A cross-section of 147 elite maize inbred lines from Europe and the USA

different populations of a single species may be due to differences in the bottlenecks experienced by them during domestication and subsequent breeding (Sect. 8.16.7). The existence of significant LD over large genomic regions, e.g., over 100 kb or more, may indicate low rates of recombination in the concerned regions. Further, selection in different populations may produce different effects on the pattern of LD. For example, recurrent selection for 12 generations in one synthetic population was accompanied with a substantial increase in LD, while there was a decline in LD in another synthetic population (Labate et al. 2000). There is some evidence that microsatellite markers may detect greater levels of LD than SNP markers. This is because polymorphism for the rapidly evolving microsatellite markers is likely to be of more recent origin than that for SNP markers. It also appears that fewer markers would be required for genome-wide association mapping in self-pollinated species than in cross-pollinated ones.

8.11 Uses of LD in Plant Molecular Biology

LD analyses provide valuable information for furthering our understanding of plant molecular biology, including genomics in the following ways: (1) determination of marker-trait associations; (2) studies in population genetics, e.g., various aspects of natural selection; (3) analysis of the effects of domestication; (4) demographic history of plants; etc. One of the chief applications of LD studies is for AM to identify markers closely linked to traits of interest to breeders (Sect. 8.12). LD studies can also facilitate the development of functional markers, which have causal relationships with specific phenotypes of the relevant traits (Sect. 2.3). These markers can be used for MAS and for positional cloning of genes. AM can be readily used in forest trees where development of mapping populations is impractical due to their perennial life cycle. LD can be used to identify genomic regions that have been targets of selection during the evolutionary process as well as during domestication. Directional selection reduces polymorphism, while balancing selection tends to maintain/increase it. When the polymorphism maintained by balancing selection is old, there will be a greater variability in the sequences of the flanking regions of the allele in question; this may be used as "signature of selection" to identify alleles subjected to balancing selection. Several studies have attempted to identify genomic regions of crop plants that were selected for during domestication, although functions of the genes present in some of these regions were not known. DNA polymorphism data pertaining to several unlinked loci spread over the entire genome provide insights into the demographic history of a species. *Demographic history* relates to changes in population size, development of subgroups within a population, and similar changes in natural populations of a species (see Gupta et al. 2005).

8.12 Experimental Designs and Models for Association Mapping

Strategies of AM were initially developed for humans, and applied to plants without much modification. Subsequently, more precise and powerful methods for unbiased AM in plants were developed. There are several different approaches for the detection of significant LD, ranging from the simple chi-square test through analysis of variance to complex likelihood-based procedures. When the LD between a marker and a QTL is strong, the various methods would give comparable results. Generally, analysis of variance is not regarded as an effective procedure for AM. Therefore, AM for quantitative traits in plants is based on regression, maximum likelihood, and Bayesian approaches; a detailed treatment of some of these procedures can be found in Oraguzie et al. (2007). The most widely used and the simplest models test for association between a single-marker locus and a single target trait at a time. More recently, models for simultaneous evaluation of multiple marker loci as well as multiple traits for association have been developed. A brief discussion of the various AM strategies is given below (Table 8.7).

8.12.1 Case and Control Approach

This is the classical method of AM based on a group of unrelated individuals, called *cases*, carrying the allele of a gene responsible for a disease (or a mutant trait phenotype) and a sample of equal number of unrelated individuals, called *control*, lacking the disease. The allelic frequencies of the concerned gene and of the markers in the case and control groups are compared, and association between the gene and a marker can be detected by a suitable test. The chief limitations of this approach are the low frequency of "cases" in the population and a strong influence of population structure and

Table 8.7 A list of experimental designs used for association mapping

Design	Features	Remarks
Case and control approach	Based on a group carrying the disease causing allele (cases) and an unrelated group of equal size lacking the disease (control)	Used in humans; modifications like HRR, genomic control
Transmission disequilibrium test	A family-based design; compares transmission versus nontransmission of the marker to the affected progeny from one heterozygous and one homozygous parent	Used in humans
Structured association	Designed to minimize the effects of population structure; one version is the general linear model (GLM)	GLM implemented in TASSEL
Mixed linear model (MLM)	Designed to minimize the effects of population structure and kinship; markers and Q treated as fixed effects, while background QTLs are treated as random effects	Uses K or both Q and K matrices; EMMA is an improved version of mixed model
Multilocus mixed model (MLMM)	Multiple loci used as cofactors in the model; uses stepwise mixed model regression for the selection of loci and an approximate version of mixed model of correction for population structure	More QTL detection power and lower FDR than single locus tests
Multitrait mixed model (MTMM)	Simultaneous analysis of two or more correlated traits using the mixed model; separates genetic and environmental correlations and corrects for population structure	More power than single trait models when the traits are correlated; otherwise, lower power
Joint linkage-association mapping	Analysis of a sample drawn from a natural population and the open-pollinated progeny from this sample	Uses both LD and linkage analysis
Nested association mapping (NAM)	LD and linkage mapping in NAM populations	Higher power than AM alone

familial relationships. Techniques like haplotype relative risk (HRR) and genomic control were developed to overcome these limitations. The case-control approach was developed for human populations to tag genes causing genetic diseases. Similar approaches have been used in some studies with plants to identify markers associated with qualitative traits, e.g., identification of SNP and InDel (insertion–deletion) polymorphisms associated with the *Y1* gene for endosperm color in maize (Palaisa et al. 2003).

8.12.2 Family-Based Designs

Transmission disequilibrium test (TDT) was the first family-based design developed to avoid limitations of the case-control approach (Spielman et al. 1993). The *TDT design* assumes linkage between the gene of interest and the marker under test. A chi-square test is used to compare transmission versus nontransmission of the marker to the affected progeny from one parent heterozygous and one parent homozygous for the concerned disease allele. TDT is widely used for unbiased mapping of genes with two alleles using biallelic markers. TDT has been modified for application with multiallelic markers, missing parental information, etc. The family-based approaches were designed for application to human populations. Similar approaches have been used for AM in plants as well, e.g., detection of marker-trait associations in radiata pine (Kumar et al. 2004). In case of both "case-control" and "family-based" approaches of AM, next-generation designs have been developed, e.g., *identity by descent mapping*, *haplotype-sharing analysis*, and *decay of haplotype sharing*.

8.12.3 Structured Association Model

Many association analysis models consider markers as linear fixed effects, in which each marker is individually examined for association with a QTL affecting the trait of interest. All QTLs affecting the trait, but not associated with the marker being tested, are treated as background; these QTLs contribute to the residual error. This approach increases the error term, reduces the statistical power for detecting true associations, and increases the risk of false negatives. At the same time, several factors like population structure, selection, admixture, etc., may cause spurious associations between markers and traits and increase the risk of false-positive associations. The method genomic control was developed to correct for population structure in case-control and TDT studies and is rarely used in plants. The *structured association* (*SA*) *model* was developed to tackle the problems due to population structure. *Population structure* describes the level of genetic differentiation among the different homogeneous groups present in the population, from which the sample was drawn for the AM study. A random mating population lacking structure would consist of a single homogeneous group of individuals. A *homogeneous group* is a group of individuals that is at Hardy–Weinberg equilibrium for all of the several random markers/loci. In contrast, a structured population is itself nonhomogeneous and is composed of two or more different homogeneous groups. Thus, population structure generates LD between unlinked loci and tends to increase the likelihood of discovery of false-positive associations. A Bayesian approach is used to detect population structure and to generate the clustering matrix Q, which is also known as "gross-level population structure." The value of Q is estimated for each individual in respect of every putative homogeneous group in the population. The Q values indicate the likelihood of an individual belonging to the different putative homogeneous clusters/groups (Pritchard et al. 2000a). The SA model uses the Q matrix to correct, by logistic regression, the false associations due to population structure. A version of the SA model is implemented in the software package *TASSEL* (Trait Analysis by Association, Evolution and Linkage; Sect. 14.3.3) as the *general linear model* (*GLM*). The GLM can include main effects, interaction effects, nested effects, and covariates. The GLM finds the ordinary least squares solution for each marker-trait pair. The significance of association is tested either by F-test or

permutation test as specified by the user. *F*-test can be used when the residual error of the trait is normally distributed. In case this requirement is not fulfilled, permutation test may be used to estimate *p*-values, or the data may be transformed to provide roughly normally distributed error term.

Programs like *STRUCTURE* (Table 8.8; Sect. 14.3.4) detect population structure and estimate the *Q* matrix, which is used as covariate in SA/GLM. When *Q* is used, the differences between subpopulations are excluded while searching for marker-trait associations. As a result, any trait variation that is responsible for the differences between subpopulations will be ignored, and the genes/QTLs responsible for this variation will not be detected. It has been shown that the *P* matrix, estimated from the more robust principal component analysis (PCA; Sect. 11.3.2), can be used in the place of *Q* matrix. The *P* matrix is derived from a relatively smaller number of component variables obtained by summarizing the total variation observed across all the markers. The principal components (PCs) can be regarded as being related to the separate subgroups present in the population, from which the sample is drawn. Further, estimation of the *Q* matrix assumes the individuals to be unrelated with each other and the population from which they are drawn to be in Hardy–Weinberg equilibrium. But PCA does not make any such assumptions. The PCs from marker data can be estimated by using the software EIGENSTRAT (Price et al. 2006). The *P* matrix is preferable since the estimation of *Q* matrix is computation intensive. Further, the use of *Q* matrix tends to overestimate the number of subgroups present in the population. In any case, neither the *Q* nor the *P* matrix works well when the population structure is complex and/or contains individuals with some degree of relatedness (Myles et al. 2009; Segura et al. 2012).

8.12.4 Mixed Linear Models

In the *mixed linear model* (*MLM*), proposed by Yu et al. (2006), the markers and the population structure (*Q*) are treated as fixed linear effects, and the additive effects of the multiple background QTLs are considered as linear random effects. Each marker allele is fitted into the model as a distinct class, and the heterozygotes are added as extra marker classes. In addition, pairwise kinship between every pair of individuals/lines is incorporated into the statistical model as random effect. Thus, the MLM model considers that the trait phenotypes of two individuals that are genetically similar are more likely to be correlated than those of genetically different individuals. The marker effects estimated from the data are tested for significance without being broken down into additive and dominance effects. In addition, the covariances due to relatedness are also incorporated into the association analysis and are represented in the marker effect estimates. The MLM uses information about both population structure (estimated as *Q* matrix) and pairwise kinship (estimated as *K* matrix or kinship coefficient matrix) or only the *K* matrix to minimize false-positive associations. *Kinship coefficient* or *coefficient of co-ancestry* indicates the degree of relatedness between different pairs of individuals/lines of the sample. *Kinship represents the probability that the alleles of a randomly chosen gene present in a pair of individuals/lines are identical by descent*. An allele is said to be identical by descent when the copies of this allele present in the two individuals/lines have originated by replication of the same ancestral copy of the allele.

The *K* matrix is estimated either by *TASSEL* or *SPAGeDi* (Table 8.8) program. Both *Q* and *K* matrices are generated from data on a set of random or unlinked markers covering the whole genome. The *K* matrix can also be estimated from pedigree information using SAS PROC INBREED (*SAS/STAT software, version 9*). The kinship estimates from random markers are likely to be more accurate than those from pedigree data (Stich et al. 2008) because they reflect the "actual" relatedness, while the pedigree data estimate the "expected" relatedness. Since the parental contributions may deviate from the expectation due to independent assortment and/or segregation distortion, the "actual" estimates of relatedness may differ from the "expected" relatedness estimated from pedigree data. But a difficulty in kinship estimation from marker data relates to the definition of unrelated

Table 8.8 Statistical software packages generally used for association mapping in plants

Software package	Brief description
Free packages	
TASSEL	LD statistic calculation and graphic visualization; sequence analysis; association mapping using logistic regression, GLM, MLM, and some other models; structure and kinship analyses; analysis of insertion/deletion, diversity estimation, etc. (http://sourceforge.net/projects/tassel; http://www.maizegenetics.net)
EMMAX	Fast computation, for large AM studies, corrects for population structure and kinship (http://genetics.cs.ucla.edu/emmax/)
GenAMap	Implements structured association mapping, employs various algorithms, good graphical presentation (http://sailing.cs.cmu.edu/genamap/)
GenABEL	GWAS for both quantitative and qualitative traits (http://www.genabel.org/packages/GenABEL)
FaST-LMM	AM based on large samples of up to 120,000 individuals (http://fastlmm.codeplex.com/)
GAPIT	Implements CMLM, R-based, fast computation (http://www.maizegenetics.net/gapit)
STRUCTURE	Population structure analysis; generates *Q* matrix; computation intensive (http://pritch.bsd.uchicago.edu/structure.html)
SPAGeDI	Kinship analysis; generates *K* matrix (http://www.ulb.ac.be/sciences/ecoevol/spagedi.html)
EINGENSTRAT	Association analysis; PCA to generate *P* matrix to be used in the place of *Q* matrix (http://genepath.med/harvard.edu/~reich/software.html)
MTDFREML	MLM analysis of animal breeding data; can be used for plants (http://aipl.arsusda.gov/curtvt/mtdfreml.html
R	Generic package; convenient for simulation work; useful for researchers with good statistics and computer programming background (http://www.r-project.org/)
Commercial packages	
ASREML	MLM analysis for animal breeding data, can be used for plants (http://www.vsni.co.uk/products/asreml)
GenStat	Implements GLM and MLM, corrects for population structure (http://www.vsni.co.uk/software/genstat)
JMP Genomics	Computation of population structure and kinship coefficient (marker-based) (http://www.jmp.com/software/genomics/)
SAS	Standard statistical package used for data analysis and methodology work (http://www.sas.com)

Based on Zhu et al. (2008) and Gupta et al. (2014)

individuals. Yu et al. (2006) considered random pairs of inbreds as unrelated. Zhao et al. (2007), on the other hand, treated those pairs of inbreds as unrelated, which do not share any marker allele. Stich et al. (2008) estimated the conditional probability (T) of marker alleles present in pairs of inbreds being identical in state using the restricted maximum likelihood (REML) approach. The T matrix is used to estimate the kinship matrix K_T, which they recommended for use in the place of K matrix. The K_T matrix can be estimated using the *SPAGeDi* program. It was concluded that the QK method is effective for AM both in cross- and self-pollinated species. Studies in maize, potato, etc. reveal that MLM performs better in reducing both false-positive and false-negative associations than the methods that use either Q or K matrix alone (see Myles et al. 2009). But when kinship is estimated as the proportion of haplotypes shared by pairs of individuals, the matrix is denoted by K^*. The K^* matrix alone seems to serve the same purpose as the combined Q and K matrices. However, in some cases, MLM may lead to false negatives due to overcompensation for population structure and kinship; in such cases, AM without Q or K would be more useful.

One of the mixed models uses Bayesian variable selection for mapping multiple QTLs and combines it with LD mapping by using estimates of population structure. Another mixed model for AM uses QTLs/candidate genes identified from earlier studies and annotated for biological functions as a priori information along with LD estimates from a separate study. This strategy

drastically reduces the total amount of marker genotyping work because the markers are selected from only those genomic regions that contain the already known QTLs. A pedigree-based MLM is applicable to populations with known pedigrees. In this method, the haplotype effects are combined with the structure of variance–covariance relatedness matrix based on pedigree information. Further, it assumes polygenic effects involved in the control of population structure to be random. The efficacy of this model depends on the size of founder population and the degree of relatedness among individuals in this population. The *founder population* is the group of individuals that formed the basis of the population under consideration. It should be sufficiently large since populations obtained from two founder individuals are grossly inadequate for this model. This model is suitable for AM in crop species because plant breeding programs have generated many populations with known pedigrees. However, detailed pedigree information may often not be available, and it may be difficult to determine population structure of elite cultivars that usually have narrow genetic base.

Myles et al. (2009) have pointed out the following two main limitations of MLM. First, kinship estimation is being increasingly based on random marker data. But it is difficult to determine whether alleles of a marker that are identical in state, i.e., have the same genotype, are also identical by descent. Stich et al. (2008) attempted to resolve this difficulty by using the REML estimates of K, the K_T matrix, in the place of K matrix, which considerably improves the power of MLM. Bernardo (2013) proposed the use of genome-wide markers to account for the effects of background QTLs (G) in the place of K. Analysis of simulated data showed that the use of G detected more true QTLs and fewer false positives. The second limitation of MLM concerns the time-consuming extensive computations needed for most large genome-wide datasets: the computation time per marker increases as the cube of the number of genotypes in the AM panel. The efficient mixed model association (EMMA) method and other modifications of the MLM method have substantially increased the speed of computation. The current implementation of EMMA is available in an R package that can be downloaded along with the documentation from http://mouse.cs.ucla.edu/emma/index.html. This website also hosts the EMMA Web server. EMMA corrects AM for population structure and kinship, and it can be used with inbred populations. EMMA uses an algorithm for deducing the phylogenetic kinship matrix applied to the linear mixed model. This kinship matrix is determined from genome-wide markers and corrects for population structure (Kang et al. 2008). The program runs in Linux, Mac, and Windows environments. A modification of LMM, called factored spectrally transformed linear mixed model (FaST-LMM), reduces computation time by using a low-rank relatedness matrix that is estimated from a few thousand SNPs in place of all the SNPs used for AM. FaST-LMM is faster than EMMA as it can analyze in few hours the phenotype data from several thousand individuals along with the genotype data for a reasonable number of markers (Lippert et al. 2011). Further, multivariate linear mixed models (MvLMM) allow testing of associations between markers and multiple correlated phenotypes and are able to control population structure. The software genome-wide efficient mixed model association (GEMMA) implements mvLMM. GEMMA has improved speed and power and can handle more than two phenotypes (Zhou and Stephens 2014). However, an effective genome-wide analysis of the traits of interest would require a sufficiently large sample size and markers distributed throughout the genome at adequate density.

The MLM, EMMA, FaST-MLM, and MvMLM procedures are described as exact methods. The modified MLM methods require reduced computation time compared to MLM, but they are still computation intensive. Therefore, several approximate methods like genome-wide rapid association using mixed model and regression (GRAMMAR), EMMA eXpedited (EMMAX), and compressed MLM (CMLM) were developed. The approximate methods work well, but the accuracy is compromised (see Gupta et al. 2014).

The choice of appropriate AM strategy for plant species depends mainly on the following: (1) the amount and evolution of LD in the concerned population, (2) the degree of population structure, (3) access to the pedigree information, (4) the complexity of the target trait, and (5) available genomic resources. MLM utilizing both population structure and kinship information, pedigree-based mixed model, and multiple QTL model perform well in the majority of the cases. It has been argued that SA and MLM models fail to take into account the effects of selection and genetic drift on LD estimates, which are the major causes of LD in plant germplasms and breeding materials. Therefore, *the pedigree-based method is considered to be the most appropriate for AM with breeding materials*.

8.12.5 Joint Linkage-Association Mapping

Wu and Zeng (2001) proposed a strategy that makes use of recombination data between two loci as well as the existence of LD between them for determining marker-trait associations. This scheme is based on joint analysis of a sample drawn from a natural population as well as the open-pollinated progeny from this sample. This strategy, called joint linkage and association mapping (JLAM), has the combined advantages of linkage mapping (power to detect a QTL) and AM (precision of the detected QTL position and effect size). The basic premise of the JLAM scheme is that recombination during meiosis leads to reduction in the intensity of linkage between a marker and a QTL and that LD is created at a historic time. Therefore, estimation of these two components would provide an insight into the basis of the significant value of LD observed between a marker and a QTL. The JLAM is expected to increase the resolution of mapping and facilitate map-based cloning of QTLs. It also provides an opportunity for cross-validation of the QTLs detected by linkage mapping through AM in the same population and vice versa (*parallel mapping*) as well as of the QTLs identified jointly by linkage mapping and AM (*integrated mapping*). There is some evidence that integrated mapping is able to identify more significant marker-trait associations than parallel mapping. The JLAM has greater power than the traditional methods of LD mapping, and it has been extended to multitrait fine mapping of QTLs in animal species (see Gupta et al. 2005). The novel multiparent populations like NAM (Sect. 8.5.2) and MAGIC (Sect. 8.5.3) populations allow the construction of populations suitable for JLAM with relative ease. In addition, JLAM can be based on a set of biparental populations or a set of biparental populations along with a panel of germplasm/breeding lines genotyped with the same set of markers. One major limitation in the implementation of JLAM is the nonavailability of a software package for the required statistical analyses.

8.12.6 Multilocus Mixed Model

In general, GWAS uses single locus tests to detect associations between individual markers and target traits. However, quantitative traits are governed by more than one QTL. As a result, single locus tests are not entirely appropriate since the test statistic can be substantially inflated, particularly when the population is structured. In case of linkage mapping, this problem is avoided by including multiple QTLs as cofactors in the model, e.g., in composite interval mapping and multiple interval mapping. *Multilocus mixed model* (*MLMM*) of GWAS was proposed to include multiple loci as cofactors in the AM model. MLMM uses a simple stepwise mixed model regression analysis combined with forward inclusion and backward elimination of loci in the model. Although MLMM is computationally demanding, it is still computationally efficient to be applicable to GWAS. It uses an approximate version of the mixed model to correct for population structure. The MLMM generally outperforms comparable single locus methods so long as the marker data include the genomic sites having QTLs affecting the target trait polymorphism. The advantage of MLMM increases with trait heritability. It performs much better than the single

locus methods when the population is structured and the traits are governed by several loci having moderate to large effects. In case of GWAS, MLMM is expected to have higher power and lower false discovery rate (FDR) than the single locus tests. The MLMM can be easily extended to Bayesian analysis as it is based on a linear model (Segura et al. 2012). Another modification of the MLM, called linear mixed model-Lasso (LMM-Lasso), uses the sparse lasso regression to increase the power of AM and to reduce FDR (Rakitsch et al. 2013).

8.12.7 Multitrait Mixed Model

Typically, GWAS is based on phenotypic data for a single target trait. But different traits measured from the same individuals may be correlated with each other due to pleiotropy and the shared environment. The power of AM may increase if these correlations were taken into account in the model for AM. The idea of multitrait models is not new, and it has been around since the 1990s. But the *multitrait mixed model* (*MTMM*) extends this idea and the linear mixed model approach of AM to analyze pairs of correlated traits in GWAS. It employs a fully parameterized model to simultaneously estimate the within-trait as well as the between-trait variance components for pairs of traits. It separates genetic correlations from environmental correlations and corrects for population structure. It seems that most traits are genetically correlated either due to pleiotropy or because of LD for the causal genomic sites. MTMM has more power than single trait analysis whenever the traits are correlated and may also discover some novel QTLs. It identifies loci affecting both the traits, loci affecting one or the other trait, and loci generating opposite effects in the two traits. The MTMM is less powerful than single trait AM when the traits are either weakly correlated or not correlated at all. MTMM does not require phenotype data from full factorial replicated trials. It can be modeled for more than two traits, but this increases the complexities of computation and result interpretation. The MTMM approach is proposed to be used as a complement to, and not as a replacement for, single trait GWAS (Korte et al. 2012).

8.13 Significance Tests for Marker-Trait Associations

False associations (*Type I error*) may result from the effects of selection, inadequate handling of the confounding effects of population structure and kinship, poor experimental design, and environmental effects on the target trait. There are primarily two approaches for testing the significance of marker-trait associations: (1) the *p*-value estimation and (2) the Bayes factor calculation. The *p-value* is the probability of Type I error or the probability of null hypothesis (i.e., lack of marker-trait association) being wrongly rejected even when it were correct. As a result, the presence of marker-trait association will be inferred even when there is no marker-trait association. This approach is termed as "classical" or "frequentist" hypothesis testing; it uses *p*-values of a test statistic like F as the measure of evidence for the presence of marker-trait association. The classical approach is the most commonly used, and it compares the likelihood of the null hypothesis (H_0) being true as compared to the alternative hypothesis (H_1).

The second approach, the *Bayesian approach*, is based on probability theory. It uses the information derived from the data to update the prior knowledge about the concerned system, and this updated information is used as evidence for the presence of association. The mathematical representation of the experimenter's prior knowledge about the system is called *prior probability distribution*, *prior distribution*, or, simply, *priors*. Thus, the priors are highly subjective, but an experienced researcher would choose priors that are a reasonable representation of the available information. In case of AM, the prior knowledge relates to the likelihood of a random marker being linked to a QTL affecting the target trait. For example, if 5,000 equally spaced genome-wide markers were tested, and the number of QTLs affecting the target trait were assumed to

be 10, the likelihood of any one of these markers being associated with a QTL would be 1/500. The updated information, on the other hand, is referred to as the *posterior probability distribution*, *posterior distribution*, or, simply, *posteriors*. The posteriors are obtained by multiplying the Bayes factor with the priors. The *Bayes factor* is the ratio of the probability of getting the observed data when H_1 is correct to that when H_0 is true. Therefore, a Bayes factor of 20 shows that the given data are 20 times more likely to be obtained when H_1 is correct than when H_0 were true. There are several techniques for estimation of the Bayes factor; it can be readily calculated by using the *ldDesign* function of the *R* package (Ball 2007).

The Bayes factor directly quantifies the strength of evidence for H_1 as compared to H_0. The strength of evidence from a given value of Bayes factor depends on the value of priors, i.e., the probability of H_1 being correct. When the priors are low, higher values of Bayes factor will be needed to provide acceptable evidence for presence of association. It may be pointed out that the availability of good prior information is quite important as it may substantially increase the priors. This, in turn, would reduce the sample size needed for generating evidence of a given strength in favor of association. In addition, the Bayes factors permit designing of experiments with the given power to detect true associations since they enable estimation of the required sample size. *The sample size increases with the Bayes factor, and the required Bayes factor is higher with lower priors* (Ball 2007). *As a general indication, the Bayes factor of an experiment should be 20 or greater to yield reliable associations*.

The Bayesian and the classical approaches yield different results from hypothesis testing, and the difference becomes larger as the sample size increases. When the sample size is sufficiently large, the strength of evidence from modest Bayes factors is comparable to that from very small *p*-values (Ball 2007). As a result, *p*-values alone tend to exaggerate the strength of evidence for association and might lead to false-positive associations. Further, making a decision based on *p*-values is problematic. For example, a low *p*-value would indicate that the H_0 is inappropriate for the given data, but the corresponding *p*-value for the H_1 may also be equally small. Generally, a threshold of 0.05 is used for deciding the rejection of H_0 on the basis of *p*-values, but this is not appropriate in AM studies (Sect. 8.14).

8.14 Controlling "False Discovery" Rate

The null hypothesis (H_0) assumes lack of difference between two treatments. For example, H_0 assumes a lack of association between a given marker and the target trait in case of association mapping. The rejection of H_0 in this case means the presence of a significant association between the marker and the trait. Therefore, when H_0 is rejected, a "discovery" (of a QTL) is made. But when the H_0 is wrongly rejected, a Type I error is committed and a "false discovery" is made. When a small sample is scored for a large number of variables, e.g., 200 plants genotyped for 1,000 markers, multiple null hypotheses are tested. In such a case, the use of methods designed for testing single hypotheses greatly increases the rate of Type I error. The classical procedures like Bonferroni correction aim to control the probability of committing any Type I error. These methods are designed to control family-wise error rate and are more stringent than the method designed to control "false discovery" rate (FDR). The methods for controlling family-wise error rate have some important limitations and have not been widely used. Benjamini and Hochberg (1995) proposed the method for controlling the FDR. The *FDR* may be defined as the expected ratio of the wrongly rejected null hypotheses to the total number of H_0 rejected in the experiment multiplied by the probability of making at least one rejection of H_0. Thus,

$$\text{FDR} = \left(\frac{\begin{array}{c}\text{Number of null hypotheses}\\ \text{wrongly rejected}\end{array}}{\begin{array}{c}\text{Total number of null hypotheses}\\ \text{rejected in the experiment}\end{array}} \right) \Pr(R > 0) \tag{8.8}$$

where Pr(*R>0*) denotes the probability of making at least one rejection. The procedure for controlling FDR is simple and more powerful than the Bonferroni method. FDR testing begins with arranging all the *p*-values for the multiple comparisons in an ascending order. After this, the following FDR control is estimated:

$$P(i) \leq \frac{i}{m} q^* \tag{8.9}$$

where *P(i)* is the *p*-value for the null hypothesis at the *i*th *rank*, *i* is the rank of the *p*-value *P(i)* when all the *p*-values are arranged in ascending order, *m* is the total number of null hypotheses being considered, and q^* is the minimum FDR at which the H_0 is to be rejected. The q^* is generally kept at 0.05, which is the same as the *p*-value at which an individual test is declared significant. All the null hypotheses having the *p*-value equal to or less than the estimate $(i/m)\ q^*$ are rejected, and all the null hypotheses are rejected when *i* represents the highest rank.

The term *positive false discovery rate* (*pFDR*) describes the expected ratio of the wrongly rejected null hypotheses to the total number of H_0 rejected in the experiment when positive findings have occurred. The quantity *q*-value, which is the pFDR analog of the *p*-value for FDR, gives a direct measure of the Type I error rate. The method using pFDR fixes the rejection region and then estimates the corresponding error rate, while the FDR approach does the exact opposite. The pFDR approach is more effective, flexible, and powerful than the FDR approach and is conceptually simpler and straightforward (Storey 2002).

8.15 Relevance of Marker Systems in LD Estimation

Molecular markers are either dominant or codominant in nature. Further, most of the markers have two alleles at a single locus (*biallelic markers*), but some of them like SSRs have more than two alleles at a single locus (*multiallelic markers*). The methods for estimation of LD were developed for biallelic codominant markers, but they have been applied to multiallelic codominant markers as well as to dominant markers. But in the case of multiallelic codominant markers, it is not easy to determine the identity by descent of the various alleles of a locus represented by multiple bands. This is particularly problematic when diverse, highly structured, and polyploid germplasms are used, and information about their pedigree is either lacking or grossly incomplete. In such cases, only those SSR loci that yield single bands may be used and scored as codominant markers. However, single band SSR markers are limited in number in polyploid species, particularly when diverse germplasm are analyzed. In such cases, the loci producing multiple bands may be treated as dominant markers, and each band is considered as a separate locus and scored as either "present" or "absent."

The dominant markers are less informative than codominant markers as they cannot identify heterozygotes. This reduces the statistical power of association analyses based on these markers. But in case of many plant species like forest trees, the use of dominant markers like amplified fragment length polymorphisms (AFLPs) and randomly amplified polymorphic DNAs (RAPDs) becomes necessary due to the nonavailability of codominant markers. In addition, SNPs are emerging as the markers of choice, and they are generally scored as biallelic dominant markers. The dominant markers can be gainfully used for clustering of individuals and populations using a Bayesian approach to analyze data for a large number of loci. They can also be used to estimate kinship coefficients between pairs of individuals. An estimate maximization algorithm allows estimation of LD from data on dominant

markers in diploid species, provided a large number of loci is analyzed and the sample size is sufficiently large. The sample size would depend primarily on allele frequencies: when frequencies of the less frequent or minor alleles are close to 0.5 and 0.1, the sample size should be $\geq$200 and $\geq$400, respectively (Li et al. 2007b).

Another aspect of marker systems relevant to LD analyses concerns mutation rates of the markers. In general, the higher is the marker mutation rate, the more rapid will be the rate of dissipation of LD. This effect will become more important at small genetic distances, particularly as the rates of recombination and mutation become comparable in magnitude. The higher mutation rates of microsatellites lead to a reduction in their power to detect LD. In contrast, SNPs have much lower mutation rates; as a result, the power of SNPs to detect LD is much higher than that of SSR markers. In view of this and, more particularly, their suitability for a very high-throughput genotyping at several-fold lower cost per data point than SSR markers, SNPs are becoming the markers of choice in LD and AM studies.

8.16 Factors Affecting LD and Association Mapping

Both LD and AM are affected by a variety of genetic and demographic factors. Some of these factors like recent occurrence of the concerned mutation, self-pollination, population structure, kinship, genetic drift, selection, admixture, epistasis, etc., increase LD. But factors, such as high recombination rate, high mutation rate, gene conversion, etc., reduce LD.

8.16.1 Mating Pattern in the Population

Populations are termed as outcrossing or selfing populations on the basis of mating pattern prevalent in them. In *outcrossing populations*, a high level of heterozygosity is maintained so that opportunity for crossing over between pairs of loci is present in each generation. As a result, the level of LD between pairs of loci declines with the number of generations. In contrast, *selfing populations* consist primarily of homozygous genotypes, and any heterozygotes arising due to natural outcrossing or mutation are rapidly resolved into homozygous genotypes. Therefore, the opportunity for crossing over between pairs of loci is limited to a small number of generations, following that generation in which they became heterozygous. The effective recombination rate (c) in a given species is estimated from the frequency of selfing (s) in that species using the following formula.

$$c = 1 - \left[\frac{s}{(2-s)}\right] \qquad (8.10)$$

The *effective recombination rate*, c, is the ratio of recombination rate in the given species to the recombination rate expected in an obligate cross-pollinated species like a self-incompatible species. Thus, in a plant species showing 95 % self-pollination, the effective recombination rate will be less than 10 % of that expected in an obligate cross-pollinated species, while it will be less than 2 % in the case of plant species having 99 % self-pollination.

In view of the above, LD blocks are expected to be much longer in selfing than in outcrossing species. It has been estimated that when $s = 0$, LD would decay within 500 bp, but when $s = 0.95$, it may extend to up to 10 kb. Therefore, considerably fewer markers would be required for genome-wide AM in a self-pollinated than in a cross-pollinated species. For example, SNP markers spaced at 100–200 bp will be needed in maize to maintain adequate power in AM, but markers spaced at 50 kb would be adequate for *A. thaliana*. There are instances where a selfing crop species seems to have originated from an outcrossing or predominantly outcrossing progenitor species. For example, cultivated soybean (*Glycine max*) shows only ~1 % cross-pollination, while its ancestor, *Glycine soja*, shows ~13 % outcrossing. In such cases, the related outcrossing species might be used for AM with a much higher resolution, e.g., ~11-fold higher resolution in *G. soja* than in *G. max*.

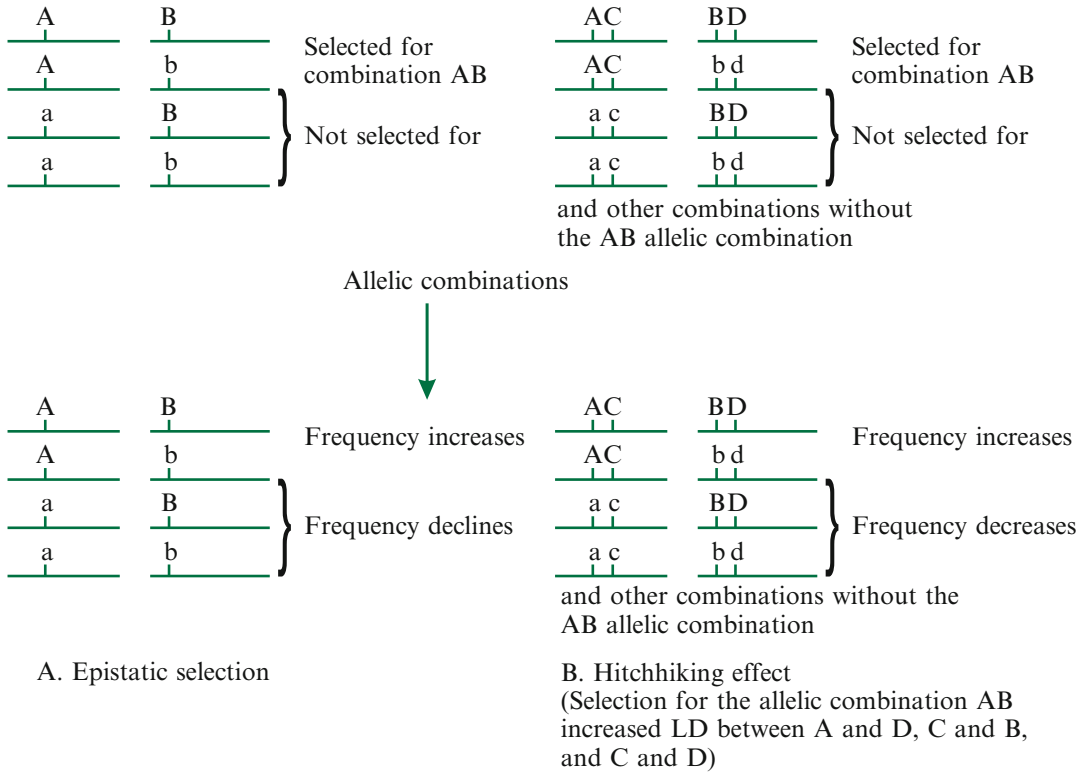


Fig. 8.8 A schematic representation of epistatic selection and the "hitchhiking" effect. There is epistatic selection for the allelic combination *AB*. The frequency of allelic combination *CD* would increase as a result of this selection because of the linkage of allele *C* with *A* and that of allele *D* with *B*, respectively, although there is no selection for the allelic combination *CD*

8.16.2 Selection

Selection may be defined as differential reproduction rates for different genotypes. Ordinarily, selection operates on the phenotypes generated by various genotypes. Selection can generate LD between unlinked genes through epistatic selection as well as the 'hitchhiking' effect. Suppose two or more interacting genes govern a trait. Selection may favor a specific phenotype of this trait. This will tend to maintain together those alleles of the concerned genes that jointly produce this phenotype even if the genes were located in separate chromosomes. This phenomenon is known as *epistatic selection*. Suppose genes *A*/*a* and *B*/*b* located on separate chromosomes control a trait that is subjected to selection (Fig. 8.8a). Further, the alleles *A* and *B* together produce the phenotype favored by selection; this will increase the frequency of allelic combination *AB*. The magnitude of this increase will depend on the intensity of selection, and may even lead to fixation of these alleles. The increased frequency of *AB* allelic combination will generate substantial LD between the unlinked genes *A*/*a* and *B*/*b*. Epistatic selection also operates when a population is subjected to simultaneous selection for multiple traits. In such cases, the favored alleles at all the loci governing the concerned traits will be selected for and maintained together irrespective of their linkage relationships. Another way of looking at selection may be to view it as creating a bottleneck for the genomic regions containing the loci subjected to selection; this would generate LD between the alleles that are selected for. The

effects of selection on LD are short-lived, i.e., for $<0.4Ne$ generations with moderate recombination rates, where *Ne* is the effective population size.

Epistatic selection can also generate substantial LD between unlinked genes by its 'hitchhiking' effect. The '*hitchhiking*' effect may be described as an increase in the frequencies of alleles at essentially neutral loci located on either side of a locus subjected to selection. Suppose the allelic combination *AB* of the unlinked genes *A*/*a* and *B*/*b* is selected for. Let us further assume that gene *C*/*c* is located close to gene *a* and gene *D*/*d* is located close to gene *b*, and that alleles *C* and *D* are linked with the alleles *A* and *B*, respectively (Fig. 8.8b). Further, genes *c* and *d* govern such traits that are not under selection. In this case, as the frequencies of alleles *A* and *B* will increase due to epistatic selection, the frequencies of alleles *C* and *D* will also increase due to their linkage with the alleles *A* and *B*, respectively. Thus alleles *C* and *D* will literally ride along or 'hitchhike' with the alleles *A* and *B* as the frequencies of latter increase due to selection. Therefore, the frequency of allelic combination *ABCD* will increase, and this will generate substantial LD between the unlinked loci *A*/*a* and *D*/*d*, *C*/*c* and *B*/*b*, and *C*/*c* and *D*/*d*. Thus, the hitchhiking effect can lead to a large increase in the frequencies of the concerned alleles and may even cause their fixation. A similar, but usually less drastic, effect is produced by selection against a deleterious allele at a locus. The hitchhiking effect will be particularly relevant in genomic regions with low recombination rates.

8.16.3 Population Structure

Population structure is ubiquitous, and arises due to geographical isolation, and natural and artificial selections. *Population structure* signifies that individuals in a population do not form a single homogeneous group, but they are distributed in few to several distinct subgroups that show different gene frequencies. As a result, the probability of sampling individuals having a specific trait phenotype from one or few of these subgroups may be higher than that from other subgroups; this might yield misleading LD estimates. Whenever a phenotypic trait is correlated with the population subdivision, the trait is likely to show significant association with even those loci that are actually not involved in expression of this trait. One approach for AM in structured populations essentially divides the population into homogeneous subgroups (putative unstructured subpopulations) on the basis of gene frequencies, and evaluates associations within the subgroups. This scheme has been modified for handling of quantitative traits as well (see, Gupta et al. 2005). From a statistical viewpoint, as the confounding effects of population structure increase, the FDR also increases. The GLM, MLM, EMMA etc. models for AM minimize the effects of population structure.

8.16.4 Admixture

Admixture refers to gene-flow between genetically distinct populations of the same species, and is the same as *migration*. It brings into the population chromosomes derived from different ancestors, gene combinations subjected to different selection pressures, different allele frequencies or even new alleles. This might generate considerable LD, which may extend to even unlinked loci. Initially, the increase in LD is proportional to the differences in allele frequencies of the two populations, and this LD does not depend on linkage relationships of the loci. However, the LD between unlinked loci disappears rapidly due to random mating, and that between linked loci also decreases due to recombination. Several approaches have been developed for estimating LD caused by admixture, and using these estimates for admixture disequilibrium mapping. However, these methods require the parental populations involved in admixture to have been relatively homogeneous and to have substantially different allele frequencies, the admixture to have preferably occurred at one fixed time in recent past, and the time of this occurrence to be known with some confidence.

8.16.5 Genomic Region

It is generally accepted that different regions of the genome of a given species show different rates of recombination that may vary >10-fold. There is evidence that gene-rich genomic regions tend to have higher rates of recombination than gene-poor regions, and that regions having repetitive DNA and retroposons show little or no recombination. Thus, the rate of LD decay would be higher in such genomic regions that show higher recombination rates, and a higher marker density would be required for LD analysis in such regions.

8.16.6 Kinship

Kinship coefficient can be estimated by using a software package like *TASSEL*, which estimates kinship coefficient as the proportion of alleles that are identical between each pair of lines/individuals in the sample. This program generates a marker similarity matrix for all the lines/individuals of the sample. The estimates in this matrix are then rescaled so that they fall between 0 and 2. Population structure and kinship coefficients are determined from data on a set of unlinked markers, often called background markers, distributed over the entire genome. It has been suggested that the number of background SSR markers should be four times the number of gametic chromosome number of the species so that two markers are located on each chromosome arm. Further, a larger number of biallelic markers like SNPs would be needed than that of the multiallelic SSR markers.

8.16.7 Genetic Drift and Bottleneck

Genetic drift is the random change in gene frequency of a population due to random sampling of gametes that unite to produce a finite number of individuals in each generation. Genetic drift occurs in small populations and consistently leads to the loss of rare allelic combinations, leading to an upward bias in LD. When genetic drift and recombination are at equilibrium, the following relationship is observed.

$$r^2 = \frac{1}{(1 + 4Ne.c)} \tag{8.11}$$

where, *Ne* is effective population size and *c* is the recombination fraction between two loci. Thus, LD becomes smaller as the effective population size becomes larger or the number of generations since the creation of LD increases. Hence, a small, stable (no change in size over generations) population is more suitable for AM because the inflated LD level increases the chance of finding markers linked to the target gene. A marked reduction in the size of a population for one or more generations is known as *bottleneck*. Bottlenecks lead to high levels of genetic drift since only few allelic combinations are transmitted to the future generations. As a result, bottleneck can generate substantial levels of LD, but in the absence of other factors like population structure, its effects are short-lived.

8.16.8 Gene Conversion

Gene conversion is a process in which a small segment of one chromosome is copied in the place of its homologous segment of the homologous chromosome during meiosis. Thus, gene conversion is an event of nonreciprocal recombination. As a result, more than 50 % of the gametes contain one allele of a gene, while the frequency of gametes with the other allele is reduced accordingly. About 50 % of the nonreciprocal recombination events do not lead to recombination between the genes/markers located on either side of the gene involved in gene conversion. In essence, gene conversion produces the same results as those generated by double crossing over. As a result, gene conversion reduces the magnitude of LD between the affected gene and the marker loci flanking this gene, but the LD between the flanking markers is not affected. Therefore, *gene conversion tends to*

reduce the correlation between LD and genetic distance between loci located close to each other. Gene conversion was discovered in *Ascomycete* fungi, but is known to occur in other eukaryotes, including *Arabidopsis*. The frequency of gene conversion seems to be rather high, and there may be gene conversion hot spots, i.e., small genomic regions with unusually high gene conversion rates. These hotspots appear to coincide with regions previously identified as crossing over hotspots.

8.16.9 Ascertainment Bias

Ascertainment bias is a systematic bias generated in a dataset by the manner in which the data were collected. This bias is important in estimation and comparison of LD among and within populations. For example, SNPs are usually identified by resequencing the genomes of a small number of individuals. The discovered SNPs are then used for analysis of a much larger sample from the same or some other population (Akey et al. 2003). This strategy may lead to ascertainment bias since the small sample used for SNP discovery would reduce the probability of identification of low-frequency SNP alleles. This would create a bias in favor of alleles with intermediate frequencies, and would tend to reduce the magnitude of LD in comparison to the actual value that would be obtained if all the SNP alleles were analyzed. The problem of ascertainment bias may also arise when SNPs are identified by one approach, say, genome resequencing, and are used for genotyping by another method like SNP microarrays. The magnitude of ascertainment bias depends on several factors, including the strategy of SNP discovery, the number of chromosomes used for SNP discovery, demographic history of the subpopulation(s), etc. *The ascertainment bias can be estimated and appropriate correction can be used for obtaining reliable estimates of LD. In any case, SNP genotyping by whole-genome resequencing of the entire sample will eliminate the risk of ascertainment bias*.

8.16.10 Marker Mutation Rate

Different marker systems may differ in their mutation rates and, as a result, in the extent of LD detected by them. SSR markers generally show much higher mutation rates than SNP markers. The high mutation rates of SSR loci are due to slippage during DNA replication that leads to generation of length variation in these loci. In general, the higher the mutation rate of a marker, the higher will be the rate of LD decay with time and, as a result, the smaller will be the extent of LD detected by this marker system. Another problem related to SSR markers is the phenomenon of homoplasy, which results from slippage during DNA replication. *Homoplasy* is the situation of two SSR alleles of identical size being different by descent. Homoplasy can be problematic for SSR alleles having high mutation rates, particularly when they are used for estimation of genetic parameters from a large sample (see, Zhu et al. 2008).

8.16.11 Errors in Genotyping

The genotyping of individuals of a sample should be, as far as possible, error-free. This is because even a low error rate of ~3 % or even less can have dramatic effects on the accuracy of LD estimates and AM. This is particularly relevant for SNPs since the rate of error varies significantly between different SNP loci even in a single assay (Ingvarsson and Street 2011).

8.17 Conclusions About LD Patterns in Plant Species

Based on the foregoing discussion about the pattern of LD in plant species, the following generalizations can be made. (1) LD decay is much more rapid in outcrossing than in selfing species. (2) The extent of LD is much higher in cultivars and breeding lines than in wild accessions and land races of a crop species. (3) Germplasm accessions and even elite breeding materials and

cultivars of both selfing and outcrossing species show smaller LD blocks than biparental populations, e.g., F_2, RIL, etc., populations. (4) Different marker systems are likely to provide different estimates of LD. (5) The extent of LD may vary markedly among the different regions of a genome. (6) Collections of germplasm accessions with narrow genetic base show longer LD blocks than those having broad genetic base. (7) The size of LD blocks and the abundance of LD determine the power and precision of AM. Finally, (8) the pattern of LD is greatly influenced by a variety of factors, including population structure and genetic drift.

8.18 LD Maps

An *LD map* depicts markers separated by distances represented by LD units (LDU). LD mapping theory extends the estimation of covariance of D for a random sample of haplotypes or diplotypes (for disomic genomes) to the association probability p so that

$$p = D/Q(1 - R) \tag{8.12}$$

where D denotes LD estimate; Q stands for the frequency of the most rare allele, which is presumed to be of the most recent origin; and R represents the frequency of the marker allele showing association (Maniatis et al. 2002). Thus, LD mapping uses the parameters D, Q, and R; the software *ALLASS* (*all*ele *ass*ociation) and *LDMAP* Version 0.1 are designed for the construction of LD maps. LD mapping has been initiated in humans, and similar efforts are expected to be initiated in plants as well.

8.19 Mapping of Expression Quantitative Trait Loci

The level of expression, called *expression value*, of a gene may be considered as a phenotype produced by that gene. The expression value may be subjected to genetic analyses in the same way as any other phenotype; this line of study has been called *genetic genomics*. Initially, gene expression values were based on RNA, but they have now been extended to proteins and other metabolites. Gene expression values have been used for mapping QTLs that affect these values; these QTLs are called *expression quantitative trait loci* (*eQTLs*). eQTLs have been mapped in plant species like *Arabidopsis*. A majority of the genes seem to be affected by eQTLs. Some of the eQTLs map in *cis*-, while others map in *trans*-position in relation to the affected genes. In general, *cis*-eQTLs have larger effect on gene expression than *trans*-eQTLs. There appear to be "hotspots" of *trans*-eQTLs that affect a surprisingly large number of genes. The available statistical tools provide only limited power to identify *cis*-eQTLs from *trans*-eQTLs, especially in outcrossing species. It has been suggested that the use of AM may provide gene-specific resolution of this trait in the outcrossing species. The second-generation RNA-Seq methods simultaneously generate both phenotype and genotype data. They provide data on the amounts of different RNA species, i.e., the phenotype data. They also generate data on the SNP alleles, which are the genotype data. But for more reliable results, the genotyping should be based on genomic DNA (Ingvarsson and Street 2011). RNA-Seq data can be used to generate SNP genotype data that can be analyzed to detect marker-trait associations; this has been referred to as *associative transcriptomics*. These markers may enable identification of functional markers responsible for the trait variations.

8.20 Power of Association Mapping

The probability of detecting "true" marker-trait associations in a sample using AM is called *power of association mapping*. The power of an AM experiment depends on several factors, including the extent and evolution of LD in the population, nature of gene effects involved in control of the target trait, sample size, experimental design, accuracy of phenotyping, type of markers, etc. The chances of detecting LD are the

greatest for mutations that are of recent origin (i.e., are in strong LD), have large effect on the phenotype, and are present in a relatively less frequent haplotype background. The power of an experiment can be improved in the following two ways: (1) by enhancing quality of the data through improved experimental design and procedures and (2) by increasing the sample size without reducing the data quality. The power of LD detection can be markedly increased by choosing a suitable study design, reducing the environmental variation, and increasing the genetic effects by selecting the extreme phenotypes. The marker genotyping work can be reduced by using such genomic regions for mapping that have known QTLs/candidate genes, selecting one marker from each haplotype of interest, etc. A large sample size would be required to enhance the reliability of associations. The required sample size for a given power can be estimated on the basis of Bayes factors (Sect. 8.13). Small mapping populations of few hundred individuals enable detection of QTLs with large effects and lead to substantial overestimation of their effect size. Most of the current AM strategies are not able to detect QTLs with ~1−2 % effect on the phenotypic variation. In many cases, all the QTLs for a trait identified by most AM studies are able to explain only 5−20 % of the phenotypic variation in the trait. This indicates the involvement of many QTLs with small individual effects in control of the quantitative traits.

8.21 Confirmation of Marker-Trait Associations Through Replication Studies

In general, a substantial proportion of positive associations detected in one study are not confirmed in subsequent independent studies. Therefore, it is important that the associations detected in one study are confirmed by an independent study with another population; such a study is often referred to as *replication study*. Replication studies help the identification of true positive associations and provide more reliable estimates of the effects of different loci/alleles on the target trait. A failure to detect significant effects in replication studies may be due to several reasons like poor experimental design in the original and/or the replication study, small sample size, inaccurate phenotyping, environmental variation, etc. However, failure to confirm a positive association in a replication study may not necessarily mean that the original association was a false positive; this failure could be due to some other reason like allelic heterogeneity at the locus in question. It may be pointed out that positive associations can also be confirmed by validation of the biological function of the concerned locus/gene by producing transgenic lines (Sect. 12.8.1; Ingvarsson and Street 2011).

8.22 The tagSNP Strategy of SNP Genotyping

The number of known SNPs present in the genomes of the different species is increasing at a rapid rate, making the task of evaluating a sample for all the known SNPs very demanding. Therefore, efforts are being made to determine a subset of a minimum number of SNPs distributed throughout the genome that can be used for effective genome-wide association analysis. Analysis of SNPs in identical chromosomal regions of different human individuals has revealed that SNPs located near each other tend to be inherited together. Further, in many chromosomal regions, only a limited number of SNP haplotypes are found. The various haplotypes for a given genomic region can often be distinguished from each other by analysis of only a small number of SNP loci from among the relatively much larger number of SNPs forming the haplotype. The subsets of SNP loci that enable reliable identification of the different SNP haplotypes present in a given genomic region are called *tagSNPs* (*tSNPs*) or *haplotype tagging SNPs* (*htSNPs*). The haplotype map (HapMap)

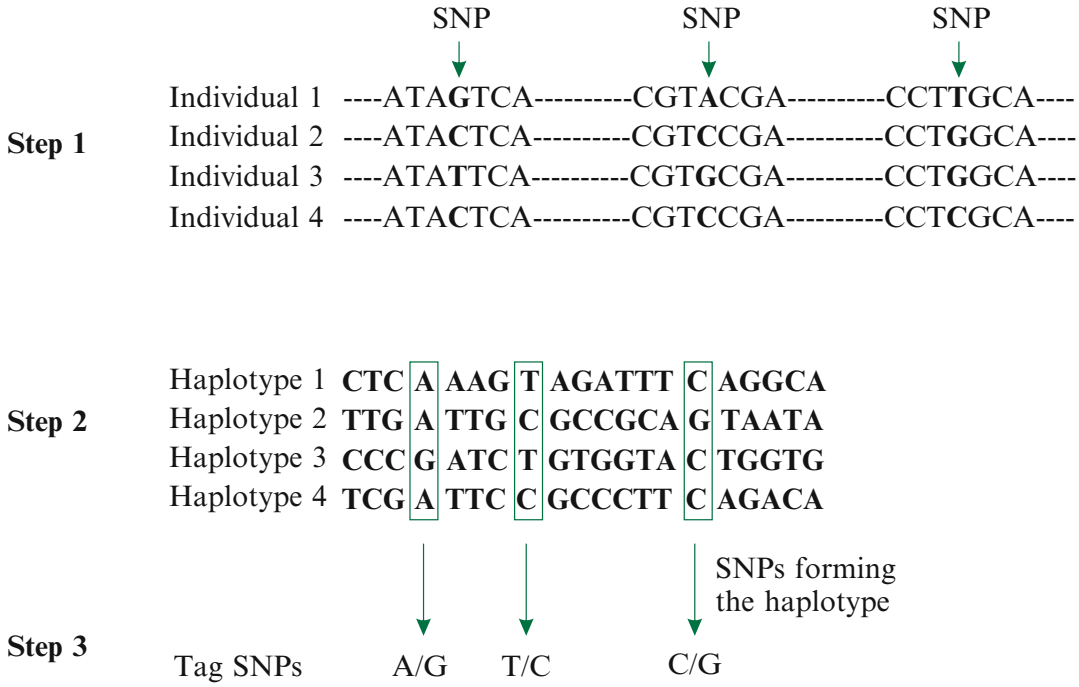


Fig. 8.9 The three-step strategy used by the HapMap project to identify tag SNPs. The first step consists of identification of the SNPs found in a genomic region. In the second step, the adjacent SNPs are organized into haplotypes. Finally, in the third step, the haplotypes are compared to identify a minimum number of SNPs that will allow dependable identification of these haplotypes (Based on http://www.hapmap.org/)

project identifies tag SNPs by a three-step process briefly outlined in Fig. 8.9. *The identification of tSNPs for different genomic regions would drastically reduce the genotyping effort required for LD analyses and AM*. Therefore, tSNP identification has been initiated in *A. thaliana*, maize, rice, potato, etc. In 2007, a catalogue of over one million SNPs was available for *Arabidopsis*, while maize HapMap had identified 1.4 million SNPs and 200,000 InDels by March, 2011. In case of rice, 259,721 polymorphic SNPs were identified by the year 2008.

8.23 Software for LD Studies

A number of software packages have been developed for LD studies and AM. Most of these packages are available for free, but some of them are commercial packages. The various packages provide a variety of functions useful for handling of animal and/or plant data, and some of them are applicable to specific designs of study. Table 8.8 briefly lists the relevant features of some of the important packages; some of them have been briefly discussed in the text as well.

8.24 Conclusions from Association Mapping Studies

1. Genome-wide AM has not been able to identify the set of genes that together will explain the total phenotypic variation in any of the quantitative traits that have been extensively investigated. The lack of knowledge about genetic basis of a large part of the heritable phenotypic variation in quantitative traits (e.g., ~75 % of heritable variation in human height) has been termed as *missing heritability*. It seems that this failure may be primarily due to the problems with design and execution of the various AM studies. For example,

analysis of a maize NAM panel having 5,000 RILs for southern leaf blight (*Cochliobolus heterostrophus*) resistance using 1.6 million HapMap SNPs identified 32 QTLs. These QTLs together explained 82 % of the phenotypic variation for the trait; this accounts for nearly all of the heritability that was estimated as 87 % (Kump et al. 2011). Thus, the use of appropriate experimental designs, sufficiently large samples for analysis, and efficient statistical methods should be able to resolve the so-called "missing heritability" issue (see also Sect. 10.7).

2. Gene interactions are involved in the control of all kinds of traits, and virtually no gene can be considered to function in isolation. Further, the expression level of a given gene is likely to affect the expression level of several other genes.
3. It was once thought that common human diseases are caused by common variant alleles of the concerned genes. But it is now suggested that each common genetic disease might be caused by many rare variant alleles rather than few common variant alleles of the concerned gene(s). But in case of plants, the variant alleles for genes controlling most of the studied traits occur in appreciable frequencies. This is not surprising since the traits studied in plants are concerned with adaptation, either in nature or under domestication. As a result, the alleles governing such traits may be expected to occur in appreciable frequencies.
4. *Surprisingly, the majority of significant associations detected by genome-wide AM studies are located in genomic regions that do not code for proteins*. For example, in plants, majority of mutations found to be associated with genetic variation in traits are located in introns, untranslated regions, and intergenic regions. There is some evidence that many of these mutations show positive associations because they are in LD with other undetected mutations in coding regions responsible for the mutant phenotypes.
5. Generally, higher levels of LD are observed in newly founded populations. Therefore, younger populations should be used for initial detection of LD. Following this, older populations can be analyzed for fine mapping of the target locus/gene.

8.25 Current Issues in Association Mapping

1. The genetic basis of the unexplained part of heritable variation in quantitative traits remains to be elucidated.
2. Considerable refinements in experimental design and statistical analyses seem to be necessary for making AM more informative and reliable.
3. Structural variations like copy number variations are now receiving greater attention in AM studies. At present, there is limited information about such variations in plants, but new generation sequencing technology is fast generating information on these variations.
4. The current experimental designs are not effective in detecting epistasis. In view of this, new algorithms for efficient detection of epistasis are being developed.
5. The issue of population structure needs to be carefully addressed during AM because a naïve use of correction for structure may lead to an increase in false negatives. In general, correction should be used when the population structure is correlated with the variation for the target trait.
6. Accurate phenotyping of individuals/lines is critical for AM. If the phenotype measurement error between individuals approaches the magnitude of phenotypic variance between them, the detection of association will be severely affected.
7. Generally, marker alleles with <5 % frequency are excluded from association analyses. This step also eliminates the chances of discovering rare alleles of genes/QTLs associated with the rare marker alleles. The possible solutions to this problem are as follows: (1) the development of biparental

populations using lines with rare marker alleles and using them for QTL mapping, (2) the use of JLAM, (3) the use of large populations, and (4) the use of novel statistical models for AM with rare marker alleles (see Gupta et al. 2014).

8.26 Future Perspectives

In view of the rapid decline in genotyping costs, the focus is likely to shift from candidate gene approach to genome-wide association studies based on complete genome sequencing of all the individuals in the sample. There will be an increasing trend of integrating gene expression data and even gene expression network information with genome-wide association studies. Increasing attention will be given to precision phenotyping, for which fast, high-throughput, and reproducible methods are being developed (Chap. 15). Further, a much greater attention will be paid to population size, and increasingly larger populations will be used to detect and map QTLs with smaller effect sizes. AM will be extended to non-model organisms, and efforts would be made to adequately address the issues like epistasis, G × E interaction effects, and phenotypic plasticity. Integration of genome-wide association, eQTL, and molecular marker data is expected to yield valuable insights into the genetic architecture of quantitative traits and also to identify genes that are likely to have been the targets of natural selection. AM may provide a much clearer picture of the architecture of QTLs and to enable identification of individual causal mutations down to the nucleotide level changes. The nucleotide changes that produce different alleles of a QTL are often referred to as *quantitative trait nucleotides* (*QTNs*). Finally, efforts should be made to develop file format and minimum data standards for deposition of all phenotype and genotype data from genome-wide AM studies in public databases for sharing among researchers and for future use.

Many quantitative traits like plant height are dynamic in that they show different patterns of development in different genotypes that may, in the end, show comparable/different values for the concerned trait. In such cases, the trait phenotype may be measured at several different time points during the development and used for AM either independently (analysis of data for a single time point at a time) or jointly (data for all the time points analyzed together). This may enable identification of different genes/QTLs governing such traits that are expressed at specific developmental stages. This approach is referred to as *functional AM* or *functional GWAS*. Finally, meta-analysis may be carried out to combine information from multiple GWASs in an effort to identify "true" positive marker-trait associations. Such analyses have been conducted in humans, and they should be initiated in plants as well.

8.27 Merits of Association Mapping

1. The populations for AM generally are samples from existing materials; this saves time, effort, and cost needed for the development of specific mapping populations.
2. The QTL-linked markers identified by AM can be directly used for MAS since they are identified from a collection of diverse germplasm/breeding lines.
3. AM has high resolution as it takes into account all the meiotic events since the origin of new allele. As a result, only those markers that are located very close to the concerned QTLs are likely to show significant LD and association with the QTLs.
4. Association analysis could identify the causal polymorphism within a gene that is responsible for the phenotypic variation in the concerned trait.
5. AM would be able to identify all the alleles of QTLs present in the population used in the study. Thus, AM would assess the entire range of diversity in the trait of interest.
6. The data on the target traits collected in earlier studies also can be used for AM.
7. The breeding populations themselves can be used for QTL discovery. Thus, QTL discovery

and mapping would be integrated with breeding activities.

8. In breeding materials, a QTL would be present in multiple genetic backgrounds. Therefore, a QTL detected in such materials would be one that is able to express itself in a range of genetic backgrounds and would be useful in breeding programs.
9. The marker genotype data generated during AM can be used for either selection of parents for hybridization or for selection of desirable segregants (pedigree selection). In *parental selection*, mixed model is used to calculate breeding values of the inbreds, which are then used as the basis for selection of superior parents for hybridization. In *pedigree selection*, the markers linked to the genes/QTLs are used for selection of the genes/QTLs being transferred, while the other markers can be used for background selection.

8.28 Limitations of Association Mapping

1. The results from AM are affected by several factors like selection history, population structure, kinship, etc., which may lead to false associations between QTLs and markers.
2. AM is based on LD, but linkage may not be the basis of significant LD between a marker and a trait (Sect. 8.16). Therefore, a joint mapping strategy that uses information about linkage between markers and genes/QTLs as well as LD between them would be more desirable.
3. In view of the much higher resolution of AM, a large number (hundreds of thousands or even millions) of markers would be required to adequately cover the entire genome. Therefore, the genotyping costs would be much greater in case of AM than in linkage mapping.
4. LD models assume constant population size over generations and the populations to be in equilibrium for genetic drift and recombination. However, these assumptions often fail leading to inaccurate estimates of physical distances between loci.
5. The rate of recombination is not uniform throughout the genome, which reduces the reliability of using LD for estimation of physical distance between loci.
6. The power of detection of marker-trait association depends on allele frequencies of the concerned gene/QTL. Low-frequency alleles have little effect on the phenotype of the concerned trait in the population as whole. Therefore, even when low-frequency alleles have large effects on the trait, they are not likely to be detected by AM.
7. The QTL detection power of AM shows a linear relationship with the magnitude of LD between the QTL and the concerned marker. A QTL with small effect can be detected only when it is in strong LD with a marker (Van Inghelandt et al. 2011).
8. In samples with strong population structure as well as relatedness, where $Q + K$ explains bulk of the phenotypic variance, the power of AM is greatly reduced.
9. The QTL effect will be underestimated when there is incomplete LD between a marker and the concerned QTL. But when functional markers become available, this problem will be eliminated (Wurschum 2012).

Questions

1. Discuss the comparative merits and limitations of linkage and association mapping.
2. Give a brief description of the procedure for association mapping, and discuss its merits and limitations.
3. Briefly describe the various populations used for association analyses, and discuss their advantages and limitations.
4. Explain the meaning of linkage disequilibrium. Briefly describe the common measures of LD and highlight their advantages and limitations.
5. "LD estimates are affected by a number of factors." Discuss the correctness of this observation in the light of relevant information.

6. Discuss the various issues relevant to testing of significance of marker-trait associations.
7. Briefly describe the various study designs proposed to tackle the problems posed by population structure and/or kinship.
8. Explain the genome-wide and candidate gene approaches for association analyses. Which of these approaches would be appropriate for a crop species with scant genomic resources and why?
9. Discuss the situations, in which association mapping will be preferable to linkage mapping.

Part IV

Applications

Marker-Assisted Selection 9

9.1 Introduction

Plant breeding involves utilization of natural as well as artificially created genetic variation for selection of superior crop genotypes that are more useful to humans. This selection is, of necessity, based on phenotype and is often referred to as *phenotypic selection*. The development of phenotype depends not only on genotype (G) but also on the prevailing environment (E) and an interaction between the two, i.e., $G \times E$ interaction (GEI). Therefore, phenotype may not always be a good indicator of genotype. The phenotypic evaluation of many traits may be either cumbersome, tedious, time-consuming (e.g., for most biochemical traits), destructive (e.g., for root traits, biomass), or dependent on specific threshold requirements (e.g., for disease resistance, lodging resistance, etc.) or may require homozygous genotypes (e.g., for recessive alleles). Further, phenotypic selection for traits like yield will not be feasible in off-season nurseries/greenhouses, which are employed for rapid generation advance. Finally, phenotypic selection for many traits, e.g., fruit and seed characteristics, has to be delayed till plant maturity. This precludes the use of selected plants for making appropriate crosses in the same generation as selection. Therefore, indirect selection for traits of interest has been a long sought after objective of plant breeders. Selection for the desirable allele of a gene/quantitative trait locus (QTL) on the basis of molecular marker(s) linked to it in the place of phenotype generated by this allele is known as *marker-assisted selection* (*MAS*). This term was first used by Beckmann and Soller (1983). *Molecular breeding*, on the other hand, is a more general term; it involves the use of molecular marker data for enhancing the effectiveness of various breeding activities, including planning and execution of breeding programs and improving selection efficiency.

In plant breeding, molecular marker data are used for the following five groups of activities: (1) characterization of germplasm; (2) diversity analyses, and selection of parents for hybridization; (3) gene introgression, gene pyramiding, and trait stacking; (4) MAS in segregating populations, including combined MAS; and (5) testing for genetic purity. MAS has been extensively used for transgene introgression (e.g., introgression of *cry* genes into maize inbred, cotton, etc., lines), germplasm conversion, introgression of genes for resistance to biotic stresses, and accumulation of QTLs having significant effects on the target trait. In addition, genomic selection (GS) scheme is designed for selection of QTLs distributed over the whole genome irrespective of their effect size (Chap. 10). There are several excellent reviews on the various aspects of MAS, including those by Hospital and Charcosset (1997), Collard et al. (2005), Kuchel et al. (2005), Eathington et al. (2007), Bernardo (2008), Collard and Mackill (2008), Jena and Mackill (2008), Xu and Crouch (2008), Gupta et al. (2010), and Jiang G-L (2013).

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_9,

9.2 Marker-Assisted Characterization of Germplasm and Genetic Purity

Molecular markers are useful in the characterization of germplasm used in breeding programs. For example, marker genotype data can be used to establish cultivar identity, select parents for hybridization, assess the genetic diversity present in germplasm collections and breeding populations, assign new inbred lines to appropriate heterotic groups, confirm the hybrid status of the F_1 seed, and determine genetic purity of seed lots. In the case of rice, SSR and STS markers are used to determine genetic purity of hybrid seed, which is much simpler and faster and less cumbersome than the standard grow-out tests. Molecular markers are being increasingly used to characterize genetic resources, and this information would assist the breeders in selection of suitable accessions as parents for hybridization programs. In the case of heterosis breeding, identification of the outstanding single crosses is based on field evaluation that is quite demanding and very expensive. The results from different studies on the effectiveness of predicting heterosis on the basis of marker genotype data range from discouraging to promising. However, marker data are useful in defining heterotic groups as well as assigning new inbreds to proper heterotic groups. Finally, marker data can be used to identify genomic regions, e.g., those harboring QTLs, that should be further analyzed and/or subjected to selection for improvement in the traits associated with these genomic regions.

9.3 Marker-Assisted Backcrossing

In *backcross breeding*, a useful trait is transferred from a donor parent (DP) into a recurrent parent (RP), which is a superior variety deficient in this trait. The trait transferred from the DP is generally referred to as *target trait* or *desired trait*. The F_1 from cross between the DP and RP and the subsequent progeny are backcrossed to the RP. As a result, the DP genome is progressively replaced by the RP genome. Each backcross to the RP reduces the amount of DP genome to one-half of that present in the previous generation. As a result, an average of only ~1.6 % of the DP genome remains in the backcross progeny after five backcrosses; this proportion would be <0.4 % after seven backcrosses (Allard 1960). These values are the average amount of DP genome retained in the progeny; the actual amounts present in the different individuals of a given generation would vary considerably. Further, these values will be obtained without any selection for the RP plant type. Therefore, the values would be higher than these when selection for the RP phenotype is practiced during the backcross generations. Generally, at the end of five to six backcrosses, the progeny are selfed (or sibmated), and the progeny plants similar to the RP and homozygous for the target gene/QTL are selected. This selected line will be almost identical to the RP, except for the transferred trait. This line may be released for commercial cultivation as an improved version of the RP. The gene/QTL being transferred from the DP must be maintained by a rigorous phenotypic selection or else it would be rapidly replaced by the RP allele. Since the net result of a backcross program is the transfer of target gene(s)/QTL(s) into the RP genotype, this process is often referred to as *gene/QTL introgression*.

Backcross breeding has been widely used because in each crop there are some varieties that are popular with the farmers. Therefore, farmers are more likely to accept an improved version of such a variety than an entirely new variety. Similarly, millers and industries using the crop produced may be reluctant to change to an entirely new variety because the processing of new varieties may need to be standardized afresh. Finally, backcross method will continue to be used for transgene introgression because in many crop species either land races or obsolete varieties have to be used for genetic transformation in view of technical difficulties. In such cases, backcross program must be used to transfer the transgenes from the agronomically inferior transgenic lines (used as DP) into the elite varieties (serving as RP).

Backcross breeding aims to achieve the following: (1) transfer of the desired trait from a donor parent into a recurrent parent, (2) maximum recovery of the recurrent parent genome, and (3) ideally, complete elimination of the donor genome, leaving only the target gene/QTL. Molecular markers can be used to achieve all the three objectives. (1) Markers linked to the target gene/QTL enable indirect selection for the gene/QTL (*foreground selection*; Fig. 9.1a). (2) Codominant markers distributed throughout the genome enable selection of plants having the highest proportion of the recurrent parent genome (*background selection*; Fig. 9.1b). (3) Finally, codominant markers located on either side of the target gene can be used to select for rare recombinants that do not have the donor genome beyond these markers (*recombinant selection*; Fig. 9.1c). A backcross program based on markers is known as *marker-assisted backcrossing* (*MABC*). *Molecular markers have been generally used for foreground selection*, *often for background profiling*, *sometimes for background selection*, *and only occasionally for recombinant selection*. When markers are used for foreground as well as background selections, the backcross scheme is often called *complete line conversion*, *full MAS*, or simply *MABC*. The markers for foreground selection should be tightly linked with the target gene/QTL. But the markers used for recombinant selection must flank the target gene/QTL beyond, but close to, the markers used for foreground selection. Finally, the markers used for background selection should be distributed over the whole genome and should be polymorphic between the DP and RP.

9.3.1 Foreground Selection

Indirect selection for the target gene/QTL on the basis of linked marker genotype was proposed by Tanksley (1983) and was called *foreground selection* by Hospital and Charcosset (1997). Foreground selection will be highly preferable to phenotypic selection when phenotypic evaluation for the target trait is problematic for any one or more of the several reasons (Sect. 9.1, Table 9.1). In addition, it will greatly facilitate multiple QTL transfer and multitrait introgression and would replace disease tests during selection. Finally, it will be indispensable for combining oligogenic and polygenic resistances to plant diseases and insect pests. The effectiveness of foreground selection will depend primarily on the genetic distance between the marker and the target gene/QTL: the closer is the marker to the gene/QTL, the greater will be the efficiency of foreground selection. The genetic distance between a marker and a gene/QTL indicates the frequency of progeny, in which the association between the marker and the target gene/QTL allele is expected to change due to recombination. For example, when a marker and a gene/QTL are separated by 5 cM, the association between their alleles would change in about 5 % of the progeny. As a result, the marker genotype will incorrectly predict the gene/QTL allele in these 5 % of plants. Therefore, the distance between the marker and the target gene/QTL should be less than 5 cM; ideally, the marker should be allele specific or, at least, gene-based (Sect. 2.3). Whenever the markers are more than 5 cM away from the gene/QTL, a pair of *flanking markers*, i.e., one marker located on either side of the gene/QTL, should be used. This would minimize the chances of incorrect prediction of the allelic state of the gene/QTL by the marker genotype. The advantage of flanking markers is as follows. Suppose, marker *A* is located at 7 cM on one side of a gene, while marker *B* is situated at 10 cM on the other side of the same gene. If either marker *A* or *B* alone were used for foreground selection, the frequency of incorrect diagnosis would be either 7 (for marker *A*) or 10 (for marker *B*) percent. In contrast, when both the markers *A* and *B* are used together, the frequency of incorrect selection would be only 0.7 % ($= 0.07 \times 0.10 \times 100$). This is because incorrect diagnosis would occur only when there is simultaneous crossing over between marker *A* and the gene and between the gene and marker *B*.

9.3.2 Background Selection

The use of molecular markers to facilitate the recovery of recurrent parent genome was proposed by Tanksley and coworkers (Tanksley

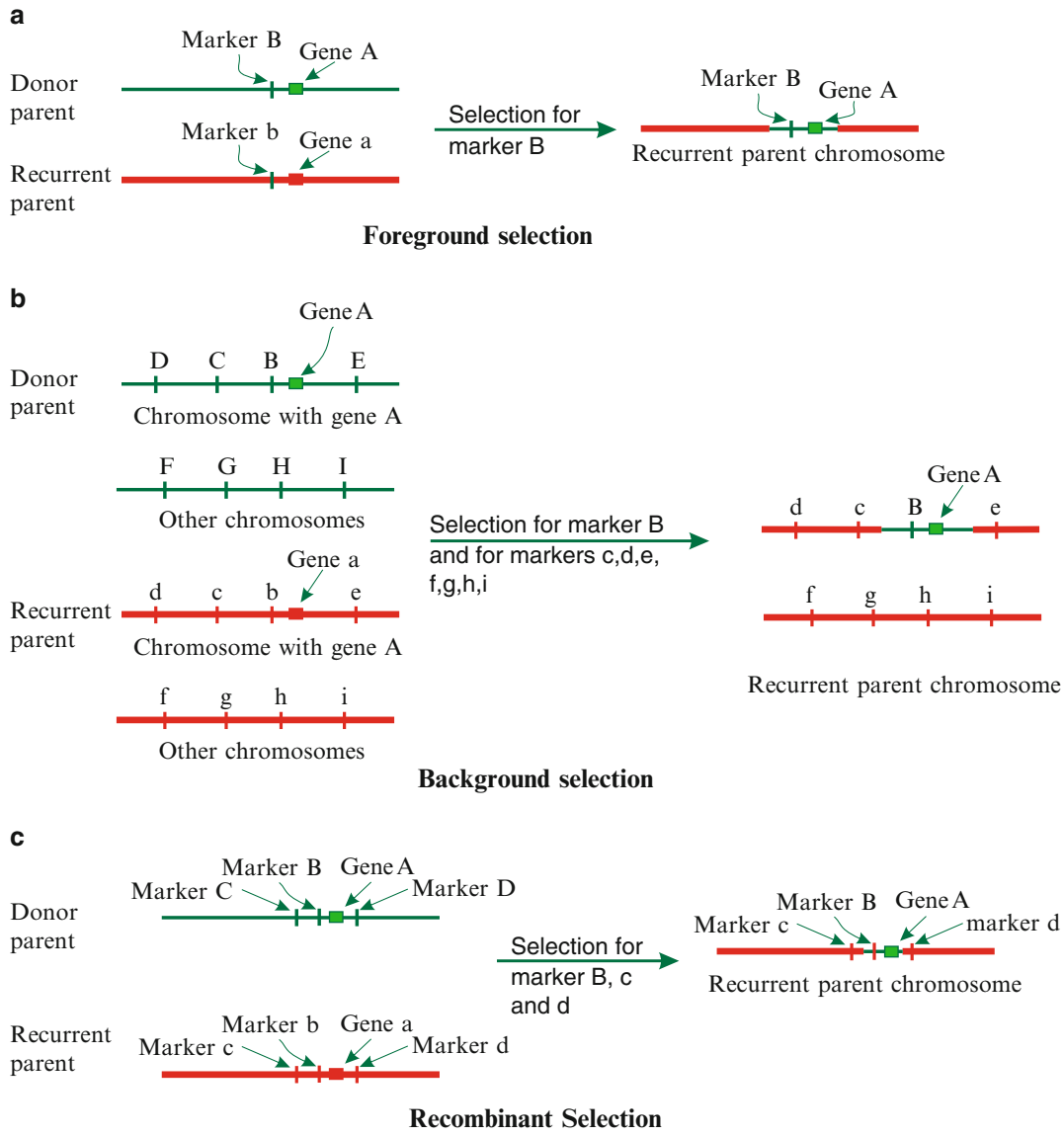


Fig. 9.1 A simplified and idealized representation of (**a**) foreground, (**b**) background, and (**c**) recombinant selections. Foreground selection focuses on selection for the target gene, i.e., the gene being introgressed. It is achieved by selection for the marker *B* tightly linked to the target gene. The background selection, on the other hand, is directed at the recovery of the recurrent parent genotype and is based on markers distributed over the entire genome. Finally, (**c**) recombinant selection aims at elimination of the donor parent genome flanking the target gene. It selects for the recurrent parent markers flanking the target gene at a short distance. Thus, foreground and recombinant selections relate to the chromosome having the target gene, while background selection is concerned with all the chromosomes of the genome. The use of symbols A, B, C, etc. and a, b, c, etc. for the marker alleles of the DP and RP, respectively, is arbitrary, and it should not be taken to indicate either dominance relationship or favorable/unfavorable action

Table 9.1 Some illustrative examples where MAS is preferable to phenotypic selection

Trait/situation	Discouraging feature of phenotypic selection	Reference
Foreground selection		
Soybean cyst nematode	Time-consuming, high cost	Young (1999), Bernardo (2008)
Cereal cyst nematode (wheat)	Slow speed, very high cost	Brennan and Martin (2007)
Crown rot resistance (wheat)	Slow speed, high to very high cost	Brennan and Martin (2007)
Small-scale quality tests (wheat)	Slow to moderate speed, medium to very high cost	Brennan and Martin (2007)
Amylose content (rice)	Reliable estimation is cumbersome	Gopalakrishnan et al. (2008)
"Quality protein" trait governed by o_2 mutant allele (maize)	Expensive biochemical assay, recessive inheritance	Babu et al. (2005)
Provitamin A (maize)	Time consuming, high cost	Muthusamy et al. (2014)
Background selection		
Recovery of recurrent parent genome	Slow progress (82 % recovery in BC_4F_7) as compared to MAS (97 % recovery in $BC_2F_{2:3}$)	Randhawa et al. (2009)
Recombinant selection	Very poor effectiveness	Young and Tanksley (1989)

et al. 1989; Young and Tanksley 1989), and it was called *background selection* by Hospital and Charcosset (1997). It has been reported that different NILs developed independently from the same cross by selection for the same marker/gene usually contain different lengths of the donor genome flanking the marker/gene. Let us assume that the length of the chromosome having the target gene/QTL, referred to here as *the "marked" chromosome*, is L_M Morgans (1 Morgan = 100 cM or centimorgans), the marker and the target gene are located in the center of this chromosome, the number of backcrosses for NIL production is b, and, finally, the progeny are selfed to obtain the NILs. The average proportion of the "marked" donor chromosome retained along with the marker in the NILs can be estimated by a formula given by Hanson (1959) when there is no background selection during the backcross and selfing generations. It can be shown that when the value of $t \times L_M$ is much larger than one, which usually will be the case, the proportion of the "marked" chromosome retained in the NILs is approximated by $2/tL_M$; here, $t = b + 1$. However, the marker and the target gene may be located at a position other than the central location of the "marked" chromosome. Further, the DP chromosome segments may be integrated at some additional position(s) of the "marked" RP chromosome. It has been shown that the above two factors have only small effect, and $\sim 2/tL_M$ seems to be a very good approximation of the proportion of DP genome retained in the "marked" RP chromosome (see Lynch and Walsh 1998).

The donor genome would also be retained in chromosomes other than the "marked" RP chromosome, i.e., in the "unmarked" RP chromosomes. Let us suppose that one "unmarked" chromosome is L_i Morgans long, and after b backcrosses the progeny are selfed to obtain NILs. The average proportion of the DP genome retained in this "unmarked" chromosome of the NILs will equal $1/2^t$, where $t = b + 1$. It can be shown that the proportion of the total donor genome retained in all the "unmarked" chromosomes of the species will equal $(L\text{-}L_M)/2^t$, where L is the total length in Morgans of all the chromosome of the species (Stam and Zeven 1981). Let us suppose that a species has ten chromosomes of 1 Morgan length each, the marker is located in the center of chromosome 5, and the progeny after the fifth backcross are selfed to obtain NILs. The expected average proportion of donor DNA retained in chromosome 5 of the NILs will be approximately $2/tL_M$ with a variance of approximately $2/(tL_M)^2$. In this case,

$t = 6$ (=5 + 1) and $L_M = 1$ Morgan. Therefore, the expected proportion of donor DNA retained in chromosome 5 will be ~0.333 (= 2/6 × 1) with variance and standard deviation (SD) of 0.056 and 0.237, respectively. In terms of cM, the chromosome length retained will be 33.3 cM (= 0.333 × 100 cM) with SD of 23.7 cM. The proportion of donor genome retained in any one of the "unmarked" chromosomes will be $1/2^t$, i.e., 1/64 = 0.016, which will come to 1.6 cM. Thus, the total donor genome retained in all the "unmarked" chromosomes will be just over 14 cM [=(10–1)/2^6 Morgans], and the total DP genome retained in all the chromosomes of the NIL will be 47.3 cM (= 33.3 + 14 cM) (see Lynch and Walsh 1998).

The above estimate of the length of donor genome flanking the target gene/QTL retained in the NILs is based on the assumption that there is normal pairing and crossing over between the DP and RP chromosomes. But in many backcross programs, particularly those involving interspecific gene transfers, chromosome pairing and recombination would be greatly reduced. In such cases, the lengths of donor chromosomes flanking the target genes/QTLs retained in the NILs will be much longer than the above estimates. Ideally, the markers used for background selection should be sufficiently dense and almost evenly distributed throughout the genome to facilitate effective selection for the RP genome. The RP, as a rule, is a popular variety of the area well known for its excellent performance, while the DP is ordinarily agronomically inferior to the RP or it may even be a related species. Therefore, it is important that the desired gene/QTL from the DP is transferred into the RP with the minimum amount of DP genome. This would minimize the risk of transferring DP genes having negative effect on the performance of the lines derived from the program. *It has been shown that two to four backcrosses coupled with background selection can recover the recurrent parent genotype to the same extent as is achieved with six backcrosses combined with phenotypic selection for RP phenotype*. Background selection may also be used in a pedigree program to ensure the recovery of a specified level of the genome from one of the parents that may have more desirable features than the other parent(s). Molecular markers closely linked to undesirable alleles of known genes/QTLs can be used to select against these alleles; this is often called *negative selection*.

9.3.3 Recombinant Selection

The term *recombinant selection* (Collard and Mackill 2008) describes a special type of background selection that aims to eliminate the DP genome flanking the target gene/QTL (Young and Tanksley 1989). Recombinant selection ensures the transfer of the target gene/QTL with a minimum of the DP genome to minimize linkage drag. *Linkage drag* is the negative effect of genes linked to the target gene/QTL on the performance of lines produced by gene transfers. Often it is very difficult to eliminate undesirable linked genes in backcross programs. A surprisingly large amount of donor genome may remain in lines derived from several backcrosses. For example, tomato cultivars developed by transfer of *Tm2* gene from *Lycopersicon peruvianum* contained DP genome segment as large as 4 cM even after 20 backcrosses. It is remarkable that one cultivar derived after 11 backcrosses had the entire chromosome arm carrying the *Tm2* gene (>51 cM). The strategy of recombinant selection is based on markers located at <5 cM, preferably ~1 cM, on either side of the marker(s) employed for foreground selection. These markers permit selection for such recombinants that have the target gene/QTL but lack the DP genome beyond the marker(s) used for foreground selection. Recombinant selection can save several generations of backcrossing without imposing a high cost.

Suppose marker *M* is tightly linked to the target gene/QTL or is a gene-based marker, and markers *A* and *B* are at 1 cM on either side of the marker *M*. Thus, the DP and RP genotypes would be *AMB* and *amb*, respectively. It may be emphasized that the use of symbols *A*, *B* and *a*, *b* for the marker alleles of the DP and the RP, respectively, is arbitrary, and it does not reflect either dominance relationship or, more particularly, favorable/unfavorable effects.

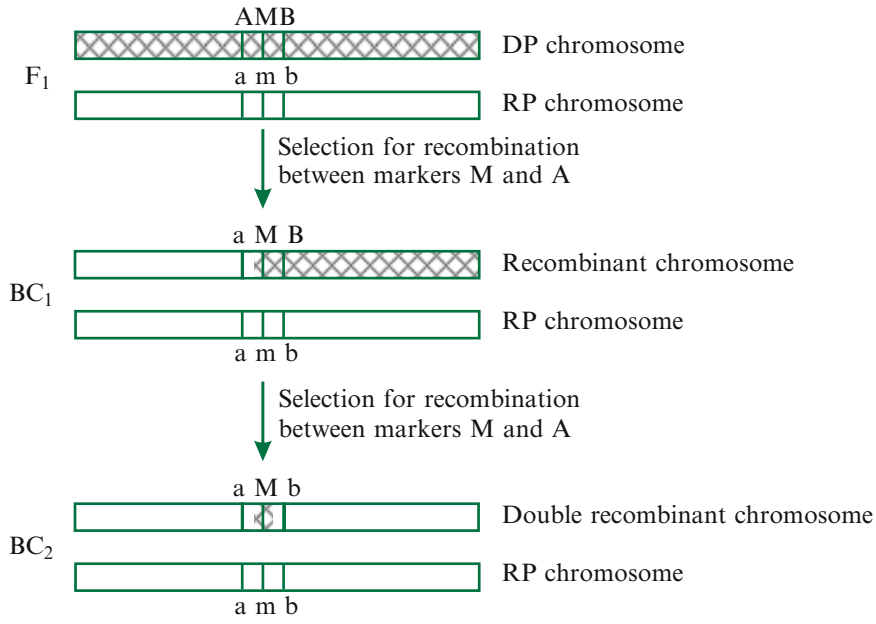


Fig. 9.2 Selection of the double recombinant *aMb* in two successive generations. In the first generation (BC_1), selection is done for single recombinants, i.e., either *aMB* (as shown in this figure) or for *AMb*, whichever is recoverable. In the next generation (BC_2), selection is done for the other single recombinant, e.g., between *M* and *b* (as depicted in this figure) or between *a* and *M*. Assuming that the markers *a* and *b* are at 1 cM from the target gene, ~185 and ~370 plants will have to be screened in BC_1 and BC_2, respectively, for a 95 % probability that at least one plant will be the desired single recombinant. In contrast, ~30,000 plants will have to be analyzed for isolation of the double recombinant in a single generation (Based on Tanksley et al. 1989)

Recombinant selection aims to isolate the double recombinant *aMb*, which would have only ~2 cM of the DP genome around the target gene/QTL. There are two approaches for recombinant selection, viz., selection of *aMb* in a single generation and double recombinant selection in two generations. *Recombinant selection in one generation* aims at isolating at least one *aMb* plant in a single, e.g., BC_1, generation. This would require screening of a very large number of backcross progeny because the frequency of two simultaneous crossovers within the 2 cM region flanking the marker *M* will be very low ($0.01 \times 0.01 = 0.0001$ assuming no interference), i.e., merely 1 in 10,000 plants. However, considerable interference may be expected within the short region of 2 cM, leading to a much lower frequency of the double crossover than expected. It has been estimated that approximately $3/(C_1 \times C_2)$ plants will have to be scored for a 95 % probability of recovering one *aMb* plant. Here, C_1 and C_2 are the frequencies of recombination between the marker *M* and the flanking markers *a* and *b*, respectively. Thus, when $C_1 = C_2 = 0.01$, about 30,000 plants (and not 10,000) will need to be scored for recovering one double recombinant with 95 % probability. It may not be feasible to produce this large number of backcross progeny in a breeding program (Lynch and Walsh 1998).

The *recombinant selection in two sequential generations* seems to be a far more desirable strategy. In this approach, at least one plant with recombination between either markers *a* and *M* (shown in Fig. 9.2) or markers *M* and *b* is selected in the first generation, say, in BC_1. Then in BC_2, at least one plant with recombination between marker *M* and the other flanking marker (marker *b* in Fig. 9.2) is isolated. Thus, two generations will be required to recover the double recombinant *aMb*, but the number of backcross progeny required to be screened will be rather small. It has been shown that about $1.85/(C_1$ or $C_2)$ plants will have to be scored in

BC_1 for finding a single recombinant plant with 95 % probability. But approximately 3.7/(C_1 or C_2) plants will have to be scored in BC_2 for finding a single desired recombinant plant with 95 % probability. Thus, in the present case, about 185 plants (= 1.85/0.01) will need to be scored in the first generation, and approximately 370 plants (= 3.7/0.01) will have to be scored in the second generation when $C_1 = C_2 = 0.01$. Therefore, a total of only 555 plants will have to be screened in the two-generation approach for finding one *aMb* plant with 95 % probability. This number is <2 % of the number of plants (30,000 plants) required in the one-generation method (Lynch and Walsh 1998).

9.3.4 A Four-Step Comprehensive Selection Strategy

Frisch et al. (1999) carried out a computer simulation study in maize ($n = 10$; markers spaced at 20 cM). The target gene was selected on the basis of either a reliable phenotype score or the genotype of a marker completely linked with the target gene. The following four-stage sampling strategy seemed to be the most efficient for both foreground and background selections:

1. Selection of plants carrying the desired allele of the target gene
2. Then, selection of plants homozygous for the RP marker alleles at loci flanking the target gene
3. Now, selection of plants homozygous for the RP alleles at the remaining marker loci in the chromosome having the target gene
4. Finally, selection of at least one plant homozygous for the RP alleles at the maximum number of marker loci

They estimated that the use of above scheme with 50–100 plants in each backcross generation would recover about 96 % of the RP genome with 90 % probability in BC_3 as against the expected 93.8 % without background selection. Further, markers spaced at 20 cM were optimum for background selection. This scheme aims to select for the double recombinant in a single generation. But it would be highly desirable to select the double recombinant in two successive generations for the reasons explained in the Sect. 9.3.3.

9.4 A Theory for Background Selection During MABC

The efficiency of background selection seems to depend primarily on marker density and distribution, population size, and selection strategy. Frisch and Melchinger (2005) developed a selection theory for MABC taking into account the combined effects of these and several other factors. The methods of their theory have been implemented in the software "Plabsoft" (Maurer et al. 2004). This theory estimates the predicted response to selection in a backcross generation by taking into consideration the number and the lengths of chromosomes, number and distribution of markers, location(s) of target gene(s), and intensity of selection. The predicted response to selection increases with marker number only up to a limit depending on the number and the lengths of chromosomes. In contrast, selection response increased with population size. Thus, for a fixed number of marker data points, selection was more efficient with larger populations than with higher marker densities. Further, the larger the genome size, the larger should be population size. The individuals having DP marker alleles at the smallest number of loci should be selected in the backcross generations for both backcrossing and selfing when the markers are evenly distributed in the genome. But when the marker distribution is uneven, plants having DP alleles at identical number of marker loci may differ for the expected proportion of DP genome present in them (*first criterion*) and/or the proportion of DP genome expected to be present in the best progeny obtained from their backcross with the RP (*second criterion*). *They concluded that in every case, plants having DP alleles at the smallest number of marker loci should be selected*. But the selected plant should have the lowest value for the first criterion if it were to be selfed, while it should have the lowest value for the second criterion if it were to be backcrossed. This theory was developed for introgression of

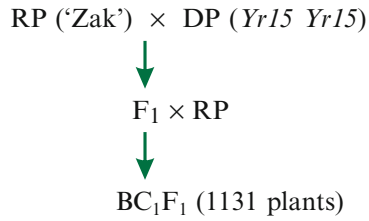
oligogenes, but the authors felt it would be equally valid for QTL introgression.

9.5 MABC for Transfer of Oligogenic Traits

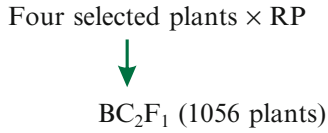
The first application of MAS involved the use of isozyme markers for transfer of oligogenic traits from unadapted germplasm into cultivated varieties (Tanksley and Rick 1980). Beckmann and Soller (1983) provided the first account of the use of DNA (RFLP) markers for crop improvement and theoretical considerations related to MABC for qualitative trait improvement. An early example of MAS for a difficult-to-evaluate trait concerns cyst nematode (*Heterodera glycines*) resistance in soybean (*Glycine max*). The phenotypic assay for this trait requires large greenhouse space and takes 5 weeks time; in addition, up to 10-h time is needed for evaluation of 100 plants. A concerted effort identified the SSR marker *Satt309* at 1–2 cM from the locus *rhg1*, which was found to be the major contributor to soybean cyst nematode (SCN) resistance. Marker *Satt309* is highly effective in identifying SCN susceptible plants/lines. But the predicted resistant plants/lines show variable resistance most likely due to the involvement of other loci in SCN resistance (Young 1999). MAS is routinely used for *rhg1* and *rhg4*, another locus for SCN resistance. The MAS for these loci requires 1–2 days and costs US $ 0.25–1.0 per sample. In contrast, the phenotypic evaluation costs US $ 1.5–5.0 per sample (Bernardo 2008). MABC has been successfully used for introgression of a large number of oligogenic traits. Some of the illustrative situations/examples where MAS is preferable to phenotypic selection are summarized in Table 9.1. In general, MAS is faster and relatively cheaper. In case a desirable trait is governed by the recessive allele of a gene, MAS allows backcrossing in the successive generations. Marker-assisted background selection is far more effective than phenotypic selection for the recurrent parent type. Finally, marker-aided recombinant selection is highly effective, while phenotypic selection was generally ineffective.

A modification of the four-step selection scheme (Sect. 9.3.4) was used to transfer wheat stripe rust (caused by *Puccinia striiformis* f. sp. *tritici*) resistance gene *Yr15* with 97 % RP genome recovery in BC_2F_3 (Randhawa et al. 2009). The optimum density of markers for background selection was 4 cM in the "marked" chromosome and ~12 cM in the gene-rich regions of the "unmarked" chromosomes. In BC_1F_1, 1,131 plants were screened for stripe rust resistance in the seedling stage (phenotypic selection), and the 156 rust-resistant plants were genotyped with 115 markers. Four plants had recombination between one flanking marker and the target gene and were homozygous for the 27 markers of the "marked" chromosome (background selection); they were used for backcrossing (Fig. 9.3). In BC_2F_1, 1,056 plants were screened for stripe rust resistance, and the 205 resistant plants were genotyped with 205 markers. Five of these plants were double recombinants, and one of them was homozygous at 188 of the 205 markers. All the BC_2F_2 progeny from this plant were selfed to produce BC_2F_3. In BC_2F_3, 37 families were homogeneous for the *Yr15* gene; they were genotyped with 251 markers, and one plant showed 97 % RP genome recovery. In contrast, selection for RP phenotype could recover only 82 % of the RP genome by BC_4F_7.

Amylose content is the key determinant of cooking and processing qualities of rice grains; this trait shows complex inheritance, but the *waxy* gene encoding granule-bound starch synthase has the major influence. The *waxy* locus was traditionally considered to have three alleles, viz., the recessive allele *wx* (produces glutinous/waxy endosperm with 0–5 % amylose) and the dominant alleles *Wxa* and *Wxb* (low to high amylose contents; *Wxa* produces higher amylose contents than *Wxb*). Phenotypic selection for this trait during segregating generations is impractical since a reliable estimation of amylose content is cumbersome. MAS for amylose content was facilitated by the development of a gene-based SSR marker, now renamed as



Step 1. Evaluation for stripe rust resistance in seedling stage.
Step 2. 156 resistant plants identified and genotyped for 115 markers.
Four plants with recombination between *Yr15* and one flanking marker selected
Step 3. These four plants were homozygous for the 27 markers of the 'marked' chromosome



Step 1. Evaluation for stripe rust resistance in seedling stage
Step 2. The 205 resistant plants genotyped with 205 markers;
5 plants were double recombinants
Step 4. One double recombinant homozygous at 188 of the 205 markers was selfed

BC_2F_3 (150 families*)

Step 1. Families homozygous for Yr15 selected (37 families*)
Step 4a. Selected families genotyped with 251 markers. Family with the maximum RP genome recovery selected
Step 4b. From the selected family, plant with the maximum (97%) RP genome recovery
nt""t

The selected line (WA8059) multiplied

Fig. 9.3 The four-step MABC procedure followed by Randhawa et al. (2009) to transfer stripe rust resistance gene *Yr15*. Step 3 consists of selection of plants homozygous for the markers on the "marked" chromosome. *, individual plant progenies

RM190, based on a (CT)n repeat present in exon 1 of the untranslated region of *waxy* gene. *RM190* has multiple alleles with 8–21 copies of the CT repeat unit; it was believed that the number of CT repeats could be used to classify most rice cultivars into the three (low, intermediate, and high) amylose classes. This marker has been used for selection of the *waxy* allele (Gopalakrishnan et al. 2008).

Molecular markers are very useful for the transfer of recessive alleles of oligogenes. For example, the *opaque-2* (*o2*) recessive mutant allele in maize is responsible for improved contents of lysine and tryptophan, the two essential amino acids deficient in cereals. Extensive efforts are being made to introgress the *o2* allele into elite maize inbred lines. But this is problematic since each backcross generation needs to be selfed as *o2* is recessive, and phenotyping is time-consuming and expensive. Three SSR markers, viz., *umc1066*, *phi057*, and *phi112*, are located within the *o2* allele and are used for

foreground selection. This permits backcrosses to be made in succession and the use of off-season nursery/greenhouse facilities for rapid advance of generations. The MABC for introgression of *o2* allele is simple, rapid, efficient, and cost-effective. MABC for *o2* allele is being used extensively in maize breeding programs (see Babu et al. 2004, 2005; Gupta et al. 2013; Babu and Prasanna 2014).

9.6 MABC for Transfer of Quantitative Trait Loci

Backcross procedure based on phenotypic selection has been used to transfer quantitative traits with high heritability (Allard 1960). Molecular markers can be used to identify and map the QTLs involved in the control of quantitative traits (Chap. 7). In general, QTL locations are not precisely known, and the confidence intervals of QTLs are, usually, 10 cM or longer. Therefore, it is necessary to use two or even more markers located within the estimated QTL region for foreground selection. Many quantitative traits of economic importance, including yield, are governed by several QTLs; in such cases, more than one QTL may need to be transferred since a single QTL may not lead to a useful improvement in the trait. This would increase the population size required in the backcross generations and would reduce the chances of background selection. Ordinarily, QTLs with large effects are used for introgression. In addition, the effects of introgressed QTLs in the end products have to be evaluated through elaborate field trials. Simulation studies revealed that at least three markers, at optimized locations with respect to the QTL position, should be used for foreground selection for each QTL, especially when QTL confidence intervals are 20 cM or longer. This approach should support introgression of up to four unlinked QTLs with population size of a few hundred. A larger number of QTLs can be handled when the QTLs are linked, larger population sizes are used, and/or the QTL positions are more precisely known. Further, background selection in the plants retained after foreground selection could save up to two generations of backcrossing. Finally, simultaneous, but separate, introgression of individual QTLs and pyramiding them at the end of the program is preferable to their introgression together in a single backcross program (Hospital and Charcosset 1997).

Application of MAS for QTLs is hampered by QTL × environment interaction, nontransferability of marker-QTL linkage across breeding populations, strong QTL × genetic background interaction, nonavailability of QTLs with major effect on the target trait, and deficiencies in QTL detection and mapping (Xu 2010; Jiang 2013). Often the level of expression of introgressed QTLs in the new genetic background may be markedly lower than expected from earlier QTL analyses. For example, MABC was used to introgress five chromosomal regions known to contain QTLs for acylsugar accumulation from wild tomato (*Lycopersicon pennellii*) into the cultivated tomato. Acylsugars are known to confer multiple pest resistance. The level of acylsugar accumulation in BC_3F_2 plants having all the five target genomic regions was lower than that in the F_1 hybrid (Lawson et al. 1997). This may have resulted from several reasons, including interaction of the introgressed QTLs with QTLs present in the RP and the presence of QTLs with negative effects in the RP. It is also likely that different QTLs may behave differently during introgression: some QTLs may show stable expression, some others may express in some genetic backgrounds/environments but not in others, while some others may fail to express in all the genetic backgrounds/environments. One possible reason for the last situation may be an incorrect identification of the concerned QTLs, i.e., the QTLs may be "false" or "unreal." Alternatively, the concerned QTL region may contain one QTL with positive effect and another one with negative effect so that their effects cancel each other out and become nonsignificant (Shen et al. 2001).

It has been suggested that QTLs governing relatively simple traits like flowering time are much more likely to show expected expression following introgression than those involved in

the control of more complex traits like yield. The logic for this suggestion is as follows: when a trait is governed by a small number of QTLs, the QTL effect size is likely to be larger, chances of most of the QTLs being detected in a biparental mapping population would be greater, and the likelihood of unexpected QTL × QTL interaction during QTL introgression would be much smaller than when the trait is governed by a large number of QTLs. In one study on QTL introgression in maize, the yield QTLs were far more sensitive to the genetic background and to G × E interaction than the QTLs for silking date (Bouchez et al. 2002). But there are reports of successful transfer of yield QTLs in barley (Schmierer et al. 2004). Thus, it may be more reasonable to suggest that QTL expression in different genetic backgrounds would largely depend on the specific QTLs and the genetic backgrounds involved rather than on whether the traits affected by them are complex or simple.

When several QTLs are being introgressed simultaneously, the screening of the segregating populations may be done in three steps as follows. In the *first step*, the population is subjected to MAS for a subset of the QTLs, e.g., for three of the five QTLs, being introgressed. The plants selected in the first step are screened with the markers for the remaining QTLs in the *second step*. The *third step* consists of background selection among the plants retained after the second step. In addition, the recurrent parent genomic regions having desirable genes/QTLs may be fixed early in the introgression program (Ribaut and Ragot 2007). This approach is expected to considerably reduce the marker genotyping work and ensure the recovery of the desired genomic regions of the RP. This scheme was used to successfully introgress five QTLs for drought resistance in maize. The resulting introgression lines showed improved drought adaptation without any yield penalty under well-watered condition.

The failure of QTLs to produce the expected phenotypic effects in a new genetic background may be due to the following reasons. The introgressed QTL is likely to exert only additive genetic effect, while its epistatic effects would be modified or even eliminated. This is more likely when a QTL is introgressed from an unadapted germplasm or from a line adapted to an environment different from that to which the RP is adapted. QTLs for complex, low heritability traits may be more prone to reduced expression than those for high heritability traits. Therefore, where feasible, phenotypic selection may be combined with MAS during introgression of QTLs for low heritability traits. In some cases, reduced/lack of QTL expression may be due to QTL × environment interaction. There is evidence that many QTLs have multiple alleles, e.g., three to four alleles of each QTL for certain fruit quality traits in tomato. The existence of multiple alleles may also contribute to the variable expression of the QTLs during introgression. Further, recombination may take place in the large genomic region representing a QTL; this may modify the effect of the concerned QTL. Finally, previously undetected QTLs may interfere with the improvement in the target trait. But in some cases, QTLs may interact with each other to produce a more desirable phenotype.

There is evidence that MAS can be used to enhance the expression of heterosis. Maize inbreds B73 and Mo17 represent the two most heterotic groups in maize. Linkage mapping in a population derived from the cross B73 × Mo17 showed that QTLs for grain yield were located on all chromosomes, except chromosome 6. Further, heterozygotes at each of these QTLs, except one, were superior to the homozygotes, suggesting that they contributed to heterosis for yield. Results from two other studies had revealed that six genomic regions in each of the inbreds Tx303 and Oh43 could contribute QTLs to B73 and Mo17, respectively, that might enhance the level of heterosis in the cross B73 × Mo17. MABC (three backcross and two selfing generations) was used to transfer the concerned genomic regions from Tx303 and Oh43 into the inbreds B73 and Mo17, respectively, to obtain "enhanced B73" and "enhanced Mo17" lines. The "enhanced B73" × "enhanced Mo17" hybrids yielded 8–10 % more than the original hybrid (B73 × Mo17) and the best "check" hybrids. The best "enhanced" inbred lines had only two to four of the genomic regions from the

DPs, and introgression of all the six regions was not desirable (Stuber et al. 1999).

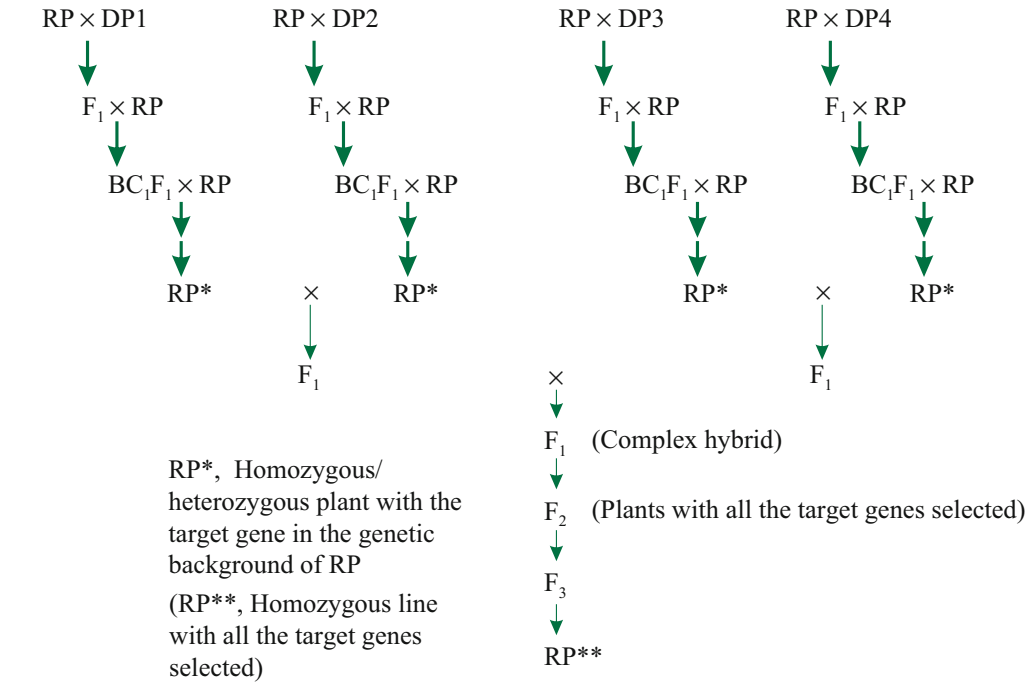
9.7 MABC for Gene Pyramiding

The concept of gene pyramiding was proposed by Nelson (1978) to develop crop varieties with durable resistance to diseases by bringing together few to several different oligogenes for resistance to the given disease. The basis for this proposal was that a host variety having two or more different oligogenes for resistance to a pathogen can be attacked by a race or pathotype of the pathogen that is virulent to all the resistance genes. This can happen only when all the concerned avirulence alleles mutate simultaneously in the same pathogen cell, successively in the same pathotype/race, or independently in different pathotypes/races, which then hybridize to generate a pathotype/race with all the virulence alleles. The probability of occurrence of any of the above events is very low. Therefore, the resistance of this variety would be far more durable than that of varieties having a single resistance gene. In addition, it was hoped that even when the pathogen was able to defeat all the resistance genes, the residual effects of these genes might still provide some protection against the pathogen; this seems to happen at least in some host–pathogen systems (see Melchinger 1990). In general terms, *gene pyramiding* may be used to describe bringing together two or more genes controlling a single trait in a single line/variety. Gene pyramiding is relatively straightforward when the same DP contributes all the genes. But when two or more DPs have to be used, relatively simple strategies can be used for gene pyramiding. Often, genes governing two or more different traits are introgressed into a single RP; this should be called *multitrait introgression* in the place of gene pyramiding.

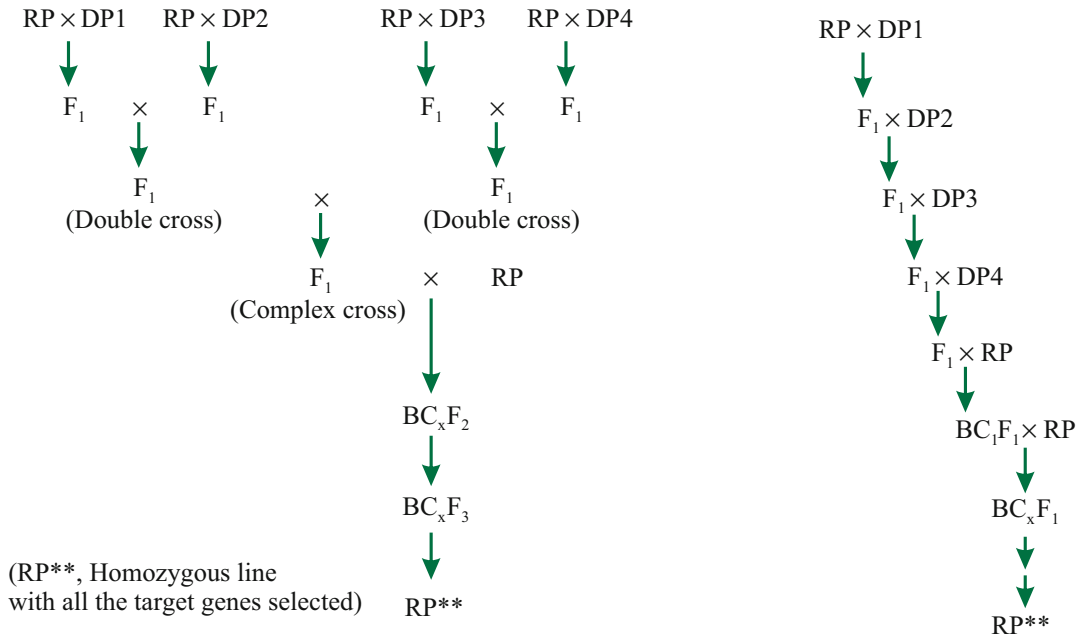
9.7.1 Strategy for Gene Pyramiding

Introgression of two or more genes from a single DP is relatively simple: the DP is crossed with the RP, and the F_1 and the subsequent progeny are repeatedly backcrossed to the RP. But when the genes to be pyramided are present in different DPs, they can be introgressed into an RP in one of the following two ways. In the first approach, each DP is used in a *separate backcross program* with the RP to recover the target gene from each DP in the genetic background of RP either in heterozygous or homozygous state. These derived lines of RP are then crossed together to produce a complex hybrid. Finally, the pyramided version of RP having all the target genes is recovered from this hybrid by selfing coupled with selection (Fig. 9.4a). In the second approach, all the DPs are ordered into a *single backcross program* according to a suitable mating scheme. In the case of *symmetrical mating scheme*, the F_1s from the crosses between different DPs and the RP are crossed in pairs to ultimately produce a complex hybrid having all the target genes from the DPs (Fig. 9.4b). But in the *tandem-mating scheme*, the recurrent parent is first crossed with one of the DPs. The F_1 from this cross is now mated to the second DP and so on till all the DPs are mated in succession to produce a complex hybrid (Fig. 9.4c). The complex hybrid obtained from either scheme is used in a backcross program with the RP to recover the pyramided version of RP (Servin et al. 2004; Ishii and Yonezawa 2007). Yet another scheme, called *stepwise scheme*, involves sequential transfer of the target genes into the RP background (Jiang 2013), but this scheme will take much longer time than the other schemes.

Theoretically, separate introgression of target genes into the RP, followed by crossing the derived RP lines according to the symmetrical mating (Fig. 9.4a), is the best strategy in terms of both efficiency and the number of backcross progeny required (Ishii and Yonezawa 2007). Sometimes, the breeder may like to develop an entirely new variety having all the target genes. In this case, the various parents having the target genes should be superior lines. These lines can be crossed in tandem (Fig. 9.4c), and the segregating generations would be handled by the pedigree method with MAS to isolate a new genotype having all the target genes.



A. Separate backcross programs



B. Single backcross : Symmetrical mating

C. Single backcross : Tandem mating

Fig. 9.4 The main schemes of backcrossing for gene pyramiding. Strategies B and C depict two alternatives for gene pyramiding through a single backcross program

9.7.2 Pyramiding of Oligogenes

Gene pyramiding has been the most widely used for development of varieties with more durable disease resistance. Host plant resistance has been classified into two broad groups, viz., vertical and horizontal resistance. Vertical resistance is governed by major genes that exhibit gene-for-gene relationship, generate hypersensitive response to specific races/pathotypes of the concerned pathogen, and usually block disease development soon after the infection stage so that plants are virtually disease-free. These genes are easy to use in breeding programs and the varieties carrying them are quite attractive to farmers. But the resistance due to oligogenes is usually not durable, and in cases like wheat rusts, it may last merely 4–5 years. This is because a resistance gene becomes ineffective when the pathogen develops a race/pathotype having the corresponding virulence gene; this often leads to disease epidemics. Horizontal resistance, on the other hand, is governed by polygenes that reduce disease development and, particularly, the pathogen reproduction rate. This type of resistance is effective against all the races/pathotypes of the pathogen (*race nonspecificity*) and for long periods of time (*durability of resistance*). Breeding for this type of resistance is difficult due to the involvement of several genes and the environmental influences on disease severity. It has been argued that genes for vertical resistance may be pyramided to effectively mimic horizontal resistance. There is some evidence that this approach may serve a useful purpose in many host–pathogen systems. Further, multiple resistance genes may increase the level of resistance to individual races of the pathogen. For example, rice lines having two or more resistance genes, each specifying resistance to a single race of the bacterial blight (BB) pathogen, consistently show smaller lesions than those having single resistance genes. The increased resistance may result from synergistic or complementary action of the resistance genes. However, several pathogens are known to have evolved races virulent to host lines having multiple resistance genes.

Transfer of a single resistance gene into a susceptible line is straightforward since the presence of this gene in a plant can be verified by its reaction to the concerned race/pathotype. But when two or more resistance, are to be transferred, identification of the plants carrying all the target genes requires progeny tests coupled with disease tests with multiple races/pathotypes. MAS for the target resistance genes greatly facilitates pyramiding since it enables easy and dependable identification of plants carrying all the target genes. There are several examples of successful pyramiding of genes, QTLs, and both genes and QTLs related mainly to disease resistance, and QTLs for yield and yield-related traits (Table 9.2). For example, the BB of rice is a serious disease worldwide, and cultivation of resistant varieties seems to be the most effective and economical approach to manage this disease. Resistance to BB is governed by over 30 dominant/recessive oligogenes that greatly reduce the disease symptoms measured as lesion size. Seventeen of these genes have been mapped, and six genes have been cloned. Many of these genes have been deployed either singly or in pyramids of two (*xa13*, *Xa21*), three (*xa5*, *xa13*, *Xa21*), and even four (*Xa4*, *xa5*, *xa13*, *Xa21*) genes. In general, these genes were transferred from single donors; foreground selection was based on linked markers; selection for recurrent parent genotype, if any, was based on phenotypic evaluation; and only background profiling of the selected plants was carried out to assess/reaffirm the recovery of recurrent parent genotype. A BB-resistant version of Pusa Basmati 1 (PB1), called "Improved Pusa Basmati 1" (IPB1), carrying the genes *xa13* and *Xa21* from the donor parent IRBB55 (Gopalakrishnan et al. 2008), and a BB-resistant version of Samba Mahsuri, referred to as "Improved Samba Mahsuri," having the genes *xa5*, *xa13*, and *Xa21* from the donor parent SS1113 (Sundaram et al. 2008), have been released for commercial cultivation.

Table 9.2 Some examples of gene/QTL introgression in different crop plants

Crop	Trait	Gene/QTL	Reference
Rice	Amylose content	*Wx*	Gopalakrishnan et al. (2008)
Maize	Protein quality	*o2*	Babu et al. (2004, 2005)
Tomato	Acyl sugar content	3–5 genomic regions	Lawson et al. (1997)
	Fruit quality	Five genomic regions	Lecomte et al. (2004)
Rice	Root depth	Four QTLs	Shen et al. (2001)
	Root depth	*QTL2*, *QTL7*, *QTL9*, *QTL11*	Steele et al. (2006)
	Rice yellow mottle virus resistance	*QTL7*, *QTL12*	Ahmadi et al. (2001)
	Submergence tolerance	QTL *SUB-1*	Neeraja et al. (2007)
Maize	Days to silking	Three QTLs	Bouchez et al. (2002)
	Yield	Two QTLs	Bouchez et al. (2002)
	Yield	Yield QTLs	Schmierer et al. (2004)
	Drought resistance	Five QTLs	Ribaut and Ragot (2007)
	Heterosis	2–4 genomic regions	Stuber et al. (1999)
Wheat	Background selection	RP genome	Randhawa et al. (2009)
Rice	Bacterial blight resistance	*xa13*, *Xa21*	Joseph et al. (2004)
		xa5, *xa13*, *Xa21*	Sanchez et al. (2000), Sundaram et al. (2008)
		Xa4, *xa5*, *xa13*, *Xa21*	Huang et al. (1997)
Barley	Barley stripe rust resistance	Oligogene *Rpsx* and *QTL4*, *QTL5*, *QTL7*	Castro et al. (2003)
Maize	Northern leaf blight and head smut resistance	Oligogenes *Ht1* and *Ht2*	Min et al. (2012)
		One major QTL	
Rice	Bacterial blight resistance	*Xa21*	Datta et al. (2002)
	Insect resistance	*Bt* gene	
	Disease resistance	*RC7 chitinase*	

9.7.3 Pyramiding of QTLs with Oligogenes Governing the Same Trait

In an ever-increasing number of cases, breeders are seeking to deploy polygenic resistance mainly to achieve wide-spectrum durable resistance. However, phenotypic selection for polygenic resistance is difficult due to the large number of genes and environmental effects. QTLs for horizontal disease resistance have been identified and mapped with the help of molecular markers. In most cases, a small number of QTLs with large effects have been identified. But a limitation of polygenic resistance is that even the resistant plants do show some disease development. Therefore, a combination of vertical resistance with a reasonably high level of horizontal resistance would be highly desirable. This combination will produce a disease-free crop so long as the vertical resistance remains effective and protect from an epidemic when the vertical resistance succumbs to new virulence. However, it is not possible to exercise phenotypic selection for horizontal resistance in the presence of effective vertical resistance. This problem can be resolved by using MAS to introgress QTLs for horizontal resistance in lines having effective oligogenic resistance.

Barley stripe rust is caused by *Puccinia striiformis* f. sp. *hordei*. The resistance to this disease is governed by both oligogenes as well as polygenes. Castro et al. (2003) combined an oligogene and one or two QTLs for barley stripe rust resistance. For example, they crossed two doubled haploid (DH) lines carrying the oligogene with one DH line having *QTL4* and *QTL7* and developed DH populations from the F_1s. They selected DH lines having the oligogene and either one or both the QTLs using SSR markers. Preliminary results from field tests suggest that DH lines having the oligogene and one or both the QTLs would show reduced disease

severity when the oligogene becomes ineffective. It was observed that when the QTLs were introgressed into a new line, some new QTLs were detected, some of the known QTLs failed to produce phenotypic effects, and the effects produced by some other QTLs were not as expected.

9.7.4 Transgene Pyramiding

Genetic transformation technology can be used for transgene pyramiding by sequential transformation, co-transformation or transformation with linked transgenes. But genetic transformation is relatively expensive and technically far more demanding than transgene pyramiding. Datta et al. (2002) pyramided the BB resistance gene *Xa21* from transgenic line TT-103 with the *Bt* gene for insect (yellow, striped, and pink stem borers and leaf folder) resistance and the *RC7 chitinase* gene for sheath blight (*Rhizoctonia solani*) resistance from the transgenic line TT-9. Lines TT-103 and TT-9 were crossed, and the F_2 plants were genotyped for the presence of *Xa21*, *Bt*, and *RC7 chitinase* transgenes using PCR and Southern hybridization. The F_2 plants were also assayed for their reaction to BB and sheath blight pathogens and to yellow stem borer. Plants having the three transgenes and showing the best bioassay performance were advanced to F_3. The F_3 plants were analyzed as above and plants homozygous for the three transgenes and with the best bioassay performance were selected and advanced to F_4. The pyramided lines were resistant to BB but showed variable levels of resistance to sheath blight.

9.8 Multitrait Introgression

In most breeding programs, attempts are made to combine two or more desirable traits present in different lines into a single line; this is called *multitrait introgression*. In such a program, each plant in the segregating generations has to be evaluated for every target trait, and those with the desired combinations of phenotypes have to be selected. Often phenotypic evaluation of one or more of these traits may present problems. Further, the breeder may like to use off-season nursery/greenhouse facilities for rapid generation advance. In such situations, MAS for the target traits will greatly facilitate their introgression. In the case of maize, northern leaf blight (*Helminthosporium turcicum*) resistance is governed by oligogenes (*Ht1* and *Ht2*), while head smut (*Sphacelotheca reiliana*) resistance is due to QTLs. Min et al. (2012) crossed an inbred line having the genes *Ht1* and *Ht2* with an inbred line carrying a major QTL for head smut resistance. The F_2 and F_3 generations were subjected to both phenotypic evaluation as well as MAS using SSR markers. F_3 lines having *Ht1* and/or *Ht2* and the QTL were resistant to both the diseases.

9.9 Combined Marker-Assisted Selection

In certain situations, a combination of MAS with phenotypic screening/selection, often called *combined marker-assisted selection* (*combined MAS*), may be more useful than either MAS or phenotypic selection alone. When the marker is not tightly linked to the target gene, combined MAS would make the foreground selection more effective. In conventional backcross programs, selection for the target gene is based on phenotype, which takes care of the effects of genetic background. But in the case of MAS, the effects of the genetic background remain unknown till the end of gene transfer program. This may lead to the development of introgression lines with a lower expression of the target trait than expected. This problem may be minimized by combining MAS with phenotypic evaluation and by rigorous phenotypic evaluation of the backcross products. In the case of QTL transfer, plants selected on the basis of marker genotype may be evaluated for the trait phenotype to assess the effect of RP genetic background on the expression of target QTLs. Phenotypic screening will also allow selection for such genetic backgrounds that tend to enhance the level of target QTL expression.

Further, when some unidentified QTLs, in addition to the QTL being transferred, are involved in control of the target trait or several QTLs are being introgressed, combined MAS may be expected to be superior to MAS alone.

MAS and phenotypic screening can be applied in tandem so that the first used criterion reduces the number of plants to be assayed for the second criterion. When phenotypic evaluation is tedious, expensive, or time-consuming, MAS may be used to reduce the number of plants for phenotypic evaluation. But when there are financial constraints, it would be desirable to use phenotypic selection to reduce the sample size for marker genotyping. For example, Joseph et al. (2004) used combined MAS for introgression of BB resistance genes *xa13* and *Xa21* from IRBB55 into the long grain scented rice variety Pusa Basmati 1 (PB 1). They used a three-tier combined MAS as follows: (1) field evaluation for BB resistance using artificial inoculation, (2) screening the putative BB-resistant plants for Basmati quality grains at maturity, and (3) genotyping the selected plants having acceptable grain quality for markers linked to the *xa13* and *Xa21* genes (Fig. 9.5). They selected 513 BB-resistant BC_1F_2 plants and evaluated them for kernel characteristics, aroma, and cooking quality to identify 57 plants producing grains with the minimum Basmati quality standards for marker genotyping. Thus, combined MAS greatly reduced the marker genotyping cost, accelerated recovery of the recurrent parent features, and permitted the isolation of desirable recombinants having some useful traits from the DP. A form of stepwise combined selection was used by Randhawa et al. (2009) for introgression of wheat stripe rust resistance gene *Yr15* (Sect. 9.5). This approach can be used when a dominant gene governs the target trait, and trait phenotyping is easy, reliable, and feasible at the seedling stage.

Some highly desirable oligogenes generate undesirable pleiotropic effects, which limit their usefulness in crop improvement. In many such cases, the pleiotropic effects can be reduced/minimized by modifying genes, which can be accumulated by a suitable selection strategy. For example, *opaque2* (*o2*) mutant allele in maize is highly desirable, but it has the undesirable pleiotropic effect of making the kernels soft and opaque. This pleiotropic effect of *o2* is

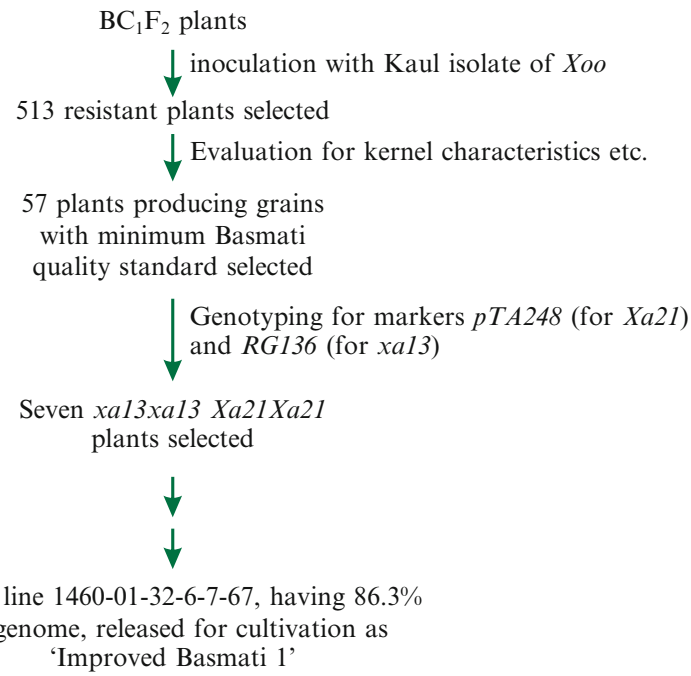


Fig. 9.5 The three-step selection strategy of combined MAS used for the transfer of BB resistance genes *xa13* and *Xa21* from the donor parent IRBB55 into Pusa Basmati1 rice. *Xoo*, *Xanthomonas oryzae* pv. *oryzae*

minimized by modifying genes. As a result, *o2o2* homozygotes placed in a suitable genetic background produce hard vitreous kernels comparable to those of normal *O2O2* genotype. Production of high-quality protein maize inbreds involves introgression of the *o2* allele along with the kernel texture modifying genes. The *o2* allele is selected using gene-based SSR markers, and the selection for modifiers is based on phenotypic evaluation of the kernels. In this example, MAS is used to select for the oligogene itself, while phenotypic selection is directed against the undesirable pleiotropic effects of this oligogene.

The results from phenotypic evaluation may be affected by several factors, including the method of assessment, the developmental stage of plants, threshold requirement for the trait, and evaluation of other traits in the same population. In the case of some traits, e.g., resistance to many diseases, different techniques used for their assessment may reveal different features of the trait, which may be governed by distinct genes/QTLs. For example, resistance to *Fusarium* head blight in wheat can be assayed by either inoculation of a single floret per spike or by spray-inoculation technique. The single floret technique assays for resistance to disease spread, which has been called type II resistance. In contrast, spray-inoculation method measures resistance to initial infection, which constitutes type I resistance. These two types of resistance are likely to involve different genes/QTLs. Thus, the conclusions drawn from phenotypic evaluation of such traits may be affected by the method used for their evaluation. Similarly, expression of many traits is affected by the stage of plant development. For example, resistance to the three rusts of wheat shows both oligogenic and polygenic control. Oligogenic resistance is usually assayed during the seedling stage, while polygenic resistance is evaluated during adult plant stage. In addition, expression of some traits may have some threshold requirements, e.g., moisture stress for drought resistance. In all such cases, the specific threshold requirement must be provided during phenotypic evaluation. Finally, evaluation of some traits may interfere with that of the other traits in the same population, e.g., evaluation of resistance to an abiotic stress would interfere with proper evaluation of yield. Therefore, when more than one trait is to be evaluated in the same population, these traits should not affect the evaluation of each other.

9.10 Marker-Assisted Recurrent Selection

In general, results from QTL analysis in one mapping population cannot be applied with confidence to other populations of the concerned species. Lande and Thompson (1990) proposed a scheme for identifying markers significantly associated with a quantitative trait in a population, and using them for MAS in the same population. In this scheme, the marker data may be used either alone or in form of a combined selection index that includes phenotype data for the trait. A *selection index* is a numerical score that combines information on all the traits associated with the dependent variable (usually, yield); generally, the value for each trait is adjusted or weighted according to its importance. The F_2 or BC_1 generation from a cross is phenotyped for the target trait and genotyped with a suitably large number of markers distributed over the entire genome. A full multiple regression analysis of the trait phenotype on marker alleles is carried out. This enables the identification of markers showing significant association with the trait. Selection in this generation is based on a selection index that combines both phenotype and marker genotype scores. In the next generation, the data on trait phenotype and genotypes of the markers showing significant association with the trait in the previous generation are used for multiple regression analysis; this yields unbiased estimates of the additive effects associated with these markers. The additive effects associated with the alleles for all the markers of a plant are summed up to obtain the net marker score (m) for that plant. The combined selection index (I)

based on both marker genotype and trait phenotype data is computed as follows:

$$I = b_z z + b_m m \tag{9.1}$$

where b_z and b_m are the weights for the trait phenotype (z) and the marker score (m), respectively. The efficiency of selection based on marker score alone relative to that of phenotypic selection at the same intensity is approximated by the formula $\sqrt{p/h^2}$, where p is the proportion of additive genetic variance for the trait associated with all the marker loci included in the marker score and h^2 is heritability of the trait. Since the association between marker loci and QTLs affecting the trait will change over generations due to recombination, this association should be reestimated after every few to several generations and, particularly, when fresh crosses are made.

Theoretically, the relative efficiency of marker index-based selection depends on trait heritability, the fraction of additive genetic variance for the target trait coupled with molecular marker loci, and the scheme used for selection. The relative efficiency of MAS based on individuals is very high for those traits that have low heritability provided the marker loci are associated with a large part of the genetic variance for the trait. For example, index selection is about 2, 1.5, and 1.2 times as efficient as phenotypic selection for a trait with heritability of 5 %, 10 %, and 20 %, respectively, when markers are associated with merely ~20 % of the additive genetic variance for the concerned trait. Further, MAS is more efficient than selection based on phenotype when the families are small. The number of markers required to detect the significant marker–trait associations will depend primarily on the total genetic length of the genome, mode of pollination of the species, and the number of generations since hybridization. For example, in a species with ten chromosomes of 1 M each (total map length, 10 M or 1,000 cM), merely 30 markers would suffice one generation after hybridization in both selfing and random mating species, but 49 and 110 markers, respectively, will be required five generations after the hybridization. An increase in the sample size increases the proportion of additive genetic variance associated with markers. In general, a sample of few hundreds to few thousands of unrelated individuals may be adequate. As a rule, the sample size should be larger for lower heritability traits (Lande and Thompson 1990).

9.10.1 MARS in Cross-Pollinated Crops

The proposal of Lande and Thompson (1990) was soon adopted as marker-assisted recurrent selection (MARS) by maize breeders, particularly in private seed companies. MARS has been mostly used for improving F_2 populations from suitable crosses before inbred isolation from them. Recurrent selection schemes were originally proposed for accumulation of desirable alleles in maize populations prior to inbred isolation from them. In these schemes, plants are selected on the basis of either their phenotype or testcross performance. The selfed progeny of the selected plants are intermated in all possible combinations to generate the population for the next cycle of selection. In this way, the selection may be continued for as many cycles as desired. The testcross parent, i.e., the tester, used in the scheme may have either narrow genetic base (*selection for specific combining ability*) or broad genetic base (*selection for general combining ability*; Allard 1960). In the case of *MARS*, the first selection cycle is based on both phenotype of a single trait or an index calculated from phenotypes of a group of traits and marker genotype data. But the next three selection cycles are based on marker scores alone. The use of marker data allows identification of superior plants much before flowering so that the selected plants are intercrossed in the same generation. Further, the use of off-season nursery/greenhouse facilities allows three to four selection cycles to be completed in a single year as against only one or two selection cycles in the conventional schemes (see Bernardo and Charcosset 2006).

Generally, the markers showing significant association with the desired trait(s) are identified afresh for each population subjected to MARS. In a simulation study in a maize F_2 population, the use of QTL-based markers increased the efficiency of MARS as compared to the use of flanking markers (Bernardo and Charcosset 2006). Further, the response to MARS was higher when the QTLs were known than when they were unknown. However, as the number of QTLs controlling the trait increased, fewer known QTLs produced the maximum efficiency. For example, when 10, 40 and 100 QTLs controlled a trait, the maximum improvement in the efficiency occurred when 10, 32, and few QTLs, respectively, were known. In theory, when markers explain a large proportion of the additive variance, MARS is much more efficient than the phenotypic recurrent selection schemes, particularly when family sizes are small. Eathington et al. (2007) have compared the effectiveness of MARS based on a multitrait index (MTI) with that of the phenotype-based selection schemes. The procedure of MARS was applied to 248 maize populations for 1 year (three cycles of MARS). These selected lines were evaluated along with the lines derived by two cycles of conventional recurrent selection schemes from the same population. The performance of the two groups of lines was compared by calculating their MTI based on the same traits that were used in the selection model for the concerned population. The average MTI values for the lines derived by conventional breeding schemes were 0.63, 0.25, and 0.76 in the years 2002, 2003, and 2004, respectively. In comparison, the values for lines developed through MARS were 1.10, 0.97, and 1.62 in the respective years. The MTI scores averaged over the 3 years were 0.50 for the conventionally derived lines as against 1.18 for the lines obtained through MARS. Similarly, in the case of soybean, 43 breeding populations were subjected to different MARS schemes and to conventional selection schemes. The MARS-derived lines showed a 37.6 kg ha^{-1} advantage over the conventionally selected lines although the MARS-derived lines were slightly delayed in maturity.

The results from several simulation and experimental studies indicate that gains per selection cycle for grain yield in maize are ~25–50 % lower for MARS than for phenotypic selection based on testcross performance. But phenotypic selection requires 2 years per cycle, while up to three cycles of MARS can be completed each year by using off-season nursery/greenhouse facilities. Therefore, gains per year are much higher for MARS than those for phenotypic selection. In the large-scale breeding programs of the private sector, MARS is becoming increasingly competitive with phenotypic selection in terms of returns per unit cost as well as per unit time (see Bernardo 2008).

9.10.2 F_2 Enrichment and MARS in Self-Pollinated Crops

Recurrent selection based on phenotypic evaluation has been used in self-pollinated crops for improvement of quantitative traits by accumulating favorable alleles of polygenes governing the target traits (Singh 2012a). The MARS scheme of Lande and Thompson (1990) can be applied to segregating generations from suitable crosses of self-pollinated crops, and the selected plants may be mated in pairs to generate the population for the next cycles of selection. In computer simulations, the efficiency of MARS for one or two generations in intercrosses among RILs or DH lines derived from specific crosses was greatly affected by the accuracy of QTL location. Further, MARS for QTL accumulation was highly preferable to QTL pyramiding (Van Berloo and Stam 1998; Charmet et al. 1999).

In the F_2 *enrichment approach*, MAS is used to eliminate from a F_2 population all plants that are homozygous for the unfavorable allele of one or more of the target QTLs. Thus, only such F_2 plants are retained that are either homozygous or heterozygous for the favorable alleles of all target QTLs. The frequency of such plants would be $(3/4)^n$ in F_2, where n is the number of target QTLs. When the value of n is 10, the frequency of the selected F_2 plants will be ~1/18. The selection will raise the frequency of the desirable

QTL alleles from 0.5 in the unselected F_2 population to 0.67 in the selected group. As a result, the frequency of lines having desirable alleles at all ten QTL loci will increase to one in every 55 ($= 0.67^{10}$) recombinant inbred lines (RILs) derived from the selected F_2 plants as compared to merely one in every 1,024 ($= 0.5^{10}$) RILs from the unselected F_2 population. Thus, *the* F_2 *enrichment approach dramatically increases the frequency of desired homozygous lines recovered from a* F_2 *population. Generally*, F_2 *enrichment is applied to the* F_2 *generations*, *but it can be used in backcrosses*, *three-way crosses*, *and double crosses as well*. MAS may be applied again in F_3 or F_4 generation, but this seems to offer little additional advantage (Howes et al. 1998; Bonnet et al. 2005).

The F_2 enrichment scheme assumes equal effect for all target QTLs, which is contrary to the real situation. If an inbred line with favorable alleles at all target loci were isolated, its genotypic value would be the same irrespective of the equal/unequal QTL effects. But when F_2 population size is small, inbred lines with favorable alleles at only some to most of the target loci are likely to be isolated. In this situation, the genotypic value of the inbred line will depend on the specific QTLs for which favorable alleles are present if the QTL effects were not equal. In such cases, a potential index may be used for selection of the homozygous lines. The *potential index* for a plant is essentially the weighted sum of genotypic values of the target QTLs present in the plant. Selection based on potential index is reported to increase the probability of isolation of superior inbred lines (Liu et al. 2004). MARS schemes require additional generations of intermating and selection before isolation of homozygous inbred lines, while F_2 enrichment scheme achieves enhanced gene frequency during the process of inbreeding itself. Thus, MARS would require more time, effort, and resources than F_2 enrichment. However, both MARS and F_2 enrichment schemes are useful in fixation of favorable alleles of up to 9–12 QTLs in the homozygous lines developed from the selected materials (Wang et al. 2007).

The above considerations do not take into account other traits, at which MAS is not directed although they are likely to be equally relevant for the usefulness of the selected lines. It may, therefore, be more practical to select plants having desirable/superior combinations of other traits and favorable alleles of most (not all) of the target QTLs. The selected plants may be mated in complementing pairs so that each mating pair together has favorable alleles at all the target QTLs. The progeny from these matings can be subjected to MAS for the target QTLs and phenotypic evaluation for the other traits; MAS may be continued for one or more cycles, if required. This approach would increase the frequency of favorable alleles of the target QTLs in a desirable genetic background, and the inbreds isolated at the end of the scheme may be expected to be more useful than those derived by selection for the target QTLs alone.

9.11 Innovative Breeding Schemes for Effective Use of MAS

The conventional breeding schemes were designed for selection based on phenotype and are not well suited to fully exploit the marker technology. This recognition has encouraged the development of several innovative breeding schemes (Table 9.3) designed to take full advantage of the markers data.

9.11.1 Inbred Enhancement and QTL Mapping

Around 1989, Stuber and coworkers proposed a breeding scheme for introgression of unidentified desired QTLs from a DP into elite inbreds and simultaneous mapping of these QTLs. In this scheme, the RP is an elite inbred/pureline deficient in some quantitative trait, and the DP is a potential source for the concerned trait. The F_1 from the cross between the DP and the RP is repeatedly backcrossed to the RP to generate a set of NILs (Fig. 9.6). Each NIL of this set would have a single distinct genomic segment from the

Table 9.3 A summary of various breeding methods, including innovative schemes, using MAS

Breeding scheme	Objective	Chief features
1. Marker-assisted backcrossing (MABC)	Introgression of genes/QTLs from one or more DPs into an RP	F_1 and subsequent generations backcrossed to RP; foreground, background, and recombinant selections, usually, based on MAS
(a) Single gene/QTL introgression	Removal of a specific defect of RP	As above
(b) Gene pyramiding	Accumulation into the RP of different genes/QTLs affecting a trait	Genes introgressed individually into RP by parallel MABC and brought together in the end
(c) Multitrait introgression	Accumulation into the RP of genes/QTLs affecting several different traits	As above
(d) Single backcross-DH scheme	Introgression of genes/QTLs from an elite DP into an RP	MAS in BC_1F_1 for target traits; haploids produced and subjected to MAS; DH produced and evaluated; no background MAS
(e) Advanced backcross QTL mapping	Introgression of genes/QTLs from unadapted germplasm	Selection against deleterious traits in the backcross generations; BC_1S_1/BC_2S_1 used for QTL identification/used in breeding programs
(f) Inbred enhancement-QTL mapping	Introgression of QTLs from DP into an elite RP deficient in the trait	Introgression line library constructed; lines evaluated for QTL detection and mapping; superior lines used in breeding/as varieties
2. Breeding by design	Development of a line with the ideal genotype created, initially, in silico for high performance	The ideal genotype designed using information on marker–trait association; this genotype is constructed by combining the target genomic regions from various DPs
3. Pedigree MAS	To ensure the presence of the desired genomic regions in the derived lines by fixing these regions	Genomic regions of interest identified from data generated in breeding activities; MAS in early segregating generations to fix these regions
4. Single large-scale MAS (SLS-MAS)	As above	Genomic regions of interest identified from appropriate crosses; MAS for fixing the target regions in F_2/F_3; subsequent generations as per pedigree scheme
5. Marker-evaluated selection (MES)	Development of genotypes for adaptation and performance in specific ecosystems	Genomic regions of interest identified by changes in marker allele frequency in the target ecosystems; MAS used for these regions
6. Marker-assisted recurrent selection (MARS)	Isolation of improved inbreds/purelines by increasing the frequency of desirable alleles in the population	Markers showing significant association with the trait(s) used for MAS; selected plants intermated and their progeny subjected to MAS; may continue for several cycles
7. Genomic selection (GS)/genome-wide selection (GWS)	Selection for all the QTLs affecting the trait irrespective of the significance of marker–trait associations	Genome-wide markers used for MAS based on genomic estimated breeding values; marker effects estimated from a suitable training population (Chap. 10)
8. Heterosis breeding	Development of superior hybrid varieties	Heterotic groups identified on the basis of marker data; complementing groups crossed to produce hybrids
		Genomic regions involved in heterosis identified; target regions introgressed into appropriate inbreds to enhance hybrid performance (Chap. 11)

DP in place of the homologous segment of the RP. The sum total of the DP segments present in the NIL set would, ideally, represent the entire DP genome; such an NIL set is known as *introgression line library* (*ILL*). An ILL is a highly powerful QTL mapping tool because it separates the DP genomic regions involved in the control of the target trait into small segments,

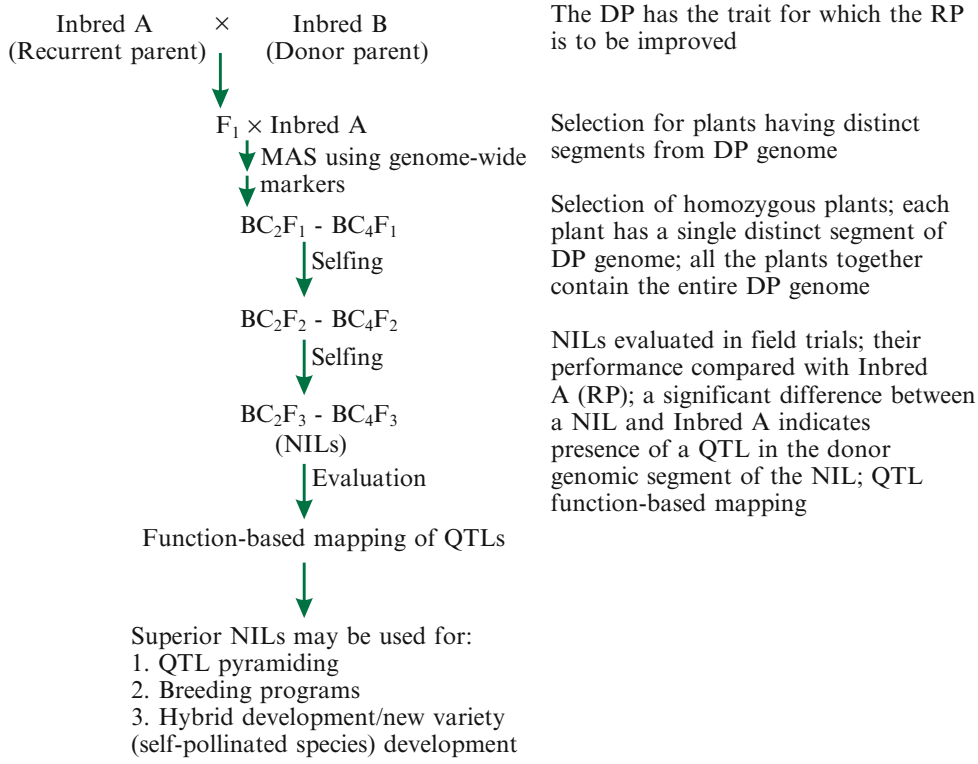


Fig. 9.6 Inbred/pureline enhancement and QTL mapping; a schematic representation

each of which is placed in the uniform genetic background of the RP. As a result, the effects of these DP genomic segments on the target trait can be estimated with greater confidence. In essence, the ILL lines are chromosome segment substitution lines (Sect. 5.11). Molecular markers covering the entire genome are used to monitor the introgression of the individual segments of the DP genome and to ensure their homozygosity in the NILs. If a sufficiently large number of plants were genotyped for markers in each generation, two backcrosses followed by one generation of selfing would be adequate to develop the ILLs (Stuber et al. 1999).

In crops like maize, each ILL line is crossed to a suitable tester, and the testcross progeny is evaluated in replicated field trials for the trait of interest. But in self-pollinated crops like wheat, ILL lines themselves are evaluated. A superior performing ILL line/testcross progeny would have received from the DP favorable alleles of one or more QTLs for the target trait. These QTLs are mapped onto the donor segment introgressed into the concerned ILL. The chief advantage of this scheme is that the DP genome segment identified to have favorable QTL allele(s) is already introgressed into an elite line. Therefore, an ILL line with superior per se or testcross performance is an "enhanced" version of the RP. Such an ILL line can be used as a parent of hybrids, as a new pureline variety, or as a parent in hybridization programs. The use of ILLs is advantageous because the evaluated genomic segments are separated from their neighboring genomic regions that might interfere with QTL detection. Further, the ILL lines with favorable QTL alleles can be used to pyramid these QTLs to further enhance the performance of the RP. However, favorable epistatic interactions, if any, between QTLs located in different genomic segments of the DP cannot be recovered in the ILLs.

The ILL scheme is similar to the advanced backcross QTL analysis (Sect. 9.11.2) in its

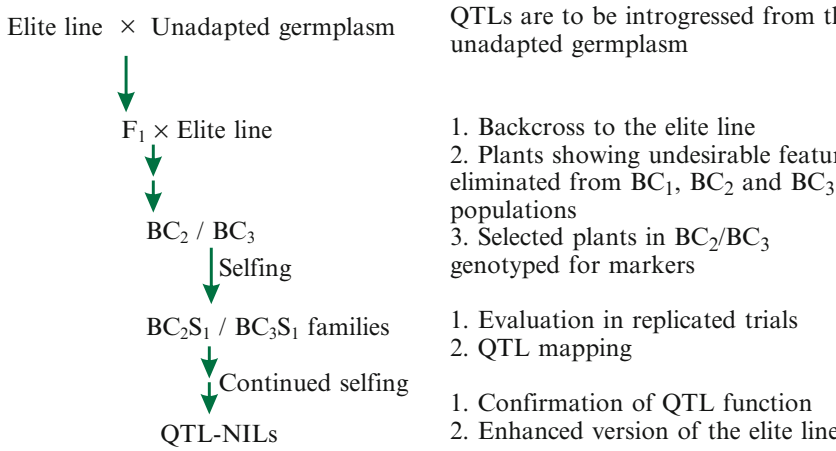


Fig. 9.7 A simple schematic representation of advanced backcross QTL analysis in a self-pollinated crop. In the case of a cross-pollinated species, the selected BC_2/BC_3 plants are crossed with a tester (in place of selfing), and testcross progeny are phenotyped. BC_2/BC_3 plants producing superior testcross progeny are selfed to isolate QTL-NILs

general approach. However, the following four features distinguish it from the latter: (1) lack of phenotypic selection; (2) use of markers to make sure that each ILL line has a distinct segment of the DP genome; (3) representation of, ideally, the complete DP genome in the ILL; and (4) function-based QTL detection and mapping using homozygous ILL lines.

9.11.2 Advanced Backcross QTL Analysis

Advanced backcross QTL analysis (*AB-QTL analysis*) was devised for introgression of favorable QTLs from unadapted germplasm like land races and related wild species into elite lines and simultaneous detection and mapping of these QTLs (Tanksley and Nelson 1996). The unadapted germplasm line is used as DP in a backcross program with the selected elite line (RP). In the backcross generations, phenotypic selection is used to eliminate plants having deleterious alleles from the DP. As a result, the selected plants have a high frequency of favorable alleles contributed by the DP. In the case of self-pollinated crops, the selected BC_2/BC_3 plants are genotyped for marker loci and selfed to produce BC_2S_1/BC_3S_1 progeny, which are evaluated in replicated trials. The genotype and phenotype data are subjected to QTL analysis to identify donor genomic regions containing favorable QTL alleles. Ultimately, QTL-NILs are extracted from the superior BC_2S_1/BC_3S_1 progeny (Fig. 9.7). A *QTL-NIL* is an NIL that contains a single segment of donor genome having, ideally, a single favorable QTL. The QTL-NILs can be used to confirm the findings from QTL mapping, to fine-map the detected QTLs, as parents in breeding programs or as new varieties.

In the case of cross-pollinated crops, the RP is an elite inbred line, e.g., inbred A involved in the outstanding single cross A × B. The selected BC_2/BC_3 plants are genotyped for marker loci and crossed with the inbred B. The testcross progeny so obtained is phenotyped, and the marker and the phenotype data are used for QTL analysis. The BC_2/BC_3 plants having favorable QTLs from the DP are identified and, ultimately, QTL-NILs are extracted from them. These QTL-NILs are in the uniform genetic background of inbred A and are used for the same purposes as in the case of self-pollinated crops, except for the use as a new variety. Instead, the QTL-NILs can be crossed with inbred B or some other inbred to assess their usefulness as parents of single cross hybrids. Simulation studies suggest that AB-QTL analysis will be effective for detection of QTLs with additive, dominant, overdominant, or partially

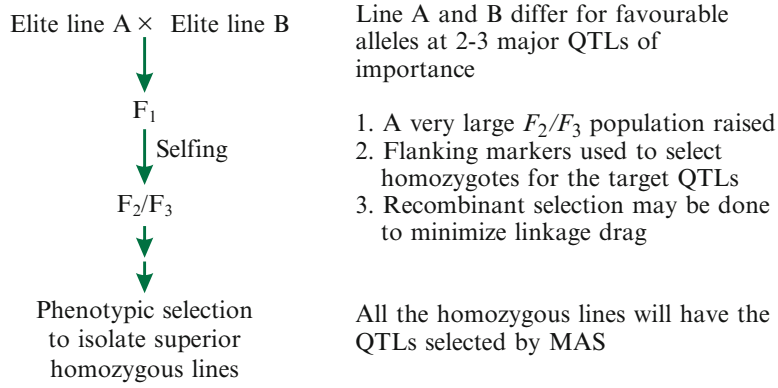


Fig. 9.8 A schematic representation of single large-scale marker-assisted selection (SLS-MAS)

dominant effects, but it will be less powerful than selfing generations for epistatic and recessive QTLs. The number of plants in BC_1 should be 100 or more, while in BC_2/BC_3 it should be large enough to leave 200 or so plants after phenotypic selection.

The AB-QTL analysis offers the following advantages: (1) mapping population is more similar in phenotype to the RP than any other mapping population, (2) frequency of deleterious alleles from the DP is greatly reduced, (3) likelihood of epistasis is reduced, (4) only one or two generations are needed for extracting QTL-NILs after QTL mapping, (5) chances of linkage drag are reduced, and (6) there is opportunity for detection of subtle pleiotropic effects of the introgressed QTLs (Tanksley and Nelson 1996). AB-QTL analysis has been effectively used for introgression of useful agronomic traits in elite line of tomato, maize, cotton, rice, soybean, barley, and wheat from unadapted germplasm, including wild relatives (see Jiang 2013).

9.11.3 Single Large-Scale MAS

The SLS-MAS strategy was proposed for the development of elite lines combining favorable alleles present in the elite germplasm at the target loci (Ribaut and Betran 1999). Each member from a set of elite lines having the trait of interest is crossed to a tester, which itself is an elite line deficient in the trait. A suitable mapping population from each cross is analyzed to identify the QTL alleles present in the concerned elite line for the target trait. The elite lines having complementing sets of QTL alleles for the target trait are identified and crossed to obtain very large segregating (F_2/F_3) populations. Markers flanking (at <5 cM) the target QTLs are used to select individuals homozygous for the favorable alleles at the target loci (Fig. 9.8). This step of MAS is limited to one generation and aims to isolate homozygotes for the target QTL alleles in F_2/F_3 so that the presence of these QTL alleles in the purelines derived from the selected plants is ensured. SLS-MAS can be used for up to three large effect and stable QTLs. Further, recombinant selection could be used, if necessary, to minimize linkage drag. It is expected that considerable variation for the rest of the genomic regions would remain in the selected population; this variation can be exploited by phenotypic selection. But genetic drift may occur at the nontarget loci due to the greatly reduced population size. For example, if homozygotes for three unlinked QTLs were selected in F_2, the proportion of selected plants will be $(0.25)^3$, i.e., merely ~1.6 %. The risk of genetic drift can be considerably reduced by SLS-MAS in F_3.

SLS-MAS is suggested to offer the following three advantages: (1) The favorable alleles of the target QTLs become fixed in F_2/F_3. (2) Considerable allelic variability is retained in the rest of the genome, which can be exploited by phenotypic selection. (3) There is opportunity for pyramiding of desirable alleles of QTLs with major effects. QTL alleles present in new germplasm can also be used, if gene-based markers were used to minimize linkage drag.

9.11.4 Pedigree MAS

The idea of pedigree MAS was proposed by Ribaut et al. (2001) for self-pollinated crops like rice and wheat that have been subjected to extensive breeding efforts. In such crops, the pedigrees of most of the elite germplasm lines used in various breeding programs are known. It was suggested that the lines frequently used in breeding programs and the materials selected by breeders from segregating generations of the crosses should be genotyped with a set of markers evenly distributed over the entire genome. The phenotypic performance data for these lines would have been collected routinely during the breeding programs. The marker genotype and the trait phenotype data can be analyzed together in an effort to identify the genomic regions that were the targets of selection by the breeders in the segregating generations of the respective crosses. The simple logic for identification of such genomic regions is that the frequencies of the marker alleles present in the regions targeted by selection will be higher than that expected on the basis of random distribution of alleles in the nonselected regions. If, for example, the segregating generations from a cross were selected for performance in a drought-prone environment, the homozygous lines developed from this program will have much higher proportion than expected of the marker alleles located in the genomic regions involved in drought resistance. This logic has been further extended to design the marker-evaluated selection scheme described in Sect. 9.11.8. Once such genomic regions are identified, MAS for them may be carried out, preferably, in the F_2 and/or F_3 generations of the concerned crosses to rapidly fix the identified desired genomics regions in the progeny.

The pedigree MAS scheme is similar to SLS-MAS in the general approach as both the schemes aim to use MAS in the early segregating generations to fix the target loci. However, the two schemes differ from each other with respect to the following features: (1) In pedigree MAS, the genomic regions of interest are identified on the basis of data generated routinely in the breeding programs, while SLS-MAS generates this information from experiments designed for the purpose. (2) Further, MAS may not be limited to a single generation in the case of pedigree MAS, while it is applied to a single (F_2 or F_3) generation in SLS-MAS.

9.11.5 Single Backcross-Doubled Haploid Scheme

In the single backcross-doubled haploid (DH) scheme for introgression of multiple genes, F_1 is backcrossed to RP to obtain about 2,500 BC_1F_1 plants (Kuchel et al. 2007). These plants are subjected to foreground selection, and around 20 plants heterozygous for the target genes (up to four genes) from the DP but homozygous for the target genes of the RP (up to three genes) are selected. About 1,000 haploids are produced from the selected plants, and plants carrying the DP target genes are selected and subjected to chromosome doubling to produce about 64 DH plants. The DH lines are multiplied and subjected to phenotypic evaluation for the target traits as well as RP phenotype. The authors suggested that phenotypic selection for RP genetic background should be adequate. They also felt that almost complete recovery of RP genetic background is a conservative approach, particularly when the DP may contribute positive alleles for some other traits as well. Thus, the key features of the scheme are a single backcross, enrichment of target alleles in BC_1F_1 using MAS, production of haploids from the selected BC_1F_1 plants, selection of haploid plants for the target alleles, phenotypic evaluation of DH lines, and a lack of background selection. The chief advantages of the scheme are short time frame (only 2 years to the DH stage), reduced cost, likelihood of improved genetic gains, and the possibility of developing DH lines superior to both RP and DP. The main limitations of the scheme relate to the choice of DP, which must be an elite line, and the inability to precisely predict the outcome from the program.

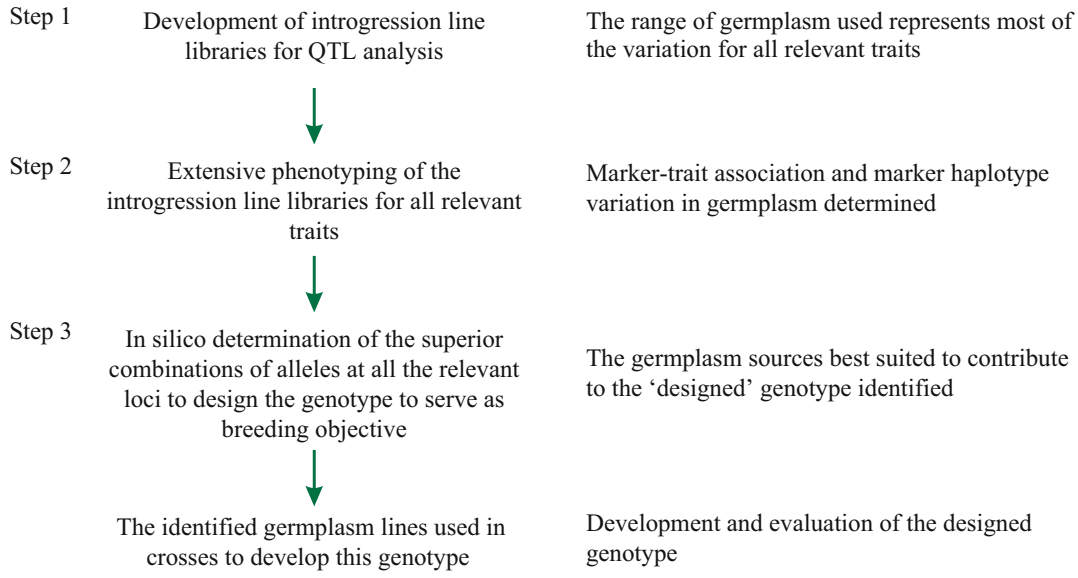


Fig. 9.9 A simple representation of breeding by design scheme

9.11.6 Breeding by Design

The *breeding by design* scheme aims to develop plant types that have the desired alleles at all the loci involved in the control of all the traits relevant to the breeding of the concerned crop species (Peleman and van der Voort 2003; Peleman et al. 2005). This strategy has the following three well-defined steps: (1) mapping, preferably by developing ILLs (Sect. 9.11.1), of all the loci governing all the relevant traits, (2) determination of the allelic effect variation present in the germplasm, and (3) in silico construction of the ideal genotype on the basis of this information (Fig. 9.9). The ILL lines are subjected to precision phenotyping for all the relevant traits and genotyped with markers providing dense genome coverage. The phenotype and genotype data are analyzed to map the QTLs governing the target traits and to identify a small set of markers tightly linked to the identified QTLs. Ideally, marker haplotypes for the various QTLs should be identified since these would be much more reliable than single markers. A set of germplasm lines are subjected to precision phenotyping and are genotyped with dense coverage of the markers. These data are analyzed to determine the allelic effect variation at all the QTLs of interest. Finally, this information is used to design in silico superior genotypes containing desirable alleles at all the target loci. Suitable breeding strategies would then be used to construct the designed genotypes, and the lines so developed will be evaluated extensively to assess their superiority.

Clearly, this approach is highly ambitious and involves enormous amount of work. In addition, appropriate algorithms and software need to be developed to support the various steps of the scheme. The breeding by design scheme is similar to the strategy of ideotype breeding (Singh 2012a). The chief difference between the two schemes lies in the basis for designing the superior genotype and in the description of the objectives to be achieved. In the case of breeding by design, the superior genotype design is based on information about the effects of the relevant QTL alleles, and the objectives are described as combinations of these alleles. But in the case of ideotype breeding, the superior genotypes are designed on the basis of expected contributions of various traits to yield, and the objectives are defined as the levels of these traits in the ideotype.

9.11.7 Mapping as You Go

The *mapping as you go* strategy is designed to take into account the fact that the QTL allele composition of the breeding population and even the QTL allele effects are likely to change with the progress of the breeding program (Podlich et al. 2004). In this scheme, QTL locations and effects are estimated from the initial set of breeding crosses. This information is used for MAS in the material used for mapping. Now, new set of crosses are made among the lines developed by MAS, and the QTL effects are reestimated. The updated QTL information is used for MAS in the new set of breeding materials. The breeding cycle may be continued as long as desired by repeating the above activities. The QTL effect estimates used in the scheme may be based on the entire breeding program or on individual crosses. In the latter case, the QTL information is used for MAS in the concerned cross, and reestimation would be done when the selected lines are used as parents for hybridization to generate the population for further selection. In general, a greater response was achieved with regular reestimation of QTL effects than with a single QTL effect estimation at the start of the program.

9.11.8 Marker-Evaluated Selection for Adaptation and Agronomic Performance

The *marker-evaluated selection* (*MES*) scheme is designed for identification of and selection for the genomic regions involved in adaptation to and agronomic performance in different agricultural environments (Steele et al. 2004). The F_2 and F_3 generations of the selected cross are grown as bulks to produce adequate quantity of seed. The F_4 and the next two to four generations are grown either in the farmer's fields or at research stations in the selected agricultural ecosystems. Farmers apply mass selection to these populations either on their own (in the farmer's fields) or in consultation with the breeders (at the research stations). The selected populations from the different environments, the two parents of the cross, and the unselected F_2 populations are genotyped with markers distributed, preferably densely, over the entire genome. The frequencies of marker alleles are compared among these populations to identify such marker alleles that differ significantly among the populations. These markers would be linked to such genes/QTLs that are important for adaptation to and agronomic performance in the different agricultural ecosystems represented in the study. The above information is then used to construct a model genotype for optimum adaptation to and performance in a given agricultural environment. MAS can now be used to create the model genotype by combining the identified genomic regions. It may be emphasized that MES does not refer to any specific trait; it is simply concerned with the genomic regions involved in adaptation to and performance in a given agricultural environment.

9.12 Integration of MAS in Breeding Programs

A successful integration of MAS in breeding programs has the following requirements: (1) A reasonably good marker system (Sect. 2.2.3) should be available in the concerned crop species suitable for the scale of the breeding program (Sect. 4.10). (2) The population used for identifying the marker–trait associations should be precisely phenotyped. (3) The genes/QTLs mapped in one population should be confirmed and validated in a range of populations/germplasm. (4) The markers to be employed for MAS should be very close to or within the target genes/QTLs. This will minimize the risk of recombination between the markers and the genes/QTLs and the consequent wrong prediction of target gene allele by the marker genotype. Where this is not feasible, a pair of markers flanking the target gene should be used for foreground selection. (5) A sufficiently dense genetic map of the crop species with well-distributed polymorphic markers is needed for background selection. (6) Since the confidence intervals of

QTLs is generally large (~10 cM or more), it is desirable to use at least three markers, two markers flanking the QTL and one marker located within the QTL, for each QTL. (7) There should be sufficient recombination between the DP and RP genomes to minimize the length of DP genome flanking the target gene/QTL and the resultant linkage drag. (8) The marker system should allow speedy, cost-effective, and accurate genotyping of a large number of plants. This is because a breeder may have to make selections from a population of, say, 100,000 or more plants in a breeding season of about 120 days. If flanking markers were used for foreground selection for just one gene, 200,000 assays would be required. (9) Finally, the large amount of data generated in breeding programs based on MAS should be processed appropriately and rapidly (Xu 2010; Jiang 2013). At present, SNPs are the preferred marker system in view of their abundance, suitability for automation, and lower costs per data point. Fully automated molecular marker fingerprinting systems starting from DNA extraction to allele identification based on fluorescent DNA reads are available. Monsanto uses a gel-free SNP detection system and a completely automated genotyping procedure (Eathington et al. 2007).

Integration of MAS in a breeding program adds to the (1) type of activities performed, (2) the types of tools and techniques employed, (3) the quantum of data generated (~seven-fold increase), (4) the need for novel statistical techniques and bioinformatics tools for data processing and decision support, and (5) the overall cost. Therefore, a gainful use of MAS in a breeding program requires (1) development of simple, rapid, cheap, and reliable large-scale plant tissue collection, DNA extraction, genotyping, and data collection protocols for routine use; (2) development and implementation of sample data tracking, management, analysis, and decision-support systems; and (3) simulation analyses for achieving various objectives like designing optimum breeding programs. In general, off-season nursery/greenhouse facilities are utilized to speed up the breeding process with the aid of MAS; this increases the complexity of breeding programs. In addition, decision making becomes more involved and frequent, e.g., three to four times each year in the case of MAS as compared to one to two times in the conventional programs. Therefore, appropriate statistical and bioinformatics resources, including databases and data mining tools, would be required. Monsanto has developed a centralized database system that allows its breeders to manage all aspects of their breeding programs, including access to the inventory of genetic materials and their pedigrees and the relevant phenotype data. This system allows, among other things, tracking of every plant from the day it was created. A similar system has been developed for marker data; this system tracks tissue samples from the field, through the genotyping process, and links the genotype data to the correct genetic material. Finally, a Web-based integrated marker decision-making system enables rapid methodology enhancement, allows breeders to submit populations for MAS, develops models for selection, and makes selection decisions (Eathington et al. 2007; Xu and Crouch 2008).

9.13 Advantages of MAS

1. Foreground selection greatly facilitate selection for such traits whose phenotypic evaluation is cumbersome, tedious, time-consuming, destructive, and/or dependent on specific threshold conditions.
2. MAS permits backcrosses to be made in succession during introgression of recessive genes.
3. MAS allows selection for the target traits in off-season nurseries/greenhouses; this allows two to four generations to be taken each year.
4. MAS can be done in the seedling stage. This permits the use of selected plants for hybridization in the same generation even when the target traits relate to fruit and seed.
5. In the case of MABC, MAS can accelerate recovery of the recurrent parent genotype (background selection).

6. In case of gene introgression from unadapted germplasm, recombinant selection helps minimize/eliminate linkage drag.
7. MAS allows selection for horizontal resistance and greatly facilitates gene pyramiding as well as trait stacking.
8. MAS enables stacking of oligogenic and polygenic resistance to obtain more durable and effective resistance to diseases. This objective cannot be achieved without MAS.
9. MAS can substitute for disease tests when the resistance gene may exhibit incomplete penetrance, the pathogen virulence may be variable, the pathogen inoculum may be inadequate or difficult to obtain, disease reaction may be affected by the environmental factors, and the disease tests may be either costly and time-consuming or can be conducted only at particular locations, seasons, or stages of plant development. For example, the isozyme marker *APS1* has been used since 1974 as a substitute for screening with nematodes to select for the *Mil* gene for nematode resistance in tomato. A more tightly linked PCR-based marker now enables a more reliable MAS for *Mil* gene.
10. MAS can be combined with recurrent selections as MARS to effectively accumulate QTLs for the target traits.
11. Genomic selection (GS) is perhaps the most ambitious specialization of MAS. GS may emerge as a highly effective strategy for the improvement of low heritability traits (Chap. 10).

9.14 Limitations of MAS

1. Tightly linked, widely applicable, and reliably diagnostic markers are available for only a limited number of target traits.
2. The marker-trait associations discovered in one population have to be independently confirmed and validated in unrelated germplasm; this is particularly relevant for QTLs.
3. QTL introgression is often problematic since relatively large genomic regions need to be selected for. This is because the confidence intervals for QTLs span ~10 cM.
4. QTLs introgressed into different genetic backgrounds generally show unpredictable expression. Often QTL introgression produces discouraging improvement in complex traits like yield, and sometimes even unfavorable effects may be obtained.
5. QTL × environment interactions complicate QTL transfer.
6. MAS increases the amount of data generated in a breeding program by about seven-fold, which increases the workload of breeders.
7. The decision making in breeding programs becomes more involved and more frequent, e.g., three to four times per year in the case of MAS as compared to one to two times in the conventional breeding programs.
8. Marker genotyping has to be accomplished in a short growing season, for which high-throughput genotyping facilities may become necessary.
9. In general, MAS increases the overall cost of the breeding program. Marker genotyping cost is the chief factor limiting the widespread adoption of MAS by plant breeding programs, especially in the developing countries.
10. The full benefit of MAS can be derived only in conjunction with dependable off-season/greenhouse facilities, which may not be available, at least to the desired extent, to many breeding programs.
11. In any case, conventional breeders are, in general, hesitant to completely replace phenotypic selection with MAS.

9.15 Present Constraints and Future Directions

At present, high marker genotyping cost is one of the major constraints in adoption of MAS, particularly for breeding of orphan crops and in the developing countries. Ribaut et al. (2010)

concluded that the limited adoption of MAS in developing countries is the result of several factors, including poor facilities for genotyping and/or phenotyping, inadequate financial resources, and limited number of well-trained staff. The consumables and labor costs, excluding sample collection and capital costs, for marker genotyping may range from US $ 0.30 to 5.0 per data point, depending on factors like marker type, the number of samples and/or markers assayed, etc. (Collard and Mackill 2008). SNP genotyping using a 384 SNP chip costs just US $ 0.09 per data point, but these costs would become US $ 34.56 and US $ 17,280.00 for one plant or sample and for a population of 500 plants, respectively. The cost per population would exceed US $ 20,000.00 when sample collection and DNA extraction costs are added to the SNP genotyping cost (Schuster 2011). If a breeding program handled 100 populations of 500 plants each year, the annual expenditure on SNP genotyping alone would be over US $ 2.00 million. DNA extraction is the single largest cost in most genotyping pipelines, particularly when the number of samples is small. The development and optimization of single seed-based nondestructive DNA extraction protocols is expected to reduce costs. CIMMYT, Mexico, has developed such a procedure for large-seeded crops.

PCR amplification is an expensive step, and detection of PCR products using gel-based assay systems is costlier than microtiter plate or dot blot assays. Multiplexing reduces PCR amplification costs, but optimization of multiplex PCR is demanding and has to be done for every cross/population. SNPs are highly scalable, extremely abundant (e.g., four million SNPs in rice in 2008), and much cheaper than other marker systems, and their technology is evolving rapidly leading to progressively lower costs. The low cost of SNPs is mainly due to total automation and large-scale application, which may be beyond most public sector breeding programs. In addition, rapid improvements in technology would necessitate frequent acquisition of state-of-the-art genotyping platforms. One approach to resolve these difficulties is to develop a centralized genotyping platform, including common markers, statistical and bioinformatics tools, etc., that is accessible to public sector breeders. Cornell University, USA, and CIMMYT, Mexico, have developed an outsourcing collaborative platform for maize gene-based SNPs for both foreground and background selections. This platform is accessible to CIMMYT collaborators worldwide and is expected to help maize breeders converge onto a common set of SNP markers. In USA, four United States Department of Agriculture regional centers for genotyping have made marker genotyping a routine activity for small grain breeders. Further, several companies are offering marker genotyping services that may be more economical than genotyping by individual laboratories. The use of commercial genotyping services would allow the public sector programs to concentrate more on breeding activities.

Keeping a track of the tissue samples collected in the field through DNA extraction, genotyping, and then back to the field for relating the marker data to the concerned plants consumes time, adds to the cost, and is error prone. Private sector has widely used bar-coding systems for sample tracking, but more efficient and error-free tracking systems are required for the increasingly large numbers of samples being handled. Several efficient Laboratory Information Management Systems are available for DNA sequencing data, but such systems are rare for marker genotyping data. Some within the laboratory sample tracking systems are, however, freely available. In general, mapping of genes/QTLs, confirmation and validation of the identified marker–trait associations, and MAS are performed in separate steps in different populations; this is considered as one of the factors limiting the use of MAS. New schemes that combine genetic diversity analysis, mapping, validation, and MAS are being developed (Sect. 9.11). In addition, efficient breeding schemes for simultaneous improvement of multiple traits should be developed. This would require development of appropriate selection indices based on an understanding of the correlations between

different traits, developmental correlations among the traits, and the genetic networks for the correlated traits. Considerable progress has been made in developing selection indices in some cases, e.g., for drought tolerance in maize and wheat, and multitrait indices are being used for MARS in crops like soybean and maize (Eathington et al. 2007). However, simultaneous selection for several loci may require impractically large population sizes. In such cases, F_2 enrichment or MARS may be used to increase the frequency of favorable alleles and, thereby, reduce the required population size. In future, MAS kits containing sets of markers, say, several thousand well-selected SNP markers (preferably functional markers), for both foreground and background selections may be developed for different crops.

Application of MAS for improvement of complex traits like yield is limited by epistasis and genotype × environment interactions (GEI). The role of epistasis in the control of quantitative traits is well recognized, but the extent of its contribution is debatable. Further, studies on epistasis are generally limited to two-locus interactions since analysis of interactions between more than two loci would require very large populations, e.g., a population of over 1,000 for three-locus interactions. When epistasis and GEI are important, the QTL effects should be regularly reestimated within the breeding populations. Computer simulation and modeling would be increasingly used to adequately address the issues relevant to the improvement of complex traits. Further, crop models, capable of predicting yields of different genotypes under various environments, should be developed so that breeders may try to create such genotypes (Xu and Crouch 2008).

As stated earlier, there is a need to develop suitable decision-support tools for various activities related to breeding, MAS, information management, breeding system design and simulation, crop modeling, etc. The software package iMAS (integrated MAS; http://www.icrisat.org/gt-bt/Imas.htm) uses open-source software that provide for some of the steps involved in MAS, e.g., experimental design, phenotype and genotype data analyses, and detection of marker–trait associations. The iMAS package needs to be integrated with the International Crop Information System (http://www.icis.cgiar.org) and modeling and simulation tools for molecular breeding. The International Crop Information System has valuable information required for breeding programs, including information on genetic resources, gene banks, and molecular breeding. A simulation tool called QuLine/QuCim uses relevant genetic data from varied sources to predict cross performance and compare selection methods. The capability of this tool has been extended to MAS including gene pyramiding. Results obtained by using these simulations would help breeders optimize breeding programs and, thereby, enhance breeding efficiency (Xu and Crouch 2008).

Plant breeding programs routinely evaluate various types of genetic materials, including germplasm lines, advanced breeding lines, experimental inbreds and hybrids, etc., often in more than one environment sampled as locations and years; most of these materials are also genotyped for markers. The huge amount of marker genotype and trait phenotype data generated in this way can be stored as per a standardized format and used for detecting marker–trait associations that would be transferable across breeding populations. However, much of this data would be highly unbalanced and structured, but suitable statistical designs may be developed for their analyses. For example, mixed models have already been used for detection of some marker–trait associations from such data in maize (Bernardo 2008).

The data from germplasm related to the cross to be subjected to MARS may be mined to obtain estimates of marker effects, from which a prior index could be computed. A *prior index* is based on marker and phenotype data from materials other than the population being subjected to selection. In contrast, an ad hoc *index* is based on data from the same population in which it will be used. For example, F_2 plants from the cross may be selected on the basis of marker data using prior index, their F_3 progenies would be phenotyped, and the F_2 marker and F_3 phenotype

Table 9.4 The relevance of genetic architecture of quantitative traits to the MAS strategy used for their improvement

	Trait governed by	
Feature	Few major QTLs	Many small effect QTLs
Activities	QTL mapping, followed by pyramiding/ introgression	Detection of marker–trait association (MTA) and MARS or genomic selection (GS) without MTA detection
QTL source	Unique germplasm	Elite germplasm
QTL localization	Precise localization is important	Not required; knowledge of only linked markers required for MARS
Significance level in QTL detection	Stringent to avoid false QTLs	Relaxed; in fact, MARS response increases with relaxation in significance levels
Markers used for MAS	Three markers for each QTL (two markers flanking the QTL and one within the QTL region)	Markers spaced at 10–15 cM for MARS[a]; more dense markers used for GS

Based on Bernardo (2008)

[a]One marker may be linked to one or more QTLs

data would be used to estimate the ad hoc index for further selection in the population. This would reduce the population size for elaborate phenotyping, and the estimates of ad hoc index would be based on superior F_3 progenies. Further, emphasis would shift from breeding schemes like MABC and MARS that require QTL mapping before MAS to those like genomic selection, which use MAS without QTL detection. This shift is more likely for traits controlled by many QTLs and having low to moderate heritability (Table 9.4).

9.16 Achievements

The first report on application of MAS concerned resistance to soybean cyst nematode (Concibido et al. 1996). Private sector has rapidly adopted MAS, particularly for crops like maize, soybean, canola, cotton, and sunflower. Efforts are being made to develop "ideal" genotypes by combining favorable segments from various sources using MAS. In the public sector breeding programs, MAS has been used mainly for gene introgression/pyramiding. In several instances, MAS has led to the development of improved varieties. The first variety developed by MAS was a maize hybrid offered for commercial cultivation in the USA in 2006 by Monsanto, USA. Two rice varieties, Cadet and Jacinto, developed through MAS have been released in the USA; they have unique cooking and processing qualities, including reduced amylose content. In Indonesia, rice varieties Angke and Conde were developed by transfer of BB resistance genes; they were BB resistant and gave 20 % higher yields than IR64. In Australia, barley varieties Sloop SA and Sloop Vic were developed by MABC to transfer multiple disease resistance into the popular variety Sloop. MAS was used to introgress three BB resistance genes (*xa5*, *xa13*, *Xa21*) into the rice variety Sambha Mahsuri (BPT-5204), and the BB-resistant version was released in India as "Improved Sambha Mahsuri." Similarly, BB resistance genes *xa13* and *Xa21* were introgressed into the Basmati rice variety Pusa Basmati 1 (PB 1) using MAS, and the BB-resistant version was released as "Improved Pusa Basmati 1" (IPB 1). IPB 1 gives 11.9 % higher yield than PB 1. MAS was used to introgress a major QTL for submergence tolerance (*Sub-1*) into some popular rice varieties. These improved versions, e.g., "Swarna-Sub 1" (recurrent parent rice variety Swarna) and "Sambha Mahsuri-Sub1," show increased submergence tolerance, which is reflected in their higher yields under submergence conditions.

MAS has been used to introgress the *o2* allele into the parental inbreds of the single cross hybrid, Vivek Hybrid Maize 9. The derived

inbreds were used to develop an extra early quality protein maize hybrid, "Vivek QPM 9." The kernel characteristics and yield of Vivek QPM9 are comparable with those of its parental version, i.e., Vivek Hybrid Maize 9. In addition, it has substantially higher lysine, tryptophan, and iron contents. A downy mildew (*Sclerospora graminicola*)-resistant pearl millet hybrid, "HHB 67-2," has been developed by MAS. MAS was used to transfer QTLs for downy mildew resistance from ICM451 into the male parent, H77/833-2, of the hybrid. MAS was used to transfer BB resistance genes *xa13* and *Xa21* into the CMS, maintainer, and restorer (PRR78) lines of the scented long grain hybrid Pusa RH10. In addition, the blast (*Magnaporthe oryzae*; Syn., *Pyricularia oryzae*) resistance genes *Pi54* and *Piz5* were introgressed into PRR78 using MAS. The "Improved Pusa RH10," obtained by crossing the parental lines modified as above, was resistant to the BB and blast and was comparable to the parental hybrid Pusa RH10 in agronomic performance, including yield.

Questions

1. Briefly discuss the various applications of molecular markers in plant breeding.
2. Briefly explain the concepts of foreground and background selections and their relevance in backcross programs.
3. "The use of molecular markers makes selection easier and more efficient and can greatly accelerate cultivar development. But it also adds to the workload of the breeders." Comment on this statement with the help of relevant information.
4. "Molecular markers greatly facilitate gene pyramiding and multitrait introgression." Discuss this observation in the light of available information.
5. Briefly describe the use of molecular markers during recurrent selection, and discuss its advantages and limitations.
6. A number of different breeding schemes have been proposed to make full use of the molecular marker data. Briefly describe any two of these schemes and discuss their usefulness in crop improvement.
7. Discuss the various applications of molecular marker technology in breeding of self-pollinated crops.
8. "The chief advantage of marker-assisted recurrent selection is the acceleration of the breeding program." Discuss this observation in the light of the available information.
9. Explain the meaning of recombinant selection. Describe the procedure for recombinant selection and discuss its potential and realized usefulness.
10. "Marker-assisted selection greatly facilitates gene pyramiding and QTL introgression." Discuss this statement with the help of suitable examples.
11. Give a brief account of the factors that limit the use of MAS in breeding programs, and indicate the areas for future improvement.
12. Explain the meaning of combined selection and discuss its relevance in marker-assisted backcrossing.
13. "Marker technology allows some such objectives to be realized that cannot be achieved through phenotypic selection." Comment on this statement with the help of suitable examples.
14. In what situations marker-assisted selection is expected to be more desirable than phenotypic selection?
15. "Marker-assisted QTL introgression often yields unexpected results." Discuss this statement with the help of suitable examples.

Genomic Selection 10

10.1 Introduction

Marker-assisted selection (MAS) is well-suited for handling oligogenes and quantitative trait loci (QTLs) with large effects. MAS has been extensively used mainly for backcross breeding, including pyramiding of useful genes/QTLs, and for marker-assisted recurrent selection (MARS; Sect. 9.10). Backcross breeding is a conservative strategy as it improves the recurrent parent only to the extent specified by the introgressed genes/QTLs. It does not generate new gene combinations that may be expected to enhance the performance potential and adaptation of the selected genotypes. It may be pointed out that the success of a genotype as a commercial variety depends on a combination of several traits, most of which are of quantitative nature. The expression of most quantitative traits is governed by one or few QTLs with relatively large effects along with several QTLs with small effects. In view of the above, MAS is not suitable for the improvement of quantitative traits as it is not designed to handle QTLs with small effects. The MARS scheme also is based on markers showing significant association with the trait(s) and for this reason has been criticized as inefficient. However, MARS attempts to take into account small effect QTLs by combining trait phenotype data with marker genotype data into a combined selection index. The genomic selection (GS) scheme was proposed by Meuwissen et al. (2001) to rectify the deficiency of MAS and MARS schemes. The GS scheme utilizes information from genome-wide marker data whether or not their associations with the concerned trait(s) are significant.

10.2 Genome-Wide Selection

GS is a specialized form of MAS, *in which information from genotype data on marker alleles covering the entire genome forms the basis of selection*. Thus, the effects associated with all the marker loci, irrespective of whether the effects are significant or not, covering the entire genome are estimated. The marker effect estimates are used to calculate the genomic estimated breeding values (GEBVs) of different individuals/lines, which form the basis of selection. The *GEBV* of an individual is the sum total of effects associated with all the marker alleles present in the individual and included in the GS model applied to the population under selection. The *breeding value* (*BV*) of an individual/line represents the expected phenotype of its progeny. *The BV is*, *therefore*, *determined by progeny testing and is based only on the additive genetic effects*. In contrast, the *genotypic value* of an individual/line is the phenotype expected from its genotype. *The genotypic value*, *therefore*, *is based on both additive and nonadditive genetic effects*. Since 1980s, phenotype data of individuals and their relatives have been used to calculate the estimated breeding values (EBVs)

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_10,

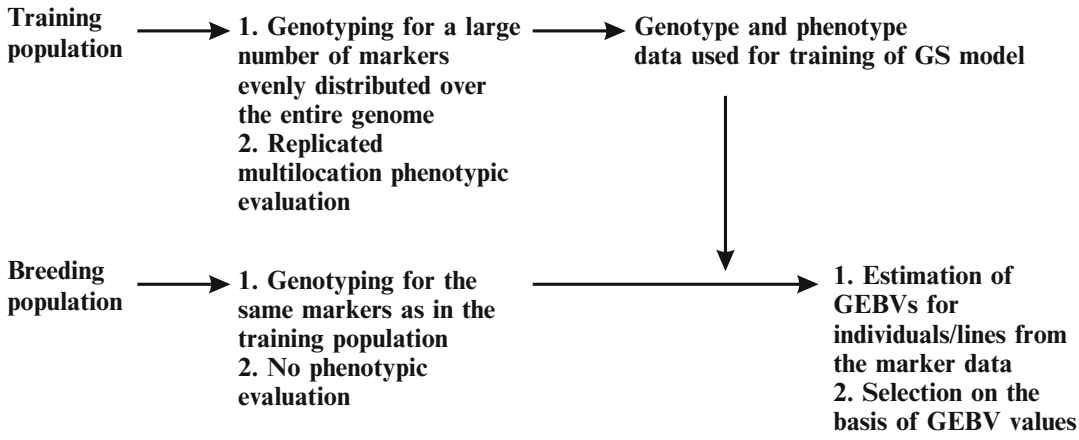


Fig. 10.1 A simple schematic representation of genomic selection (GS) scheme. The GS model training would need to be repeated regularly over time as new *lines* become available and are included in the breeding and training populations (Based on Heffner et al. 2009). *GS* genomic selection, *GEBV* genomic estimated breeding value

of individuals. These EBVs have been used for selection in animal and, more recently, plant breeding programs. When data on markers known to be linked to known QTLs were combined with the phenotype data for computing EBVs, the gains from selection showed substantial increase in animal breeding experiments. Similarly, when known QTLs were included in the GS model, the targeted QTLs were accumulated in much higher frequencies than when the standard ridge regression was used. A GS model that uses information about known QTLs has been termed as *gene-assisted genomic selection* (see Rutkoski et al. 2010).

10.3 A Generalized Procedure for Genomic Selection

The GS method is based on two separate, but related, populations, viz., a training population and a breeding population (Fig. 10.1; see Heffner et al. 2009). The *training population* is used for training of the GS model and for obtaining estimates of the marker-associated effects needed for estimation of GEBVs of individuals/lines in the breeding population. The *breeding population*, on the other hand, is the population subjected to GS for achieving the desired improvement and isolation of superior lines for use as new varieties/parents of new improved hybrids.

1. The first step in a GS program is to create a training population suitable for the concerned breeding population (Sect. 10.4).
2. The individuals/lines in the training population are genotyped for a large number of markers evenly distributed over the entire genome at adequate density.
3. The individuals/lines in the training population are subjected to extensive phenotypic evaluation for the target trait(s) in replicated trials over locations and, preferably, years.
4. The phenotype and marker genotype data are used for computing the GS model parameters; this is called *model training*. Model training can be performed repeatedly to include data on new markers and additional traits. The estimates of GS model parameters are retained for subsequent application to the breeding population.
5. The breeding population is evaluated for the same set of markers that was used for estimation of the model parameters in the training population. *There is no phenotypic evaluation of the breeding population*.
6. The GEBVs of individuals/lines of the breeding population are calculated from their marker genotype data and the marker-

associated effects estimated from the training population.

7. The superior individuals/lines are selected from the breeding population on the basis of their GEBV estimates (Fig. 10.1).

This GS procedure can be combined with an appropriate breeding scheme to achieve the desired objectives (Sect. 10.8).

10.4 Training Population

The training population must be representative of the breeding population. It should maximize the proportion of trait variance associated with the markers. This can be achieved by including in the population such individuals/lines that have divergent GEBVs. The training population should exhibit low collinearity between markers. Colinearity between markers is disturbed by recombination; therefore, the individuals/lines included in the training population should have undergone several rounds of recombination. Low colinearity between markers is needed since high colinearity tends to reduce prediction accuracy of certain GS models. Finally, the training population should adequately represent the genetic diversity present in the breeding population. This could be achieved by selecting individual/lines from the breeding population on the basis of some form of cluster analysis and including them in the training population (Jannink et al. 2010).

10.4.1 Genetic Composition

The composition of training population remains perhaps one of the most difficult decisions of the GS procedure. The training population should be large and should consist of the parents or very recent ancestors of the concerned breeding population. Alternatively, the training population may be unrelated to the breeding population. In this case, the two populations are likely to differ for marker and/or QTL alleles and their frequencies and the effects of their genetic backgrounds. As a result, the accuracy of GEBV predictions of the individuals of breeding population on the basis of marker effects estimated from the training population may be much lower than expected. This would happen even when sufficiently large number of markers and a large training population is used. The available evidence indicates that *the training population should include individuals related to the breeding population for high GEBV accuracies* (see Rutkoski et al. 2010).

A training population may consist of historical data or it may be a real population consisting of existing individuals. In the case of dairy cattle, training populations usually consist of historical data. In the case of plants, several types of real training populations can be created, e.g., biparental crosses, doubled haploid testcrosses, and intermated inbred lines. *Ideally*, *a new training population should be generated for each breeding population*. This approach will lead to high accuracy in GEBV prediction because the breeding population would be directly related to the training population. As a result, the two populations will share genetic background, QTL effects, moderate allele frequencies, etc. But this will necessitate precision phenotypic evaluation of a separate training population for each breeding population in the target set of environments. This will delay the progress of the breeding program and will add to the cost.

Alternatively, there may be a single training population for the entire breeding program. This population would consist of samples of individuals/lines drawn from all the breeding populations being handled in the breeding program. GS models trained on such a populations would be able to accurately predict GEBVs of individuals from each breeding population represented in the training population. Several simulation studies indicate that GS models trained on such populations are able to predict GEBVs of the concerned populations with high accuracy, particularly when very high marker densities are used. This approach would reduce the cost as well as the duration of selection cycles. Improvements in marker technology may be expected to allow the use of extremely dense markers for genotyping very large training populations composed of representative samples

of lines/individuals from different populations of the breeding program. If it were assumed that QTL effects are conserved across populations, i.e., QTL × genetic background interaction is not significant, the use of extremely high marker densities and very large training populations should enable accurate prediction of GEBVs of individuals distantly related to the training population. However, this assumption may be unrealistic, and QTL × genetic background interactions have been reported in several cases. Therefore, a greater research effort is needed to determine the optimum composition of training populations for maximizing GEBV prediction accuracies with the available resources.

In plant breeding programs, large amounts of phenotype data are routinely generated. These data could be used as hypothetical training populations for training of GS models without the cost incurred for precision phenotyping. Generally, evaluations in breeding programs are done in two stages. In the first stage, a relatively large number of lines are evaluated in a smaller number of environments. But in the second stage, a smaller number of lines are evaluated in a large number of environments. The heritability of traits is higher in the second evaluation stage than in the first evaluation stage, making the second stage data more desirable for use as training population. But unidirectional selection may generate considerable bias in the genetic composition of such a training population, which may adversely affect the accuracy of GS models (Zhao et al. 2012b). In one study, phenotype and marker genotype data on 788 testcross progenies derived from segregating populations of crosses between elite maize inbreds were used to construct five different types of training populations. The GS models trained on these populations were used for GEBV estimation of the concerned individuals. There was a substantial loss in the GEBV accuracy when the training population was derived by unidirectional selection. However, there was only a marginal loss in the accuracy when the training population was derived by bidirectional selection. *Thus, training populations derived by bidirectional selection applied to phenotype data generated in plant breeding programs would reduce costs and lead to only marginal reduction in prediction accuracy.*

In conclusion, *it is desirable that the training and breeding populations have similar genetic composition, and the two are closely related.* More importantly, the training and the concerned breeding populations should have either equal or, at least, comparable linkage disequilibrium (LD; Sect. 8.6) decay rates. If, for example, the rate of decay of LD between markers and QTLs is greater in the breeding population than in the training population, the strength of association between the markers and QTLs will decline, and this will reduce the accuracy of GEBV estimates. This situation will arise, for example, when the breeding population is genetically divergent from the training population or when the two populations are related, but separated by several generations from each other. Since the allele frequencies and LD structure in the breeding population would change due to selection, *the training population should be updated by including individuals/lines selected from the breeding population. In addition, the GS model should be retrained on each updated version of the training population.* In fact, training of the GS model should be carried out for more than one generation to achieve the maximum prediction accuracy.

10.4.2 Population Size

In general, the accuracy of GEBV prediction increases almost linearly with the population size. Therefore, the training population should be adequately large, particularly when it is not closely related to the breeding population. In a simulation study, the GEBV prediction accuracy was 0.848 when the training population consisted of 2,200 individuals/lines, but it declined to 0.708 when the population size was reduced to 500. Further, the size of training population should increase with an increase in the effective breeding population size. Finally, a larger training population should be used for traits with low heritability than for traits with higher heritability. In general, the ratio of training to breeding

population should be higher when genetic diversity is greater, the size of breeding populations is smaller, the traits have lower heritability, and the number of QTLs involved in control of the trait is larger. Further, the appropriate training population size would be smaller for self-pollinated species than for cross-pollinated species.

10.4.3 Marker Density

Ideally, the number of markers should be sufficiently large so that the maximum number of QTLs affecting the trait is in strong LD with at least one marker. Therefore, marker density will depend on the extent of LD in the concerned species: in general, marker density should be much greater in cross-pollinated than in self-pollinated species. Further, marker density would be much greater for traits with low heritability than for those with high heritability. It is, however, not easy to determine the adequate marker density for a given crop species since different populations of a single species and even different genomic regions of a single individual tend to show considerably different LD estimates. *Generally*, *GEBV accuracy improves with marker density up to a point, beyond which there is little improvement*. In most cases, however, it may neither be feasible nor affordable to achieve the ideal marker density, and useful estimates of GEBV may be obtained even without marker saturation. Further, evenly spaced, low-density markers may predict GEBVs with lower accuracy than those selected on the basis of their additive-effect size on the concerned trait.

10.5 Computation of Genomic Estimated Breeding Values

In the GS scheme, genotype data on genome-wide markers are analyzed with trait phenotype data to estimate the trait phenotypic effects associated with each of these markers. A strong LD between markers and QTLs governing the trait is assumed to ensure a consistent linkage between them across different families of the breeding population. But when a large number of marker effects, called *predictors* (designated as p), are to be estimated from a much smaller number of phenotypic observations (denoted by n), the degrees of freedom available for the predictors is not sufficient. In addition, there may be a high degree of colinearity among the marker effects. Therefore, GS prediction models use information from all the markers so that the estimates of marker effects would be unbiased and without exaggeration. A number of marker effect estimation models have been developed, some of which are as follows: (1) stepwise regression, (2) ridge regression, (3) Bayesian estimation models, (4) semi-parametric regressions like kernel-based regression and neural networks, and (5) machine learning methods like random forest and support vector machine models (Jannink et al. 2010).

10.5.1 Stepwise Regression

The *stepwise regression* approach treats marker effects as fixed and has been the basis for traditional MAS. It avoids the large p, small n problem by fitting the markers into the model either singly or in small groups. In this process, only those markers that are associated with significant effects on the trait are retained, while other markers are discarded. The experimenter arbitrarily selects the level of significance. The markers having nonsignificant effects are assigned "zero" effect value, and the effects of significant markers are estimated. This approach is generally followed in QTL mapping; it tends to overestimate marker effects, and only a portion of the genetic variance is accounted for by the markers. A GS model based on this approach detects a limited number of QTLs and the accuracy of GEBV is low. In general, when significance thresholds are less stringent, there is an increase in the accuracy of GEBV prediction (see Heffner et al. 2009).

10.5.2 Ridge Regression

The *ridge regression* method was proposed by Whittaker et al. (2000) for MAS in biparental populations. Meuwissen et al. (2001) proposed the use of this method for calculating the best linear unbiased predictor estimates simultaneously for all the markers by treating the markers as random effects. In this model, all the marker effects are considered to belong to a normal distribution with mean zero and variance σ_g^2, where σ_g^2 is obtained by dividing the genetic variance by the number of marker effects. Thus, ridge regression model shrinks all marker effects toward zero although they are likely to differ from each other substantially. The assumption of equal marker variance is, therefore, unrealistic, and shrinking of marker effects toward zero over-shrinks large marker effects. Yet this approach is superior to stepwise regression as it avoids the bias introduced by the selection of the markers with significant effects, and is more appropriate when there are many QTLs with small effects (see Heffner et al. 2009). Results from simulation studies show that GEBV estimates from ridge regression approach show greater accuracy than those from stepwise regression and phenotype-based best linear unbiased predictor approaches. This superiority of ridge regression is higher for traits with lower heritability.

10.5.3 Bayesian Approach

The *Bayesian approach* relaxes the assumption of equal marker effects and a common variance; it estimates a separate variance for each marker and accommodates marker effects of different sizes. Meuwissen et al. (2001) proposed two Bayesian models called BayesA and BayesB. In the case of *BayesA*, the marker variance distribution is an inverted chi-square distribution. The degrees of freedom and the scale parameters of this distribution are chosen in such a way that mean and variance for the distribution match the mean and variance for the marker. In a modification of BayesA, more marker effects and variances are shrunk close to zero, but not zero. The modified BayesA detected large effect QTLs and provided better estimates of QTL effects and locations than the original BayesA model. In contrast, the *BayesB* model allows some markers to have zero effects and variances, while other markers may have effects greater than zero and an inverted chi-square distribution for their variances. In a simulation study, the BayesB model had superior GEBV prediction accuracy than BayesA, stepwise regression, and ridge regression approaches. In addition, BayesB model is the least demanding in terms of computation, and the GEBV accuracy does not decline with an increase in the number of markers. Thus, the Bayesian methods appear to be better than other approaches when there is high density of markers and limited number of phenotypic records. However, at higher marker densities, computational issues may arise that need to be resolved by an improvement in the statistical methods (see Heffner et al. 2009).

Some other Bayesian regression methods have also been developed. A Bayesian shrinkage procedure based on the idea of GS has been further modified to obtain more accurate estimates of QTL positions and effects. These models can be regarded as a form of GS and can be used for prediction of GEBVs. Similarly, Bayesian LASSO (least absolute shrinkage and selection operator), an additive linear regression model, has been implemented in the BLR (Bayesian linear regression) package of R (Park and Casella 2008; de los Campos and Perez 2010), now as BLR version 1.3. Bayesian LASSO, like ridge regression, is a regularization method, in which the parameters do become zero on reduction; in contrast, they do not become zero in the case of ridge regression. LASSO tends to capture a small number of QTLs with large effects, while ridge regression will capture many QTLs with small effects (Heffner et al. 2009). The Bayesian LASSO model is similar in overall GEBV prediction accuracy to some of the semi-parametric regression models (Sect. 10.5.5) and is not affected by redundant interactions between markers.

10.5.4 Semi-parametric Regression Methods

There is considerable evidence for the involvement of epistasis in control of quantitative traits in plants. The use of appropriate contrasts allows linear regression models to accommodate interactions between two or more marker loci, but this is not feasible for the large number of markers used in GS. This makes parametric modeling of complex epistatic interactions impractical. But semi-parametric regressions (Gianola et al. 2006) like kernel-based and neural network methods permit the inclusion of higher-order epistatic interactions in GS models. The *reproducing kernel Hilbert spaces* (*RKHS*) method can capture more complex epistatic interactions than linear regression models. Both simulation and empirical studies indicate RKHS to be superior to linear models in prediction of phenotypes of some quantitative traits (Crossa et al. 2010; Gonzalez-Camacho et al. 2012). But a potential limitation of RKHS method is the necessity to define a priori, by selection of the kernel, the basis functions used for regression. The selection of an inappropriate kernel may limit the ability of RKHS to capture complex interactions. In contrast, the basis functions used for regression in the case of *neural networks* are estimated from marker data. The regression in the *standard single hidden layer neural networks* procedure may be viewed as being performed in two stages. In the *first stage* (*the hidden layer*), the basis functions are estimated from a linear combination of marker genotypes, and the scores are transformed using nonlinear activation function. In the *second stage* (*the output layer*), the phenotype is regressed on the basis functions employing a nonlinear procedure. This gives neural networks a great flexibility in capturing complex interactions. However, the computational load of the procedure is extremely high and there is a risk of over-fitting the training data.

A class of neural networks, called *radial basis function neural networks* (*RBFNNs*), is closely related to RKHS. The RBFNN methods combine features of both neural networks and RKHS, and have greatly reduced computational burden as compared to that of the standard single hidden layer neural networks. In an RBFNN, a predetermined number of radial basis functions are used as the basis functions, each of which is indexed by parameters to be estimated from the data. The output layer of RBFNN may be subjected to regularized regression, shrinkage regression, or ordinary least squares procedure. Generally, all the markers are assigned equal weights, but differential weights may also be assigned. Simulation studies indicate that RKHS and RBFNN can capture epistatic interactions. But the inclusion of redundant interactions between markers can reduce their accuracy. This situation is quite likely when high-density markers are used. In contrast, linear additive regression models are not affected by the inclusion of redundant interactions between markers. The accuracies of GEBV prediction by additive Bayesian LASSO, RKHS, and RBFNN procedures were compared using a very large dataset from maize (300 tropical maize inbred lines, 21 trait–environment combinations, 55,000 SNP loci). The three models were found to be comparable with slight superiority of RKHS and RBFNN over the Bayesian LASSO model (Gonzalez-Camacho et al. 2012).

10.5.5 Machine Learning Methods

The *machine learning methods* were developed for resolving classification problems. They have been extended to regression analysis of data with large p and small n conditions. The *support vector machine model* maps samples from the predictor space to a high-dimensional feature space via a nonlinear mapping function. Then, it performs linear regression in the multidimensional space. *Random forest* is a complete predictor that consists of a collection of predictors structured like trees. In turn, each tree is grown on the basis of a bootstrapped sample of the training dataset and predicts the target response. The "forest," on the other hand, predicts the target response as an average of the tree

predictors. These models will be particularly useful when epistasis makes significant contribution to the genetic variance for a trait (see Jannink et al. 2010).

10.6 Factors Affecting the Accuracy of GEBV Estimates

The *accuracy of GEBV estimates* may be defined as Pearson correlation between GEBV and true breeding value estimates. This measure of accuracy is directly proportional to gain from selection based on GEBV. As a result, $R = ir\Sigma A$, where R is response to selection, i is selection intensity, r is accuracy of GEBV estimates, and ΣA is square root of additive genetic variance of the true breeding value estimates. Thus, for any given value of i and ΣA, the accuracy of GEBV estimates would determine the response to selection, i.e., effectiveness of GS. The effectiveness of GS procedure may be assessed on the basis of theoretical considerations, simulation studies, and empirical findings. The following factors would affect the effectiveness of GS: (1) the method used for estimation of marker effects, (2) polygenic effect term based on kinship, (3) the method of phenotypic evaluation of the training population, (4) marker type and density, (5) heritability of the trait and the number of QTLs involved, and (6) breeding population.

10.6.1 The Method of Estimation of Marker Effects

Theoretical studies have shown that GS methods estimate marker allele effects on the basis of the following: (1) strong LD between markers and QTLs and (2) genetic relationships between individuals as indicated by their marker genotype data. For example, ridge regression captures genetic relationships between individuals more effectively than BayesB, while the latter is more effective in capturing the strength of LD between markers and QTLs. Simulation studies show that 39 % and 21 % of the accuracy of GS based on ridge regression and BayesB, respectively, is due to genetic relatedness of individuals. But the current estimations of GS accuracy take into account only such markers that are in strong LD with QTLs. Simulation studies indicate superiority of BayesB over ridge regression when additive gene action is assumed (Meuwissen et al. 2001; Habier et al. 2007). But analysis of actual data from two-row barley suggested a slight superiority of ridge regression over BayesB (Zhong et al. 2009). It has been concluded that GS provides greater accuracies than predictions based on pedigree information alone. Therefore, GS can be used to accelerate breeding cycle by effecting several cycles of marker-based selection in year round nurseries/greenhouses prior to phenotypic evaluation. Further, BayesB may be expected to work better than ridge regression only when markers are in strong association with the QTLs; this can be expected only when the QTL effects are large (Jannink et al. 2010).

10.6.2 The Polygenic Effect Term Based on Kinship

The breeding value estimated from phenotypic data varies according to the additive relationship (A) matrix. The *A matrix* or *polygenic effect term* contains the proportion of alleles that are identical by descent for each pair of individuals. The markers used for estimation of GEBV would not be able to capture the whole of genetic effects for the trait, and the residual portion of effects is expected to vary according to the A matrix. The A matrix can be calculated from either pedigree data (estimates of *expected relationships*) or from marker data (estimates of *realized relationships*). *When marker density is adequate, the A matrix obtained from marker data will be a better reflection of the true relationships than that derived from pedigree data*. This is because the realized relationships are likely to diverge from the relationships deduced from pedigree data due to segregation and recombination, segregation distortion, selection, and errors in pedigrees records. The usefulness of inclusion of the polygenic effect term for GEBV

estimation is greatly affected by marker density. At lower marker densities, the inclusion of polygenic effect term tends to increase the accuracy of GEBV estimates. But at higher marker densities, almost all the genetic effects will be captured by the marker data since all the QTLs affecting the trait would be in LD with at least one marker. Further, even such markers that do not show LD with any QTL will provide information on genetic relationships among the individuals; this will increase the accuracy of GEBV estimates (see Heffner et al. 2009).

10.6.3 The Method of Phenotypic Evaluation of Training Population

Theoretically, GS accuracy would be greater when a large number of individuals of the training population are evaluated in un-replicated trials than when a small number of individuals are evaluated in replicated trials. Simulation studies show that the appropriate method of phenotypic evaluation depends primarily on the relatedness of the training population to the breeding population. When the breeding population was only one generation removed from the training population, evaluation of a smaller number of individuals with a larger number of replications gave higher accuracy than that of a larger number of individuals without replication. The increased number of replications leads to higher estimates of heritability, which enhances the accuracy of GS predictions. However, when the breeding population was removed from the training population by four generations, an increase in the number of genotypes rather than the number of replications resulted in higher accuracies of GEBV estimates (Zhong et al. 2009).

10.6.4 The Marker Type and Density

In general, dominant markers have lower LD detection power than codominant markers. Further, the LD detection power of dominant markers improves when three loci are used for LD analysis. Therefore, marker haplotypes used for GEBV prediction should be based on dominant markers. Simulation studies show that marker density (number of markers per Morgan) should increase with effective breeding population size to achieve comparable prediction accuracies from populations of different sizes (Solberg et al. 2008). The increased marker density seems to be better exploited by some GS models, e.g., BayesB, than by some other models, e.g., ridge regression. Further, for a given breeding population size, accuracy increases with marker density, but the advantage becomes smaller as the marker density becomes higher. Marker type has a marked influence on marker density requirement, e.g., the density of SNP markers should be two to three times higher than that of SSR markers for achieving comparable accuracies (Meuwissen 2009). In addition, a greater accuracy is obtained when all the markers are considered separately than when pairs of markers are grouped into haplotype blocks (Solberg et al. 2008). Simulation studies reveal that in biparental populations, the optimum marker densities are very low (eight markers per Morgan), and higher marker densities tend to depress accuracy. Lower marker densities can also be used if the individuals being subjected to GS are the progeny of the training population. In such a case, the use of only one marker per 10 cM is expected to result in a loss of only 4–6 % in GEBV prediction accuracy. Finally, information about the rest of the markers of an individual may be deduced from marker genotypes of its parents on the basis of co-segregation.

10.6.5 Trait Heritability and the Number of QTLs Affecting the Trait

The accuracy of GEBV prediction declines with trait heritability, but an increase in the training population size can compensate for this decline. At high heritability, the accuracies of both GEBV estimates and phenotypic selection will be high; therefore, phenotypic selection may be

as effective as GS. However, *GS is expected to perform much better than phenotypic selection when trait heritability is low*, *particularly when the GS model is trained with at least two generations of the training population* (see Rutkoski et al. 2010).

10.6.6 The Breeding Population

In most theoretical and simulation studies on GS, the training population is designed to capture population-wide LD. Here, *population* may mean the entire breeding program on a crop; a market class of a crop, e.g., vegetable common bean; or a single segregating population of the crop being subjected to GS. Many breeding populations, particularly of self-fertilizing species, are biparental populations derived from a cross between two homozygous parents. Simulation studies on ridge regression method of GS in biparental populations show that the accuracy of GS is only slightly (0–8 %) lower with the assumption of all marker variances being equal than the assumption of true marker variances being known. Further, GS outperforms phenotypic selection even with rather small, e.g., 35, effective population size. Thus, for GS in biparental populations, separate model training would be needed for each cross. But the need for phenotypic evaluation would slow the breeding cycle down in comparison to the program-wide GS model training (Jannink et al. 2010).

10.7 Effects of Genomic Selection on Genetic Diversity

The accuracy of breeding values estimated from phenotype data increases when information from relatives is included. This increases selection gain as well as correlation between the predicted breeding values of relatives. As a result, the probability of selection of related plants/lines also increases leading to a decline in genetic diversity in the selected populations. However, the information from relatives does not indicate the values of alleles each individual received from its parents. In contrast, GS utilizes the effects associated with the marker alleles due to their strong LD with QTL alleles affecting the trait. Therefore, the correlation between the predicted breeding values of relatives will be lower under GS than that between their traditional breeding value estimates. The findings from a simulation study show that with low marker density (400 markers) and small training population size (400 individuals), the major part of GS accuracy was due to genetic relationships; this is contrary to the theoretical expectations. However, at a high marker density (4,000 markers), most of the GS accuracy was due to LD, particularly when the number of QTLs was large (200 QTLs; Jannink et al. 2010).

It was concluded that the assumption of an infinite number of QTLs, each with extremely small effects, governing quantitative traits seems to be closer to the real situation than the assumption of few QTLs, each having large effect. Even if there were few loci at which some alleles produce large effects on the phenotype, these alleles may have very low frequencies so that each such allele would generate only small phenotypic variance. If all loci had several low-frequency alleles with large effects, the trait in question would show high heritability and considerable genetic variability. However, the extent of LD detected between the markers and the QTLs will generally be low; this would have a negative impact on the LD component of GS accuracy. It would also give rise to the situation where even for a trait with high heritability, only a small part of the genetic variation would be explained by markers in association studies, giving rise to the cases of the so-called missing heritability (Sect. 8.24; Jannink et al. 2010). Further, extensive mapping studies on flowering time in maize support the hypothesis that many variants that affect flowering time are located in clusters at a small number of loci. This organization of QTLs would generate high heritability, but would lead to low association between markers and QTLs. In addition, ridge regression would perform better than BayesB for such traits.

10.8 Integration of Genomic Selection in Breeding Programs

GS is a special form of MAS, but it differs from the latter in several important features (Table 10.1). GS can be easily integrated in any suitable breeding scheme. In a conventional breeding program, selection is based on phenotypic evaluation of the population subjected to selection. But in a breeding program based on GS, a training population is used to estimate marker effects. The estimated marker effects are used to predict GEBVs of individuals in the breeding population and selection is based on the GEBV values. Further, the selection of parents for hybridization is also based on their GEBV estimates. Thus, phenotypic evaluation is limited to the training population, and the lines are selected from the breeding population for release as varieties (Fig. 10.2). A simple GS-based recurrent selection scheme for self-pollinated crops is as follows. First of all, recombinant inbred lines (RILs) are developed from a cross between two homozygous lines. These RILs are evaluated in replicated multilocation trials for 1 or 2 years, genotyped for the markers to be used in GS, and the GS model is trained on these data. Superior RILs are selected on the basis of their phenotypic performance only, crossed with each other, and the F_2 generation from each cross is raised and subjected to GS. The F_2 plants with the best GEBV estimates are selected, crosses are made among the selected plants, and the F_2 generation from each cross is raised and subjected to GS. These steps can be repeated several times before homozygous lines are isolated from the population. In a crop like barley, three generations can be grown each year by using off-season nursery/greenhouse facilities. Thus, one selection (in F_2 generation) – intermating (among the selected F_2 plants) – selfing (raising of the F_1 generation) cycle can be completed every year (Bernardo 2010).

The above recurrent selection scheme may be modified as follows. The breeding population can be maintained indefinitely as *recombination*

Table 10.1 Main differences between GS and MAS

Feature	GS	MAS
Targeted QTLs	All QTLs affecting the trait	QTLs with significant and large effects
Basis of selection	GEBVs estimated from marker genotypes	Marker genotype
Number of markers used	Large number of genome-wide markers	Few markers linked to the targeted QTLs
QTL discovery, confirmation, and validation	Not required; QTL effects associated with the markers are estimated	Necessary for successful MAS
Model training	Necessary; based on a suitable training population	Not required
Population used for model training/QTL discovery	Related to the breeding population subjected to GS	Generally, not related to the population used for MAS; in fact, unrelated germplasm used for QTL confirmation and validation
Perpetuation of the population used for model training/QTL discovery	Population is maintained and regularly updated by addition of new lines	Population is generally not maintained and is certainly not updated
Phenotypic evaluation	Confined to the training population	During QTL discovery, confirmation, and validation
Overall objective of the breeding program	Improvement in the targeted quantitative traits	Introgression/accumulation of the targeted QTLs
Selection for multiple traits	The same set of markers are used for selection for all the traits	Different sets of markers must be used for each QTL

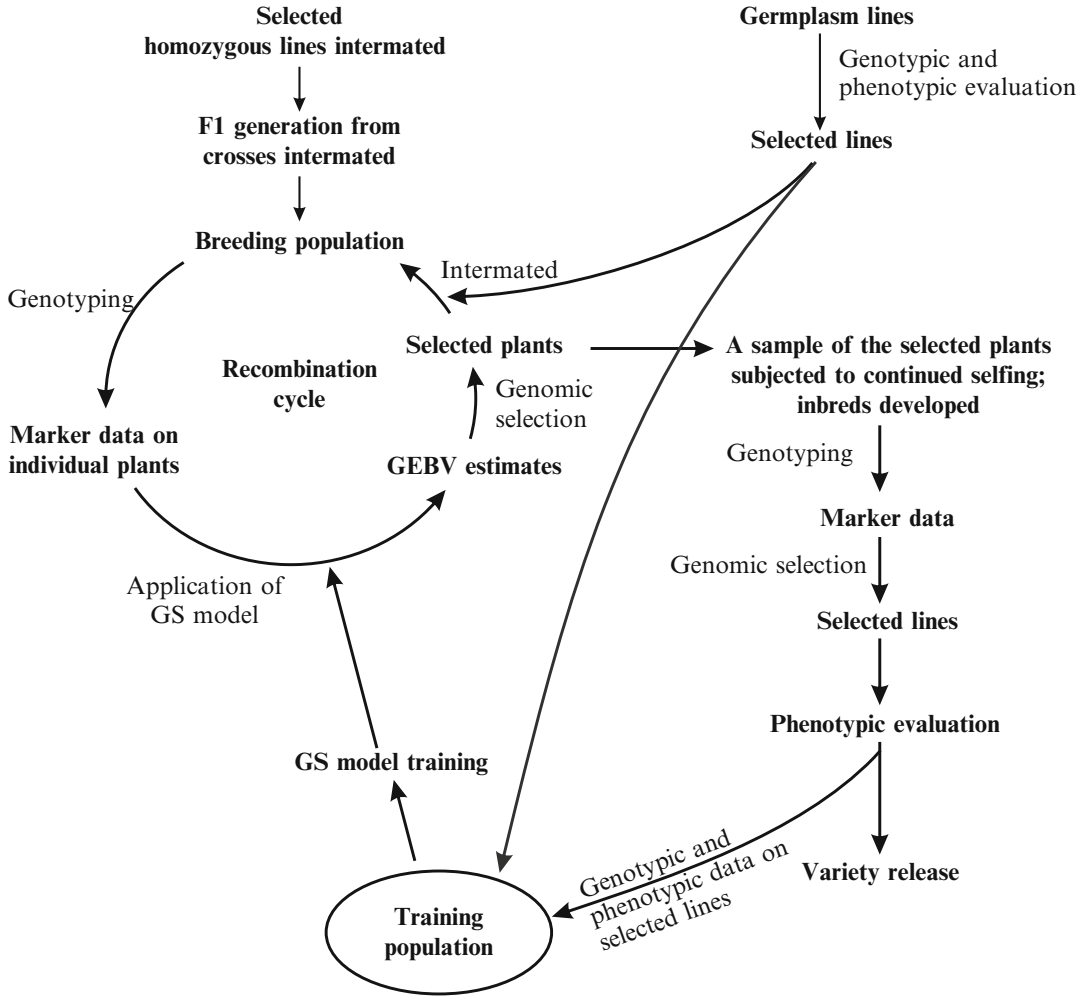


Fig. 10.2 A scheme of recurrent selection for GS in a self-pollinated crop. The scheme provides for inclusion of new germplasm, periodic development of purelines for variety release, continuous recombination cycle, and regular updating of training population and model training for maintaining GEBV accuracy (Modified from Rutkoski et al. 2010)

(selection–intermating–selfing) *cycle population*. Further, new germplasm lines can be included in the breeding and training populations to maintain diversity in the former and to update the latter. Samples of selected plants may be periodically taken out of the breeding population and subjected to continued selfing and GS to produce superior purelines that can be used as new varieties (Fig. 10.2). These selected lines would also be included in the training population. The use of such a regularly updated training population is expected to maintain the accuracy of GEBV estimates in the advancing generations of the breeding population. However, the inclusion of unadapted germplasm would require the use of very high marker densities in view of the low population-wide LD in a population that is composed of diverse individuals/lines. Further, a larger effective population size (*Ne*) would be needed to represent the increased genetic diversity. It has been estimated that the number of markers should be about $10 \times Ne \times$ genome size in Morgans for achieving GEBV accuracy of ~0.9 (Meuwissen 2009). In a species like

wheat (genome size ~35 Morgans), the number of markers required for a modest *Ne* of 100 would be $10 \times 35 \times 100 = 35{,}000$, which seems impractical at present. *Ne* values can be estimated from marker data using a computer program, but they are not in common use in plant breeding programs. Therefore, estimates of genome-wide LD may be more useful (see Rutkoski 2010).

In crop species like maize, a *two-step GS-based breeding scheme* may be used. In the *first step* of this scheme, plants of a segregating population, e.g., F_2, are genotyped for a suitably large number (e.g., ~250–500) of markers covering the entire genome. These plants are also evaluated for testcross performance in multilocation replicated yield trials. The marker genotype and testcross performance data are used to estimate breeding values associated with each marker allele; these estimates are used for the prediction of GEBVs of the F_2 plants. In the *second step*, the F_2 plants with the highest GEBV estimates are selected and crossed in all possible combinations to produce the population for the next cycle of selection. Thus, GS follows a simple select-and-intercross procedure, which allows one selection cycle to be completed in a single generation. As a result, up to three generations can be raised each year by using off-season nurseries/greenhouse facilities, and the time required for completing a given number of selection cycles can be greatly reduced (Bernardo and Yu 2007).

The introgression of exotic germplasm in breeding programs of a crop like maize requires pre-breeding involving 10–20 years of recurrent selection. This has severely limited the utilization of exotic germplasm. GS can facilitate the use of exotic germplasm in breeding programs in the following manner. An exotic inbred can be crossed with an adapted inbred and the F_2 and subsequent generations from this cross can be handled by the two-step procedure described above. This strategy would allow 7–8 cycles of GS to be completed in about 3 years time (assuming three generations per year). In case it is desired to continue GS beyond 7–8 cycles, the plants in cycle 7 or 8 should be genotyped as well as evaluated for testcross performance, marker effects should be estimated afresh, and the new estimates should be used for further GS. It was concluded that the F_2 population should be large (~300), testcross performance should be evaluated in replicated trials over locations, and F_2 generation is superior to BC_1 and BC_2 generations for the initiation of recurrent selection (Bernardo 2009).

A futuristic GS-based breeding program is envisioned to consist of two separate cycles of activities, viz., (1) model training and (2) line development cycles (Fig. 10.3). The *model training cycle* pertains to the training population and aims to continuously improve the accuracy of GEBV estimates. For this, the training population is continuously improved by inclusion of superior lines derived from the line development cycle. The newly added lines are evaluated phenotypically, but their marker genotype data are taken from the line development cycle. In the *line development cycle*, superior parents are selected for hybridization on the basis of their GEBV estimates, and the segregating generations from their crosses are advanced. The advanced generation lines are genotyped for the markers and subjected to GS, and the lines with the highest GEBV estimates are selected. The selected lines can be subjected to phenotypic evaluation to assess their suitability for release as new varieties. In any case, these lines are used for hybridization with each other to generate the population for the next cycle of GS. New germplasm can be inducted in the line development cycle: the promising germplasm lines are genotyped for markers and phenotypically evaluated, their GEBVs are calculated, and the lines with the highest GEBVs are selected as parents for hybridization. Further, MAS for important QTLs may be done in F_2 and F_3 generations to accumulate these QTLs, and GS may be applied in the F_5 generation for isolating superior lines. This procedure eliminates the plants/lines lacking the important QTL alleles, which reduces population size, marker genotyping work, and off-season nursery/greenhouse space requirement (Heffner et al. 2009).

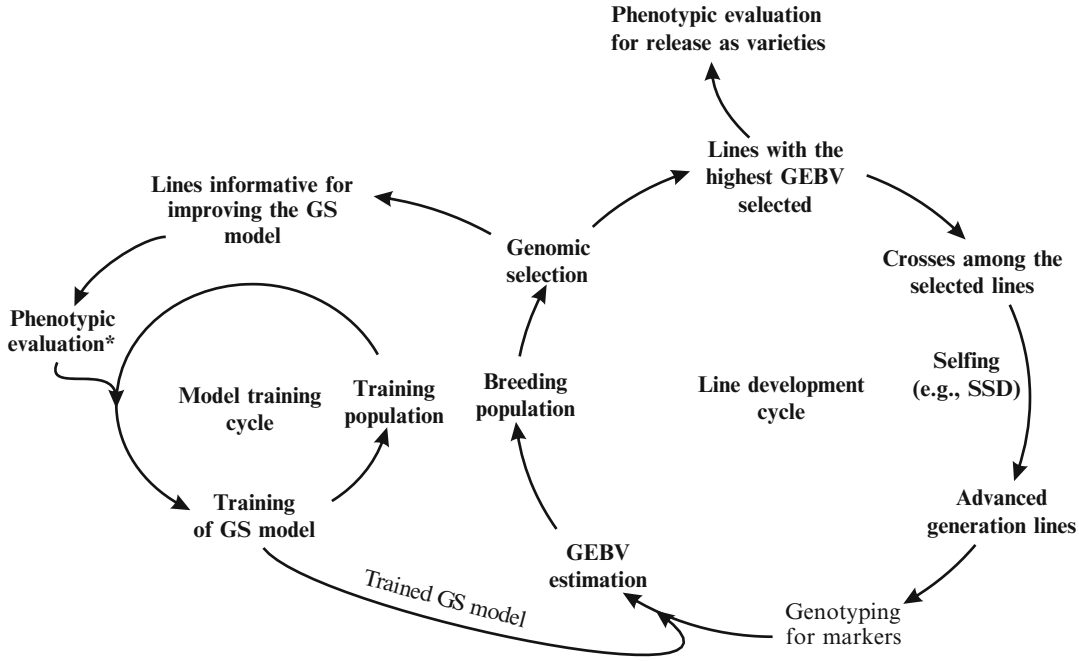


Fig. 10.3 Schematic representation of a breeding program based on genomic selection (GS): a futuristic view. *SSD* single seed descent, * marker genotype data taken from the line development cycle (Based on Heffner et al. 2009)

10.9 Effectiveness of Genomic Selection

The effectiveness of GS has been amply supported by empirical results in dairy cattle, and cattle breeding is being revolutionized by a rapid adoption of GS. In the case of plants, empirical evidence for the effectiveness of GS has been obtained mainly by *cross-validation*, in which real datasets existing for various plant populations are used for assessing the accuracy of GEBV prediction. For example, studies on cross-validation are available for biparental families of maize, barley, and *Arabidopsis* (Lorenzana and Bernardo 2009), for diverse panels of wheat (Crossa et al. 2010, 2011), for testcross progenies of maize inbred lines (Albrecht et al. 2011; Zhao et al. 2012a, 2013), and for inbred lines of maize (Gonzalez-Camacho et al. 2012; Table 10.2). Results from these studies indicate the effectiveness of GS as a breeding strategy as evidenced by high accuracies of the predicted GEBVs. Findings from some studies highlight the importance of genotype × environment interaction. Results obtained in some other studies, e.g., in maize, indicate that the accuracy of GEBV prediction is markedly influenced by the specific cross (see Nakaya and Isobe 2012; Zhao et al. 2012a). For example, findings from maize, barley, and *Arabidopsis* show that GEBVs based on significant markers only have consistently lower accuracy than those based on genome-wide marker data (Lorenzana and Bernardo 2009). Similarly, simulation studies with maize revealed 18–43 % larger selection gains under GS than with MARS (three cycles of selection), depending mainly on the number of QTLs (20, 40, and 100) involved in the control of the trait(s) (Bernardo and Yu 2007).

10.10 Advantages of Genomic Selection

1. The marker effects are estimated from the training population and used directly for GS in the concerned breeding population, and QTL discovery, mapping, etc. are not required.

Table 10.2 Some examples of empirical studies on GS in different crop species

Crop	Population size		Number of markers	GEBV accuracy[a]	GS model[b]	Reference
	Breeding population	Training population				
Maize	119	95	1,339	0.40–0.50	BLUP	Lorenzana and Bernardo (2009)
	349	28, 35, 70	160	0.59–0.72	BLUP	
A. thaliana	415	50–133	69	0.90–0.93	BLUP	
Barley	150	54, 96, 120	223	0.64–083	BLUP	
Maize	208	208	136	1.00	Several[c]	Piepho (2009)
Wheat	599	60	1,279	0.48–0.61	PM-RKHS[c]	Crossa et al. (2010)
Maize	300	270	1,148	0.42–0.79	LASSO	
Wheat	209	24, 48, 96	399	0.32–0.84	RR-BLUP	Heffner et al. (2010)
Wheat	174	24, 48, 96	574	0.41–0.73	RR-BLUP	
Maize	25 populations of 126–196	25–157 for each population	1,106	0.26–0.57	RR-BLUP	Guo et al. (2012)

Based on Nakaya and Isobe (2012)
[a]Correlation between observed phenotypic values and GEBVs
[b]*BLUP* best linear unbiased prediction, *PM-RKHS* pedigree information combined with molecular – reproducing kernel Hilbert space regression, *LASSO* Bayesian LASSO, *RR-BLUP* ridge regression-best linear unbiased prediction
[c]GS for only one trait; in the remaining studies, GS was practiced for three or more traits

2. Both simulation and empirical studies reveal that GS produces greater gains per unit time than phenotypic selection. For example, a simulation study in maize showed GS to be superior to MARS, particularly for traits having low heritability (Bernardo and Yu 2007). Further, GS is able to predict the performance of breeding lines more accurately than that based on pedigree data, and GS seems to be an effective tool for improving the efficiency of rice breeding (Spindel et al. 2015).
3. The selection index approach integrates appropriately weighted data from multiple traits into an index that serves as the basis for simultaneous selection for the concerned traits. The genome-wide marker data can be integrated into a selection index either alone or along with phenotype data on one or more traits. Simulation studies show that the above combined selection index approach of GS increases the effectiveness of selection, particularly for low heritability traits (Dekkers 2007; Heffner et al. 2010).
4. GS would tend to reduce the rate of inbreeding and the loss of genetic variability in comparison to selection based on breeding values estimated from phenotype data; this would be achieved without sacrificing selection gains. This may be particularly important in species that show severe inbreeding depression.
5. In the case of GS, phenotyping for every selection cycle in the breeding population is not required. This greatly reduces the length of breeding cycle, particularly in perennial species. For example, GS was estimated to reduce the selection cycle time from 19 years to merely 6 years in case of oil palm (*Elaeis guineensis*). Further, GS was estimated to outperform MARS and phenotypic selection even with a population size of 50 when selection gain was considered on per unit time and cost, but not on per selection cycle, basis. The selection cycle is reduced because GS does not require evaluation of testcross performance of the plants being subjected to selection, which is necessary in the case of phenotypic selection. In perennial species, GS is expected to facilitate commercialization of improved genotypes at much shorter intervals of time than phenotypic selection (Wong and Bernardo 2008).

6. GS may allow breeders to select parents for hybridization programs from among those lines that have not been evaluated in the target environment. This selection would be based on GEBVs of these lines estimated for their adaptation to the target environment. This would facilitate germplasm exchange and their utilization in breeding programs.
7. Genotype × environment interaction is an important component of phenotype and its estimation is quite demanding. GS can utilize information on marker genotype and trait phenotype accumulated over time in various evaluation programs covering a variety of environments and integrate the same in GEBV estimates of the various individuals/lines. This would allow GEBV estimation even for traits for which they have never been tested.
8. Theoretically, GEBV estimates can be used for the selection of parents for hybridization programs and, possibly, for the development of hybrid varieties. These applications, however, must await validation of the concept in practice.

10.11 Limitations of Genomic Selection

1. GS has still not become popular with plant breeding community primarily due to insufficient evidence for its practical usefulness. In fact, most discussions on its usefulness are largely statistical treatments and simulations that are not easily appreciated by plant breeders.
2. The potential value of GS should be assessed with caution because GS has been mostly evaluated in simulation studies. There is an urgent need to evaluate GS in breeding situations to demonstrate its practical usefulness.
3. The marker effects and, as a result, GEBV estimates may change due to changes in gene frequencies and epistatic interactions. This would necessitate updating of the GS model with every breeding cycle so that the gains from GS are not reduced.
4. The accuracy of GEBV estimates has been evaluated using simulation models based on additive genetic variance. These models ignore epistatic effects, which does not seem to be realistic. It has been argued that since epistasis makes only a small contribution to the breeding value, the use of only additive genetic models for GS may be expected to maximize selection gains (Heffner et al. 2009). However, this argument will be fully valid only for self-fertilizing species, where homozygous lines are used as parents as well as varieties. But in other species, progeny performance will depend on dominance and epistatic gene effects as well. The GS, therefore, is not very effective for traits with low narrow sense heritability (see Nakaya and Isobe 2012).
5. Our knowledge about the genetic architecture of quantitative traits is severely limited, which limits our ability to develop appropriate models of GS to achieve the maximum prediction accuracy.
6. The selection response declines at a faster rate under GS than with pedigree selection. This can be minimized by continually including new markers for the prediction of GEBVs. The long-term response under GS can also be increased by placing higher weights on the low-frequency favorable alleles, particularly in the beginning of GS program (see Nakaya and Isobe 2012).
7. GS is more effective than phenotypic selection on per unit time basis only when off-season/greenhouse facilities are used to grow up to three generations per year. The usefulness and the cost-effectiveness of GS would be doubtful where such facilities are not available.
8. The need for genotyping of a large number of marker loci in every generation of selection adds considerably to the cost of breeding programs. It has been projected that, in the future, a greater emphasis will be placed on the use of marker data than on phenotype data.

Such a shift, however, would require the cost of a single marker data point to be merely 1/5,000 the cost of phenotyping a single entry.

9. Implementation of GS would require considerable infrastructure and other resources, which may be beyond the reach of most moderate size public sector breeding programs, particularly in the developing countries. In addition, planning and execution of GS is quite demanding and the breeders would be required to reorient their approach to the breeding programs.

10.12 Future Directions

The technology concerning molecular markers is in a constant flux, and new marker types, viz., copy number and epigenetic variation, and genotyping platforms are being developed. The new marker systems would generate quantitatively as well as qualitatively different information making it necessary to develop appropriate statistical tools and procedures for their proper utilization for increasing the accuracy of GEBV estimates. At the same time, suitable user-friendly software packages for implementing these statistical innovations need to be developed to facilitate practical implementation of GS. At present, an R-Package for GS is available on http://www.r-project.org/. In addition, the vast amounts of data generated on various aspects, e.g., the lines being evaluated, the test environments, genotype and phenotype data, etc., in the plant breeding programs need to be stored in appropriate formats in robust databases. Appropriate data management systems need to be established for efficient utilization and integration of the data stored in various databases. Some such public sector databases have already been implemented, e.g., The Hordeum Toolbox of the Barley Coordinated Agricultural Project (http://hordeumtoolbox.org/), GDPDM Database Scheme linked with the software TASSEL (http://brain.uni-hohenheim.e/eng/indexeng.html), and the Canadian COOL-DUDE. Linking of these and other similar tools with GS, and the development of user-friendly software packages for GS will greatly facilitate practical application of the GS procedure (Heffner et al. 2009).

There is need to develop suitable guidelines for the construction of appropriate training populations, inclusion of new lines in them, and the most appropriate population for a given breeding program/objective. The training population should be designed to support accurate prediction of GEBVs over time with the minimum demand on resources for phenotyping. The population could be updated periodically by inclusion of new lines developed in the breeding population and limiting the phenotyping to the new inclusions. Different methods for estimation of marker effects may capture different aspects of the relationship between marker genotype and the trait phenotype. Therefore, some of these methods may complement each other, and a synthesis of such methods may improve GEBV prediction accuracy. In a simulation study, parametric methods like ridge regression and BayesB were superior to the nonparametric methods and the machine learning methods. Surprisingly, simple mean of all the methods performed much better than any individual method, suggesting that a meta-predictor may turn out to be the most accurate. These issues need to be examined closely using simulation and empirical investigations (Jannink et al. 2010).

GS is expected to cause less inbreeding and retain more genetic diversity than selection based on breeding values estimated from pedigree information. However, GS does capture some information on genetic relationships, which increases the chances of inbreeding. The rate of inbreeding during GS, therefore, may need to be managed, for example, by using marker information. One approach for this is to vary the weight given to marker information as a function of allele frequency at each marker locus (Goddard 2009). Alternatively, the marker information may be used to implement a selection scheme similar to within family selection. For example, the individuals may be grouped into several clusters on the basis of marker data, and selection may be restricted within these clusters (and not between clusters). This approach may reduce the

short-term selection gain, but would increase the long-term gain; simulation results tend to support this expectation (Jannink et al. 2010).

The current GS models generally take into account only the additive gene effects. It is likely that the dominant gene effects would be satisfactorily accommodated in GS models in the near future, but the satisfactory inclusion of epistatic interactions remains a challenge. The semi-parametric GS models and the machine learning methods are being developed to enable the inclusion of epistatic interactions in the regression models. The available evidence indicates that inclusion of epistatic gene effects greatly enhances the accuracy of GS. Further, the present GS models do not take into account the effects of genetic background on QTL expression. There is almost complete lack of information on the effects of such interactions, present in the training and/or breeding populations, on the accuracy of GEBV estimates. It may be expected that significant QTL × genetic background interactions will reduce the accuracy of both QTL effects and GEBV estimates.

The GS scheme was regarded as a crazy idea when it was proposed in 2001 (Meuwissen 2009). But the concept of GS has now been validated, and it seems to be a potent and attractive breeding strategy. GS is likely to facilitate designing of ideal genotypes and creating them through targeted GS. The chief weakness of the GS approach seems to be its disregard for an understanding of the biological phenomena underlying the development of the concerned phenotypes. This is because the GS algorithms do not take into account the various findings from genetics and genomics, including the identification of QTLs and the genes represented by them. It may be rewarding to develop such GS algorithms that take advantage of the findings from genetics and genomics fields. GS appears to be the starting point for the next phase of MAS. GS is likely to be integrated into many plant breeding programs as the various issues related to it are adequately resolved and the discussions on GS become less and less theoretical and mathematical. Further, the progressive reduction in marker genotyping costs will accelerate the adoption of GS in regular plant breeding programs.

Questions

1. Briefly describe the procedure for genomic selection and compare it with marker-assisted selection.
2. Discuss the relevance of training population in genomic selection and describe briefly the important considerations during creation of a suitable training population.
3. Briefly describe the various approaches for the estimation of genomic estimated breeding values. Which of these approaches is in common use and why?
4. "Genomic estimated breeding values are affected by several factors". Comment on this statement in the light of the available relevant information.
5. Discuss the integration of genomic selection in plant breeding programs and its effects on genetic diversity.
6. Discuss the effectiveness, merits, and limitations of genomic selection.
7. "Genomic selection is a breeding scheme of the future." Analyze this statement giving appropriate reasons and evidence in support of your arguments.

Phylogenetic Relationships and Genetic Diversity

11

11.1 Introduction

The sum total of genetic differences present among different individuals, genotypes, strains, clones, or populations of a species is called *genetic diversity*. Genetic diversity among populations originates as a consequence of either geographical separation and/or genetic barriers to crossability. The concept of genetic diversity differs from that of variability with respect to the following: variability is expressed as phenotypic variation, while genetic diversity may or may not be expressed at the phenotypic level. Genetic diversity can be studied using a sample of inbred lines, pure lines, clones, or populations, often termed as *entities*, of a species. Diversity analysis may use data from pedigrees, qualitative and/or quantitative traits, isozymes, and DNA markers. These data are analyzed either by a single statistical method or by a combination of methods. Genetic diversity analysis may serve one or more of the following purposes: (1) It reveals the amount of genetic variation existing among the varieties of a crop. (2) It facilitates identification of diverse lines that could be used for hybridization to produce either hybrid varieties or superior segregating populations to be subjected to selection. (3) It helps avoid the use of closely related germplasm lines in hybridization programs since this would narrow down the genetic base of the derived varieties. (4) It enables a reliable categorization of germplasm accessions and the determination of core collections for specific breeding applications. Finally, (5) it may help in the introgression of desirable genes/alleles from diverse germplasm into elite germplasm of a crop. The following discussion on genetic diversity analysis, including phylogenetic relationship determination, is based largely on Mohammadi and Prasanna (2003), but other sources have also been used.

11.2 Estimation of Genetic Distance/Similarity

Generally, *genetic diversity analysis* involves estimation of genetic similarity or dissimilarity between pairs of entities and use of these estimates for grouping of the entities. In simple terms, *genetic distance* (*GD*) is a quantitative estimate of the genetic differences between two entities in terms of differences in their DNA sequences and/or gene frequencies. Thus, *genetic similarity* between two entities equals 1–GD. Some of the common methods for estimation of GD from morphological and molecular marker data are briefly described in the following sections.

11.2.1 Estimation of Genetic Distance from Morphological Trait Data

Diversity studies are generally based on two types of morphological data, viz., qualitative and quantitative trait data. Qualitative trait data

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_11, © Author(s) 2015

are discrete and may contain nominal measures, which differentiate between entities based on the names of the deviant traits. Here numbers can also be used to differentiate entities, but these numbers have no numerical value or relations. Quantitative data, on the other hand, exhibit continuous variation. Another type of data, called ordinal data, can also be used; these data contain a rank order but without any degree of difference associated with the ranks. Because of the fundamental differences in the measurement scales, qualitative and quantitative data cannot be combined for diversity analysis unless they are processed following specific procedures.

Generally, genetic distance from quantitative trait data is estimated as the Euclidean or straight-line measure of distance as follows:

$$D_{ij} = \left[(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2\right]^{1/2} \tag{11.1}$$

where D_{ij} is GD between entities i and j; p is the number of morphological traits scored; $x_1, x_2, \ldots x_p$ are observations for morphological traits 1, 2, … p, respectively, for entity i; and $y_1, y_2, \ldots y_p$ are observations for traits 1, 2, … p, respectively, for entity j. Gower (1971) proposed to award "0" and "1" scores for a match and mismatch, respectively, between two entities for a qualitative trait, i.e., nominal, data. This approach can be extended to ordinal data as well. But in the case of a quantitative trait, the difference between values for any two entities is divided by the range for the concerned trait in all the entities included in the study. This procedure converts the differences for quantitative traits into scores ranging from 0 to 1. Thus, this procedure allows the use of both qualitative and quantitative trait data for estimating the *Gower's measure of distance* or the *average taxonomic distance* between entities i and j (DG_{ij}) by the following formula:

$$DG_{ij} = \frac{1}{p}\sum w_k \, d_{ijk} \tag{11.2}$$

where p is the number of characters, $w_k = 1/R_k$, R_k is the overall range (range for the sample used in the study) for the kth trait, and d_{ijk} is the difference between the values for the kth trait for the entities i and j.

11.2.2 Estimation of Genetic Distance from Molecular Marker Data

There are several approaches for estimation of genetic distance from both protein- and DNA-based molecular marker data, some of which are briefly considered here. In case of codominant markers, allele frequencies can be readily estimated and used for estimation of the genetic distance as follows:

$$D_{ij} = \text{Constant}\left(\sum\nolimits_{a=1}^{n} \left|X_{ai} - X_{aj}\right|^r\right)^{1/r} \tag{11.3}$$

where X_{ai} and X_{aj} are frequencies of allele a for entities i and j, n denotes the number of alleles per locus, and r is a constant derived from the coefficient used. In its simplest form, $r = 1$; in this case, the genetic distance (D_{ij}) would be estimated by the following formula:

$$D_{ij} = \frac{1}{2}\sum\nolimits_{a=1}^{n} \left|X_{ai} - X_{aj}\right| \tag{11.4}$$

But when the value of $r = 2$, the above formula becomes

$$D_{ij} = \frac{1}{2}\left[\sum\nolimits_{a=1}^{n} \left|X_{ai} - X_{aj}\right|^2\right]^{1/2} \tag{11.5}$$

and this estimate of GD (D_{ij}) is known as *Roger's measure of distance* (RD; Rogers 1972).

In general, a binary matrix (scores of "1" and "0" for the presence and absence, respectively, of electrophoretic bands) generated from molecular marker data is used for the estimation of GD by one of the following formulae:

$$GD_{NL} = 1 - \left[\frac{2N_{11}}{(2N_{11} + N_{10} + N_{01})}\right] \tag{11.6}$$

$$GD_{J} = 1 - \left[\frac{N_{11}}{(N_{11} + N_{10} + N_{01})}\right] \tag{11.7}$$

$$GD_{SM} = 1 - \left[\frac{(N_{11} + N_{00})}{(2N_{11} + N_{10} + N_{01} + N_{00})}\right] \tag{11.8}$$

$$GD_{MR} = \left[\frac{(N_{10} + N_{01})}{2N}\right]^{1/2} \tag{11.9}$$

where GD_{NL} is Nei and Li's coefficient (Nei and Li 1979); GD_J is Jaccard's coefficient (Jaccard 1908); GD_{SM} is simple matching coefficient (Sokal and Michener 1958); GD_{MR} is modified Rogers' distance (Wright 1978); N_{11}, N_{10}, N_{01}, and N_{00} are the numbers of bands present in both the entities i and j, in entity i only, in entity j only, and in none of the entities, respectively; and N is the total number of bands scored in the sample. The *Jaccard's coefficient* considers only matches in the bands, while *Nei and Li's coefficient*, also called *Dice coefficient*, places a greater emphasis on the bands that match as it multiplies the number of matches by two. *Simple matching coefficient* is a Euclidean measure of GD that takes into consideration the bands absent in both the individuals. Another Euclidean measure, the *modified Rogers' distance*, treats every scored locus as an orthogonal distance. GD_{MR} *is perhaps the most widely used measure of GD*. In case of codominant markers like SSR, the marker bands are scored "1" and "2" to denote the presence of one and two bands, respectively, of different lengths. The simple matching coefficient is the most suitable measure of GD for such data since it takes into consideration all the four combinations of bands. Sometimes, both the entities may fail to show amplification for some of the SSR markers. Ordinarily, it is very difficult to determine if this failure were due to the presence of "null alleles" or a consequence of experimental error. Whenever the "null allele" status of a "missing band" is doubtful, it should be treated as missing data for the estimation of GD. *In such situations, Jaccard's coefficient, Nei and Li's coefficient, and modified Rogers' distance are widely used for estimating GD*.

11.2.3 Estimation of Genetic Distance from Populations

The diversity analysis among populations involves sampling of several (~50) individuals from each population. This sampling is complicated by a variety of factors, including inbreeding, population structure, and migration. In general, when a choice is to be made between scoring a larger number of individuals in each population and a larger number of loci per individual, the latter should be preferred. In addition, the same set of marker loci should be scored in all the populations. When p_{ij} is the frequency of jth allele at the ith marker locus, n_i is the number of alleles at the marker locus i, and m is the number of marker loci scored; the *total gene diversity* (H) or *average expected heterozygosity* within each population is estimated as follows:

$$H = 1 - \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i} p_{ij}^2 \tag{11.10}$$

When p_{ij} and q_{ij} are allele frequencies of jth allele at ith marker locus in the two entities being compared, the Euclidean measure of genetic distance, also known as *Nei's geometric distance* (GD_N), between two populations or individuals is estimated as given below:

$$GD_N = \left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(p_{ij} - q_{ij}\right)^2\right]^{\frac{1}{2}} \tag{11.11}$$

Similarly, the modified Roger's distance (GD_{MR}) between two populations or individuals is estimated from the following formula:

$$GD_{MR} = \left[\frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{ni}\left(p_{ij} - q_{ij}\right)^2\right]^{\frac{1}{2}} \tag{11.12}$$

11.2.4 Choice of the Genetic Distance Measure

The selection of appropriate measure of GD is an important issue in genetic diversity analysis. The

GD among a set of inbred lines measured as GD_{NL} and GD_J reveals identical rankings of the GD estimates. However, their rankings may differ for GD estimated among entities like hybrids and random mating populations since they have heterozygous loci. The statistical distributions of both GD_J and GD_{NL} are not known. Therefore, the bootstrap method may be used to calculate their confidence intervals. GD_{MR} is widely used since the genetical and statistical properties of this measure of GD are excellent. GD_{SM} has Euclidean metric properties; therefore, it can be used in hierarchical clustering strategies. But GD_{SM} assigns equal weights to 0–0 as well as 1–1 matches, although 0–0 matches might not be due to the loci in the two entities being identical by descent, which is the case for 1–1 matches. When two or more different measures of GD are estimated from the same data, the correspondence between the different GD matrices may be evaluated by Mantel's test. The significance of Mantel's test is assessed by permutation test.

11.3 Genetic Diversity Analysis: Phylogenetic Relationships

In simple words, *phylogenetic analysis* involves grouping of the various entities included in a study on the basis of their genetic relationships so that the groupings indicate the degrees of genetic similarities-dissimilarities among them. These groupings are also expected to reflect the pattern of evolution of the various entities, often from a common ancestral entity. The various methods used for grouping of entities use either a GD matrix or the original datasets as inputs and generate as output either a graphic or a textual representation of the groupings. These representations are often termed as *phylogenetic trees*, which describe the evolutionary relationships among various entities; these relationships are determined on the basis of similarities and differences in the physical and genetic characteristics of these entities. *The various entities represented in a phylogenetic tree are presumed to be descendents of a common ancestor*. In general, the methods used for this purpose analyze data on multiple traits for each entity; therefore, they are usually referred to as *multivariate methods*. These data may pertain to either morphological and/or biochemical traits, molecular marker genotypes, or a combination of two or more types of data, including pedigree information. The most common multivariate methods are (1) cluster analysis, (2) principal component analysis (PCA), (3) principal coordinate analysis (PCoA), and (4) multidimensional scaling (MDS; Table 11.1). Menu-driven statistical packages NTSYS-pc (F.J. Rohlf, State

Table 11.1 A summary of the main features of various multivariate methods used for grouping entities, i.e., individuals, inbred lines, pure lines, clones, and populations

Multivariate analysis method	Input	Output	Remarks
Cluster analysis			
(a) Distance-based	Genetic distance	Graphic/textual phylogenetic tree or grouping of entities	Tree may be rooted or unrooted, UPGMA is generally used for tree construction
(b) Model-based		Groups of the entities	Maximum likelihood, Bayesian methods
Principal component analysis (PCA)[a]	Variance-covariance matrix or correlation matrix	Two- or three-dimensional scatter plot	The first 2 or 3 PCs should explain most of the variation; also for PCoA
Principal coordinate analysis (PCoA)	Similarity-dissimilarity matrix	Two- or three-dimensional scatter plot	Preferable to PCA in case of missing data, and with more traits than entities
Multidimensional scaling (MDS)[a]	Similarity-dissimilarity matrix	Two- or three-dimensional map	Better reflection of differences between close entities than PCA and PCoA

[a]Can be used to determine the optimum number of groups in a study

University of New York, Stony Brook, USA), PHYLIP (J. Felsenstein, University of Washington, Seattle, USA; Sect. 14.5.8), and DARwin (X. Perrier and J.P. Jacquemoud-Collet, CIRAD, France) use diverse datasets for the analysis of genetic diversity, including that of gene and genotype frequencies, cluster patterns, and implementation of resampling methods like bootstrap and Jackknife.

11.3.1 Cluster Analysis

Cluster analysis groups the entities with similar features into the same cluster. As a result, GD between entities within the same cluster is much smaller than that between any two clusters. The clustering methods may be either GD-based or model-based. In *GD-based clustering methods*, the input for clustering consists of a pair-wise distance matrix, and a graphical representation as a tree or dendrogram is generated as output. This allows an easy visualization of the clusters and their member entities. These methods group the entities on the basis of their similarities/distances and do not use an evolutionary model for tree construction. In case of *model-based clustering methods*, standard statistical methods like maximum parsimony and maximum likelihood (often within a Bayesian framework) methods are used to draw inferences about each cluster and the cluster membership on the basis of some parametric model. The maximum parsimony method implies a model of evolution, i.e., parsimony, while the maximum likelihood method applies an explicit evolutionary model for construction of trees. The software package STRUCTURE (Sect. 14.3.4) can be used for model-based grouping of entities; in this case, the number of specified groups may be varied from 1 to 10.

The dendrograms or phylogenetic trees may be either rooted (Fig. 11.1a) or unrooted (Fig. 11.1b). In a rooted tree, the root represents the most recent (usually, deduced and unknown) common ancestor of all the entities present at the ends of the branches. Generally, trees are rooted by including one or, preferably, more well-known out-group entities that are clearly separated from the other entities, but are close enough to converge at the root of the tree. In contrast, no assumption is made about ancestry in the case of unrooted trees, and only the relatedness of various entities is depicted. Rooted trees can be readily converted into unrooted ones simply by omitting the root. However, the conversion of an unrooted tree into a rooted one requires additional assumptions/data. An unrooted tree is useful to depict a very large number of entities, in which case a rooted tree will appear too crowded to be readily comprehendible. It may be pointed out that the cluster patterns obtained from both the types of trees are essentially comparable.

In general, the most widely used clustering methods are distance-based methods. These methods can be either hierarchical or nonhierarchical, the former being more commonly used for genetic diversity analyses in crop species. Some hierarchical methods start with each group having a single entity; then they carry out a series of consecutive mergers of the most similar groups till the optimum number of groups is generated. Unweighted pair group method using arithmetic average (UPGMA) is the most commonly used algorithm for this type of clustering, but the Ward's minimum variance method is also widely used. Another algorithm used is unweighted pair group method using centroids (UPGMC). The softwares DARwin and GENES (C.D. Cruz, Universidade Federal de Viçosa, Brasil) calculate GD and implement hierarchical grouping methods, including simple linkage, complete linkage, and UPGMA. Some other hierarchical methods, however, begin with a single group with all the entities and successively divide the groups to obtain the optimum number of clusters. The *nonhierarchical clustering procedures* do not generate dendrograms; they assign individuals to specific clusters with the help of specific approaches, provided the number of clusters is specified beforehand. Statistical packages like SAS [FASTCLUS] and SPSS [QUICK CLUSTER] can be used for nonhierarchical clustering. *Nonhierarchical clustering methods are rarely used for the analysis of genetic diversity within crop species*. For

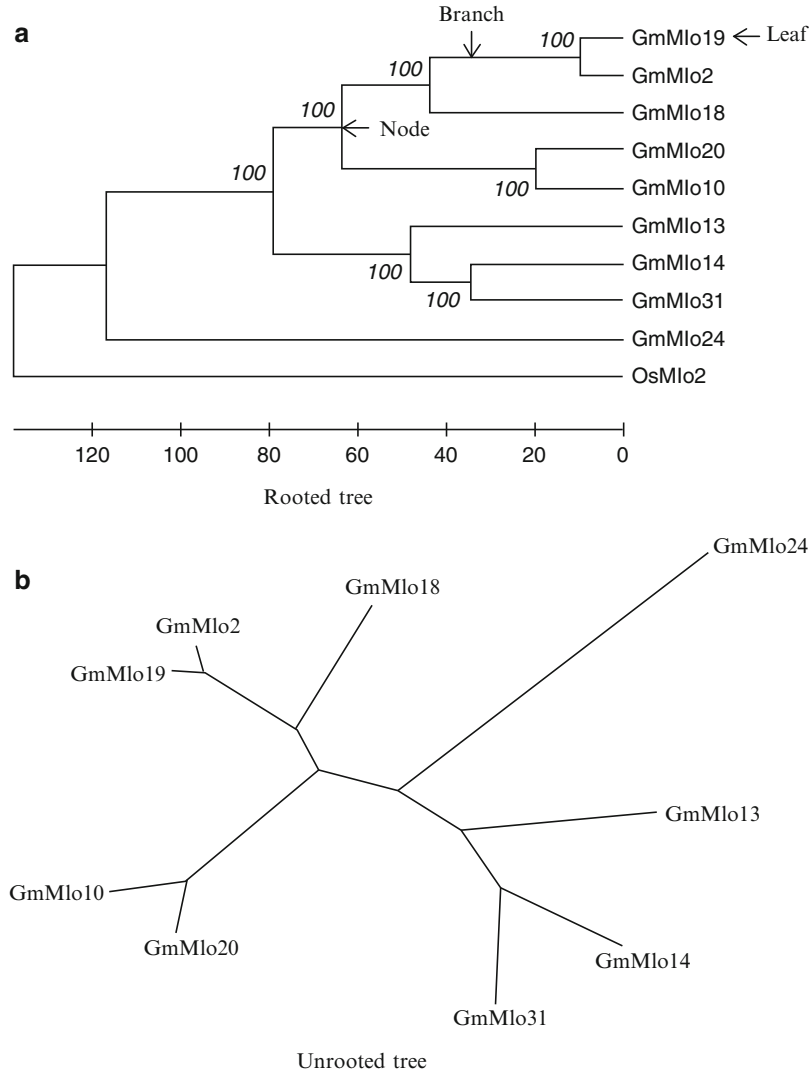


Fig. 11.1 (**a**) A rooted dendrogram or phylogenetic tree depicting the relationships among nine members of soybean (*Glycine max*) *Mlo* (*GmMlo*) gene family. A rice (*Oryza sativa*) *Mlo* (*OsMlo2*) gene has been included as an out-group to allow the depiction of the root. The terms "branch," "node," and "leaf" describe the parts of the tree as indicated in the figure. The numbers at the nodes reveal the percent bootstrap support for relationship shown at the concerned node. The numbers at the *bottom line* (below the X-axis) specify the number of amino acid differences per sequence. (**b**) An unrooted tree for the same set of genes, except the *OsMlo* (Courtesy, Reena Deshmukh, Varanasi)

quantitative traits, Mahalanobis distance (D^2) can be used for clustering.

11.3.2 Principal Component Analysis

Both principal component analysis and principal coordinate analysis (PCoA) can be used to generate two- or three-dimensional scatter plots of the entities included in the study (Fig. 11.2). In these plots, the geometrical distance between any two entities reflects the GD between them, and the grouping of entities reveals the sets of genetically similar entities. PCA is based on a variance-covariance matrix when the different traits have the same scale, but a correlation matrix is preferable when the traits have different scales. PCA implements linear transformation to reduce the original variables to a small number of new variables that are uncorrelated and cumulative; these variables are called *principal components* (*PCs*). Each PC discloses different properties of the original variables. Most of the variation existing in the original data is represented in the first PC. The second PC represents the next largest portion of the variation that is not explained by the first PC and so on. The proportion of variation due to a PC is

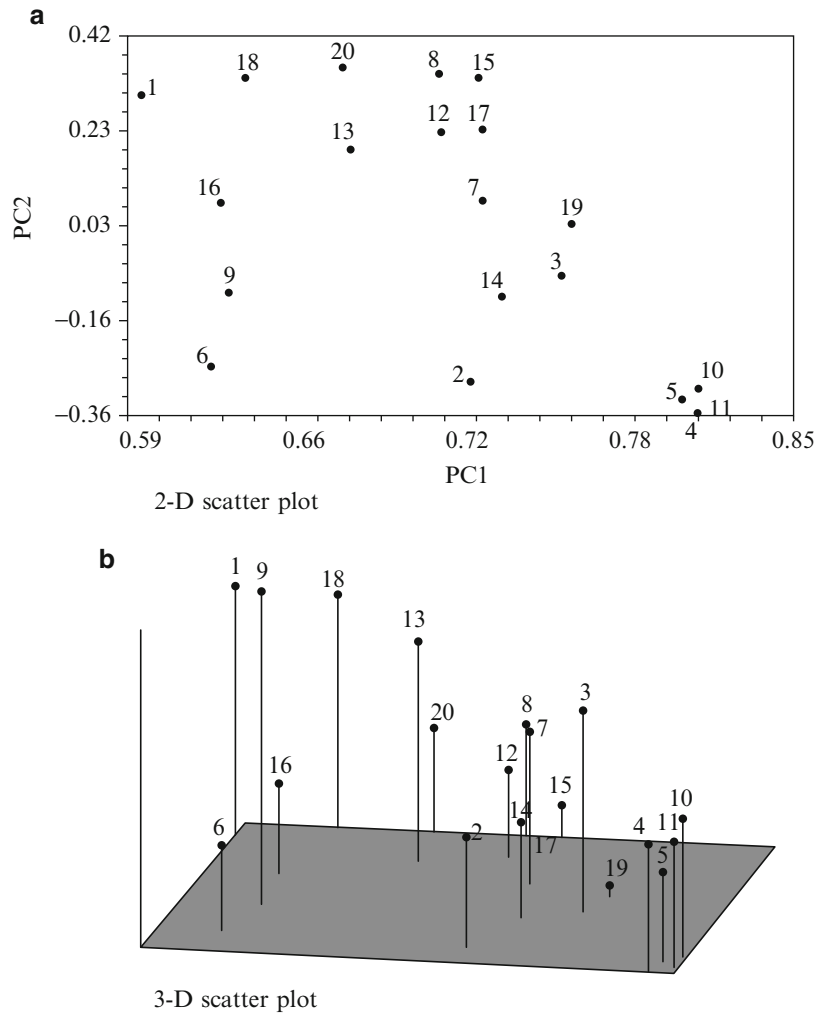


Fig. 11.2 Principal components (PC) scatter plots derived from RAPD profiles of the 20 genotypes of pea; names of the genotypes are listed in Fig. 3.4. (**a**) 2-D scatter plot. The first two PCs explained 51.5 % and 8.9 % of the variation. (**b**) 3-D scatter plot. The first three PCs explained 51.5 %, 8.9 %, and 7.7 % of the variation (Courtesy Kusum Yadav, Lucknow)

given by the eigenvalue for the concerned PC divided by the total of all the eigenvalues. In case of molecular marker data, very low and negative eigenvalues should preferably not be used, and such values may be eliminated by a suitable transformation of the similarity matrix. In addition to generating scatter plots, PCA can also enable the determination of the optimal cluster number for a study. At the start, all the entities are placed in a single cluster. This cluster and the subsequent clusters are successively split till the magnitude of the second eigenvalue of each of the clusters is lower than the value selected by the user. The second eigenvalue may be set at 0.75 to ensure that the first PC accounts for most of the variation.

11.3.3 Principal Coordinate Analysis

Principal coordinate analysis or *classical scaling* is an ordination or scaling method that uses as input a similarity or dissimilarity matrix to generate a graphical representation in a small number of dimensions. In this plot, the distances between any two points are comparable to the extent of original dissimilarity between the concerned entities. In contrast, PCA uses the original data matrix as the input. The outputs from PCA and PCoA (Fig. 11.3) are similar when the number of characters is small and there are no missing data. When lots of data are missing or the number of characters is more than that of entities included in the study, PCoA is

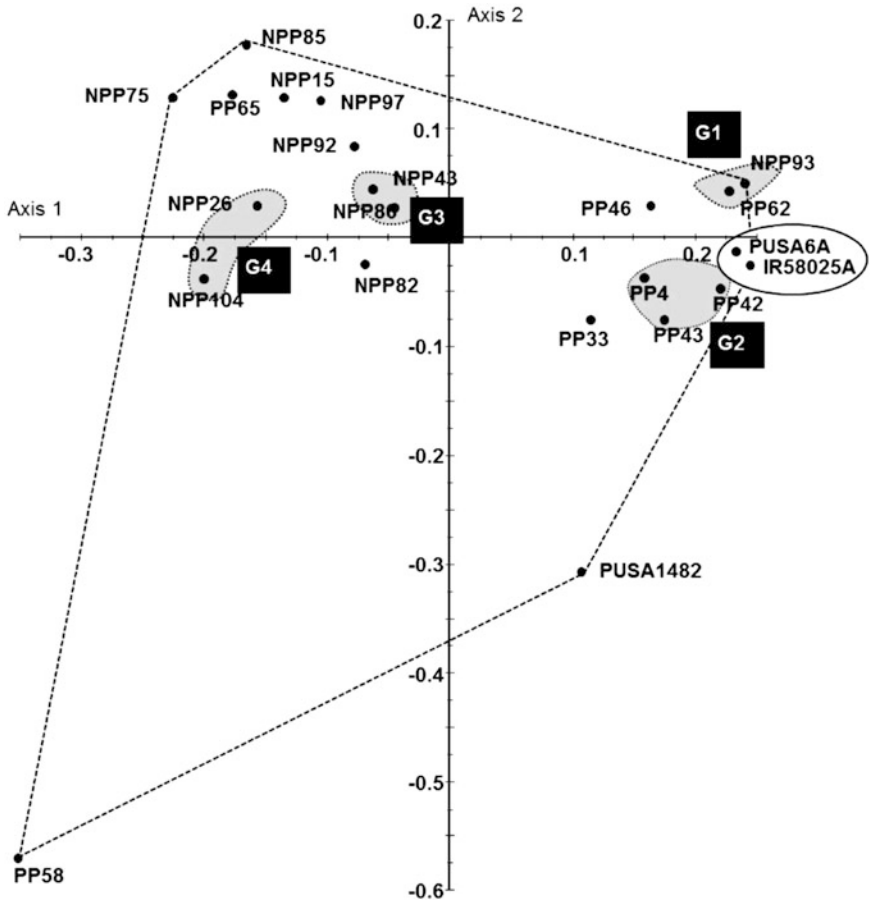


Fig. 11.3 Dispersion of rice genotypes (parents of a set of 20 test rice hybrids) by principal coordinate analysis (PCoA) using genetic distances estimated from data on molecular markers linked with QTLs for yield and its components. The genotype groupings have more than 70 % bootstrap support. *Dotted lines* show convex hull. Dispersion of parental lines between first two axes revealed one group consisting of female parents, IR58025A and Pusa6A. The male parents formed four groups G1, G2, G3, and G4, while 11 male parents did not form any group (Courtesy, K.K. Vinod, Aduthurai)

preferable to PCA. PCA and PCoA are particularly useful when most of the variation in the original data is explained by the first two or three PCs or principal coordinates; this situation is likely to be encountered in cases of highly correlated original variables.

11.3.4 Multidimensional Scaling

Multidimensional scaling uses a similarity/distance matrix between a set of entities to represent them in a map with few (usually, two to three) dimensions. The distances between the entities in this map nearly match the original distances. The MDS representation can be metric or nonmetric depending on whether the magnitudes or the rank orders of the original similarities-distances are used to generate the map. MDS can be used to convert the original distance matrix into two- or more dimensional coordinates for visualizing genetic relationships. The distance matrix may be obtained from morphological and/or molecular markers. The actual number of groups that would be generated by cluster analysis can be estimated from the pattern produced by the MDS procedure.

In general, PCA, PCoA, and MDS generate similar groupings of the entities. However, MDS is superior to PCA and PCoA in reflecting the differences between close entities. Further, when the number of entities is very large, MDS is preferable to PCA and PCoA. Finally, PCA should be used only when there is no missing data and there are many more entities than the number of characters scored.

11.3.5 Determination of the Optimal Number of Clusters

An *acceptable cluster* is a group of two or more such entities, for which the magnitude of within cluster GD is smaller than the overall mean GD for the dataset. At the same time, GD between any two clusters is greater than the within cluster GD for either of the two clusters. The *optimal* or *acceptable number of clusters* in a study comprises the minimum number of "acceptable clusters" for the dataset. In case of a dendrogram, the acceptable number of clusters is provided by a "cut" that separates the "true" or "natural" groups from each other. The D^2 and the upper tail approach are two relatively simple methods for determining the optimal number of clusters. In the case of D^2, the best "cut" point for a dendrogram yields the largest D^2 between centroids (vectors of means) of the groups created by the cut. In the upper tail approach, the optimal number of clusters is calculated from the mean of the standard deviation of GD estimated at the fusion points. In addition, the optimal number of clusters can be calculated by using bootstrap, multivariate analysis of variance (MANOVA), or discriminant analysis. In case of MANOVA, the dendrogram is cut at different points. The clusters generated at each "cut" point are regarded as treatments. Further, the entities occurring within each of these clusters are treated as replicates for the concerned treatments. The MANOVA is carried out separately for each of the cut point. The cut point that yields the highest value of F will generate the optimal number of clusters.

11.3.6 Choice of Clustering Method

The efficiency of different clustering methods can be compared by estimating their cophenetic correlation coefficients. This coefficient measures the concurrence between dissimilarity-similarity from a phenogram-dendrogram (the output) that is derived using the distance-similarity matrix as input for cluster analysis. A high ($\geq$0.8) cophenetic correlation coefficient indicates the method to be appropriate for the concerned analysis. In several studies, UPGMA method had a higher cophenetic correlation coefficient than UPGMC, single linkage, and the Ward's method. In cases of hierarchical and reticular patterns of diversity, ordination methods like PCA, PCoA, and MDS may be preferable as they do not assume linearity. In case of molecular marker data, PCA and PCoA seem to provide faithful depiction of relationships between major groups. However, in case the first two or three PCs or principal coordinates account for <25 % of the total variation, the relationship between closely related lines may be distorted; in such cases, cluster analysis is more reliable. When more than three dimensions have to be used, both PCA and PCoA become impractical. Further, nonlinear relationships among variables (a common problem with allele frequency data) and linkage disequilibrium may adversely affect the results from PCA and PCoA.

The UPGMA clustering algorithm yields relatively consistent grouping of genotypes when their relationships are estimated from diverse types of data. Results from simulation studies reveal poor performance of UPGMA over a broad range of tree space. Therefore, its use is not recommended, although it continues to be used in many publications. It is not possible to objectively define an optimal tree or dendrogram in the cases of UPGMA, UPGMC, single linkage, complete linkage, and the Ward's methods. The use of alternative clustering methods, such as neighbor joining (NJ) and Fitch-Margoliash (FM) methods, can eliminate systemic errors that are likely to occur during cluster analysis

reconstructions. NJ is the fastest procedure, yields a tree close to that from minimum evolution (ME) method, and has been widely used for phylogenetic studies, including those within crop plants. However, NJ yields only one tree. In simulation studies, FM and ME appear to be the best procedures for clustering and seem to perform almost equally well. The reliability of the nodes of dendrograms needs to be assessed by the bootstrapping of allele frequencies.

11.3.7 Use of Diverse Datasets

When traits with different scales are used as input for cluster analysis, the data can be standardized by using either the standard deviation or, preferably, the range for the concerned traits. This standardization eliminates the scale effects and gives equal weightage to the contributions of all the traits to the final output. But standardization is not needed for binary data like qualitative trait and molecular marker data. In case of quantitative traits, PCs may be used as input for clustering in the place of actual data, particularly when the traits are correlated. Data from different datasets may be analyzed separately or used for combined analysis. However, different datasets should be combined only after evaluating the agreement among the results obtained from the different individual datasets. Further, caution needs to be exercised when qualitative and quantitative trait data are to be combined; this is facilitated by a modified location model. But a combination of pedigree information with genetic marker data can yield better estimates of genetic relationships. It is preferable to analyze data for the traits with different modes of inheritance separately and, where possible, in many different ways to be able to draw reliable conclusions.

11.3.8 Resampling Techniques

Resampling techniques like bootstrap and jackknife are very useful in genetic diversity studies based on marker genotype data. These techniques also help determine the smallest set of markers needed for a correct appraisal of genetic relationships among the entities of a given sample. The *effective number of markers* is that number of markers with which the standard deviation of the various estimates is not significantly affected by either an increase or decrease in the number of analyzed loci/bands. Resampling methods are used to generate the measures of statistical accuracy of genetic diversity analyses. In the bootstrap method, the first step is to estimate the parameter of interest from the original sample. After this, a very large number of "bootstrap samples" of the same size as the original sample are produced by repeated sampling coupled with data replacement. The parameter of interest is then estimated from each one of the above "bootstrap samples." Finally, the variance for these estimates of the concerned parameter is calculated. Alternatively, for a random sample of N markers distributed over the genome, the standard error (SE) of Roger's distance (RD) between any two homozygous inbred lines is estimated as follows:

$$\mathrm{SE} = \mathrm{RD}(1 - \mathrm{RD})/N \qquad (11.13)$$

This SE estimate is the same as that obtained by the jackknife method.

Bootstrapping is widely used to estimate the statistical support to the branches emanating from different nodes of a dendrogram. The bootstrap support is the percent of times a particular branching pattern is obtained following bootstrap resampling. As a general rule, the tree branches having bootstrap support of over 70 would have a probability of 95 % to be correct. In the jackknife method, resampling does not involve data replacement. This resampling method estimates the bias in the estimates of the genetic parameters as well as the variance for these estimates. But jackknife method imposes a restriction on the number of resampling units and provides only slight information on the distribution of the estimates.

11.4 Genetic Diversity Analysis: Conservation of Genetic Resources

The *germplasm* or *genetic resources* of a crop species comprise, in theory, the sum total of genetic information, i.e., all the alleles of the different genes, present in the given crop species, and its related species. In practice, however, the germplasm of a species consists of a large collection of different accessions of the concerned crop species and the wild species related to this crop species. The germplasm accessions of a crop would comprise land races or primitive varieties; obsolete varieties; varieties currently in commercial cultivation; breeding lines/populations; special genetic stocks, including mutants and transgenic lines; and wild forms and wild relatives. The wild species from which a crop species is considered to have evolved directly is known as *wild form* of the concerned crop species. But the *wild relatives* of a crop species include all the wild species that are phylogenetically related to the crop species in question. The germplasm of a crop species is essential for its improvement through breeding since it provides the genes/alleles necessary for generating the improved gene combinations. The genetic resources of all crop species are depleting rapidly primarily due to human activities, including modern agricultural practices and development projects, and deployment of genetically homogeneous high-yielding varieties.

11.4.1 Germplasm Conservation

Since attempts to curb human activities would be futile, the only feasible option is to conserve the genetic resources either in situ or ex situ. *In situ conservation of germplasm* involves the establishment of gene sanctuaries or biosphere reserves in areas of high variability within the centers of origin of the concerned crop species by protecting these areas from human interference. However, this approach is applicable to only the wild relatives of crop species. Further, often it may be nearly impractical to properly implement and sustain the reserves, particularly in the developing world where such reserves need to be located. But *ex situ germplasm conservation* involves maintenance of germplasm accessions in germplasm or gene banks located away from their natural habitats. Ordinarily, the germplasm accessions are stored as seed samples since they can be stored for relatively long periods under a suitable low-moisture and low-temperature regime. But the germplasm of asexually reproducing species and of species producing recalcitrant seeds cannot be conserved as seed; they have to be maintained either in the field as growing plants, in tissue culture laboratories as slow-growth shoot cultures, or in liquid nitrogen at −196 °C as cryopreserved tissues/organs. The various activities in germplasm conservation are as follows: (1) collection or procurement, (2) conservation or storage, (3) evaluation or characterization, (4) cataloging or data storage and retrieval, and (5) multiplication and distribution for their utilization.

11.4.2 Applications of Molecular Markers in Germplasm Conservation

The main concerns in germplasm conservation may be summarized as follows: (1) maximization of the genetic diversity covered in the collection with the minimum number of accessions, (2) minimizing the inclusion of duplicate accessions in the collections, (3) minimizing the risk of genetic drift during storage/conservation in the case of heterogeneous accessions, and (4) eliminating the risk of genetic variation arising as a result of mutation (Brown and Kresovich 1996). Molecular markers can be used to address most of these issues. In particular, they are very useful for analysis of the genetic diversity present in the existing germplasm accessions of a crop species included in a collection.

11.4.2.1 Acquisition of Germplasm

Ideally, the accessions maintained in a good germplasm collection should provide

Table 11.2 A summary of the DNA barcodes commonly used for identification of fungal, bacterial, and viral plant pathogens

Pathogen	DNA bar code	Features/remarks
Fungi	Internally transcribed spacer region of the nuclear ribosomal RNA operon	Rapidly evolving genomic region; good discrimination between biologically distinct species; presents problems in some groups of fungi; in use for over two decades
Bacteria	16*S ribosomal RNA* gene	Robust universal PCR primers available; reliable up to genus level; multi-locus sequence analysis for more precise identification of species and possibly strains
Viruses	Usually, coat protein or polymerase genes; in case of DNA viruses, comparison of genome sequence	No universal PCR primers; primers available for genus level application; primers available for some of the largest and economically most important groups of viruses
Pathovars of a single pathogen species	Diagnostic PCR primers designed for specific genes differing between pathovars	Target genes identified by in silico analysis of the genome sequences of the concerned pathovars; primers developed for the pathovars of *Xanthomonas oryzae*
Races and pathotypes	SNP markers identified form NGS genome sequence data	Can differentiate very closely related species; successful in case of some viruses

Based on Geering (2013)

comprehensive and representative coverage of the genetic resources of the concerned crop species. Molecular marker data can be used to assay the extent of genetic diversity present in the germplasm collection of a crop species, and to identify the existing gaps in the collection and the under-represented germplasm categories. Then acquisitions and collections/explorations may be planned to rectify the gaps and under-representations. Similarly, the populations from which collections are to be made can be analyzed using molecular markers to enable a more rationalized collection strategy. This is because it is virtually impossible to characterize the various plant populations for all the traits of interest and to identify the genes and their alleles present in them. Molecular markers can be used to determine the genotypes of the populations assuming that the marker diversity is highly correlated with genetic diversity. This approach will be particularly useful in the cases of wild relatives of crops since their populations are highly heterogeneous as well as heterozygous.

Germplasm accessions are often imported from other countries, and such materials must be quarantined to prevent the entry of pathogens, insects, and weeds. Germplasm quarantine is a potential area, in which molecular markers may find application to facilitate the detection of new species/pathovars/races of fungal, bacterial, and/or viral pathogens. Pathogen identification is based on DNA barcodes and on single-nucleotide polymorphism (SNP) markers. In this context, a *DNA barcode* is a standardized genomic DNA sequence of over 400 bp length that is used for a reliable identification of organisms (Geering 2013). The genomic sequences used for bar coding differ from one group of organisms to the other, and their reliability may vary considerably (Table 11.2). DNA barcoding is a common practice in identification of microorganisms, but its use in germplasm quarantine is limited by the availability of suitable barcodes for plant pathogens. It may be expected that rapid developments in this field will soon enable the use of DNA barcodes for germplasm quarantine.

11.4.2.2 Storage or Maintenance of Germplasm

One of the major concerns of germplasm storage is the presence of duplicate accessions in the existing germplasm collections since duplicate accessions consume resources without contributing to the

worth of the collection. According to an estimate, the extent of duplication in the different germplasm collections may range from 30 to 70 %, and the average may be well over 50 %. Molecular markers can be used for unambiguous identification of duplicate accessions, which can be safely removed from collection holdings. Another issue relates to the genetic changes that may occur during germplasm storage due to genetic drift and/or spontaneous mutations. *Genetic drift* describes changes in gene and genotype frequencies of a sample/population entirely due to chance; it is usually associated with small population size. Genetic drift may occur during germplasm regeneration since usually only a small population size (50–100 plants) is grown for this purpose. Molecular markers can be used to keep a track of the genetic compositions of germplasm accessions and, thereby, minimize the risk of random drift. Old seeds with reduced viability show increased frequency of spontaneous mutations; often a loss of 50 % in seed viability may be associated with a mutation frequency in the surviving seeds equivalent to that due to 100 Gray (Gy) X-ray treatment. Mutations may also occur in slow-growth cultures and in cryopreserved materials. Molecular markers have been used to monitor the genetic fidelity of in vitro conserved materials. These studies show that there is little risk of increased mutation rate due to the storage conditions.

In addition, there is always a risk of errors in labeling of germplasm accessions and of mechanical mixtures during their regeneration. Detection and rectification of these errors on the basis of phenotypic evaluation are not only time consuming, but they are also tedious and not fully reliable. In contrast, the use of molecular markers would greatly facilitate the detection of such errors. Finally, each germplasm collection attempts to define a "core" collection that has a much smaller number of accessions, but is representative of the entire collection. Core collections are developed in order to maximize the utilization of germplasm collections and to minimize the efforts required for their handling. Molecular markers are being used to assess the genetic diversity of germplasm accessions and to define the core collections more accurately than that based on phenotype data.

11.4.2.3 Characterization of Germplasm

The value of a germplasm accession depends on the genes and gene combinations it contains; this is ordinarily determined by the evaluation of the accessions for various traits. The data recorded at the time of germplasm collection include a limited amount of critical information about each sample, which does not indicate their usefulness. Therefore, the accessions have to be evaluated in field trials by various specialists to generate data for a standard set of descriptors (trait descriptions) developed for each crop species. The inclusion of various traits in the descriptor set is based primarily on their usefulness in the exchange of information about germplasm collections and the germplasms themselves. Evaluation is the most critical step for germplasm utilization, but it is also the most demanding activity. As a consequence, genetic diversity present in many germplasm collections remains unexplored or only partially explored. In addition, phenotypic evaluation is quite effective for qualitative traits, but is unable to reveal the real value of various accessions for the improvement of quantitative traits. Further, the inferior performing accessions like land races and accessions of wild relatives of crops possess genes/alleles for improving the performance of segregants derived from their crosses with superior cultivars/breeding lines.

Molecular markers are being used to develop linkage maps for almost all the crop species. This should facilitate the use of markers for characterization of germplasm accessions. Genome-wide marker data may be expected to provide much more detailed information about the usefulness of different accessions than that available from the passport data alone. However, the usefulness of marker data in indicating the presence of specific traits in an accession would depend on our knowledge of the marker-trait associations in the concerned crop species. In addition, the reliability of this information when extended to accessions of unadapted germplasm and related

wild species would be a critical issue. The generation of this information would require considerable effort and resources. However, once this information is generated, it will greatly facilitate the identification of germplasm accessions possessing desired traits, including favorable alleles for quantitative traits.

11.4.2.4 Utilization of Germplasm

Germplasm utilization will be greatly facilitated if the users are able to identify easily and quickly the germplasm accessions that have either the traits (mainly qualitative traits) of their interest or that are expected to best serve their desired objectives (mainly improvement in quantitative traits). A germplasm collection well characterized for simply inherited traits and molecular marker genotypes would greatly facilitate quick and easy selection of the desired accessions. In addition, molecular markers associated with specific traits will be useful in marker-assisted transfer of the desired genes/gene combinations into the breeding materials/well-adapted cultivars. It may be emphasized that phenotypic evaluation will not reveal the usefulness of germplasm accessions, particularly unadapted accessions, for improvement of quantitative traits. In such cases, the accessions may be selected on the basis of divergence in their molecular marker genotypes. As a general rule, the more distinct is the marker genotype of an accession from that of the elite germplasm/breeding line, the greater is the likelihood that the accession would contribute useful quantitative trait loci (QTLs) to the progeny from its crosses with the elite line. Further, each new unadapted germplasm accession selected for use in a breeding program should have as divergent marker genotype as possible from the earlier used unadapted accessions. The marker genotype may relate to specific genomic regions when the locations of the desired QTLs are known, or it may cover the whole genome. This approach, for example, has been quite successful in the use of exotic germplasm for improvement of certain quantitative traits, including yield, of rice and tomato.

Innovative breeding approaches like advanced backcross QTL (AB-QTL) method may be used for simultaneous identification and introgression of useful QTLs from the poor-performing unadapted germplasm accessions into the elite germplasm representing the breeding materials (Tanksley and Nelson 1996; Bernacchi et al. 1998). This approach has been successful in the introgression of favorable QTLs governing fruit yield, total soluble solids content of fruits, and fruit color from *Lycopersicon hirsutum*, a wild relative of tomato, into cultivated tomato. This program has been able to isolate lines exhibiting 48, 22, and 33 %, respectively, improvement in these traits. These lines have shown improved performance in evaluation trials conducted under different environments around the world. Similarly, the AB-QTL scheme was used to transfer two QTLs for yield from the wild species *Oryza rufipogon* into the cultivated rice. Thus, a planned use of molecular markers in innovative breeding schemes would greatly increase the utilization of genetic resources of related species for the improvement of crop species. It may be pointed out that there is considerable evidence that wild relatives of crops can provide many favorable QTLs that can be useful for improving agronomic as well as economic traits, of crop plants, including their yields (Singh 2012a).

11.4.2.5 Curatorial Issues in the Long-Term Conservation of Germplasm

There are four main curatorial issues in the long-term conservation of germplasms of plant species: (1) identification of the accessions, (2) determination of the relationships within and among accessions, (3) assessment of the genetic structure of the collection, and (4) determination of the locations of the genes/gene complexes in the collection (Brown and Kresovich 1996). *Identification of accessions* relates to the correct cataloging of the accession identity, e.g., name of the variety/genotype. It is important to ensure that the accession is in fact the same strain/genotype as indicated by the label and that it is properly maintained in an adequately viable state. The *degree of relationships* among individuals within an accession, e.g., whether an accession is

homozygous and homogeneous or heterogeneous or it is heterozygous and heterogeneous, and the degree of relationship among the accessions should be known to facilitate suitable strategies for their conservation. The *structure of collection* describes the partitioning of variation among individuals, accessions, populations, and species included in the collection. The genetic structure of a plant species is influenced by in situ demographic factors like population size, mating patterns, mode of pollination, and migration. The *location of traits* in a germplasm collection relates to the presence of specific traits in particular accessions or even individuals. Further, whenever available, the knowledge of genomic locations of genes/gene complexes governing specific traits would facilitate utilization of the genetic resources.

11.4.3 Conservation of Wild Species

The conservation of natural populations of plant species involves the estimation of the level and the relevance of several biological processes to short- and long-term survival of small populations of threatened plant species. Small populations show varying levels of inbreeding and the consequent inbreeding depression, decrease in genetic diversity, and reduction in adaptive potential. Heterogeneity in the prevailing environment may influence the distribution and the frequencies of different genotypes of a species by affecting gene flow/admixture and reproductive success. Finally, the identification of evolutionary significant units is a major challenge since they are in a dynamic state. In addition, the degree of divergence of such units is confounded with the issues of limited adaptive capacity of the individuals surviving in small populations. Further, often there is a loss in the fitness of progeny from intermating between divergent populations/provenances due to the breakdown of adaptive gene complexes following segregation and recombination (Rymer and Rossetto 2013).

A variety of molecular markers are used to seek answers to the above questions relevant to the decision making for conservation of natural populations. Allozymes were widely used in the past to identify the associated life history traits and distribution patterns of different genotypes. Arbitrary dominant DNA markers like RAPD, inter-simple sequence repeat (ISSR), and AFLP can be used with any plant species. Of these, AFLPs are the most reliable, and they can be used for most studies, except investigation of inbreeding and outbreeding depressions. But AFLPs are the most suited for studies on gene flow/admixture and identification of evolutionary significant units. SSRs are the markers of choice and have applications similar to those of dominant markers, except that SSRs are the most informative for the study of inbreeding as well. Next-generation sequencing (NGS) technologies allow detection and genotyping of a large number of SNP markers; they are the most useful for studies on loss of genetic diversity, gene flow/admixture, environmental processes, identity of evolutionary significant units, and various temporal processes. However, analysis of inbreeding/outbreeding depression is possible only with the transcriptome data, which are the most informative for studies on loss of adaptive potential, loss of diversity, etc. (Rymer and Rossetto 2013).

11.5 Genetic Diversity Analysis: Prediction of Heterotic Pools and Heterotic Combinations

The phenomenon of heterosis is almost universal, is used at commercial scale worldwide, and has been both extensively and intensively investigated, but its genetic and molecular bases remain, at best, less than clear. Occasionally, the terms "true heterosis" or "euheterosis" are also used to describe the phenomenon of heterosis. When the F_1 hybrid from a cross is superior to both the parents for yield or some other trait, this phenomenon is called *heterosis* (Fig. 11.4). Generally, heterosis leads to increased vigor, size, growth rate, yield, etc. of the concerned crop, but some heterotic F_{1s} may be inferior to the inferior parent for a specific trait. For example, some tomato hybrids flower earlier than their early flowering parent; they are thus inferior to

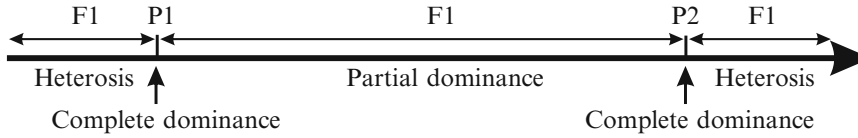


Fig. 11.4 A schematic representation of the concepts of partial dominance, complete dominance, and heterosis. P1, P2, and F_1: first parent, second parent, and their F_1 hybrid, respectively

Table 11.3 Some of the terms in common use to describe different features of heterosis

Term	Meaning	Remarks
Heterosis	F_1 mean lies outside the parental range	The most widely used term
True heterosis, euheterosis	F_1 mean lies outside the parental range	Rarely used
Heterobeltiosis	F_1 superior to the superior parent	Sometimes used
Hybrid vigor	F_1 superior to the superior parent	Generally used as synonym of heterosis
Average heterosis, mid-parent heterosis, relative heterosis	F_1 superior or inferior to the mid-parent value[a]	Extensively used in quantitative genetic analyses
Economic heterosis, standard heterosis, useful heterosis	F_1 superior to the best commercial variety of the crop	The only estimate of heterosis that has commercial utility
Mutational heterosis	Heterosis produced by dominance gene action	Little relevance in plant breeding as it cannot be estimated
Balanced heterosis	Heterosis due to overdominance gene action	Little relevance in plant breeding as it cannot be estimated

[a]Mid-parent value is the average of the two parental values

their inferior parent for days to flowering. This situation can be viewed as a faster completion of the vegetative phase of development in the hybrid than in its parents. The term *hybrid vigor* is generally used as a synonym of heterosis, but this term may not cover those situations where the F_1 is inferior to its inferior parent. There are several other terms that describe one or the other feature of heterosis (Table 11.3). Of these, the terms *economic*, *standard*, and *useful heterosis* denote that the concerned F_1 hybrids are superior to the best variety of the concerned crop that is in commercial cultivation. This estimate of heterosis is of great practical value since an F_1 hybrid has to be superior to the best/ruling commercial varieties, including hybrid varieties, of the concerned crop to succeed at the commercial level. In this context, performance of the parents of the hybrid assumes relevance as it would be reflected in the superiority of the heterotic hybrid over the prevalent varieties. Heterosis is trait-specific, i.e., some traits of F_1 may show heterosis, while other traits may not. Heterosis is generated in crosses between genetically distinct individuals/lines. It seems to involve many loci and a variety of different mechanisms. In addition, and different genes for heterosis may affect different traits of the hybrids. Further, there may be substantial differences in the genetic bases of heterosis in different crop species.

The extent of heterosis is much higher and its expression is more widespread in maize than in other plant species. One reason for the greater expression of heterosis in maize may be the large amount of diversity in genome sequences of different inbred lines of maize. The sequence variations in maize include single-nucleotide polymorphisms (SNPs), insertions and deletions (InDels), presence/absence of genes, genomic

segment rearrangements, copy-number variations, and insertion of transposable elements. On an average, one polymorphism would be observed at every 100 bp of the genomes of any two maize inbred lines chosen at random. In contrast, the genome sequence variations in other plant species are less frequent and consist mainly of SNPs and small InDels.

11.5.1 Genetic Basis of Heterosis

It has been known for long that outcrossing generally improves vigor and fertility (*hybrid vigor*), while inbreeding reduces them (*inbreeding depression*). Mating between individuals related by descent is known as *inbreeding*; it increases homozygosity in the progeny in proportion to the amount of inbreeding. In general, inbreeding depression is common for characters associated with fitness, while other traits show either little or no inbreeding depression. The degree of inbreeding depression observed in natural populations of different species ranges from very low (or even zero) to very high. This variation apparently depends on the level of inbreeding prevalent in the concerned populations. In general, naturally cross-pollinated and asexually reproducing crops show more severe inbreeding depression than self-pollinated crop species or such cross-pollinated crops that have been subjected to various degrees of inbreeding during their domestication. In short, inbreeding increases homozygosity, while outcrossing generally reduces homozygosity and increases heterozygosity. Therefore, heterosis and inbreeding depression have been considered as the opposite aspects of the same phenomenon.

A number of different hypotheses have been proposed to explain heterosis and inbreeding depression. These hypotheses can be grouped into the following two categories: (1) specific gene-based hypotheses and (2) hypotheses based on genome-wide mechanisms (Kaeppler 2012). The two most popular gene-based hypotheses were proposed in the year 1908. Davenport proposed that heterosis results from the masking of the deleterious effects of recessive alleles by their respective dominant alleles in the heterozygous F_1 hybrids. But these alleles are expressed in homozygotes leading to inbreeding depression (*dominance hypothesis*, also called *complementation hypothesis*). Complementation may involve variation resulting from single-base changes, altered gene expression, altered epigenetic regulation, or whole-gene presence/absence variation. In contrast, East and Shull postulated that the heterozygotes at certain loci perform better than the homozygotes for the respective loci, which leads to heterosis (*overdominance hypothesis*). The dominance and overdominance hypotheses generally lead to similar expectations, but some of their consequences are notably different. For example, according to the dominance hypothesis, the heterozygote should be comparable to the homozygote for dominant alleles of all the concerned genes. Therefore, it should be possible to isolate homozygous lines comparable to the F_1 hybrid in vigor and fertility. However, as per the overdominance hypothesis, F_1 hybrid will always be superior to the homozygote for the dominant alleles of all the genes involved. Therefore, inbred lines comparable to the F_1 hybrid in vigor and performance cannot be isolated. Much later, Gowen (1952) suggested that *epistasis* may also be involved in heterosis and inbreeding depression.

Considerable experimental evidence has been accumulated in an effort to determine the roles of dominance, overdominance, and epistasis in heterosis and inbreeding depression. The heterozygotes for some genes are definitely known to be superior to the concerned homozygotes, but the number of such genes seems to be rather small. It may be pointed out that many of the cases of heterozygote superiority may be the result of repulsion-phase linkage (*pseudo-overdominance*; Fig. 11.5), epistasis, or both. As a result of repulsion-phase linkage, unfavorable allele of a locus would be retained even in the face of strong and persistent selection by breeders for enhanced performance (Schnable and Springer 2013). Such slightly deleterious alleles will be maintained especially in the genomic regions with low recombination rates, e.g.,

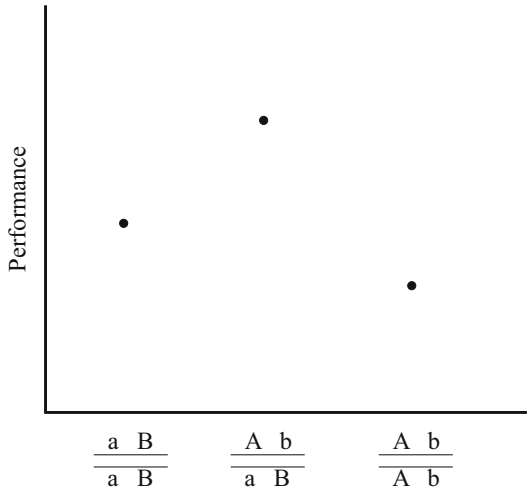


Fig. 11.5 Repulsion-phase linkage leads to heterosis and mimics overdominance (*pseudo-overdominance*). Gene *A* affects one trait, and its allele *a* has unfavorable effect. Similarly, gene *B* affects another trait and its allele *b* has unfavorable effect. If the genes *A* and *B* were closely linked in repulsion phase, the heterozygote *Ab*/*aB* will be superior in performance to both the homozygotes

regions surrounding centromeres (often termed as *recombination deserts*). This situation would give rise to *pseudo-overdominant heterosis QTLs* (*hQTLs*). Epistatic gene action can also lead to heterosis in a manner that resembles overdominance. For example, when a functional protein is a heterodimer, e.g., A-B, two inbred lines may produce the protein molecules A1-B2 and A2-B1, respectively. These molecules may be less efficient, say, in responding to changing environmental conditions, than the molecules A1-B1 and A2-B2, which will be produced by the hybrids in addition to the two parental molecules. As a result, the hybrids would tolerate stress better than the parental inbred lines and in this respect will be superior to the parents. It may be added that numerous studies in several species have characterized hQTLs for many traits. Many reports show that the hQTLs exhibit dominance, while other studies found apparent overdominance, which may be real or a consequence of pseudo-overdominance.

It is generally accepted that dominance gene action is the major contributor to heterosis, but overdominance and epistasis are also involved. The relative importance of epistasis and overdominance may vary depending on the species, the cross, and even the trait concerned. In case of maize, dominance seems to be the primary cause of heterosis, and the reported cases of overdominance could result from epistasis and/or repulsion-phase linkage. In the case of rice hybrid Shanyou 63, partial-, full-, and overdominance gene effects and digenic epistasis, particularly, dominance × dominance epistasis, were involved in heterosis. In contrast, an analysis of sister-line maize hybrids revealed that heterosis could be produced without nonadditive gene effects and in the absence of substantial genome-wide heterozygosity. The parental inbreds of these sister-line hybrids were derived from the same cross and were 47–77 % identical by descent (Lee et al. 2006).

Recent studies reveal that in some cases of heterosis, dominance, overdominance, and epistasis are unable to explain the observed results. Therefore, the following *genome-wide mechanisms* have been proposed for heterosis: (1) gene dosage balance; (2) epigenetic mechanisms, including DNA methylation; and (3) increased energy efficiency hypotheses (Kaeppler 2012; Schnable and Springer 2013). These mechanisms are gene/allele independent and trait-nonspecific, i.e., would produce heterosis for all the traits to a similar degree (Sect. 11.5.2). Thus, it seems that there is no simple unifying theory for heterosis, and the mechanisms involved in heterosis may vary with species, cross, and even trait.

11.5.2 Molecular Basis of Heterosis

In general, the growth of heterotic hybrids is faster, they flower earlier, they have higher leaf area index, and they accumulate greater biomass than their parents. However, the harvest index of these hybrids is usually comparable to that of their parents. A variety of hypotheses have been advanced to explain the phenomenon of heterosis. For example, a superior growth regulator balance and/or complementation for levels of different rate limiting enzymes in the parents

have been suggested as the possible causes of heterosis. Thus far, the following nine types of mechanisms have been proposed to explain heterosis: (1) an intermediate amount of a biochemical (governed by a single gene) as the optimum level (e.g., *p*-amino benzoic acid content in *Neurospora crassa*), (2) the protein products of different alleles with different properties/functions (e.g., sickle-cell hemoglobin in humans, *Adh1* in maize), (3) generation of "hybrid" products (e.g., "hybrid" transcription factors), (4) effects produced in two different tissues, (5) enhanced expression levels of some genes, (6) expression of a greater number of genes, (7) superior gene dosage balance, (8) increased energy efficiency, and (9) increased connectivity in the metabolome networks. In case of maize, the numbers of genes expressed in hybrids are usually up to 30–60 % higher than those in their parents, and heterotic hybrids express a much larger number of genes than do nonheterotic hybrids. This enhanced gene expression may be due to the transcription factors encoded by QTLs showing overdominance (Kaeppler 2012; Krishnan et al. 2013; Schnable and Springer 2013).

The above mechanisms for heterosis are based on function of coding regions of genes. Some mechanisms involving changes in noncoding regions of genes have also been proposed to explain heterosis. In case of heterotic hybrids of rice, about 50 % of the genes differentially expressed in the hybrids and their parents had InDels in their predicted promoter regions. Often these InDels either generated or disrupted some predicted regulatory sequences, to which transcription factors would bind. It has been proposed that different combinations of the functional/nonfunctional states of such binding sites and of the genes encoding the corresponding transcription factors could lead to additive, dominance, or overdominance gene actions (Zhang et al. 2008). For example, one parent of a hybrid may produce a functional transcription factor, but it may lack a functional binding site for the factor in the promoter(s) of the concerned gene(s). The other parent of the hybrid may have the functional binding site in the promoter(s) of the concerned gene(s), but it may not produce the transcription factor in functional state. In such a case, the F_1 hybrid will have functional binding site in the concerned promoter(s) as well as produce functional transcription factor. This will lead to overdominance gene action since the F_1 hybrid will be able to express the concerned gene, while its two parents will fail to do so. In such a case, the observed overdominance is, in fact, based on epistatic interaction between the genes encoding the transcription factor and the gene(s) activated by this factor (Fig. 11.6). Comprehensive analyses of gene expression reveal that most of the genes that are expressed at different levels in the parents show intermediate levels of expression in the hybrid suggesting additive action. But there is also evidence for nonadditive, i.e., dominant and overdominant, gene action in hybrids, suggesting that *trans*-acting factors like transcription factors play an important role in heterosis. The proportion of genes exhibiting nonadditive expression is mostly small, but in some studies it is relatively higher; this variation depends on species, parental genotypes, tissues analyzed, and the experimental design used. In addition, hybrids show nonadditive expression patterns for proteins as well, and the proportions of such proteins and the type of nonadditive effects observed vary substantially among different studies.

It has been proposed that heterosis may be the result of a *greater cellular energy efficiency of hybrids* due to selective protein synthesis and metabolism (Goff 2011). About 1–5 % of the genes of different species have alleles that encode unstable or inefficient proteins. It has been postulated that a quality control mechanism in the hybrids detects the alleles encoding unstable or inefficient proteins and suppresses their expression. This selective suppression of gene expression would lead to an overall saving of energy that would have otherwise been spent on the synthesis and degradation of these unstable/inefficient proteins. This energy saving would become substantial when it is summed up over all such loci. As a result, there would be more rapid cell division in the hybrids leading to a growth advantage and, ultimately, improved

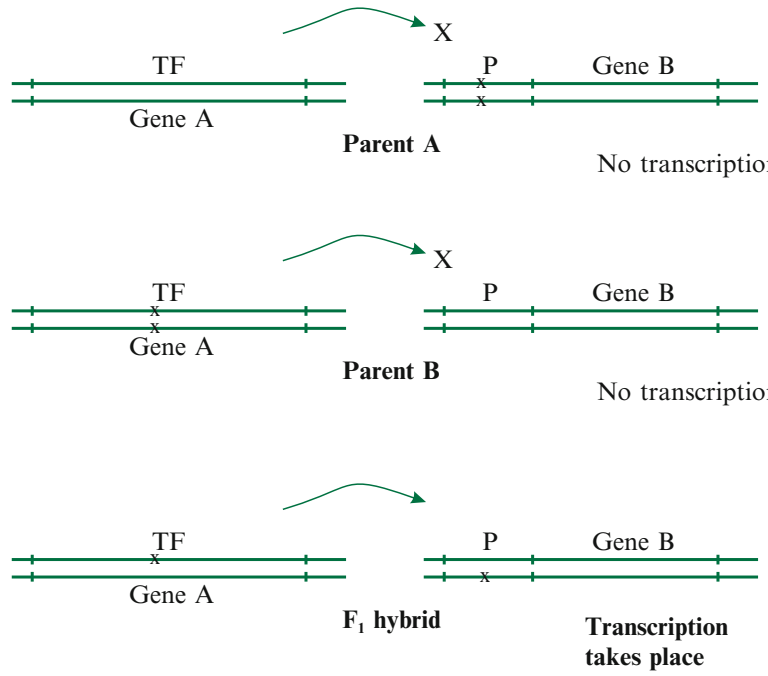


Fig. 11.6 Interaction between a transcription factor (TF) encoding gene *A* and the promoter binding site (*P*) of another gene (gene *B*) produces heterosis. The TF and *P* marked with X are nonfunctional. The *arrow* indicates that TF binds to the promoter to enable transcription, while the cross in front of the *arrow* signifies failure of this binding

performance. It has been argued that this model of heterosis is consistent with the three specific gene-based theories of heterosis and can explain a number of findings where these theories prove inadequate. Another mechanism proposed to explain heterosis involves the epigenetic mechanisms, including DNA methylation. Several studies have revealed the presence of epigenetic variation within plant species, e.g., DNA methylation or histone modification differences in *Arabidopsis*, maize, and rice genotypes, that can exhibit stable inheritance. There is also evidence for natural variation in small-RNA abundance among individuals of the same species. Studies with maize and other crops reveal that, in general, genomic DNAs of inbreds are more methylated than those of hybrids and that DNAs of heterotic hybrids are less methylated than those of less heterotic and nonheterotic hybrids. In addition, different genomic regions may be hypermethylated in different inbreds. It is believed that in plants the level of DNA methylation is negatively associated with gene expression. Therefore, the reduced level of DNA methylation in hybrids may lead to increased gene expression and, consequently, heterosis (Baranwal et al. 2012).

In general, aneuploidy reduces vigor, while polyploids are more vigorous than diploids. This indicates that organisms have mechanisms to sense gene dosage balance. The dosage sensing is likely to be a genome-wide mechanism that is independent of specific alleles/loci, and this mechanism could contribute to heterosis. For example, there is considerable evidence for the existence of the following two types of variations in the genomes of different genotypes/lines of a crop species. First, some genes may be present in one line, but they may be absent from another line (*presence/absence variation*). Second, some genomic regions may be present in more than one copy in the genome of one line, while the genome of another line may have additional copies of some other genomic regions (*copy-number variation*; *CNV*). In case two inbred lines differ for the genomic regions showing presence/absence and CNV, their hybrid would have average copy number of all the genes involved in these variations. In comparison, the parental inbred lines would show the two extremes of

copy numbers possible for them. Suppose that one inbred line has three copies per genome of one genomic region and the other inbred line has four copies per genome of another genomic region. These inbred lines, therefore, will have six and eight copies, respectively, of the concerned genomic regions, while their hybrid will have only four and five copies, respectively, of these two regions. As a result, the hybrid would show a greater *gene dosage balance* for the two genomic regions and would be more vigorous than the two parental inbred lines. The gene dosage effects over the entire genome will add up to generate heterosis. The findings of a study involving diploid and triploid inbred lines of maize indicate that the main part of heterosis results from dosage-sensitive mechanisms. Further, there is evidence that the mechanism of heterosis is significantly affected by dosage of the different cotton genomes (Yao et al. 2012).

11.5.3 Identification/Prediction of Heterotic Pools and Heterotic Cross Combinations

Heterosis is best utilized by hybrid varieties. When the F_1 generations from crosses involving generally two or sometimes more parents are used for commercial cultivation, they are called *hybrid varieties*. The parents of hybrid varieties can be inbreds, pure lines, clones, or other populations that are genetically dissimilar and combine well with other. In case of sexually reproducing crops, inbreds or pure lines are generally used as parents of hybrid varieties since this approach offers several advantages over the use of other types of parental materials. The different inbreds/pure lines of a crop show marked differences in their ability to combine well with each other and produce superior hybrids. Thus, some inbreds/pure lines combine well and produce excellent hybrids, while many others produce either average or poor hybrids. Therefore, it is extremely important to identify such inbreds that would produce superior hybrids. But this is the most expensive operation in a hybrid breeding program. In case of maize, this involves phenotypic evaluation of inbreds to reject weak and inferior inbreds, topcross test to assess the general combining ability (GCA; based on this test, ~50 % of the poor inbreds are discarded), and evaluation of single crosses among the remaining inbreds to identify those producing excellent single crosses. The single-cross evaluation is the most demanding step as it involves the production of $n(n-1)/2$ single crosses among n inbreds, and their evaluation in replicated field trials is conducted under multiple environments (Singh 2012a).

11.5.4 Molecular Markers in Resolution of the Genetic Basis of Heterosis

Several researchers have used molecular markers in studies with several different species in an effort to unravel the genetic basis of heterosis. Some studies have reported the importance of overdominance, others have found dominance to be predominant, while some others have observed extensive epistasis. For example, in one of the early studies, molecular markers linked to QTLs were used to analyze the expression of heterosis in an elite rice hybrid (Zhenshan 97 × Minghui 63). These QTLs showed overdominance, but the overall heterozygosity was not relevant for heterosis. Digenic epistasis was common and even such pairs of loci that did not exhibit overdominance generally showed epistatic interaction. Thus, in this rice hybrid, the main contributors to heterosis were overdominance at individual loci and epistasis, while overall heterozygosity was not that important (Yu et al. 1997). But it may be pointed out that a single QTL is likely to comprise more than one gene; these genes may be linked in repulsion phase or they may show epistatic interactions, which would easily mimic overdominance. In a recent study, Shi et al. (2011) crossed several doubled haploid (DH) lines of *Brassica napus* in random pairs to produce a population of F_1s, which were evaluated in three environments, and observations were recorded on 15 yield traits. The F_1 hybrid performance was not always positively associated with heterozygosity, but it was

explained by additive, dominance (partial, complete, and overdominance), and epistatic (additive × additive, additive × dominance, and dominance × dominance) gene actions. Similar results have been observed in several other crops, including rice, cotton, and *Brassica rapa*.

A set of introgression lines may be generated by repeated backcrossing to a single recurrent parent (RP), and these lines may be crossed back with the RP to produce several different F_1 hybrids. Each of these hybrids will be heterozygous for a different genomic segment introgressed from the donor parent. Evaluation of these F_1s for performance and for the heterozygous genomic segments would allow identification of such genomic regions that have the highest impact on heterosis. A line derived from the interspecific cross *Brassica oleracea* × *B. rapa* was crossed with the variety Express of *B. napus*, and 250 DH lines were isolated from the resulting hybrid. These DH lines were backcrossed to the variety Express. Evaluation of the resulting backcross progeny revealed positive overall mid-parent heterosis, but negative overall superior parent heterosis. Complete dominance or overdominance was observed only for seed yield, while only partial dominance was observed for the other traits. In addition, epistatic interactions were observed in many cases. These observations prompted the conclusion that epistasis along with partial dominance, complete dominance, and overdominance is responsible for heterosis in *B. napus* (Radoev et al. 2008; Kaeppler 2012).

11.5.5 Molecular Markers for Identification/Prediction of Heterotic Pools and Heterotic Cross Combinations

If one could identify the potential parents of superior hybrids on the basis of evaluation of the parental lines, it would save considerable time, effort, and other resources required for field evaluation of single crosses. Phenotypic performance of maize inbreds is a good indicator of hybrid performance for traits with high heritability. However, inbred yields show only a small (usually ~0.2) positive correlation with the yields of their hybrids, which is too small for predicting hybrid yields. The quantitative genetics theory predicts that mid-parent heterosis is a function of the genetic divergence between parents of a hybrid; this suggests a possible basis for predicting heterotic cross combinations. When epistasis is absent and there are only two alleles per locus, the mid-parent heterosis in maize population crosses is a function of the square of modified Rogers' distance (GD_{MR}) (Moll et al. 1965; Melchinger et al. 1990). It was concluded from an analysis of hybrids from a diallel mating among US maize populations that mid-parent heterosis increases linearly with GD_{MR}. But the findings from a study based on tropical maize populations collected from diverse geographic regions suggested that mid-parent heterosis increases with GD (inferred from geographic origins of the populations) up to an optimum level, beyond which it declines as highly divergent parents are crossed. This decline might result from fertility distortion and epistatic interactions. In a later study with seven tropical late maize populations, the GD was determined using SSR markers (Reif et al. 2003). In this study, the mid-parent heterosis increased linearly with GD possibly because the seven maize populations were not highly divergent. Thus, the positive correlation between GD and heterosis is the clearest when hybrids with similar adaptation and selected for productivity are compared. But as the GD increases and hybrids differing in adaptation are included, the relationship between GD and heterosis is lost.

The relationship between GD and heterosis can be useful in heterosis breeding in the following two ways. First, the GD may be used to classify the lines/populations into distinct heterotic groups. A *heterotic group* comprises such lines/populations, which when crossed with each other show little or no heterosis. But when lines/populations from different heterotic groups are crossed, there is moderate to high heterosis, which is often termed as *heterotic pattern*. In case of maize, heterotic groups are well established (Melchinger 1999); typically, the inbred lines of a heterotic group have high

kinship coefficient (relatedness by descent), are similar in their main characteristics, and have high general combining ability (GCA). In the beginning, heterotic group classification was based on pedigree information and performance of the hybrids in field trials. Later, combining ability analysis was used for the classification of the heterotic groups. More recently, molecular marker data have also been used for this purpose, but conclusive classification still depends on combining ability analysis. The maize inbreds/populations were initially grouped into two heterotic groups, the Lancaster and the Reid Yellow Dent groups. Subsequently, new germplasm introductions and gene/gene combination introgressions from them into the existing germplasm gave rise to new heterotic groups. There is considerable evidence that GD estimates from molecular markers are able to successfully assign new inbreds/populations to specific heterotic groups. In addition to maize, there is some evidence for heterotic groups in rice: the restorer lines and the male sterile/maintainer lines of the three-line *indica* hybrids constitute two main heterotic groups, viz., medium-maturity *indica* rice of South Asia and Southeast Asia and early-maturity *indica* rice of South China. The male sterile lines of the two-line hybrid system possibly form the third heterotic group. *Brassica napus* is another crop that may have heterotic groups (Krishnan et al. 2013).

The second application of GD estimates would be for predicting the level of heterosis expected in hybrids from specific cross combinations. However, this expectation has not been realized to any satisfactory extent in any of the crop species, including maize (Lee et al. 2006). As a result, GD cannot be relied on as a criterion for selection of parents for the production of superior hybrids. Therefore, in maize, several parental lines belonging to different heterotic groups, known to exhibit favorable heterotic patterns, are usually selected for further evaluation. Crosses are made between pairs of the selected inbred lines, and their hybrids are evaluated in trials, usually under multiple environments, to identify the parents producing superior hybrids. In view of this, the following criteria have been recommended for the selection of heterotic combinations of maize populations: (1) the hybrid population should have high mean performance as well as large genetic variance; (2) the populations themselves should have high per se performance and show good adaptation to the target region(s); and (3) they should show low inbreeding depression in case inbreds are to be used for the production of hybrids. Finally, (4) GD based on molecular markers may also be used as an additional criterion in the selection of cross combinations. Marker information complements field trials in finding out groups of genetically similar germplasm and helps in more efficient planning of field trials for identifying promising heterotic patterns. When heterotic groups are to be established for a large number of germplasm accessions, the accessions may be classified into heterotic groups on the basis of GD estimated from marker data. For example, heterotic groups are not clearly defined in tropical maize germplasm. The use of GD estimated from SSR marker data allowed the classification of seven tropical maize populations into the established as well as some new heterotic groups (Reif et al. 2003). The above procedure would allow the breeding effort to be concentrated on a relatively smaller number of predicted heterotic combinations. This approach would be particularly useful when it is applied in the initial stages of a hybrid program.

The failure of GD to predict the performance of individual hybrids was thought to be due to the following two reasons. First, the GD was estimated from data on phenotypic traits, whose expression is affected by the prevailing environment. Second, most of the traits measured for estimating GD may not be involved in the expression of heterosis. Therefore, once molecular markers became available, it was expected that they would provide a more dependable estimate of GD since they detect differences at the DNA sequence level, which is not affected by the environment. However, since DNA markers are generally random, relating this GD to the divergence in gene function remains a challenge. A number of investigators have estimated GD from molecular marker data and used them to predict

Table 11.4 A summary of the number of different studies on correlation between genetic divergence and heterosis in different crop species and the model plant *Arabidopsis thaliana*

	The strength of correlation between genetic divergence and heterosis		
Crop species	High	Moderate	Low
Maize[a]		1	3
Rice	2[b]	1	1
Wheat			6
Brassica		1	4
Cotton		1[c]	3
Soybean		1[d]	1
Arabidopsis (model plant species)			2 (no correlation)

These studies were based on large number of experimental hybrids evaluated in multiple environments (Based on Krishnan et al. 2013)

[a]Molecular marker data enable effective classification of new inbreds into the established heterotic groups

[b]Elite rice lines, including parental lines of commercial hybrids

[c]In the case of interspecific hybrids

[d]Based on a small sample of hybrids evaluated in a single environment

heterotic cross combinations. In the first such study, GD estimated from restriction fragment length polymorphism (RFLP) data was used for the prediction of hybrid performance in maize (Lee et al. 1989). The authors evaluated 28 single cross hybrids from eight inbreds for grain yield at two locations for 2 years. It was concluded that RFLP-based GD estimates could be used in the place of field evaluation data for assigning maize inbreds to different heterotic groups (Lee et al. 2006).

Krishnan et al. (2013) have critically analyzed the findings from recent studies with maize, rice, wheat, *Brassica* spp., cotton, soybean, and *A. thaliana* (Table 11.4). In each of these studies, several F_1 hybrids were evaluated in more than one environment. These hybrids were derived by mating the following types of parental lines: ecotypes; germplasm accessions; open-pollinated populations; RILs or DH lines mated in random pairs or backcrossed to one of the parents, cultivars, CMS and restorer lines; and parental lines of commercial hybrids. The GD between parental lines was assayed from data on one or more of the following marker systems: isozymes, RFLPs, AFLPs, CAPSs, SSRs, SNPs, STSs, RAPDs, metabolic markers, and proteome/transcriptome analysis data (Table 11.5). These markers were used to assay genome-wide genetic divergence, divergence at yield QTL-linked markers, divergence for haplotype blocks, or heterozygosity for specific marker loci.

The salient features of the findings from above studies are as follows. (1) The crop species had a strong influence on the results obtained from a study. For example, the existence of heterotic groups in maize was confirmed, rice and *B. napus* seem to have heterotic groups, and there is little evidence for heterotic groups in other plant species. It has been argued that the distinct heterotic groups in maize may be the result of its unique domestication history and other features, including mode of reproduction. On the other hand, several factors, including domestication history, mode of pollination, ploidy level, and narrow genetic base may be responsible for the absence of heterotic groups in the other plant species. Finally, differences in genome organizations of maize and other plant species (Sect. 11.5) may also contribute to the lack of distinct heterotic groups in the other species. (2) The specific germplasm of a given species used for a study seems to have a marked effect on the findings. For example, in one study with rice, the highest correlation between marker diversity and hybrid performance was observed in a set of southern US long-grained varieties, followed by that in the set of Chinese rice hybrids. In contrast, a much lower correlation

Table 11.5 The usefulness of various types of molecular markers used so far for the prediction of heterotic groups and heterotic hybrid combinations

Type of marker	Application/remarks
Random markers (located in anonymous genomic regions)	Extensively used for investigation; in maize, assigns populations/inbreds to the correct heterotic groups; little or no value in predicting hybrid performance; comparable results from sparsely and densely distributed markers
QTL-linked markers (markers linked to QTLs for yield-related traits)	Used in some studies (particularly in rice); effectiveness in predicting hybrid performance not much better than that of random markers; to be effective, >30 % of the trait QTLs should be covered and <20 % of the markers should be random
Transcriptome/proteome analysis	Many genes upregulated and many others downregulated; expression level shows additive, dominance, and overdominance effects; may be more effective in predicting hybrid performance than markers; results affected by many factors, including genotype, tissue, developmental stage, environment, and statistical analysis
Markers linked to heterosis loci (HLs; detected by using the level of heterosis for QTL analysis)	Most of the HLs different from the QTLs for yield-related traits; expected to predict hybrid performance more effectively than other markers; effectiveness yet to be demonstrated

was seen in the set of *indica* and *japonica* mixtures. (3) In case of maize, estimates of GD from marker data are able to successfully assign exotic germplasm accessions and inbreds to specific heterotic groups. (4) But in none of the plant species studied, including maize, GD was able to reliably predict the performance of individual hybrids. The observed correlations between GD and heterosis were low in bulk (18) of these studies and moderate and high in a small number (6 and 2, respectively) of studies. It is noteworthy that the high correlations were confined to rice, and that in these studies, parents of elite hybrids were used for hybridization (Table 11.4). (5) There is some indication that prediction based on marker haplotype blocks (in the place on individual markers) and per se performance of the inbreds may improve the prediction power. (6) Finally, GD estimates from transcriptome/proteome data seem to be more precise than those from marker data in predicting hybrid performance. The above general picture remained the same whether sparsely or densely distributed markers were used for the assessment of the GD (Table 11.5).

There is some evidence that gene expression profiles can be used for the prediction of heterosis for yield in *A. thaliana* and maize and that transcript abundance of parental inbred lines of certain genes is a good predictor for hybrid vigor (Table 11.5). However, there is no consensus set of genes that is differentially expressed between all inbred/hybrid combinations. The results from gene expression analyses are affected by differences in genotype, plant material (including the specific tissue), developmental stage, environmental conditions, experimental design, and statistical procedure used in the study (Lippman and Zamir 2007; Kaepler 2012). In one study, more genes showed negative heterosis for gene expression than those showing positive heterosis. Positive heterosis was observed for genes involved in DNA replication and repair, while genes for other metabolic pathways either showed negative or both negative and positive heterosis for the individual pathways. However, the findings from gene expression analyses are inconclusive, and the cost involved and the benefits offered need to be critically assessed.

It has been proposed that GD estimated from markers based on yield-related genes would be more dependable for predicting yield heterosis. One of the conditions identified by Bernardo (2001) for a meaningful use of yield gene-based markers is as follows: at least 30–50 % of the QTLs for the trait must be linked to the molecular markers and the proportion of random markers, i.e., markers not linked to QTLs, should be less

than 20–30 %. There is little evidence, at least in rice, that markers linked to yield QTLs have useful predictive power for hybrid performance (Table 11.5). This is not surprising since it seems that the genomic regions involved in the expression of heterosis, i.e., *heterosis loci* (*HLs*) or *heterosis QTLs* (*hQTLs*), are mostly distinct from those harboring QTLs for yield-related traits. The terms *heterotic loci* and *heterotic QTLs* have also been used; however, these terms are not appropriate since the concerned loci themselves do not exhibit heterosis, but they are involved in the expression of heterosis. *HLs are detected by using the magnitude of heterosis as a trait for QTL analysis*. Several HLs have been detected and described in rice; some of these HLs were located in the same genomic regions as the QTLs affecting the trait, while most of them were located in other genomic regions. For example, in an immortalized F_2 population developed from a hybrid rice variety, 33 HLs were identified for four yield-related traits. Only ten of these HLs were detected by QTL analysis based on trait data, indicating that most of the HLs were different from QTLs governing trait performance (Hua et al. 2003). Similarly, 42 HLs were discovered for six yield traits in a set of 265 introgression lines (representing 81.5 % of genome) of *O. rufipogon*, the progenitor of cultivated rice, into a high-yielding *indica* cultivar. Fifteen of these HLs were located in the same/similar genomic regions reported to have QTLs, and two HLs (*hyp2* and *hsp11*) were co-localized with yield QTLs *qGY2-1* and *qGY11-2*, respectively (Luo et al. 2011). It may, therefore, be expected that GD estimated from HL-based markers may be more efficient than random or trait QTL-based markers in predicting heterosis (Table 11.5).

It is reasonable to assume that the parents of a heterotic hybrid would differ for a proportion of the HLs present in the concerned species, and different sets of HLs would be involved in different hybrids. Therefore, evaluation of several different heterotic hybrids will be necessary to detect and map most, if not all, the HLs present in a given crop species. If and when this were achieved, markers based on only HLs may be used to estimate GD for predicting the hybrid performance. It may be expected that this estimate of GD will have much greater heterosis prediction power than that from other marker types (Table 11.5). But until this is achieved, random and gene/QTL-based markers will have to be used for this purpose. In this situation, if each marker were linked to an HL, the magnitude of GD might show strong association with heterosis. This association would become progressively weaker as the proportion of markers linked to HLs declines. In case a very large number of HLs were evenly distributed throughout the genome, the GD of parental lines estimated from a large number of random markers densely distributed over the entire genome may be expected to show strong association with heterosis. This is because in such a situation, most of the markers may be expected to be linked to an HL. However, the available evidence does not support this expectation since higher marker densities do not lead to improved prediction of heterosis. Therefore, it seems likely that HLs are located in some limited number of genomic regions. It is also likely that only some of the HLs would have major effects, while the remaining ones would have minor effects. These suggestions derive some support from the patterns of genomic locations and effect size distribution of QTLs for various quantitative traits. Therefore, diversity for only such genomic regions that have groups of HLs and/or major effect HLs would be relevant for heterosis prediction. Further, genome-wide marker diversity is not likely to closely reflect the diversity for these genomic regions, and is likely to weaken the relationship between GD and heterosis.

Questions

1. "Heterosis is a common phenomenon, the genetic and molecular bases of which are not well known". Critically analyze this statement in the light of relevant information.
2. Discuss the usefulness of molecular markers in furthering the understanding and utilization of heterosis.
3. "Marker data successfully assign inbred lines to appropriate heterotic groups, but they are

unable to predict heterotic patterns". Evaluate this statement in the light of available relevant information.

4. Explain the meanings of genetic diversity and genetic diversity analysis. Briefly describe the genetic diversity analysis using morphological and molecular marker data.
5. Briefly describe the various methods used for phylogenetic studies, and highlight their strengths and weaknesses.
6. Discuss the various applications of molecular markers in the conservation and utilization of plant genetic resources.
7. Explain the meaning of gene-based and genome-wide mechanisms of heterosis and discuss their significance in explaining the phenomenon of heterosis.

Fingerprinting and Gene Cloning 12

12.1 Introduction

The availability of molecular markers has facilitated studies designed to gain insights into a variety of biological phenomena. These markers were initially developed for construction of saturated linkage maps of crop species since such maps were helpful in genome sequencing efforts and plant breeding activities. Soon, DNA markers were explored for their usefulness in indirect selection for traits that were simply inherited, but difficult or cumbersome to evaluate. This approach has been so encouraging that marker-assisted selection (MAS) has become an integral part of every comprehensive breeding program. These successes prompted the use of molecular markers in other plant breeding endeavors, including diversity analysis and heterosis prediction (Chap. 11). Molecular marker profiles can also be used for unequivocal identification of germplasm lines and crop varieties much in the same way as human DNA fingerprinting. In addition, high-resolution mapping enables the identification of markers very close to the mutant genes that produce detectable changes in phenotype. These closely linked markers can be used to identify large DNA fragments containing the mutant genes. Ultimately, these genes can be identified, cloned, and characterized using appropriate molecular techniques. In this chapter, the use of DNA markers for variety identification and for identification and isolation of mutant genes will be discussed in some detail.

12.2 DNA Fingerprinting

In 1985, Jeffreys and coworkers used the term *DNA fingerprinting* to describe a multilocus assay based on minisatellite DNA for unequivocal identification of human individuals. But at present, polymerase chain reaction-based (PCR-based) single locus assays using microsatellite or simple sequence repeat (SSR) DNA are used for human DNA fingerprinting. Initially, DNA fingerprinting studies in plants were based on human minisatellite probes and synthetic simple sequence repeat oligonucleotide probes. In addition, single and double locus probes have also been used for discriminating plant varieties. The marker systems available for DNA fingerprinting of plant varieties include amplified fragment length polymorphism (AFLP), DNA amplification fingerprinting (DAF), randomly amplified polymorphic DNA (RAPD), restriction fragment length polymorphism (RFLP), sequence characterized amplified regions (SCARs), SSRs, and sequence-tagged sites (STSs). The discrimination among plant varieties is required for intellectual property rights (IPR) protection, and for the assessment of genetic purity of parents of hybrids, different categories of seeds, and to verify the identity of plant produce offered for sale.

In general, the discrimination power of a marker system increases with the number of polymorphic bands in a gel lane. However, the scoring of bands becomes increasingly

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_12, © Author(s) 2015

complicated as the number of bands per lane increases. Further, in the cases of marker systems like DAF and RAPD, which use arbitrary primers, DNA fragments of similar size can be amplified from different genomic regions. This fact makes it extremely difficult to score the alleles of different bands with any degree of confidence. One problem with AFLP markers is that, at least in cereals, they tend to cluster in those genomic regions that exhibit low recombination frequency, e.g., the regions around the centromeres. As a result, AFLPs do not provide sufficiently uniform genome coverage. The use of methylation-sensitive restriction enzymes like *Mse*I would get around this difficulty since such enzymes would preferably cleave in the genomic regions deficient in highly methylated repetitive DNA. The enzyme combination *Pst*I/*Mse*I is reported to generate AFLP markers that provide good genome coverage in barley. The SSR markers are the preferred marker system due to their abundance, high level of polymorphism, genome-wide coverage, codominant nature, and ease and speed of genotyping. These markers are frequently used for several purposes, including identification of plant varieties/hybrids.

The ability of a molecular marker to reveal genetic relationships among inbreds and varieties depends on the following factors: number of polymorphic loci scored, degree of genome coverage, average number of alleles present at each marker locus, and the linkage relationships among the polymorphic marker loci. It has been shown that for n number of varieties to be distinguished, the total number of polymorphic bands scored should be between n and $2n$. When the optimal number of bands is scored, increasing the stringency of distinctness criterion does not markedly enhance the rate of discrimination among varieties. The *stringency of distinctness criterion* may be defined as the number of bands required to differ between a pair of varieties/lines for them to be accepted as distinct varieties/lines (Law et al. 1998). In a study with related maize inbreds, the associations detected with RFLP, DAF, RAPD, and AFLP markers were similar and they reflected the pedigrees of these inbreds. In contrast, isozyme data were less reflective of pedigree relationships, but the number of good isozyme loci was rather limited (Smith and Helentjaris 1996).

12.3 Characterization of Lines and Hybrids for Intellectual Property Rights Protection

Plant varieties and hybrids are considered as intellectual property since their development involves careful and scientific planning and years of hard work both in the laboratory and the field. Further, although they are developed from lines/strains occurring in nature, they usually represent a considerable reorganization of the existing gene combinations and involve a skilled selection work. An *intellectual property* (*IP*) may be defined as an idea, a design, an invention, a manuscript etc., which can eventually be converted into a useful product or an application. One of the chief limitations of intellectual properties is that others can copy, imitate, or reproduce them. These activities will greatly reduce the economic benefits to the inventors of the intellectual properties. Therefore, almost all the nations have enacted appropriate laws to protect the rights of the creators of IPs to gain exclusive economic benefits out of their inventions. These legal mechanisms together constitute the *intellectual property rights* (*IPRs*). Therefore, the individual(s) and/or institution(s) involved in the development of a plant variety/hybrid can use a suitable IPR protection for safeguarding their financial interests. The various methods available for IPR protection are trade secret, patent, copyright, and plant breeder's rights. In many countries, including India, plant varieties can be protected by either patent or, more generally, plant breeder's rights.

12.3.1 Plant Breeder's Rights

A government grants plant breeder's rights (PBR) for plant varieties (including hybrids) to the concerned plant breeders, originators, or the owners of the respective plant varieties/hybrids.

Table 12.1 A comparison among UPOV Act (1991), PPVFR Act (2001)[a] and patents

Feature	UPOV Act (1991)	PPVFR Act (2001)	Patent
Protection coverage	Varieties of all plant genera and species	Varieties of plant species specified by the nation	Inventions
Requirements for protection	Novelty[b]	Novelty[c]	Novelty
	Distinctness	Distinctness	Inventiveness
	Uniformity	Uniformity	Non-obviousness
	Stability	Stability	Industrial application and usefulness
Duration of protection	Minimum 20 years	15 years for varieties of crop species; 18 years for varieties of trees and vines	17–20 year
Scope of protection	Commercial use of all material	Commercial use of all material	Commercial use of the protected subject matter
Breeder's exemption	Yes, except for parental clones/ inbreds of hybrids, and essentially derived varieties	Yes, except for parental clones/ inbreds of hybrids, and essentially derived varieties	No
Farmer's privilege	Left to the national laws	Yes, as farmer's right; more extensive than UPOV Act (1991)	No

[a]The protection of Plant Varieties and Farmer's Rights Act (2001)
[b]A variety not in commercial use for more than 1 year
[c]A variety not in commercial use for more than 1 year in India or 4 years (6 years in case of trees and vines) outside India

A PBR protection gives the breeder/originator of a variety/hybrid the exclusive right to produce and commercialize the propagating material of that variety/hybrid and to exclude others from doing the same. The PBR is granted for a 15–20 year period. The person/institution to whom the PBR title for a variety/hybrid is awarded can authorize other interested persons/ organizations for producing and marketing the propagating material of this variety/hybrid. The PBR protection is not available to the parents of a variety and the breeding procedures. However, the parental inbreds of hybrid varieties are protected by the PBR laws. The PBR systems, in addition, have provisions for breeder's exemption and farmer's privilege in one or the other form. The development of PBR systems in different countries has been considerably aided by the efforts of UPOV (Union Internationale pour la Protection des Obstentions Vegetales; International Union for Protection of New Plant Varieties). The UPOV convention generates the broad framework of the PBR system, which is adopted by the member states to the convention and by other interested nations. At present, the UPOV 1991 Act is in vogue, and the protection offered by this act is more comparable to patent protection than that by the earlier UPOV Act of 1978. India has developed its own PBR system that combines specific features of both the UPOV Acts and has some unique features of its own (Table 12.1).

The protection available under UPOV 1991 Act includes production for commercial purpose and sale of all materials of the varieties/hybrids for 20 or more years. The farmer may be allowed to save a portion of his harvest and use it as seed for planting his next crop (*farmer's privilege*), but he can neither sell nor exchange this produce. The protected variety can be freely used for scientific purposes and for creation of genetic variability in plant breeding programs (*breeder's exemption*). However, essentially derived varieties are not exempt, and they are subject to the PBR protection granted to the initial variety. An *initial variety* is the variety that was used for the development of a new variety using a suitable breeding procedure. An *essentially derived variety*, on the other hand, has been defined as a variety that has been "predominantly derived from the initial variety." As a result, expression of the important characteristics of an essentially

derived variety is primarily determined by the genes or the gene combinations contributed by the initial variety. Thus, for example, a variety developed by either mutation of a single gene or transfer into an initial variety of a single gene through backcross method or by means of genetic transformation can be treated as an essentially derived variety. Obviously, such a new variety will be covered by the PBR protection granted to the initial variety. Therefore, permission from the holder of the PBR title to the initial variety will be needed for commercialization of the new variety. The development of transgenic varieties has to be based on the best existing variety of the concerned crops. In view of this, extension of the PBR protection granted to an initial variety to such varieties that are predominantly derived from this variety provides an incentive to the breeder for continued efforts to generate genetic variation for developing new varieties with further improvements in performance, including yielding ability. Thus, the risk of narrowing down the genetic base of cultivated varieties of a crop is minimized, and the efforts for developing new varieties with progressively higher yields are continued.

A plant variety must be registered with the appropriate authority for IPR protection under the provisions of the concerned PBR Act. For protection under the provisions of the UPOV Act (1991), every plant variety has to meet the following four requirements: (1) novelty, (2) distinctness, (3) uniformity, and (4) stability. These requirements are generally referred to as DUS (distinctness, uniformity and stability) criteria. The criterion of *novelty* requires the variety to have been in commercial cultivation for less than one year prior to the claim for its PBR protection. *Distinctness* requires the new variety to be clearly identifiable from all other varieties of the concerned crop species on the basis of at least one morphological, physiological, or some other character. Further, the new variety has to be sufficiently uniform in appearance on plant-by-plant basis when it is grown under the environment to which it is specified to be adapted (*uniformity*). The uniformity criterion will be particularly relevant for the traits that are used to establish its distinctness. Finally, in order to satisfy the criterion of *stability*, the appearance and the clonal characteristics of the new variety must not vary over generations grown under the environment of its adaptation.

12.3.2 Description of Plant Varieties

For PBR protection, each variety must be described in terms of such characteristics that distinguish it from other varieties of the crop. These characteristics should be little affected by the environment; they should be interpretable in genetic terms and they should reflect the pedigree and genetic makeup of the concerned varieties. In addition, the process of variety registration, including verification of the distinctness, uniformity, and stability should involve as little time, effort, and expenditure as possible. In general, the PBR regulations consider only morphological and physiological traits, including disease and pest resistance as indicators of distinctness, and DNA marker data are generally treated as additional data (Arens et al. 2010). For example, the DUS guidelines being used in India mainly rely on morphological traits despite the fact that these criteria are not able to distinguish closely related genotypes. The data on morphological traits are often considered indispensable in view of the ease in scoring, well-known genetic basis and universal availability of these traits. But often the expression of many of these traits is affected by the environment, in some crops these traits may have poor discriminating power, collection of relevant data is time-taking and becoming increasingly expensive, and for many traits the genetic basis may not be well understood. In addition, morphological data are usually not amenable to genetic distance estimation; therefore, they are not suited for decision on essentially derived status of varieties.

A working group was established by UPOV to make recommendations on the use of molecular markers for determination of distinctness of plant varieties. This group has recommended that molecular markers closely linked to oligogenes and even quantitative trait loci (QTLs) governing

morphological traits may be used for DUS testing (Button 2006). The markers that are to be used for DUS testing should be preferably gene-based or even functional markers so that there is 100 % correspondence between the marker and trait data. In addition, the markers should be transferable between laboratories, especially when they are based on PCR. Such markers usually present problems in transfer between laboratories due to a variety of factors, especially the PCR machine and the reagents (mainly *Taq* polymerase) used. UPOV seems to be favorable to the use of markers linked to oligogenic traits (Arens et al. 2010). It may also be feasible to use markers linked to major QTLs that show stable expression over environments.

The environment has little effect on data pertaining to electrophoretic patterns of isozymes and seed storage proteins, and the genetic basis of these patterns is well known. These patterns are used as supplementary evidence of distinctness and are included in the category of simply inherited characteristics in France. Isozyme data are also used to verify pedigrees and to establish identities of seed lots. However, these data alone are not sufficient for variety identification and discrimination, and they must be used in combination with morphological data. Further, they cannot be used for determining the status of essentially derived varieties for the same reason as the morphological traits. DNA markers overcome virtually all the limitations of the morphological markers. They are not affected by the environmental factors, their genetic basis is clearly known, and marker profiles can be interpreted as presence or absence of specific alleles. They are abundant, cover the whole genome, and can discriminate even between very closely related varieties. Some of the marker systems are amenable to extremely high-throughput genotyping, which increases the data acquisition speed and minimizes the cost per data point. DNA markers are helpful in the determination of distinctness, and the essentially derived status of varieties, in establishing pedigrees, and for assays of uniformity. In addition, they are extremely helpful in the detection of unauthorized use or misappropriation of protected materials in breeding of new varieties, e.g., the use of inbreds as parents of hybrids or for developing essentially derived inbred lines. Private sector breeders have been routinely profiling their protected inbred lines with DNA markers to monitor and guard against misappropriation by other breeders/seed companies (Smith and Helentjaris 1996).

12.3.3 Limitations of Molecular Markers

In some crops like rapeseed, molecular marker profiles may reveal some degree of heterogeneity, while morphological traits may not exhibit detectable heterogeneity. This may create difficulties in variety registration on the basis of molecular marker profiles, particularly in crops that show some amount of cross-pollination. In addition, DNA markers would distinguish even very closely related varieties. This discrimination would allow registration of even such varieties that are genetically much closer than they would have been if they were distinguished on the basis of morphological traits. This would tend to narrow down the genetic base of cultivated varieties and increase the risk of genetic vulnerability. Therefore, a suitably stringent criterion of distinctness in terms of marker profiles needs to be established to minimize the risk of narrowing down of the genetic base. The residual heterozygosity in the varieties will be revealed by marker data, especially when codominant markers are used, and its elimination would require additional selection effort. Since verification of uniformity would require marker data from several dozen plants of a given variety, the breeders will have to do considerable amount of additional work.

It is argued that marker data do not conform to the UPOV requirement of a variety being "defined by expression of the characteristics resulting from a given genotype or a combination of genotypes." The DNA markers usually represent sequence variation in noncoding genomic regions. Therefore, it is generally difficult to accept the variation in marker profile as an

"expression" of the genotype. But it may be argued that the markers are in fact an "expression" of the genotype, the only difference being that the "expression" in this case is visualized as band patterns in gel images rather than trait phenotypes of the whole plants. Further, the noncoding DNA is not entirely irrelevant in the expression of plant characteristics. In any case, the meaning of "expression" needs to be more clearly stated so that the use of marker data for variety description and registration becomes legally sound. In addition, standardized, fast, user-friendly, and acceptable methods of marker data generation, acquisition, and comparison need to be developed and agreed upon. Finally, the minimum genetic distance needed to establish distinctness has to be defined keeping in view the risk of narrowing down of the genetic base. It may be pointed out that a single morphological trait summarizes the actions of many loci participating in the development of the concerned trait even when one or few major genes may appear to control the visible variation in the trait. In contrast, a DNA marker represents variation in nucleotide sequence at a single genomic site, and this variation may or may not affect the agronomic performance of the concerned individuals. At the same time, the amount of variation reflected by one morphological trait may not be the same as that by a single DNA marker. Therefore, the relevance of molecular marker data may be limited in deciding issues like the relatedness and distinctness of the varieties.

The concept of essentially derived varieties is not easy to implement since the conditions for deciding about the "essentially derived" status are not clearly defined. In simple terms, a variety is considered to be essentially derived if it were "predominantly derived" from another variety. In addition, the "essential characteristics" of this variety depend on "the genotype or combination of genotypes" of the variety from which it was derived (see Bernardo and Kahler 2001). The above description does not provide an objective measure for determining the "predominantly derived" feature. In addition, it is not clear whether the phrase "essential characteristics" refers to all the characteristics relevant to varietal performance, including yield and yield traits, or to only those traits that are listed as the distinguishing features of the initial variety. In case this phrase refers to "all the characteristics," then the question arises whether the same criterion is used for registration and protection of all the varieties. One view considers modification of the initial variety by mutation, or transfer by backcross or genetic transformation of one or few genes to develop a new variety would generate an essentially derived variety. It has been suggested that genetic distance estimated from marker genotype data covering the whole genome may be used to develop an objective criterion for determining the "essentially derived" status (Smith and Helentjaris 1996). This may be fine enough for varieties developed by genetic transformation and mutation particularly when the genes involved are themselves ignored. However, it may not work well in the case of varieties developed by backcross breeding because such a variety may be expected to retain over ~1 % of the donor parent genome even after five backcrosses with rigorous selection for the recurrent parent type. Thus, the new variety would retain around 9 Mb, 25 Mb and 160 Mb of the donor parent genome in case of tomato, maize, and wheat, respectively. A major part of this DNA would be confined to the chromosome into which the target gene has been transferred, but the rest of it would be distributed randomly over the genome. This donor DNA would contribute to the genetic distance based on molecular markers, but the effects of this DNA may not be reflected at the phenotypic level. Therefore, the genetic distance criteria may have to be defined separately for the different methods used for the development of the new varieties.

12.4 Assessment of Genetic Purity of Lines and Hybrids

The benefits from improved crop varieties and hybrids can be realized by their commercial cultivation, which requires the production of quality

seed that is often marketed as certified seed. Seed certification requires, among other things, prescribed standards of genetic purity. In addition, genetic purity of crop varieties and hybrids is essential for achieving the full potential of improved performance inherent in their genotypes. In a crop like rice, the plants of maintainer lines are the major off-types present in the CMS (cytoplasmic male sterile) lines and their F_1 hybrids. One percent contamination of the F_1 hybrid seed by the female line (maintainer and CMS lines) seed is estimated to reduce the yield of the concerned hybrid crop by about 100 kg/ha. The hybrid rice seed lot purity should be 98 % and 96 % in India and People's Republic of China, respectively. Maintenance of the purity of parental CMS lines would ensure the desired purity of F_1 hybrids. In order to achieve the above standard of hybrid seed purity, the genetic purity of seeds of the parental lines of the hybrid should be 99 % or more. The maintainer line plants present in the CMS lines are identical to those of the latter in morphological characteristics. Therefore, they can be distinguished from the CMS line plants only at the heading stage by monitoring pollen fertility or, at a later stage, by the seed set rates. This detection is not only too late, it is prone to yield variable results due to the environment and the skill of worker recording the data. Therefore, simple, fast, reliable, and efficient methods should be developed for determining genetic purity of the seed lots of the parental CMS lines (Sang et al. 2006; Rajendrakumar et al. 2007).

The seed lot genetic purity is assessed by evaluation of morphological and biochemical features of the plants and/or their seeds. Generally, grow-out test is regarded as the most reliable and elaborate genetic purity test, but this is time taking, cumbersome, and expensive. Assessment of genetic purity on the basis of molecular marker genotypes would be highly reliable, rapid, and convenient. Since a CMS line and its maintainer line have identical nuclear genotype and they differ only for their cytoplasm, the DNA marker has to be based on either mitochondrial or chloroplast genome. PCR-based DNA markers capable of distinguishing the plants of a maintainer line from those of the corresponding CMS line have been developed. Most of these markers are based on mitochondrial genome polymorphism, but one marker based on chloroplast genome also has been described. For example, a set of two pairs of PCR primers, viz., one primer pair designated as the *cms* primers and a second pair of primers denoted as RG136F and R, successfully distinguishes each of the several wild abortive CMS lines from their cognate maintainer lines. The *cms* marker amplifies a fragment of 386-bp from a region of mitochondrial DNA of the CMS lines only. The other primer pair, however, produces a single monomorphic fragment of ~1.1 kb from both CMS and the maintainer lines and serves as control (Yashitola et al. 2004). Subsequently, Rajendrakumar et al. (2007) developed the *drrcms* marker based on a mitochondrial SSR (an AT-repeat) that is amplified by the primer pair drrcmsF and drrcmsR. This primer pair consistently amplifies a single 130-bp fragment in the CMS line, while a single 142-bp band is produced in the maintainer line enabling their unambiguous identification. The *drrcms* marker is able to distinguish the CMS lines from their cognate maintainer lines in the case of not only the wild abortive CMS system, but also in the cases of CMS systems, based on *O. nivara* and *O. rufipogon* cytoplasms. The marker *drrcms* of Rajendrakumar et al. (2007) has the advantage of using a single primer pair as compared to the use of two primer pairs in a multiplex PCR in the case of the *cms* marker of Yashitola et al. (2004).

The DNA markers based on the chloroplast genome may be expected to be more stable and reliable in the long run than those based on the mitochondrial genome since the former is more conserved and undergoes much less reorganization than the latter. One AFLP fragment based on the rice chloroplast genome was found to be polymorphic between five rice CMS lines and their corresponding maintainer lines. The sequence of this fragment was determined; this revealed one more copy of a 6-bp tandem repeat in the AFLP fragments from the CMS lines than in those from their corresponding maintainers.

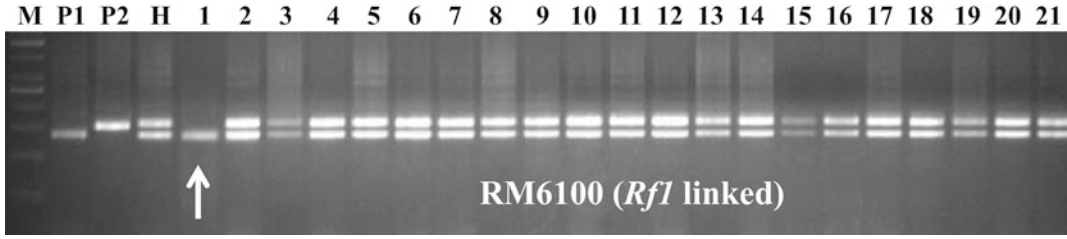


Fig. 12.1 Discrimination between hybrid and parental seeds by their SSR profiles. P1-Pusa 6A, P2-PRR78, H-PusaRH10, 1–21 – Test plants from the hybrid seed lot of PusaRH10, *white arrow* (lane # 1): contaminant (P1)

The sequence of this AFLP fragment was used to design a set of three PCR primers, viz., one forward primer and two reverse primers. This set of PCR primers generated a marker that was able to differentiate between rice CMS lines and the corresponding maintainer lines in the case of sporophytic CMS lines (three abortive cytoplasms from *O. rufipogon*, *O. sativa* ssp. *indica*, and *O. sativa* ssp. *japonica*). However, this marker could not distinguish the gametophytic CMS lines from their maintainers. The genetic purity of two sporophytic CMS lines and the corresponding maintainer lines were analyzed using this marker. The findings from this analysis agreed well with the results of genetic purity assay based on the grow-out test (Sang et al. 2006).

The genetic purity of seed lots of hybrid varieties can be monitored on the basis of the molecular marker profiles of commercial hybrids (Fig. 12.1). The purity of parental line and hybrid variety seed lots can be assayed by screening bulked DNA samples with SSR markers. The genetic purity of hybrid seed lots can, in addition, be assessed by using a molecular marker linked to the concerned restorer gene. It has been proposed that a national database of molecular fingerprints of all commercially grown hybrids may be created to curb marketing of unscrupulous hybrids, protection of IPR, and to authenticate the genetic purity of seed lots of commercial hybrids. Ideally, the parental lines of the hybrids should be used for this purpose, but they are generally not available; this is particularly true in the case of private sector hybrids since their parental lines are maintained as trade secrets. In view of this, it is necessary to develop a hybrid identification system that uses the DNA marker profiles of the commercial hybrids themselves. In one study, 25 commercially grown rice hybrids were fingerprinted with 40 SSR markers. It was found that data on at least 20 SSR markers were needed for distinguishing the 25 hybrid varieties from each other. Further, one hybrid could be reliably identified with a single SSR marker, while two or more SSR markers were needed for a clear cut identification of the remaining hybrids (Anand et al. 2012).

12.5 In Silico Gene Prediction

Identification of specific genes is basic to their isolation and cloning, elucidation of their function, and their utilization for the development of products and/or services, if any, for human welfare. Prior to the era of genome sequencing, gene detection and isolation involved a series of cumbersome and technically demanding experiments using living cells and organisms. These methods used genomic DNA clones and cDNA libraries for analysis with a variety of sophisticated techniques and were suitable for detection of individual genes. But as complete and virtually sequencing error-free genome sequences became available, the technology for genome-wide in silico search for genes was rapidly developed and refined. These endeavors have resulted in the creation of powerful computational resources, which have greatly facilitated gene identification by analyzing genome sequences. Some of the commonly used tools and database servers dedicated to gene prediction are listed in Table 12.2. The development of these efficient

Table 12.2 A list of some important and widely used gene prediction servers and tools

Name	Description/function
ATGpr	Identification of translation initiation sites in cDNA sequences
AUGUSTUS	Prediction of genes in eukaryotic genome sequences
BGF	A program for hidden Markov model-based ab initio gene prediction
GENIUS	The predicted genes from complete genome sequences are linked to the known protein 3D structures listed in the database
GENEID	This server predicts genes, signal sequences, and exons
GENEPARSER	Detection of introns and exons in the genes predicted from genome sequences
GeneMark	A family of gene prediction programs; based on a modified GeneScan algorithm
GeneMark.hmm	A gene prediction program for genome sequences of prokaryotes and eukaryotes
NIX	Web tool gene prediction based on combining results from different programs
VEIL	A server using hidden Markov model for finding genes in vertebrate DNA
Splice Predictor	This program identifies potential splice sites in plant pre-mRNA using Bayesian methods
GENESCAN	Gene prediction using Fourier transform
Fgenesh	The fastest and most accurate ab initio gene prediction program for eukaryotic genome sequences
NNPP	Promoter prediction by neural networks
NNSPLICE	Splice site prediction using neural network
GENOMESCAN	Prediction of the locations of exon–intron boundaries in genome sequences
ORF FINDER	A graphical analysis tool for prediction of open reading frames
GrailEXP	Predicts exons, genes, promoters, poly-As, CpG islands, and repetitive elements within DNA sequences
EuGène	Gene detection in eukaryotic genomes; uses probabilistic models to discriminate between coding and noncoding sequences and to distinguish between effective splice sites and false splice sites
tRNAscanSE	Prediction of tRNA encoding genes

Based on Singh et al. (2014b), and other sources

computer programs for gene prediction is considered as one of the most important single developments that have facilitated functional analysis of genomes.

Gene prediction, *gene hunting*, or *gene finding* refers to identification, by analysis of genome sequences, of such genomic regions that function as genes, i.e., encode proteins or various types of RNA species. Gene prediction is the first step in genome annotation taken up after the genome sequence has been assembled and checked for errors. *Genome annotation* is the process of identifying genes, their 5′- and 3′-regulatory sequences, as well as their functions. In addition, mobile genetic elements and repetitive sequence families are also identified and characterized. Thus, genome annotation involves not only the identification of protein and RNA encoding genes and their regulatory sequences, but also the detection and description of such other functional elements that have regulatory functions or are relevant in some other way for genome organization and function. In short, in silico gene prediction is one of the first and most important steps in the quest for understanding the genome organization and function of a species with the help of a detailed analysis of its genome sequence. The findings from the in silico analyses are subsequently validated by suitably designed in vitro and in vivo studies.

The first step in the identification of a protein-coding gene using a DNA sequence is the determination of the correct reading frame. A *reading frame* is the arrangement of sets of three bases, each representing a codon, beginning at a specific nucleotide in a DNA sequence. Therefore, three reading frames are possible for each strand of a DNA molecule. In view of this, the correct reading frame is determined by carrying out a six-frame translation of a given DNA sequence. The longest reading frame that is not interrupted by a translation termination or nonsense codon (TAA, TAG, or TGA) is presumed to be the *correct reading frame*; generally, such reading

frames are known as *open reading frames* (*ORFs*). *An ORF has an initiation codon* (*typically*, *ATG*) *at its beginning and at least one of the termination codons at its end*. The determination of the 3′-ends of ORFs is relatively easier than that of their 5′-ends since the ATG codon can occur at internal sites of the genes as well. Therefore, additional criteria have to be used to locate the 5′-ends of ORFs, e.g., the presence of a Kozak sequence (CCGCC*ATG*GG) that includes the *ATG* codon. The 5′-ends of many vertebrate genes have characteristic CpG islands, and analysis of codon usage may provide helpful indications. However, sequencing errors may hamper the correct identification of ORFs.

Protein coding genes are usually identified by using a computer program for inspecting the genome sequence for such features that are specific to genes. For example, protein-coding genes, as a rule, comprise ORFs, and their detection is very effective in gene identification in the case of bacteria. In general, the longer is an ORF, the greater is the chance that it represents a gene. However, several features of eukaryotic genes make a direct search for genes on the basis of ORFs very difficult. For example, most eukaryotic genes comprise alternating exons (coding regions) and introns (noncoding regions) in the place of continuous ORFs. Further, the genes in humans and other eukaryotes are often widely spaced; this feature increases the chances of finding "false" genes in the long intergenic regions. The newer versions of ORF scanning software for eukaryotic genomes account for these features and enable an efficient scanning for genes in eukaryotic genomes.

There are mainly two strategies for the detection of genes from genome sequences. The first strategy is based on the nucleotide sequences of already identified genes, cDNAs and ESTs, and the amino acid sequences of known proteins available in various databases. These sequences are used for searching homologous sequences present in the given genome sequence using tools like BLAST. The sequences used for homology search may belong to the same species, a related species, or even a distant species. The reason for this relaxed requirement is that the coding sequences have usually been highly conserved during evolution. For example, sequences of *Mlo* gene family from *Arabidopsis thaliana* have been used for detecting genes in the genome sequences of soybean, rice, sorghum, wheat, etc. This approach can be used for identification of specific genes and genes belonging to particular gene families, but it cannot be used for a search of all the genes present in the genome of a given organism.

In the second approach of gene prediction, specialized software are used to search the genome sequences for the presence of genes; this is termed as *ab initio gene prediction*. This is relatively easy and quite efficient in the case of prokaryotes. Computer programs like GeneMark.hmm and GLIMMER are capable of identifying all types of genes in the prokaryotic genome sequences; these programs can detect even overlapping genes. GeneMarkS is a self-training program and is suitable for gene prediction from novel genomes. MetaGeneMark is designed for analysis of metagenomic sequences. Several sophisticated tools for gene prediction from eukaryotic genome sequences, e.g., GeneMark-E, GeneMark.hmm-E, AUGUSTUS, GENESCAN, EUGENE, Fgenesh, etc., are now available (Table 12.2). Some of these programs are designed for gene hunt in a specific species or group of species; e.g., the program EUGENE was developed for *A. thaliana*. GeneMark-ES is a self-training program and suitable for use with novel eukaryotic genomes. Fgenesh is perhaps the fastest program for gene prediction from eukaryotic genomes, and it is also considered to be the most accurate of such programs. Some programs serve specific functions, e.g., NNPP performs promoter prediction using neural networks, Splice Predictor identifies potential splice sites in plant pre-mRNA using Bayesian methods, GENEPARSER detects introns and exons in the genes predicted from genomic sequences, etc. (Table 12.2).

The gene prediction programs search for gene-specific features, *such as promoters*, *splice sites*, *and polyadenylation sites or for pertinent gene contents like ORFs*. Many of the currently available gene search programs combine

different search criteria and their sensitivities vary widely. The identification of ORFs, usually, exceeding 300 nucleotides, is sufficient to find most genes in prokaryotic genomes. However, such a simple search criterion will miss smaller genes and overlapping genes. These problems are resolved by using algorithms that consider differences in base composition between genes and noncoding DNA, e.g., in programs like GeneMark. The gene prediction programs used in eukaryotes use the output from several algorithms to generate a whole gene model. In this model, a gene is defined as a series of exons that are coordinately transcribed. The various features of eukaryotic genes including transcriptional and translational controls like TATA box, cap site, Kozak sequence, and polyadenylation sites are recognized during the gene detection process. But problems arise as TATA box is missing in ~70 % of human genes, and polyadenylation signal sequences can differ considerably from the consensus sequence AATAAA. Further, the above criteria identify only the first and the last exon of a given gene. Therefore, additional features have been included in the modern gene search tools; these features include 5′- and 3′-splice sites, differences in base composition between coding and noncoding DNA, etc.

Once the genes are predicted, their functions can be determined as follows. The simplest method for identifying the function of a new gene is to translate its base sequence into the amino acid sequence it is expected to encode. This protein sequence is then compared with a protein database like PDB (the Protein Data Bank); a program like tBLASTx will perform both these operations. If the predicted protein is homologous to a protein in the database, it suggests the gene function and confirms the identification of a new gene. Alternatively, the gene sequence may be compared with the genes present in the syntenic genomic region of a related species that has rich genomic resources. Homology with a gene in the syntenic region would indicate the most likely function of the gene. Several types of RNA species are noncoding, e.g., rRNA, tRNA, and a variety of small RNA species, etc. Of the various types of RNA species, the genes encoding rRNA are the easiest to detect; this is done by sequence similarity search since their sequence is highly conserved across species. The program tRNAscanSE searches for tRNA encoding genes.

12.6 Chromosome Walking

Chromosome walking is used for characterization of large DNA fragments. Generally, a cosmid library is used for chromosome walking. Each clone in such a library may be expected to have a DNA insert of ~50 kb. In chromosome walking, one begins with a DNA fragment that contains a known gene/DNA marker shown to be linked to a gene that is to be isolated and cloned. The sequence located at one end of this DNA fragment is used as probe to enable the identification of a new clone having DNA insert that partly overlaps the first fragment. Now the other end (the non-overlapping end) of the new fragment is used as a probe to identify overlapping fragments. In this way one continues to move step-by-step (*each fragment represents one step*) toward the gene of interest located close (usually, <1 cM) to the known gene/genetic marker. This technique, therefore, is known as *chromosome walking* as each new overlapping clone takes the researcher one step closer to the gene of interest (Fig. 12.2).

1. The first step in chromosome walking comprises the isolation of a DNA fragment (fragment 1 in Fig. 12.2) having a known gene/marker located close to the gene of interest. The only information available for the target gene is the phenotypic effect produced by this gene. This DNA fragment provides the starting point for the "walk" and a point of reference on the genetic map.
2. A restriction map of this DNA fragment (fragment 1) is prepared. A small segment representing one end of fragment 1 is isolated and cloned; this is called *subcloning*. The subcloned segment should be located close to one end of the fragment. This segment is used as probe for identification of those clones of the genomic library that hybridize with this

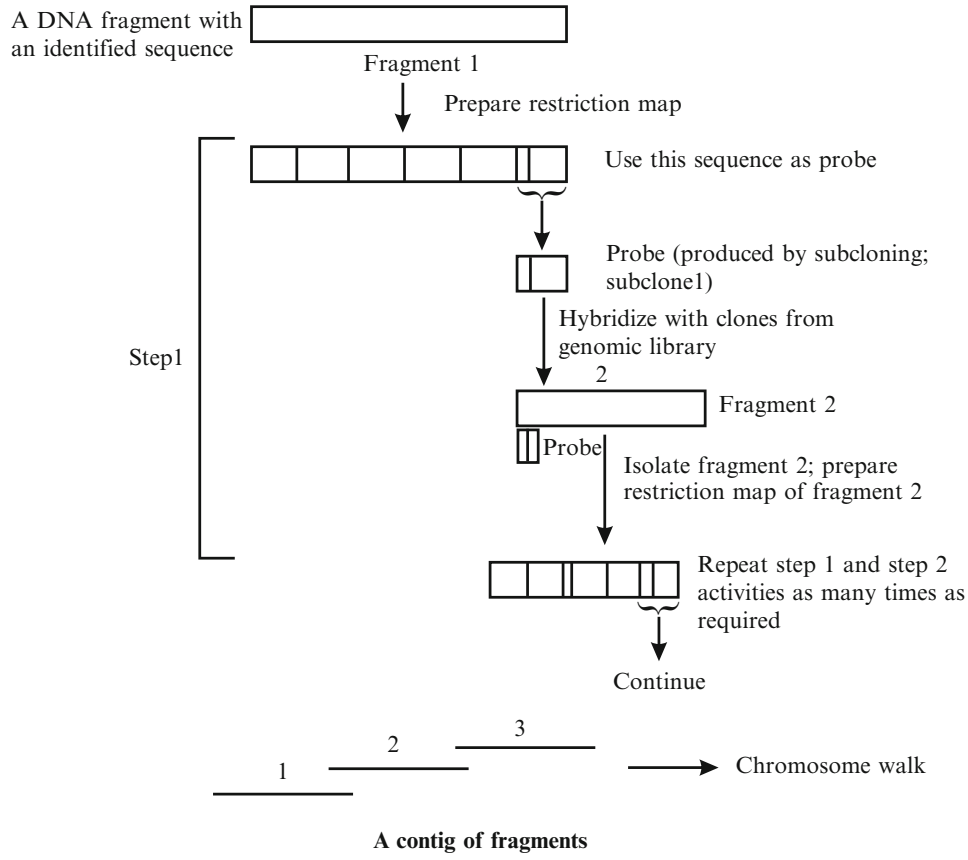


Fig. 12.2 A schematic representation of the technique of "chromosome walking." In this approach, the sequence at one end of a fragment is used as probe to identify the next overlapping fragment and so on

segment and, as a result, are expected to overlap fragment 1.

3. Restriction maps of DNA inserts of the clones identified in this way are prepared, and the DNA insert overlapping fragment 1 at one end is selected and labeled as fragment 2. The sequence at the other end of fragment 2 is subcloned and used as probe for hybridization of the genomic library clones to identify clones having DNA inserts overlapping the fragment 2.
4. The DNA inserts from clones identified in the step 3 are processed in the same way as described in item 2 above; these steps are repeated till the target gene is reached.

The other end of the fragment 1 may also be used as probe to identify overlapping clones; this strategy would permit "chromosome walk" in the opposite direction. As far as possible, the same genomic library should be used for the successive screenings with the fragment end sequences so that one is able to distinguish new clones from those clones that were identified and used previously. This is necessary to avoid using the same set of clones repeatedly leading to "walks" back and forth without any real progress in the walk. The procedure described above requires subcloning of the end sequences of the fragments, which involves additional work. Some vectors like λDASHII and λFIXII permit the generation of probes from DNA insert endpoints of the selected clones and make subcloning unnecessary; this also facilitates "walk" in the same direction.

12.7 Chromosome Jumping

Contigs of large DNA fragments of up to several hundred kilo bases can be created by using the chromosome jumping technique. In contrast, the technique of chromosome walking (Sect. 12.6) is applicable to DNA fragments of a much smaller size. A relatively simple method of chromosome jumping is based on specialized genomic libraries called "jumping" and "linking" libraries. These libraries are produced by using a rare cutter restriction enzyme, such as *Not*I for humans (Fig. 12.3). In the case of a *jumping library*, the DNA insert of every clone has the genomic sequences present on a single side of two adjacent recognition sites, e.g., the regions marked as 2 and 3 in Fig. 12.3, for the restriction enzyme used to construct the library. In contrast, the DNA insert in each clone of a *linking library* contains the genomic regions located on both the sides of the same restriction site for the concerned restriction enzyme, e.g., the regions denoted as 1 and 2 in the clone 1 shown in the Fig. 12.3. The construction of a *Not*I jumping library is described here in simple terms. *Not*I

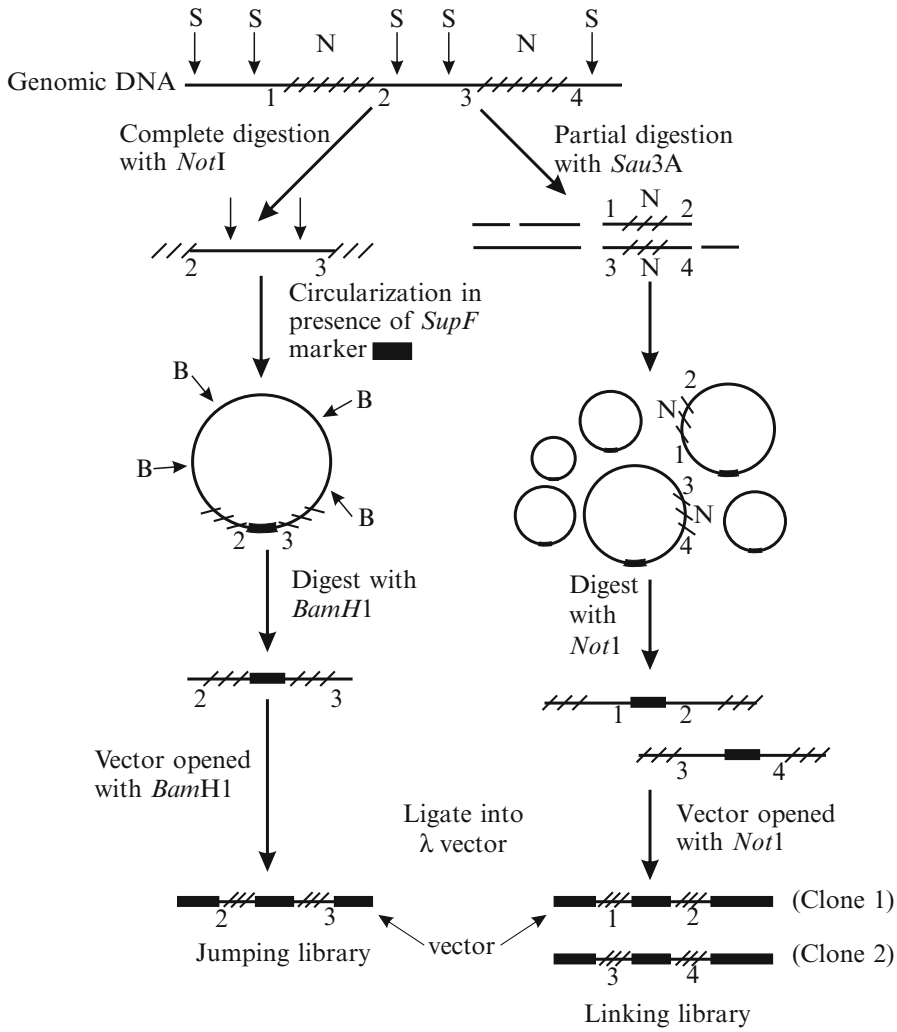


Fig. 12.3 A schematic representation of the construction of *Not*I "jumping" and "linking" libraries. In the genomic DNA, the *Not*I sites are cross-hatched and *Sau*3A sites are marked by *arrows*. The regions on either side of *Not*I sites are sequentially numbered as 1,2,3,4 to depict the chromosome regions represented in the clones of "jumping" and "linking" libraries

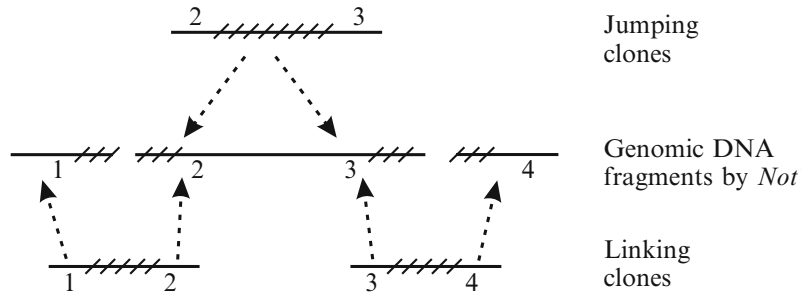


Fig. 12.4 Use of *Not*1 "jumping" and "linking" library clones to determine the correct order of DNA fragments produced by *Not*1 digestion

restriction enzyme is used to completely digest purified, high molecular weight genomic DNA. The fragments generated in this way are circularized in such a way that the *SupF* marker is included within the circularized fragments. The circular molecules are now digested using a frequent cutter restriction enzyme like *Bam*HI. This digestion removes bulk of the genomic DNA present beyond the first cutting site located on the two sides of the *SupF* marker. This step makes all fragments linear, which are integrated into an appropriate λ vector and propagated in a suitable strain of *Escherichia coli* (Fig. 12.3). The *SupF* marker is, in fact, a mutant tRNA gene that has a mutated anticodon. This mutant anticodon recognizes a nonsense codon produced in an essential gene by a suppressor-sensitive mutation. As a result, when *SupF* marker is introduced into such a strain of *E. coli* that has this suppressor-sensitive mutation in one of its essential genes, it functions like a selectable marker.

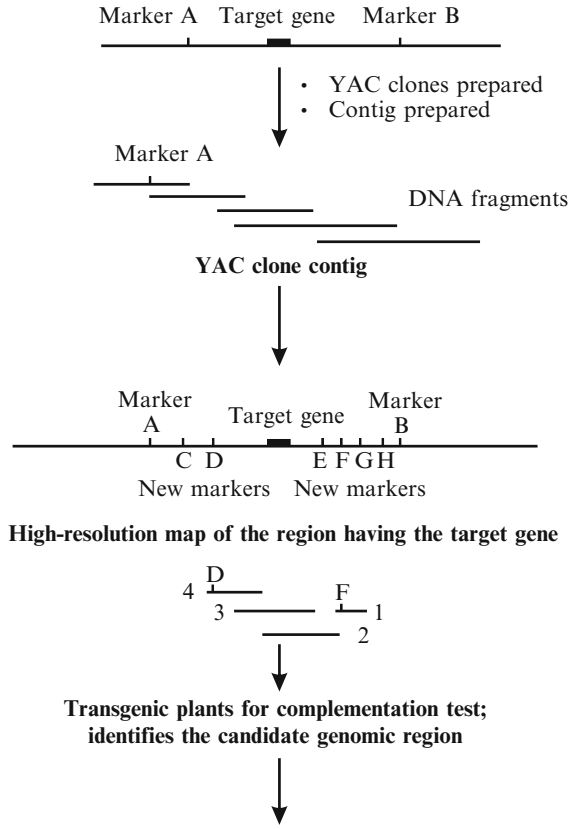
Similarly, the *Not*I linking library is prepared by partial digestion of the genomic DNA using a frequent cutter restriction enzyme, such as *Sau*3A. The fragments so obtained are made circular, and the *SupF* marker is integrated within them. The circularized molecules are now cut open with *Not*I, the enzyme that was used to create the jumping library. The circular molecules having at least one restriction site for *Not*I will become linear; these fragments are inserted into an appropriate λ vector and cloned into a suitable strain of *E. coli*. It may be noted that only those circularized molecules that contain a *Not*I recognition site will become linear and will be cloned (Fig. 12.3). The jumping and linking libraries are used to create contigs. A *contig* comprises a set of clones whose DNA inserts overlap each other and they together cover a specific genomic region. Generally, YAC (yeast artificial chromosomes) and cosmid clones are used for the construction of contigs. The first step in the physical mapping of a chromosome is to prepare YAC clones from this chromosome using, for example, the restriction enzyme *Not*I for digesting the chromosomal DNA. In order to create a contig of the YAC clones, jumping and linking libraries are made from DNA of the same chromosome. As a rule, every clone of a linking library will hybridize with two separate clones of the corresponding jumping library, but it will hybridize with only one YAC clone. Similarly, each clone of the jumping library will hybridize with two distinct clones of the corresponding linking library, as well as with two separate YAC clones (Fig. 12.4). All the YAC clones generated from the chromosome of interest are separately hybridized with every clone of the linking as well as the jumping libraries. The data acquired from all the hybridization assays are pooled and analyzed together to find out the order in which the DNA inserts of the YAC clones would be located in the target chromosome. This information is used to create the contig representing the target chromosome. Thus, the main use of chromosome jumping and chromosome walking techniques is to create contigs representing entire chromosomes and relatively large (~50 kb) DNA molecules, respectively.

12.8 Positional Gene Cloning

A clear-cut determination of function of a given gene involves cloning and isolation of the gene, introduction of this gene into a suitable host by genetic transformation, and analysis of the phenotypic effects produced by the gene in the resulting transgenics. Transgenic plants are not only useful in basic studies on gene function and plant development, they also serve as improved varieties with novel and valuable traits like insect resistance, improved quality, etc. (Sect. 1.5). Mutagenesis can be used to induce a large number of mutations, each of which produces an easily detectable change in phenotype. Ordinarily, neither the base sequence nor the protein product of the genes involved in these mutational changes will be known, and the mutated genes are detectable only by the specific phenotypic effects produced by them. The gene responsible for a mutant phenotype can be easily isolated if it were produced by *insertional mutagenesis*, i.e., induction of mutations by insertion of either a transposable element or *Agrobacterium* T-DNA sequence into the concerned gene. The gene into which these sequences become integrated is unable to function normally, and a loss of function mutant allele is produced. The genomic region having this mutant allele can be readily isolated by using the transposable element/*Agrobacterium* T-DNA sequence as probe for screening the genomic library (Appendix 2.2) produced from the concerned mutant. This approach is limited to such species, for which suitable transposon constructs or *Agrobacterium*-mediated transformation protocols are available. Alternatively, positional cloning can be used to identify and isolate the gene involved in a mutation. There are two requirements for positional cloning as follows: (1) individuals of a population should differ for the trait in question, and (2) at least one DNA marker closely linked to the gene responsible for these differences in the target trait should be known. It is highly desirable that two markers located as close to the gene as possible and on either side of the target gene (*flanking markers*) are known. These requirements can be fulfilled for most of the traits, including those governed by multiple genes.

12.8.1 The Three Steps of Positional Cloning

Positional cloning or map-based cloning involves the following three steps: (1) isolation of a mutant strain with a recognizable change in phenotype and mapping of the mutant allele within a short genomic region to identify a pair of markers flanking the mutant allele, (2) using these markers for the identification and isolation of the DNA fragment containing the mutant allele, and finally (3) determining the function of the concerned gene (Meyer et al. 1996). Once a mutant allele of interest is identified, linkage mapping using a large mapping population allows the identification of molecular markers showing linkage with the mutated gene (Fig. 12.5). It is desirable that a pair of flanking markers located at <1 cM from the gene of interest is identified. This constitutes *the first step of positional cloning*. The genomic DNA is then isolated from the concerned mutant. A high-capacity vector like YAC, BAC (bacterial artificial chromosome), or PAC (P_1-derived artificial chromosome) vector is then used for creating a genomic library from this DNA. A type of BAC vectors, called transformation-competent artificial chromosome (TAC) vectors, has elements of the *Agrobacterium* T-DNA. As a result, TAC vectors facilitate positional gene cloning and library screening. This genomic library is screened with the marker closest to the mutant gene to identify the clones that have DNA inserts with this marker. The library screening either involves hybridization with labeled probes or is PCR-based. The identified DNA insert forms the starting point for chromosome walking (Sect. 12.6) till the clone having DNA insert with the marker located on the other side of the mutant gene is reached. This procedure identifies a set of DNA inserts that together form a contig of overlapping fragments that spans the genomic region known to have the desired mutant allele,



Linkage mapping of the target gene; marker A is closer to the gene

YAC clones of a genomic library screened with the marker A; chromosome walk from this point creates a contig of YAC clones

Markers derived from the contig clones used for linkage mapping; this identifies a shorter genomic region having the target gene

Cosmid clones from the genomic region between markers D and E; contig created. Fragments 1-4 used for genetic transformation; clones 2 and 3 complement the mutant phenotype

Identification of the gene present in fragment 2 and determination of its function

1. **Screening of cDNA library; cDNA clones used for high resolution mapping**
2. **Sequencing of cDNAs and the fragment 2**
3. **Expression pattern of the gene correlated with phenotype**
4. **Mutant lines probed with cDNAs to identify the genes showing mutational changes**
5. **Candidate genes present in fragment 2 used separately for complementation test**

Fig. 12.5 A generalized procedure for positional cloning in plants

i.e., *candidate genomic region*. This completes the *second step of positional cloning*.

Effort should be made to minimize the size of candidate genomic region; this may be achieved as follows. A large mapping population can be screened with new markers developed from the end segments of the DNA fragments making up the contig developed in the second step. The population size needed for this purpose is proportional to the estimated physical distance in kb per cM in the target region; roughly, it is about three times as large as this estimate. The minimum population size for maize and *A. thaliana* has been estimated as 1,000. Linkage mapping using the new markers allows their placement on the linkage map relative to the known markers flanking the gene. This step not only identifies new markers located closer to the target gene, but also ensures that the walk progresses in the correct direction. The fragments making up the contig representing the new relatively smaller candidate region are used for complementation test to identify the fragment(s) having the desired gene. In

complementation test, the candidate region from the nonmutant plant is introduced into the mutant line by genetic transformation, and the phenotype of concerned trait of the transgenic plants is evaluated. *If the transgenic plants have the wild type or nonmutant phenotype*, *the candidate genomic region is said to complement the mutant allele and is concluded to have the nonmutant allele of this gene*. The fragment identified by complementation test is cut into smaller fragments, which are cloned in a suitable cosmid/phage vector; this is called *subcloning*. The DNA inserts of these clones are analyzed to prepare a contig spanning the genomic region represented by the above fragment. Finally, the DNA fragments comprising this contig are evaluated in complementation test to identify the fragment(s) that has the target gene.

The *third step of positional cloning* consists of identification of the mutated gene and confirmation of its role in the mutant phenotype. The candidate genomic region remaining at the end of the second step may sometimes contain a single gene; this would clearly establish the role of this gene in the mutant trait. For example, maize gene *ra3* was narrowed down to a candidate genomic region of 6 kb or 0.2 cM by linkage mapping of additional markers derived from BAC clones in a population having 1,700 mutant plants. This candidate region was sequenced, and it was found to contain a single predicted gene; obviously this gene is responsible for the *ra3* phenotype (Bortiri et al. 2006). But in most cases, the candidate genomic region would be several kilo bases long and would contain more than one gene. For example, a 40 kb candidate region may contain >3 genes. In such a case, a suitable strategy will have to be used to identify the gene of interest. In order to achieve this, cDNA libraries are screened with the candidate genomic region and all the cDNAs hybridizing with it are identified. A suitably sensitive method for detection of the cDNAs should be used to ensure that all the cDNAs encoded by the candidate genomic region are detected. However, in case one of the candidate genes encodes a low abundance RNA, this gene may not be represented in the cDNA library. Identification of the particular cDNA that corresponds to the target gene can be achieved in one of the following seven ways:

1. The cDNA clones are used as markers for high-resolution mapping in order to identify the cDNA clone that always co-segregates with the concerned mutant phenotype.
2. The expression patterns of the candidate genes are analyzed to identify the gene whose expression pattern is consistent with the expression of the concerned phenotype; e.g., the gene is transcribed in the appropriate tissues, during the expected developmental stage and/or induced by the appropriate stimuli.
3. Several independent mutant lines with the same mutant phenotype can be probed with the cDNA clones to isolate and sequence the positive clones. The gene and/or the mRNA that shows change(s) in base sequence in every mutant line in comparison to its sequence in the wild type strain is identified. This approach is particularly useful in such species, where transgenics cannot be produced. For example, analysis of the mRNA from several lines with different alleles of the locus helped identification of the *ts4* gene of maize. But in some cases, at least, several independent mutant alleles of a gene may not be available.
4. The DNA sequences of the genes present in the candidate region can be used for homology search in the databases. This analysis may permit the identification of homologous genes with known functions, which might help determine the gene involved in the concerned mutation.
5. The candidate region may be compared with the corresponding chromosomal region of related species that are known to have syntenic relationships; e.g., maize and rice genomes have considerable synteny. This comparison may allow identification of a gene that may be involved in a similar mutation/function.

The above approaches are useful in such cases where production of transgenic plants is not feasible; these approaches are often termed as the *candidate gene approach*.

6. Each of the candidate genes from the wild type may be introduced separately into the mutant line by genetic transformation. The transgenic plants expressing a candidate gene would show the wild type phenotype only if this gene were involved in the concerned mutation. However, this approach is applicable to recessive mutations only. In case of dominant and partially dominant mutations, the candidate genes from the mutant lines can be introduced into the wild-type strains. In this case, the candidate gene involved in the mutation would generate the mutant phenotype in the transgenic plants. For example, the *ABI1* gene of *A. thaliana* was identified by this approach (Meyer et al. 1996). *The most direct and conclusive proof for the function of concerned gene is provided by this approach.*
7. When it is either not desirable or feasible to test for complementation, the expression of a candidate gene in the wild-type strain may be blocked by a suitable strategy, e.g., antisense RNA technology, RNA interference, etc. In case a given candidate gene were involved in the mutant phenotype, the concerned transgenic plants would show the mutant phenotype.

12.8.2 Positional Cloning of Some Plant Genes

The mutant allele *ABI1*, of the gene *abi1* of *A. thaliana*, is dominant and specifies insensitivity to abscisic acid. The *ABI1* gene was mapped, using a population of 300 F_2 plants, between two RFLP markers located at 0.16 cM (marker B) and 11.3 cM (marker C) from the gene (Fig. 12.6). The marker B was used to screen YAC genomic libraries of *A. thaliana*, including a library of the mutant *ABI1*. The YAC clones identified in this way were used to initiate chromosome walk to create contigs covering the *ABI1* region. The end sequence of DNA insert of each YAC clone comprising the contig was used as a marker for linkage mapping using the F_3 generation derived from the above F_2 population. A new marker E together with marker B was found to define the 150 kb candidate genomic region having the *ABI1* locus. The YAC clone representing the candidate genomic region was subcloned in the binary cosmid vector pBIC20 (this vector can be used for *Agrobacterium*-mediated genetic transformation), and the DNA inserts from the cosmid clones were aligned into a contig. The members of this contig developed from the mutant line were used for genetic transformation of the wild-type plants. Two DNA fragments of the contig produced ABA insensitivity in the transgenic plants. A comparison of these two fragments with the other contig fragments helped localize the *ABI1* locus to a short region common to the two fragments. A comparison of the sequence of this short region from the wild type and the mutant lines identified a single point mutation in the *ABI1* gene to be involved in the mutant phenotype (Meyer et al. 1996).

The pollen fertility in BT CMS lines of *O. sativa* var. *japonica* is restored by the nuclear gene *Rf-1*. Linkage mapping identified two markers flanking the *Rf-1* gene. One of these markers was located at 0.1 cM (= ~28 kb DNA) from *Rf-1*, and chromosome walk was started from this marker. A segregating population of 4,103 plants was screened with these two markers to identify plants that showed recombination between *Rf-1* and one of the markers. A contig of λ clones spanning the *Rf-1* genomic region between the two markers was created, and the DNA inserts from these clones were used to develop new markers. The recombinant plants identified as above were screened with these new markers to fine-map the *Rf-1* region. The contig clones spanning the *Rf-1* region were used to produce transgenic plants for complementation test. One 15.6 kb fragment out of the 12 fragments in the contig sponsored variable restoration of fertility. Three probes were produced from this fragment and used to screen a cDNA library to identify cDNA clones derived from the *Rf-1* region. The sequences of these cDNA clones were compared with that of the above 15.6 kb genomic fragment; this enabled the determination of *Rf-1* gene structure. This

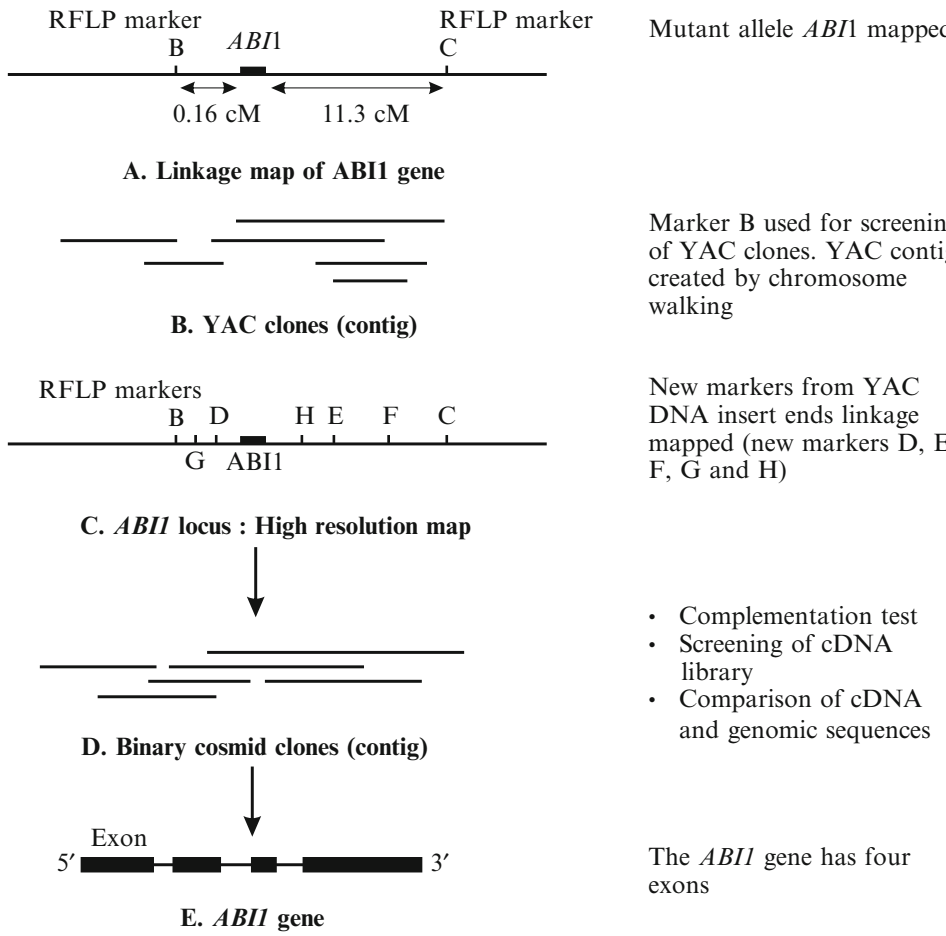


Fig. 12.6 Schematic depiction of positional cloning of *A. thaliana ABI1* gene (Simplified from Meyer et al. 1996)

gene was predicted to encode a protein having 16 tandemly repeated copies of the pentatricopeptide repeat (PPR) motif, which comprises 35-amino acid residues. Therefore, the protein encoded by this gene was designated as PPR791 (Komori et al. 2004).

12.8.3 Some Useful Tips for Positional Gene Cloning

It is important that the walk progresses in the same direction; this can be ensured by the following approaches: (1) use of the same genomic library for successive screenings, (2) construction of the genomic library using a suitable vector like λDASHII and λFIXII, (3) using only such probes in chromosome walk that exclusively contain unique sequences, and (4) simultaneous use of the probes for chromosome walk and in situ hybridization with polytene chromosomes, if available. In case of plants, the probes can be used as new markers to screen a large mapping population. This would not only help maintain the direction of walk, but also define a smaller candidate genomic region for further analysis. The technique of positional cloning is simple in principle but is technically demanding. In addition, what seems to be tight linkage, e.g., 1.0 cM genetic distance, may actually represent up to 1–1.5 Mb DNA. Therefore, it is desirable that the DNA markers linked to the target gene be situated at <1.0 cM from the target gene. This necessitates additional experiments for fine

mapping (Sect. 6.13) of the target gene. However, in many cases a tightly linked marker may not be available. In such cases, long-distance walks have to be performed; these walks should be performed with genomic libraries constructed in vectors like BACs and YACs that can accommodate very large DNA inserts. This would reduce remarkably the number of steps required for reaching the target gene. Before these vectors became available, scientists resorted to ingenious strategies like using genomic libraries constructed from special cytogenetic stocks like those having inversions and translocations. The inversions and translocation bring together distant regions of chromosomes next to each other. Such strains are characterized in detail to identify the distant genomic regions brought adjacent to each other. Such stocks serve the same purpose as the chromosome jumping technique (Sect. 12.7), which was designed to facilitate long-distance jumps.

12.8.4 Problems in Positional Cloning

Map-based cloning presents certain practical difficulties. (1) The eukaryotic genomes are enormous in size. Therefore, preparation of saturated linkage maps of molecular markers is a difficult and expensive task. For example, the genome of wheat is ~16 billion base pairs. Even if 16,000 markers were equally spaced in such a genome, the physical distance between two neighboring markers will still be ~1,000 kb. As a result, it may often be difficult to find a marker located close enough to the gene of interest. For this reason, a second round of mapping using a suitably large mapping population and new markers derived from the contig clones is often taken up to reduce the size of the candidate genomic region. (2) Further, the correspondence between one map unit (cM) estimated from recombination data and the physical distance determined as base pairs between two genes is affected by several factors. For example, the genetic distance of 1 cM may correspond to merely 140 kb in *A. thaliana* through 750 kb in tomato to 4,600 kb in wheat (Table 6.1). In addition, there is considerable variation in the physical distance per centimorgan in different genomic regions of the same species. For example, the physical distance corresponding to 1 cM in the different regions of tomato genome may vary by up to 100-fold. Walking such large distances is now more practicable than before due to the availability of high-capacity vectors like BACs, PACs, and YACs. (3) Chromosome walk is confused by the highly repetitive DNAs dispersed throughout the eukaryotic genomes. This problem can be resolved by using suitable strategies, including chromosome jumping. In addition, (4) chromosome walks are time consuming, (5) it is a demanding and time taking task to identify the DNA insert having the gene of interest from among those making up the contig created by chromosome walks, and (6) some of the large insert libraries like YAC libraries contain chimeric or modified DNA inserts.

12.9 Chromosome Landing

The chromosome landing strategy is a variant of positional cloning. In this approach, high-resolution mapping is used to identify one or more markers located very close to the target gene. The average size of the DNA inserts in the genomic library used for chromosome landing should be greater than the physical distance separating the marker and the gene of interest. For a YAC library the marker should be located within 200 kb of the target gene (average insert size assumed as 200 kb). Such a marker would allow direct identification of the clone with the target gene in a single step, and for this reason, the term *chromosome landing* is used to denote this procedure (Tanksley et al. 1995). This approach is greatly facilitated by the availability of high-density marker linkage map of the concerned species. More importantly, high-resolution mapping (Sect. 6.13) can be used to identify markers located very close (at 0.1 cM or less) to the gene of interest. For example, the tomato gene *Pto* specifies resistance to *Pseudomonas syringae* pv. *tomato*. Tanksley et al. (1995)

identified 18 markers showing close linkage to the *Pto* gene. An F_2 population of 1,200 plants was analyzed to identify one DNA marker, which co-segregated with the *Pto* gene. The use of this molecular marker enabled the identification of such YAC clones that had DNA inserts with this marker and, presumably, the genomic region containing the gene *Pto*. The end sequences of the YAC clones were used as markers for linkage mapping using the above mapping population. This allowed the identification of one YAC clone that spanned the *Pto* locus. The DNA insert from this clone was isolated and radioactively labeled. The labeled DNA insert was used as probe for screening of a cDNA library. This screening classified the cDNA clones into two classes. High-resolution linkage mapping using these cDNAs as markers showed that one of the two cDNA classes always co-segregated with the *Pto* gene. The *Agrobacterium*-mediated gene transfer technique was used to introduce this candidate cDNA into susceptible tomato plants. The transgenic tomato plants produced in this way were resistant to bacterial pathogen *P. syringae* pv. *tomato*.

It would be seen that chromosome landing requires markers tightly linked to the gene of interest. In addition, a high-resolution linkage map needs to be constructed to identify the candidate cDNAs quickly and effectively. This approach can be used for cloning of genes, including QTLs, in most sexually reproducing plant species.

12.10 Positional Cloning of Quantitative Trait Loci

Polygenes governing quantitative traits were finally given physical location as QTLs by linkage mapping using molecular markers. However, it could not be clarified whether a single QTL represented one or more than one gene. Further, the identity and the functions performed by the genes located in the QTL regions also remained to be elucidated. These questions can be answered by precisely mapping the QTLs, followed by cloning and characterization of the genes present in the QTLs.

1. The QTLs are ordinarily mapped within a confidence interval of 10–20 cM, which is too large for positional cloning. Therefore, QTLs have to be fine-mapped within short genomic regions with the help of a suitable approach, e.g., by using backcross-derived lines (BILs: Sect. 5.12).
2. Once a QTL has been fine-mapped, the genes present in the QTL region can be identified by either chromosome walking or chromosome landing. A large segregating population can be analyzed with chromosome-region-specific markers to identify the candidate genomic region for further analysis. The candidate genomic region should, preferably, be of ~50 kb. Chromosome region-specific markers can be obtained as follows: ESTs identified to hybridize with the concerned genomic region may themselves serve as markers, or sequence (where available) of the concerned genomic region may be used to develop suitable markers.
3. The candidate genomic region may be sequenced, followed by gene prediction (Sect. 12.5), to identify the candidate genes. Alternatively, gene expression profiles may be compared with specific phenomena in plant growth and development, e.g., response to an abiotic or biotic stress, for the identification of candidate genes. The genes showing altered expression patterns in response to the concerned phenomenon and are co-localized with the concerned QTL(s) offer themselves as candidate genes for further analysis (Sect. 7.11.7).
4. The candidate genes may be evaluated in one of several ways to identify the gene involved in QTL function (Sect. 12.8.1). The findings from this evaluation are confirmed by genetic complementation analysis. But in case of some quantitative traits, a line expressing the null allele, i.e., the allele showing complete-loss-of-function, may not be obtainable. In the case of such traits, the following two approaches may be used: (1) A sense construct of the gene may be expressed in a line

with low phenotypic mean value for the target trait. But sometimes this may produce unexpected effects on plant phenotype, making it difficult to confirm the candidate gene function. (2) Alternatively, an antisense construct of the candidate gene or RNA interference (RNAi) can be used to reduce the function of the target gene. The effects of reduced gene expression on the target function or phenotype are critically evaluated. But sometimes the suppression of gene expression may not be stable, making the conclusions less than reliable.

The positional cloning of major photoperiod sensitivity QTL, *Heading-date1* (*Hd1*), of rice involved high-resolution mapping in a population of 1,505 plants. This enabled direct landing onto the candidate genomic region of 150 kb present in a single PAC clone. This DNA insert was sequenced, and the sequence was used to develop cleaved amplified polymorphic sequence (CAPS) markers. These CAPS markers were used for linkage mapping, which reduced the candidate genomic region to merely 12 kb predicted to contain two genes. One of these genes exhibited high similarity to the *A. thaliana* photoperiod response gene *CONSTANS*. Complementation analysis confirmed the function of *Hd1*. Further, base sequence analysis of *Hd1* revealed it to be allelic to the *Se1* gene of rice, which is the major gene governing photoperiod sensitivity in rice (Yano 2001).

Linkage mapping with a very large mapping population located the wheat (*Triticum monococcum*) *VRN1* gene in the candidate genomic region of 0.03 cM, representing ~324 kb DNA. This candidate genomic region contained two genes, *AP1* and *AGLG1*; this was comparable with situations for the syntenic regions of rice and sorghum. The *AP1* and *AGLG1* genes of wheat are similar to the *AP1* and *AGL2* genes, respectively, of *Arabidopsis*. The *Arabidopsis* genes are known to be involved in the determination of meristem identity. Vernalization regulated the transcription of gene *AP1* in wheat shoot apices as well as leaves. Further, the level of transcription of *AP1* was positively correlated with the degree of effect exerted by vernalization on the flowering time. In contrast, *AGLG1* was transcribed during differentiation of spikes, but its transcription was not detectable in shoot apices or leaves subjected to vernalization. These observations suggest that *AP1* is expressed before *AGLG1*, and the expression of *AP1* leads to the expression of *AGLG1* in response to vernalization. Finally, in the wheat genotypes having different alleles of *VRN1* gene, the promoter regions of their *AP1* genes had three independent deletions. Thus, gene *AP1* seems to be a much stronger candidate for the gene *VRN1* than the gene *AGLG1*. However, the role of gene *AP1* in vernalization needs to be confirmed by producing transgenic plants (Yan et al. 2003).

12.11 cDNA Sequencing in Positional Cloning

The sequencing of full-length cDNA is quite useful in positional cloning of genes in the following two ways: (1) The analysis of full-length cDNA clone sequence corresponding to a candidate gene permits the deduction of amino acid sequence of the encoded protein. This information can be used for homology search of protein databases to deduce the putative function of the concerned candidate gene. (2) The analysis of sequences of full-length cDNAs is the most reliable and accurate method for determining the exon–intron organizations of the concerned genes. For example, the sequence of nearly full-length cDNA clone corresponding to the *ABI3* gene of *A. thaliana* was analyzed to determine the amino acid sequence of the protein encoded by this cDNA. This protein was found to be truncated by 40 % in the case of *abi3-4* allele, which produces the most severe phenotypic effect. This truncation was the result of a point mutation, which generated a chain termination codon within the coding region. The complete cDNA sequencing can be achieved by primer walking, concatenated cDNA sequencing, and shotgun sequencing. The first step in *primer walking* consists of sequencing of the two ends

of a complete cDNA clone using the Sanger method. After this, a series of primers are synthesized from the sequences so generated, and used for sequencing. Primer walking is used for sequencing of larger clones, but is problematic, costly, and time taking.

Concatenated cDNA sequencing involves isolation of multiple cDNA clones. The cDNAs are pooled, enzymatically concatenated, and shotgun sequenced. The sequence reads are assembled and then analyzed using a suitable computer program to deduce individual cDNA sequences. However, this approach presents several difficulties. Alternatively, a mixture of several cDNA clones may be subjected to shotgun sequencing by the Sanger method, and the sequence reads may be assembled into individual cDNA sequences. A standard reference genome sequence is generally used for aligning the short reads to avoid ambiguity in contig/scaffold production (*reference assembly*). In the case of reference assembly, differences in the reference genome and the cDNA sequences due to InDels and mutations, and the presence of exons shorter than 36 bp interfere with the correct assembly of the short reads. It may be pointed out that about 14 % of the human cDNA clones were found to have one or more exons that were less than 36 bp in size. The short reads can also be assembled de novo without the help of a reference genome (*de novo assembly*). But de novo assembly would require the reads to be distributed more uniformly than in the case of reference assembly, or else the sequencing coverage should be higher. A new hybrid assembly approach integrates the features of both de novo and reference assembly strategies (Kuroshu et al. 2010).

Some computer programs have been developed to facilitate full-length cDNA sequencing. The program GeneMark is used to analyze cDNA sequences to detect putative frameshift sequence errors (Hirosawa et al. 2000). This analysis reliably detects such cDNA clones whose coding regions have been disrupted due to the various types of spurious splits. The Multiclone Shotgun Integrated cDNA Assembler 1 (MuSICA 1) is a reference genome assembly algorithm. This assembler needs a reference genome to align the shotgun reads against this sequence. After this, it assembles these reads into contigs. Similarly, MuSICA 2 assembles the short reads generated by an Illumina GA sequencer into the full-length cDNA sequences. The MuSICA 2 uses an assembly algorithm that combines the features of both reference and de novo assembly algorithms, and performs better than either of these two. In this program, an external de novo assembler like Velvet is used to de novo assemble the short reads into contigs. These de novo contigs are then improved by using the reference assembly approach. Either one or both ends of the cDNA clones are sequenced by the Sanger–Coulson method, and these sequences are used to associate each finalized contig with a specific full-length cDNA clone (Kuroshu et al. 2010).

12.12 Achievements

The DNA markers can discriminate between even such lines/varieties that are closely related and cannot be easily differentiated from each other using data on morphological traits and isozymes. In addition, the theoretical bases for the use of marker data for this purpose are also being elucidated. However, the acceptance of marker profile as the sole basis of variety registration for IPR protection has been disappointingly slow most likely due to the implications of the provisions of PBR laws. It is encouraging that in some countries marker data are accepted as additional information for variety registration. The use of marker profile for assessment of the genetic purity of seed crops and seed lots is being extensively investigated. It may be hoped that with the development of simpler and less expensive marker profiling procedures, this approach would become the method of choice for assessing the genetic purity of seed lots. There are several successful examples of identification and cloning of mutant genes, including QTLs, with the help of map-based cloning. Some of these examples are listed in Tables 12.3 and 12.4, and a couple of them have been described in the preceding sections.

Table 12.3 Some examples of positional cloning of oligogenes

Crop/species	Trait	Gene	Function/remarks
Rice	Fertility restoration	*Rf-1*	Restores pollen fertility in BT CMS lines
	Lax panicle	*lax panicle*	Encodes a basic helix–loop–helix protein
Maize	Barren stalk 1; tassel branches and spikelets, and ears absent	*ba1*	Synteny with rice *lax panicle* provided a candidate gene
	Thick tassel dwarf	*td1*	Orthologous to Arabidopsis gene *clv*
	Ramosa 1	*ra1*	Highly branched tassel and branched ear; codes for a transcription factor
	Inflorescence	*ra2*	Orthologous to *LOB* gene of rice
	Tassel morphology	*ra3*	Markers derived from BACs placed *ra3* in to a 6 kb region, which contained a single predicted gene
	Seeds produced in tassel	*ts4*	Analysis of RNA levels in multiple alleles helped find the correct gene
Arabidopsis	Insensitivity to ABA	*ABI1*	Dominant, pleiotropic
	Insensitivity to ABA	*abi3*	Encodes a protein similar to maize Viviparous-1 protein
Tomato	Resistance to *Pseudomonas syringae* pv. *tomato*	*Pto*	First gene to be cloned, in 1993, by positional cloning; encodes a protein kinase
Medicago sativa	Nodulation	*NORK*	NBS-LRR-receptor kinase
Glycine max	Nodulation; nodule autoregulation	*NARK*	LRR-receptor kinase
	Disease resistance	*Pst1*	–
Lotus japonicus	Hyper-nodulation	*HAR1*	LRR-receptor kinase
	Nodulation	*SymRK*	NBS-LRR-receptor kinase
	Nodulation	*NFR1*	LysM-receptor kinase
	Nodulation	*NFR5*	LysM-receptor kinase
Medicago truncatula	Nodulation; early infection	*LYK3*	LysM-receptor kinase; NFR1 like

Based on Gresshoff (2005), Bortiri et al. (2006), and other sources

Table 12.4 Some examples of genes present at QTLs that have been cloned and characterized

Crop	QTL for the trait	Gene	Function/remarks
Tomato	Soluble solids content of fruit	*Brix9-2-2*	Encodes apoplastic invertase
	Fruit size	*fw2.2*	Similar to human oncogene *c-H-ras p21*
Rice	Photoperiod sensitivity	*Hd1*	Similar to *A. thaliana* photoperiod response gene *CONSTANS*; allelic to the rice photoperiod-sensitivity gene *Se1*
	Photoperiod sensitivity	*Hd6*	Encodes a casein kinase II -subunit
	Grain number	*Gn1*	Encodes cytokinin oxidase/dehydrogenase
Maize	Teosinte glume architecture	*tga1*	First gene to be cloned in maize by positional cloning; affects hardness of seed coat
	Teosinte branched 1	*tb1*	Cloned by transposon tagging
Wheat	Vernalization	*VRN1*	Most likely due to mutations in the promoter region of gene *AP1*
	Vernalization	*VRN2*	Most likely due to mutations in the gene *ZCCT1*; confirmed by RNA interference

Based on Yano (2001), Bortiri et al. 2006, and other sources

Questions

1. Discuss the relevance of molecular markers in the implementation of plant breeder's rights.
2. Discuss the use of molecular markers in monitoring of genetic purity of seed lots of plant varieties, hybrids and the parental lines of hybrids.
3. Explain the three main steps of positional cloning of plant genes. Briefly describe the cloning of one or two plant genes.
4. Discuss the positional cloning of plant quantitative trait loci.
5. Briefly describe the technique of chromosome walking and discuss its relevance to positional cloning of genes.
6. "Chromosome landing is a specialized method of chromosome walking". Comment on this statement giving full justification.
7. Discuss the relevance of chromosome jumping and cDNA sequencing in positional cloning of genes and QTLs.

High-Throughput SNP Genotyping 13

13.1 Introduction

The development of next-generation sequencing (NGS) methods has helped the emergence of single nucleotide polymorphism (SNP) as the marker of choice. The SNPs are becoming increasingly popular in view of their abundance, ease in discovery, and the extremely high-throughput SNP genotyping at relatively low cost per data point. In simple terms, *throughput* means the number of assays, e.g., SNP genotyping, carried out by an assay system in a unit time. *High-throughput genotyping* may be defined as simultaneous genotyping for few to several hundreds or thousands of markers in hundreds to thousands of individuals. A variety of SNP genotyping strategies have been developed, some of which have been described in Chap. 4. The methods described in that chapter are applicable to already identified SNP loci with known sequences of their flanking regions. In any case, these methods can simultaneously genotype a small number of SNP loci usually in a small number of samples. But the chief advantage of these methods is that they can be used in small to moderate size laboratories without expensive instrumentation. This chapter deals with techniques for high to very-high-throughput SNP genotyping. These methods require moderate to considerable sophistication and expensive instrumentation, and most of them have been commercialized. Therefore, most of them are closed technologies available to the user through the genotyping systems offered by the concerned companies. The various methods for high-throughput SNP genotyping can be grouped into two broad categories: (1) genotyping of already identified SNP loci and (2) simultaneous identification and genotyping of SNP loci.

13.2 High-Throughput Genotyping of Known SNP Loci

The methods commonly used for high-throughput genotyping of already identified SNPs are as follows: (1) invader technology, (2) pyrosequencing, (3) KASP™ genotyping assay, (4) TaqMan OpenArray genotyping system, (5) SNP analysis by MALDI-TOF MS, (6) nanofluidic dynamic arrays, (7) array tape technology, (8) Illumina GoldenGate platform, (9) molecular inversion probe (MIP) technology, and (10) whole-genome microarray platforms. These technologies differ in the level of multiplexing permitted by the system and can be grouped on this basis as follows. *Single-plex or low multiplex* SNP genotyping platforms include invader technology, pyrosequencing, and KASP™ genotyping assay. The TaqMan OpenArray genotyping system, SNP analysis by MALDI-TOF MS, nanofluidic dynamic arrays, and array tape technology permit *moderate multiplexing*. *High-level multiplexing* is achieved with the Illumina GoldenGate and Affymetrix's targeted whole-genome genotyping (MIP) platforms. The

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_13, © Author(s) 2015

whole-genome-based array platforms include Illumina Infinium HD assay, Affymetrix's GeneChip and Axiom® genome-wide arrays, Agilent SurePrint arrays, and Beckman Coulter's SNPstream genotyping system. Most of these techniques use PCR for amplification of the target genomic regions. This step increases genotyping cost since PCR amplification is relatively expensive and is not amenable to automation.

13.2.1 The Invader Technology

The single-stranded 3′ end of a DNA molecule can invade a homologous DNA duplex and pair with its complementary strand. In this reaction, the strand having the same sequence as the invading strand is displaced from the homologous duplex, and a specific *invasive nucleic acid structure* is formed (Fig. 13.1). The invader technology (Lyamichev et al. 1999) gets its name from this phenomenon as it uses the formation of invasive nucleic acid structure for detection of SNPs, InDels, etc. This technology exploits the ability of certain enzymes to specifically recognize the invasive nucleic acid structure formed by two oligonucleotides (oligos) when they pair with the concerned target DNA strand/RNA molecule and cleave one of these oligos at a specific site (Fig. 13.1). The cleavage enzymes used for the DNA-based assay are usually derived from the family of flap endonuclease (FEN-1) of the thermophilic archaebacteria, while RNA-based assays use the 5′-exonuclease domain of DNA polymerase I obtained from thermophilic eubacteria (Kahl et al. 2005).

The invader technology is based on three oligonucleotides, namely, an invader oligo, a primary probe, and a FRET (fluorescence resonance energy transfer) oligo (Fig. 13.1b–d; Kahl et al. 2005). Let us consider these oligos in the context of DNA target-based detection of allele *A* at a given SNP locus (Fig. 13.1a). The *invader oligo* is complementary to the target sequence on the 3′ side of this SNP locus and includes the base T, which is complementary to the SNP allele *A* at its 3′ end. Therefore, the 3′ terminal base, termed as *position 1*, of this oligo is complementary to the SNP allele it is designed to detect (Fig. 13.1b), and this base overlaps the target-specific region (TSR) of the primary probe. The *primary probe oligo* has two regions, the 3′ TSR and the 5′ flap region. The TSR is complementary to the target sequence located on the 5′ side of the SNP locus and includes the SNP locus, while the flap region has a universal sequence (Fig. 13.1c). The *FRET oligo* also has two regions: its 3′ region is complementary to the flap region of the primary probe oligo, while its 5′ region forms a hairpin loop. The 5′ terminal nucleotide of the 5′ region overlaps the 3′ terminal base of the flap region of primary probe oligo and has a fluorophore attached to it. In addition, a quenching dye molecule is attached to a neighboring nucleotide on the 3′ side of the cleavage site (Fig. 13.1d), as a result of which the FRET oligo does not generate fluorescence as long as it retains its original structure.

For detection of the SNP allele, the target DNA is denatured and invader and primary probe oligos are allowed to anneal to it. In case the 3′ terminal nucleotide of the invader oligo is complementary to the SNP allele present in the target DNA strand, it will pair with the SNP allele by displacing the TSR segment of the primary probe oligo. As a result, an invasive nucleic acid structure very similar to a replication fork will be formed (Fig. 13.1e). The cleavage enzyme will recognize this structure and cleave the primary probe oligo immediately on the 3′ side of its base corresponding to the SNP locus. This cleavage releases the flap region of the primary probe oligo, which now pairs with the 3′ region of the FRET oligo. The released flap region also functions as an invader oligo and displaces the 5′ terminal base of the FRET oligo forming an invasive structure (Fig. 13.1g). The cleavage enzyme now cleaves the 5′ terminal base of the FRET oligo. The terminal base with its attached fluorophore now moves away from the quencher dye (Fig. 13.1g). This fluorophore is, therefore, able to generate fluorescence. After incubating the sample for 4 h at 63 °C, a standard fluorescence plate reader is used to detect this fluorescence. Primary probe oligos specific for both wild type and mutant alleles are evaluated

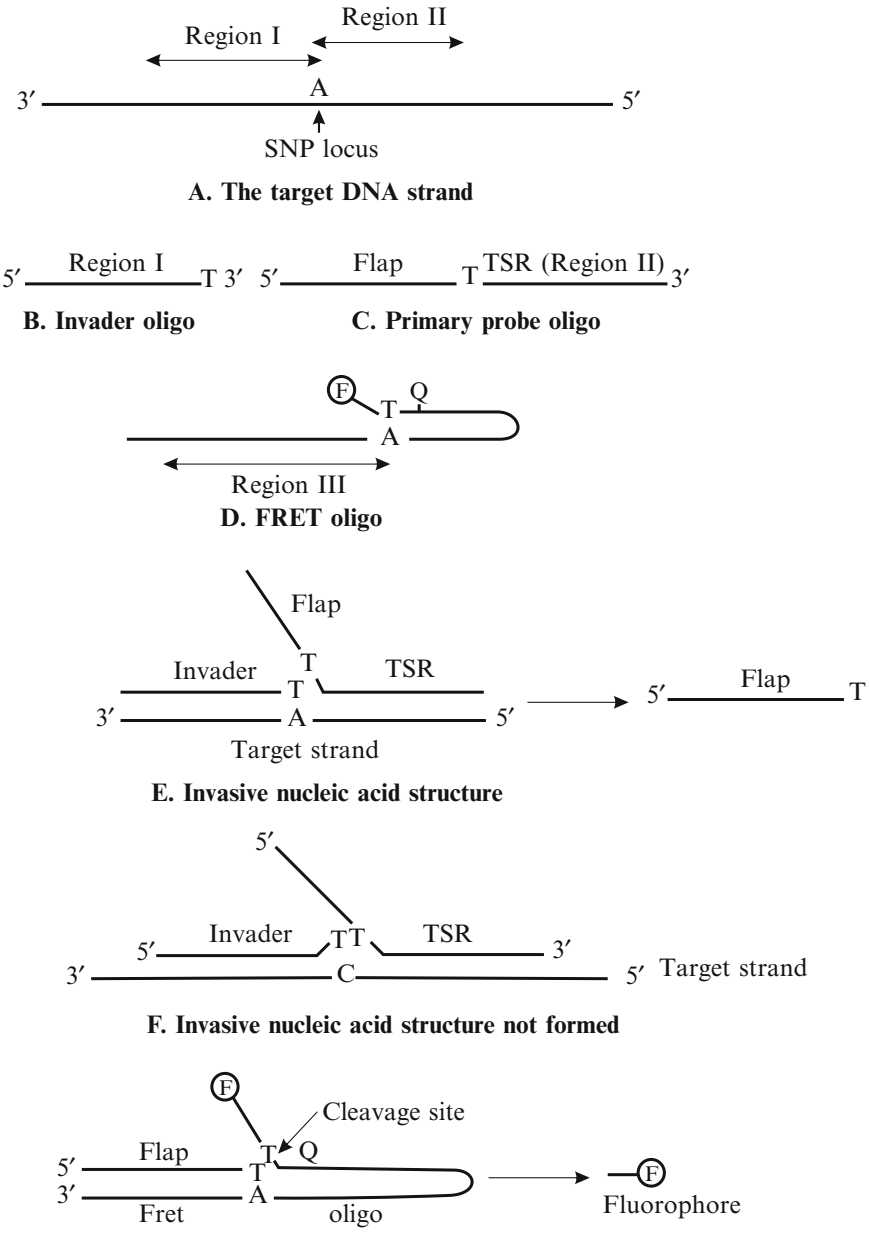


Fig. 13.1 A schematic representation of the invader technology. The region III in FRET oligo is complementary to the flap region of primary probe oligo. Invasive nucleic acid structure is formed only when the 3′ terminal base of the invader oligo is complementary to the SNP allele. *TSR* target-specific region, *FRET* fluorescence resonance energy transfer (Based on Kahl et al. 2005)

for each SNP locus. The SNP allele present in the sample DNA is determined from the ratio of signals generated by the two probe oligos for each SNP locus. In case the 3′ terminal base of the invader oligo is not complementary to the SNP allele present in the target DNA strand, the terminal base will not pair with the target strand. At the same time, the "position 1" base of TSR

segment of the primary probe oligo will also not pair with the target strand, and the invasive structure will not be formed because a gap will be present at the SNP locus (Fig. 13.1f). Therefore, cleavage of the flap segment of primary probe oligo will not take place, and there will be no fluorescence (Kahl et al. 2005).

The TSR of primary probe oligo is so designed, and its concentration and the reaction temperature are so adjusted that after the flap region is cleaved, the TSR dissociates from the target strand. Now, a new intact primary probe oligo pairs with the freed target strand. In this way, a single target strand pairs with several primary probe oligos, and from each of them, one flap segment is liberated. Similarly, the flap segment dissociates from the FRET oligo after the base with the fluorophore is cleaved. The flap segment now pairs with an intact FRET oligo to liberate another fluorophore-labeled base. Thus, the signal generated by perfect pairing between an invader oligo and a target DNA strand is amplified by the following two reactions. (1) The cleaved primary probe oligo is repeatedly replaced by new intact molecules so that multiple copies of the flap segment are generated from a single target DNA molecule. Similarly, (2) a single flap segment sequentially associates with several FRET oligo molecules to liberate as many fluorophore-labeled bases. This amplification affords an easy and reliable detection of the signal.

A limited multiplexing is possible by carrying out two reactions in a single well; this would require the use of two different easily distinguishable fluorophores to label the two flap sequences as well as their corresponding FRET oligos. However, a high level of multiplexing is possible by using hundreds of especially designed electrophoretically distinguishable fluorescent reporter molecules called eTag reporters. The eTag reporters are resolved and quantified by capillary electrophoresis of the automated first-generation DNA sequencers. The detection of the 5′ flap of the primary probe oligo can be done on the basis of size, sequence, charge, or fluorescence by using detection strategies like mass spectrometry, capillary electrophoresis, microfluidics, universal array chips, time release fluorescence, etc. In case of DNA-based detection, both the primary probe and the FRET oligo are simultaneously cleaved, while these two cleavages take place sequentially in the RNA-based detection format.

The invader technology is highly specific due to the strict need for an invasive structure for cleavage of the primary probe oligo, which depends on perfect pairing of the invader oligo and the target DNA. This technique is highly accurate in distinguishing heterozygotes from the two homozygotes. The technology is easy to use, flexible, and scalable to ultrahigh-throughput by using standard microtiter plate formats, e.g., 96- or 384-well format. It requires minimal hands-on time, and the assays are completely homogeneous. In a *homogeneous reaction*, all the steps are carried out in a single vessel, and there is no need to transfer the reactants from one vessel into another during the assay. The assay does not involve PCR amplification of the target DNA. But target sequence information is needed for designing the invader oligo and the TSR of the primary probe oligo. The amount of DNA required for reliable genotyping of a large number of SNPs is 50 ng or more, which is relatively quite high. The DNA requirement can be reduced by PCR amplification of the target regions for the invader assay; this would make the assay more robust, but would add to the cost and operational activities (Kahl et al. 2005).

This technology can be used for detection of SNPs, InDels, transgenes, and gene copy number variations (CNVs). It can also be used for detection of infectious agents and for studying gene expression. This assay can be performed with genomic DNA, RNA, cDNA, or PCR products. Custom-made invader assays are available for wheat, barley, oat, maize, soybean, cotton, rice, tomato, tobacco, etc.

13.2.2 Pyrosequencing

Pyrosequencing (Sect. 4.2.2.2) is suitable for both SNP discovery and SNP genotyping

(de Vienne et al. 2003). It has been commercialized by Biotage AB (earlier Pyrosequencing AB), Sweden for detection of SNPs and InDels and for estimation of allele frequencies from PCR products of the target genomic regions. Biotage AB provides a range of systems capable of medium (96-well format) and high (384-well format) throughput, the necessary software, and reagent kits for rapid sample preparation and processing. The 384-well format system can score several thousands to few tens of thousands of genotypes per day. The pyrosequencing assay is homogeneous and amenable to automation. The software allows multiplexing, as it is capable of detection of multiple SNPs in a single template using the same primer and in even different templates in the same assay. In this technology, the 3′ end of the primer does not have to be placed next to the SNP locus, which could be an advantage in case of some plant species (Kahl et al. 2005).

The general procedure for SNP genotyping is as follows (Kahl et al. 2005). The panel of SNPs to be genotyped is first identified, and sequences flanking the SNP loci are determined. The sequence information is used to design specific PCR primer pairs for each SNP locus. A short genomic region including a given SNP locus is amplified using the specific primer pair, which also serves as sequencing primer. One of the two PCR primers for each SNP locus is biotinylated so that each PCR product has one biotinylated strand. Streptavidin-coated magnetic beads are used to separate the biotinylated strands, which are then used for pyrosequencing of 20–40 bases surrounding the SNP locus. This is an accurate SNP genotyping method. The SNP alleles in a sample can be called on the basis of analysis of the concerned sample without the need for comparison of the signal with the signals from other samples/controls.

Pyrosequencing assay is quantitative since the sizes of peaks reflect the numbers of molecules of the concerned nucleotides incorporated during a single reaction. As a result, heterozygotes can be reliably distinguished from homozygotes, and even allele frequencies can be determined from pooled DNA samples. This technology enables the identification of several adjacent SNPs, as a result of which it is very useful for the SNP haplotype construction. In one strategy, long-range allele-specific PCR (AS-PCR; Sect. 4.6.1) is combined with the pyrosequencing technology for determination of haplotypes from DNA samples/pools.

The pyrosequencing method seems to be the most suitable technique for analysis of SNPs in polyploid plant species. For example, it was able to correctly identify all the five possible different allelic combinations (*aaaa*, *aaaA*, *aaAA*, *aAAA*, and *AAAA*) for a SNP locus in potato, a tetraploid crop species. Refinements in the pyrosequencing technology, e.g., the use of single-strand binding proteins, software-assisted improved designing of primers, and reduced reaction volumes, may be expected to increase the success rate and lower the assay cost. The main limitations of this technology include the need for PCR amplification of the target regions and the preparation of single-stranded templates for pyrosequencing.

13.2.3 KASP™ Genotyping Assay

The *KASP™* (*Kompetitive Allele-Specific PCR*; LGC Genomics, UK) *genotyping assay* (earlier KASPar) is a high-throughput allele-specific PCR-based assay. It uses two forward primers, one reverse primer, and two FRET oligos. The 3′ terminal base of one of the forward primers is complementary to one of the two SNP alleles, while that of the other primer is complementary to the other allele of the SNP locus. Both the forward primers consist of two distinct regions: their 3′ regions correspond to the sequence on the 3′ side of the SNP allele, but their 5′ regions represent the sequence of one of the two FRET oligos (Fig. 13.2). One FRET oligo carries the FAM™ fluorescent dye at the 5′ end of one strand, while the other FRET oligo has the HEX™ dye. The second strand of both the FRET oligos has a quencher dye at its 3′ end; the quencher corresponds to the fluorescent label at the 5′ end of the complementary strand. As a

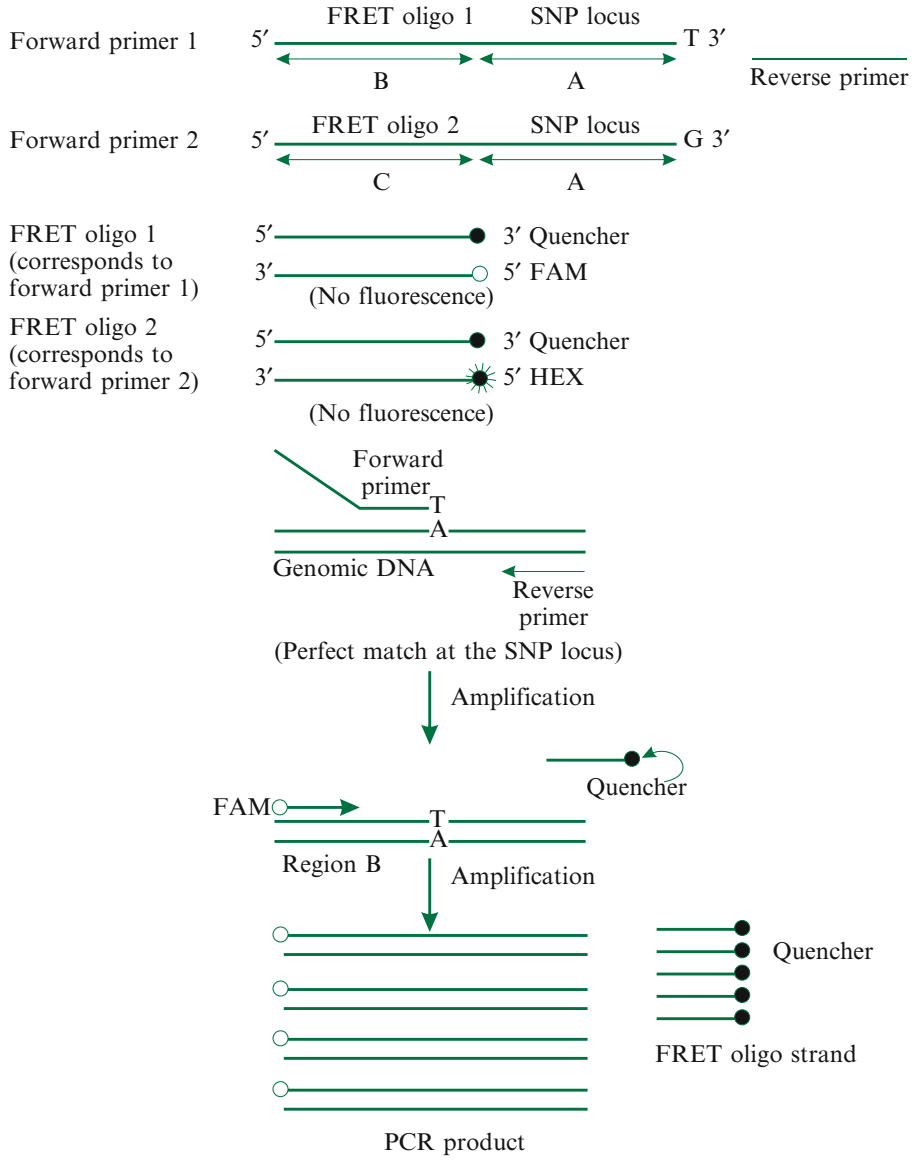


Fig. 13.2 A schematic representation of the KASP genotyping assay. Region A of the forward primers is specific to the SNP locus, Region B corresponds to FRET oligo 1 (the FAM carrying strand), while Region C has the sequence of FRET oligo 2 (the HEX carrying strand). The single reverse primer is common to both the forward primers. First round of PCR is based on the forward primer pairing perfectly to the SNP locus. This PCR product has the sequence of the corresponding FRET oligo. FRET oligo strand serves as primer in the second and subsequent rounds of PCR. As a result, the fluorophore is separated from the quencher, and fluorescence is generated. In case there is mismatch at the SNP locus, there will be no amplification, and fluorescence will not be generated

result, the two FRET oligos are unable to generate fluorescence by themselves.

The sample genomic DNA, the two forward and one reverse primers, and the two FRET oligos are mixed and used for PCR. The forward primer having at its 3′ end the base complementary to the SNP allele present in the sample DNA and the reverse primer would amplify the

genomic region flanking the concerned SNP locus. The PCR product generated in this way will also have the sequence corresponding to one of the two FRET oligos. As a result, in the second and the subsequent PCR cycles, the FRET oligo strands will be used as primers. In this way, increasing amounts of the fluorophore will be removed from the vicinity of the quencher dye and will now generate fluorescence. The fluorescence is captured and processed to deduce the SNP allele present in the DNA sample. Fluorescence due to the FAM™ or HEX™ alone will indicate the sample to be homozygous for the corresponding SNP allele, while fluorescence due to both the dyes will reveal the sample to be heterozygous.

The KASP™ genotyping assay is homogeneous and a single-plex reaction. It can be run in 94-, 384-, or 1,536-well PCR plates and can generate up to 500,000 data points per day. It provides accurate SNP and InDel genotype data (accuracy, 99.8 %). It permits the maximum flexibility in assay design, repeat assays, etc. and saves cost by using labeled universal FRET oligos in the place of labeled specific primers for each reaction. This assay uses only 0.1–10 ng DNA per sample, and PCR amplification of the template is not required.

13.2.4 TaqMan OpenArray Genotyping System

The *TaqMan OpenArray genotyping system* (Applied Biosystems, USA) is a low-cost, medium- to high-throughput SNP genotyping platform based on TaqMan technology (Sect. 4.6.2; http://www.ikerkuntza.ehu.es). In this system, the TaqMan™ probes are combined with a minor groove binder (MGB); this increases the melting temperature (*Tm*) of probes without an increase in their lengths. As a result, the matched and mismatched probes exhibit greater difference in *Tm* values, which increases the assay accuracy. The assay for each SNP locus uses two allele-specific MGB-TaqMan probes and two primers (one forward and one reverse primer) for PCR amplification. One of the probes is specific to one SNP allele and is labeled with VIC® dye linked to its 5′ end, while the other probe is specific to the second SNP allele and is labeled with FAM™ dye. The 3′ ends of both the probes are linked to corresponding nonfluorescent quencher molecules. The assay procedure involves real-time PCR to liberate and amplify the fluorescence signal, which is captured and processed to deduce the SNP allele at the concerned locus.

The OpenArray technology uses nanofilter fluidics to achieve high throughput and to reduce cost. The 63 × 19 mm OpenArray Genotyping Plate is divided into 48 subarrays. Each subarray has 64 through holes of 33 nL each, which act as isolated reactors. Each through hole is preloaded with a unique TaqMan assay, i.e., reagents, including the probes and the primers. The system allows multiple sample loading without cross-contamination. The minimum project size is 480 samples assayed for 16 SNPs. One person can test 1,500 samples for up to 256 SNPs in one day without the use of robotics. The assay time from purified DNA to genotyping results is ~8 h, and the overall call rate is 99 %. The genomic DNA needed per subarray (64 through holes) is 125 ng, i.e., ~2 ng per reaction. The TaqMan OpenArray SNP genotyping assays consist of one mouse and two human assay collections. In addition, custom-made assays can be created to suit individual research needs.

13.2.5 SNP Analysis by MALDI-TOF MS (The Homogeneous MassEXTEND Assay)

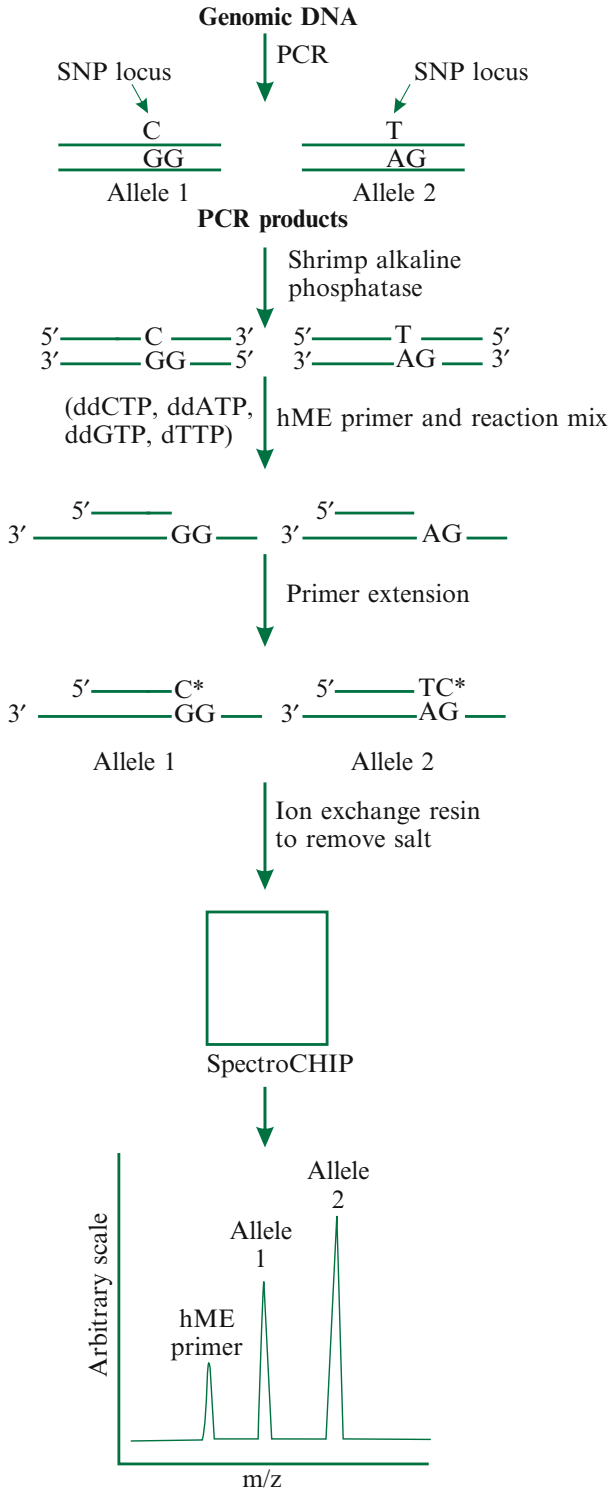
In simple terms, the process of *MALDI-TOF MS* (*matrix-assisted laser desorption ionization time of flight mass spectrometry*) is as follows. A laser is used to induce desorption of biomolecules with the help of a matrix substance, an excess amount of which is co-crystallized with the target biomolecules. A *matrix substance* is an organic molecule that has the same spectrum of energy absorption as the selected laser wavelength and does not interact chemically with the target biomolecule. The matrix substance absorbs energy

from the laser, due to which it evaporates into the vacuum of the mass spectrometer. The target biomolecule is also desorbed along with the matrix and is ionized by proton transfer, and the resulting ions are accelerated in an electric field. Finally, the time of flight (TOF) of the ions through a drift region that is electric field-free is determined. The time of flight of an ion increases with the molecular mass of the ion. MALDI-TOF MS enables direct analysis of DNA on the basis of molecular mass, which is determined very accurately. MALDI-TOF MS has evolved as a versatile high-throughput genotyping platform that is suitable for rapid SNP discovery, screening for mutations, and quantitation of gene expression (de Vienne et al. 2003; Kahl et al. 2005).

Generally, mass spectrometry-based SNP genotyping methods use primer extension (Sect. 4.6.6), e.g., homogeneous MassEXTEND (hME) assay. Usually, 384-well microtiter plates are used for performing the hME assay (Rodi et al. 2002). The genomic region having the SNP is first subjected to PCR amplification, and the unused dNTPs are dephosphorylated by a treatment with shrimp alkaline phosphatase. Then, the following are added to the reaction mixture: a hME primer, DNA polymerase, three dideoxyribonucleotides, and one deoxyribonucleotide. The hME primer is designed to pair with the target genomic region in such a way that its 3′ terminal base is placed immediately adjacent to the SNP locus. The deoxyribonucleotide used in the reaction should be complementary to the SNP allele to be detected, e.g., dTTP for SNP allele *A*, while the other three nucleotides will be used as dideoxyribonucleotides (ddATP, ddCTP, ddGTP).

The DNA polymerase will extend the hME primer and will generate two types of products. (1) In case the first nucleotide incorporated into the hME primer is one of the dideoxyribonucleotides, there will be no further extension of the primer; this will yield primer molecules extended by one nucleotide only. (2) But if the normal deoxyribonucleotide was the first nucleotide to be added to the hME primer, primer extension will continue till one of the ddNTPs was incorporated; this will produce primer molecules extended by two or more nucleotides (Fig. 13.3). (3) In addition, the reaction mixture will have some unused primers as well. The masses of these three types of molecules will differ by at least 300 Daltons (Da), while this system can differentiate between molecules differing by merely 3 Da. After the primer extension step, Na^+ and K^+ ions are removed, and very small volumes (10 nL) of the samples are dispensed by a specialized Pintool device onto highly precise silicon chip arrays (SpectroCHIP) having the appropriate matrix. The masses of the DNA molecules present in each sample are now determined by MALDI-TOF MS. The data acquisition and processing are fully automated: the data quality is evaluated within microseconds, and the SNP genotype is assigned for a sample spot. MALDI-TOF MS facilitates estimation of the relative abundance of alleles of a SNP locus in pools of DNA samples; this is termed as *allelotyping*. Allelotyping also allows rapid separation of true SNPs from those due to sequencing errors/false positives of in silico selection procedures. Allelotyping has been used in association studies to identify genomic regions/candidate genes for more detailed investigation.

MALDI-TOF MS can discriminate oligos having 17–30 nucleotides. Some degree of multiplexing can be achieved by designing the hME primers for two or more SNP loci in such a way that their masses and the masses of the extension products of these primers do not overlap. Such primers can be used together in a single reaction. Specific computer programs like MassARRAY Assay Design (available from SEQUENOM) may be used for designing hME primers for multiplexing of up to 15 reactions. The MassEXTEND assay is simple, fast, and inexpensive. Both PCR and hME extension steps are carried out as single-tube reactions so that they are amenable to automation. The analysis of a 384-element chip array takes ~30 min. Thus, about 9,000 genotypes can be obtained per hour at a cost of less than US $ 0.10 per genotype. The MALDI-TOF MS has emerged as a powerful tool for rapid generation of SNP assay panels for genetic studies.



PCR amplification of the genomic region having the SNP locus. The two SNP alleles are *G* and *A*, and the next base is G in the case of both the alleles. Remaining dNTPs are dephosphorylated by shrimp alkaline phosphatase

The mix has hME primer, enzyme, three ddNTPs and one selected deoxynucleotide (dTTP in this figure)

In case of allele 1 addition of ddCTP stops further extension. But in case of allele 2, normal dTTP is added so that termination occurs only when the next nucleotide (ddCTP) is added

Samples are transfered onto a silicon chip (SpectroCHIP) and subjected to MALDI-TOF-MS analysis

Fig. 13.3 A simple representation of MALDI-TOF MS analysis for SNP genotyping. * indicates dideoxynucleotide, which terminates primer extension (Based on de Vienne et al. 2003; Kahl et al. 2005)

13.2.6 Nanofluidic Dynamic Array-Based Assays

The TaqMan® assay (Sect. 4.6.2) has been adopted for a medium-multiplexing, high-throughput SNP genotyping assay based on nanofluidic dynamic arrays (Wang et al. 2009). Each dynamic array has 2,304 reaction chambers of 6.75 nL each. Each dynamic array can test each of 48 different DNA samples for 48 SNPs, and 50–60 ng DNA is required for each sample. In case the quantity/quality of the genomic DNA is not adequate, the genomic regions with the target SNP loci may be amplified by PCR and used for the assay. The loading of all the reaction chambers in the chip takes ~45 min. After loading, the chip is placed on a thermal cycler for PCR. Following PCR, fluorescence image data is captured and processed for SNP genotype calls. This system is simple to use and has call rates of over 99.5 % with accuracy of >99.8 %. This assay system takes merely 3 h from sample to result. The extremely small reaction volume reduces the genotyping cost to only US $ 0.05 per data point, while other genotyping platforms cost at least twice as much.

The Fluidigm EP1™ System is an efficient high-throughput nanofluidic dynamic array-based platform for SNP genotyping and copy number variation analysis. The system comprises various integrated fluidic circuits (IFCs) and other equipment needed for streamlined PCR and detection of the assay results and software for data acquisition and analysis. The system is flexible as it can be expanded as per the requirement. The system was originally designed for 5′ nuclease assays, but it is quite flexible and can use any assay procedure. The SNP genotyping in the EP1™ System is based on allele-specific PCR and requires minimum experimental setup time. This technology incurs extremely low operating costs, has very easy workflow, and permits low to moderate multiplexing. The software collects and analyzes the data and offers several result output format options, including scatter plots, tables, heat maps, etc.

13.2.7 The Array Tape Technology

A novel approach for using array tape or Microtape® in the place of microplates provides a miniaturized, flexible, and accurate PCR-based SNP genotyping. The Array Tape™ can be used for any fluorescence-based SNP assay that is homogeneous and single step, and can be performed in a closed tube, e.g., TaqMan, Invader Plus, KASP, etc. The Array tape is a continuous plastic tape, into which 96-, 384-, or 1,536-well formats are embossed. A single reel of tape enables simultaneous processing of assays equivalent to hundreds of microplates. The miniature well size (700–800 nL reaction volume) leads to a saving of up to 90 % on reagents and consumables. The system is completely automated, including pipetting, drying, dispensing, cover tape sealing, data acquisition, analysis, SNP calling, etc. The overall error rate of the system is 0.023 %. One person can generate up to 307,200 data points (equivalent of 800 microplates) in one working day. This method can also be used for high-throughput genotyping of SSR markers (Chudyk 2006). More data points can be generated by using barcoded magnetic carboxyl beads (from Applied Biocode). Each SNP assay is assigned to a unique barcoded bead. A single well may be used for assaying up to 4,096 SNPs; this permits a very high level of multiplexing.

13.2.8 The Illumina GoldenGate SNP Genotyping Platform

The microarray/DNA chip technology was initially used for hybridization reaction-based genotyping (Sect. 4.6.4). But the methods based on allele-specific primer extension are much more specific than hybridization-based techniques due to the high accuracy of DNA polymerase. Therefore, single-base extension method was used for microarray development by attaching the primers to a solid phase like glass slides. But these microarrays had to be

developed for every species and for every new set of markers. Subsequently, microarrays were designed with "tag" oligonucleotides so that a chip could be used universally for any species with any set of markers. A *tag oligonucleotide* is a unique oligonucleotide unrelated to the sequences of the loci to be genotyped. Each tag sequence is attached to a known position on the microarray/chip. The sole function of the tag sequences is a specific separation of the primer extension products on the basis of hybridization with the 5′ regions of the primers used for the extension reaction. Therefore, tag sequences are often referred to as *capture oligos* (Kahl et al. 2005).

The 5′ region of each primer used in the primer extension reaction is complementary to a single tag sequence attached to the chip, and these sequences hybridize during the SNP detection step. The 3′ region of the primer is complementary to the sequence on the 3′ side of a given SNP locus and is involved in allele-specific primer extension. The "tag" array approach offers the following advantages. (1) The DNA chips having the tag oligos are generic in that a chip can be used with any species and any set of markers; this greatly reduces the chip production cost. (2) The primer extension is performed in solution and is separated from the signal detection step. Finally, (3) the DNA chip formats are more flexible since the tag oligos are not required to have a free 3′ -OH residue.

The Illumina GoldenGate genotyping platform (http://www.illumina.com/support.ilmn) is based on a specially designed chip having "tag" sequences attached to microbeads. It achieves allele discrimination as follows: the extension of an allele-specific oligo (ASO) enables the ligation of this oligo with a locus-specific oligo (LSO). The ligation product is PCR amplified and the PCR product is hybridized with the tag oligo immobilized on the array matrix. In addition, it combines a miniaturized matrix of miniature arrays (*Sentrix Array Matrix*), a high-resolution confocal scanner (*BeadArray Reader*), and a highly multiplexed genotyping assay (*GoldenGate assay*). The above combination improves data quality, reduces cost per data point, and makes the assay user-friendly (Kahl et al. 2005; Fan et al. 2006; Lin et al. 2009; Thompson et al. 2012).

13.2.8.1 The Sentrix Array Matrix

The *Sentrix Array Matrix* consists of 96 (8 × 12 format) individual 1.2 mm diameter miniature arrays (Fig. 13.4a). Each miniature array consists of ~50,000 fiber optic strands fused together as a bundle (Fig. 13.4b). The end of each fiber of the array has a well of ~3 μm in diameter. In the well of each fiber, one 3-μm diameter bead is self-assembled; this bead has "tag oligos" of a single type covalently attached to its surface (Fig. 13.4c). The universal Sentrix Array Matrix has 1,624 unique capture oligo sequences, and several molecules of a single oligo are attached to a given bead type. The tag oligo forms the *address sequence* for a given bead type. Thus, one miniature array would contain, on average, ~30 beads of each type; this increases the accuracy of assay results. A hybridization-based decoding procedure using "decoding" oligos is used to determine the type of bead present in every fiber of a given miniature array and to check the quality of every bead of the array.

13.2.8.2 The BeadArray Reader

The *BeadArray Reader* is a laser confocal scanner for the Sentrix Array Matrix. It has resolution of ~0.8 μm, has two excitation lasers (532 and 635 nm), and simultaneously captures images in two colors for all the 96 miniature arrays. Its software registers the images, extracts, and saves the light intensity data, and the genotyping software deduces SNP genotypes from these data.

13.2.8.3 The GoldenGate Assay

In the *GoldenGate assay*, discrimination between SNP alleles, identification of different SNP loci, and signal amplification steps are separated from each other. Therefore, each reaction can be individually optimized. This feature has improved the assay accuracy and enabled very high degree

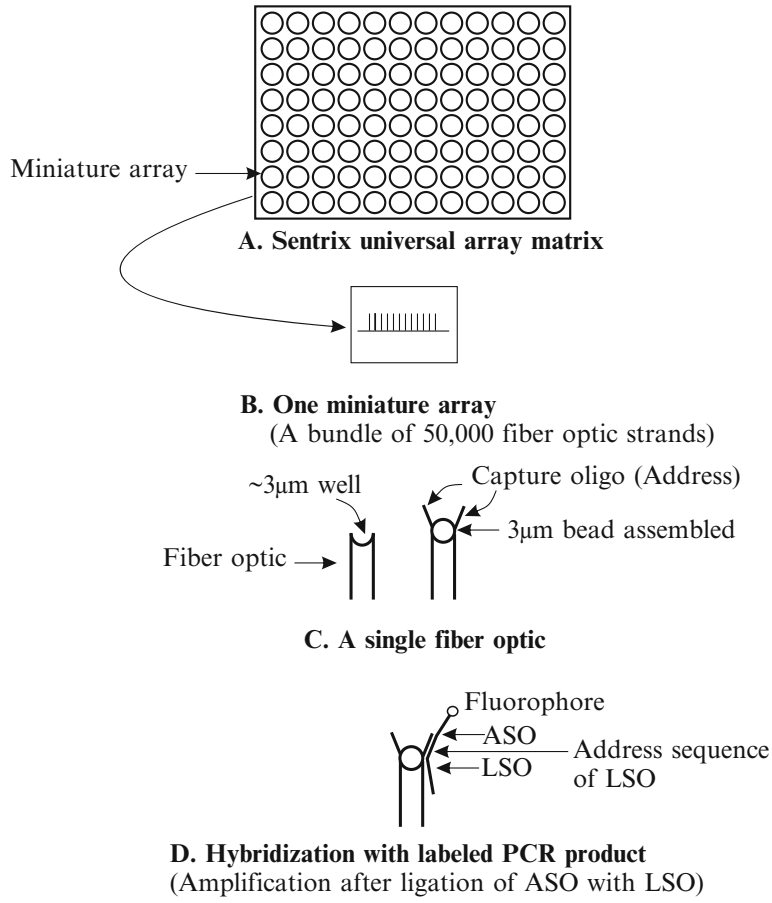


Fig. 13.4 A schematic representation of the organization of Sentrix Array Matrix. *ASO* allele-specific oligo, *LSO* locus-specific oligo (Based on Kahl et al. 2005)

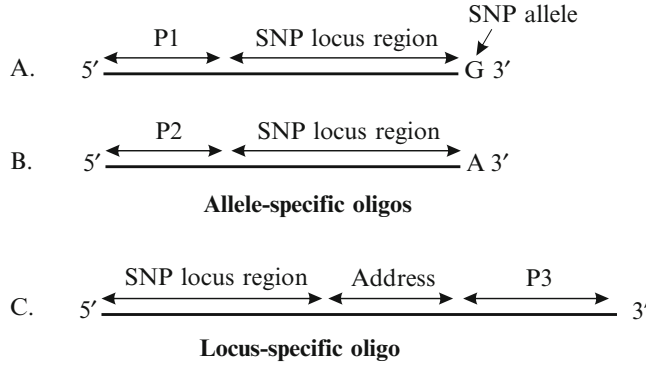


Fig. 13.5 The three oligos used in Illumina GoldenGate SNP assay. The address is a unique sequence for each SNP locus; it is complementary to one of the oligos fixed on the beads of Sentrix Array Matrix. P1, P2, and P3 are sequences complementary to the three universal primers used for PCR amplification (Based on Kahl et al. 2005)

of multiplexing. The GoldenGate assay uses two ASOs and one LSO for each SNP locus (Fig. 13.5). The 5′ region of one ASO is complementary to the universal PCR primer P1, while its 3′ region is complementary to the sequence on the 3′ side of the SNP locus and its 3′ terminal base is complementary to one of the two alleles at the SNP locus. Similarly, the 5′ region of the

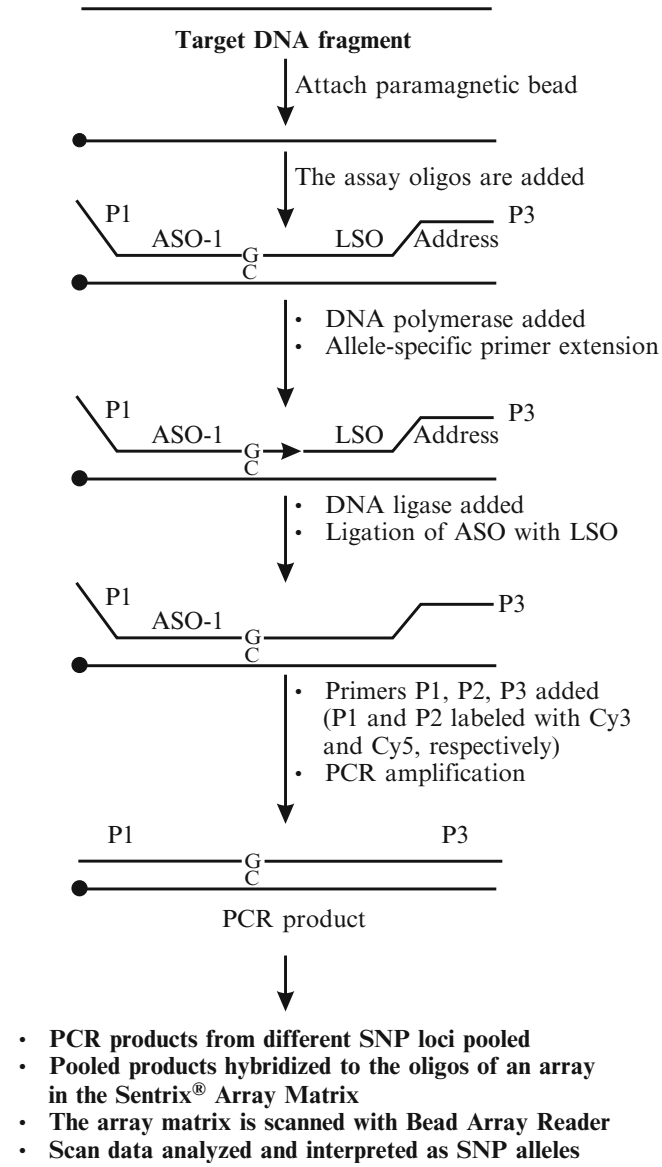


Fig. 13.6 A simplified schematic representation of the reactions in the Illumina GoldenGate genotyping platform (Based on Kahl et al. 2005)

other ASO is specific to the universal PCR primer P2 and its 3′ region is also specific to the SNP locus, but its terminal base is complementary to the other allele at the SNP locus. The *LSO* has three regions: (1) its 3′ region is specific to the sequence on the 5′ side of the SNP locus, (2) the middle sequence is complementary to one of the capture oligos of the Sentrix Array Matrix, and the 5′ region is specific to the universal PCR primer P3. The 5′ terminal nucleotide of the LSO is phosphorylated.

The genomic DNA to be analyzed is fragmented, and the fragments are attached to paramagnetic particles; this helps in purification of the assay oligos properly hybridized to the target DNA (Fig. 13.6). The GoldenGate assay can analyze 1,536 SNP loci in a single reaction. The three assay oligos (two ASOs and one LSO) for each of the 1,536 SNP loci are pooled and used for hybridization to the fragments prepared from the genomic DNA. The non-hybridized and improperly hybridized assay oligos are

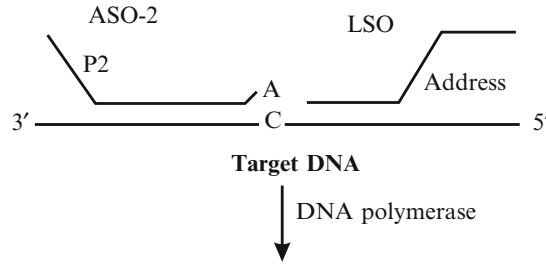


Fig. 13.7 The GoldenGate assay. When the SNP allele is not complementary to the 3′ terminal base of the allele-specific oligo, there is no extension of ASO and, as a result, no ligation with the LSO (Based on Kahl et al. 2005)

- **Mismatch at the SNP locus**
- **As a result, ASO-2 is not extended, and it cannot be ligated to LSO**
- **There will be no amplification by PCR**
- **There will be no fluorescence at the corresponding address in the array matrix**

removed by washing. DNA polymerase is then added to the reaction mixture having the assay oligos properly hybridized to the target DNA fragments. In case the 3′ base of an ASO is complementary to the SNP allele, it will base pair with the SNP allele. The DNA polymerase will extend this ASO up to the 5′ end of the LSO hybridized on the 5′ side of the SNP locus (Fig. 13.6). A ligase is now added to the reaction mixture to ligate the 3′ end of extended ASO to the 5′ end of LSO.

The above ligation product serves as template for PCR using the universal primers P1, P2, and P3. The primers P1 and P2 are labeled with Cy3 and Cy5 fluorophores, respectively. PCR amplification will occur only when a given ASO is extended and ligated to the LSO. In Fig. 13.6, ASO-1 having P1-specific 5′ region and G at its 3′ terminus properly pairs with the allele *C* present at the SNP locus. As a result, ASO-1 is extended and ligated to LSO to yield the PCR product labeled with Cy3. But if the SNP allele were *T* in the place of *C*, the ASO-2 (having P2-specific 5′ region and A as the 3′ terminal base) will properly pair at the SNP locus and will ultimately yield PCR product labeled with Cy5. The above results will be obtained in the cases of individuals homozygous for the SNP alleles. In the case of heterozygous individuals, both the ASOs will properly pair with the concerned SNP alleles and ultimately give rise to PCR products labeled with Cy3 and Cy5, respectively. But when the 3′ terminal base of the ASO is not complementary to the SNP allele, the two will not pair, the ASO will not be extended, and, ultimately, there will be no PCR product. As a result, there will be no fluorescence at the corresponding address in the array matrix (Fig. 13.7).

The PCR products are denatured to make them single stranded prior to their hybridization to a single miniature array of the Sentrix Array Matrix. It may be pointed out that the address sequence of LSO is unique for every SNP locus being analyzed. Therefore, PCR products for each SNP locus will hybridize to the complementary capture oligo sequence bound to the beads located at known positions in the miniature array. Thus, the 94 miniature arrays of the matrix array can be hybridized with reactions representing 94 different individuals, i.e., 94 × 1,536 different SNP loci. The array matrix is scanned with the BeadArray Reader, and the fluorescence color and intensity data are used to deduce the genotype at each SNP locus. A given bead will give either Cy3 signal (homozygous for the concerned SNP allele), or Cy5 signal (homozygous for the other SNP allele), or both Cy3 and Cy5 signals (heterozygous at the SNP locus). The mean intensity in each of the two colors is determined for each bead type and used for deducing the SNP genotype.

The GoldenGate assay system is efficient, accurate, cost-effective (per data point cost US $ 0.03), very high throughput, and one of the most popular platforms. The assay system is quite flexible and can genotype a large number of SNP loci specified by the user across a large number of samples. This assay has been used in barley, soybean, maize, wheat, etc. The

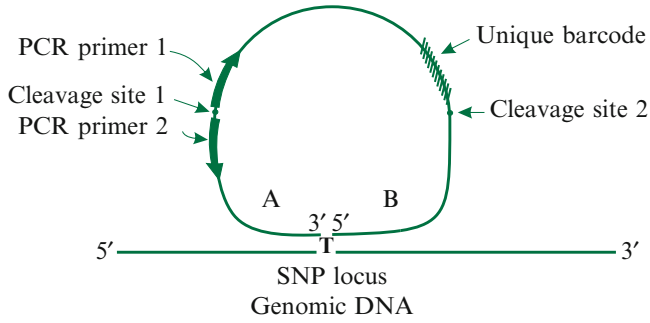


Fig. 13.8 The molecular inversion probe (MIP) design and function. A and B are the 3′ and 5′ terminal sequences of the probe; they are complementary to the genomic sequences flanking the SNP locus. The gap at the SNP locus is filled, making the MIP circular. MIP is opened by cleavage at site 1, amplified using PCR primers 1 and 2 and the amplification product is cleaved at site 2. The barcode sequence-containing fragment is hybridized with the array containing the unique barcode sequences as probes (Based on Ji and Welch 2009)

VeraCode technology (coupled with BeadXpress Reader) of the GoldenGate assay is fluidics based and costs less per sample than the fixed array, but it can genotype only 384 SNP loci per sample.

13.2.9 Molecular Inversion Probe Technology

Molecular inversion probe (MIP) technology was developed by ParAllele Bioscience and subsequently acquired by Affymetrix. The MIP is a robust proprietary technology for candidate gene (targeted whole-genome) approach for large-scale SNP genotyping and CNV analysis. MIP analyzes the genetic polymorphisms by examining unique barcode tags rather than the sequence polymorphism itself. A MIP is a single 120 nucleotide (nt) long oligonucleotide that hybridizes to a specific sequence of the genome. It has two terminal inverted sequences of 20–30 nt each; these sequences are complementary to the sequences flanking the targeted SNP locus, but the SNP locus itself is excluded. The MIP has a cleavage site on either side of which PCR primer sequences are located. After, one of these primer sequences is located a barcode that has unique sequence for each SNP locus, and a second cleavage site is placed beyond this sequence.

The MIP is hybridized to the target genomic DNA; it forms a circle that has a single base pair gap at the SNP site between its two termini (Fig. 13.8). Now, a mixture of dATP and dTTP (or dGTP and dCTP) is added, the gap is filled, and the two ends of the probe are ligated producing a complete circular molecule. Each of the two pairs of dNTPs (dATP/dTTP and dGTP/dCTP) has attachment sites for specific dyes that are used during the staining step. The circular MIP is then made double stranded, opened up by cleavage at the site 1, and amplified using the two PCR primers. The barcode sequence is liberated from the probe by cleavage at site 2 and used for hybridization with the barcode sequences attached to the microarray. The stained and washed arrays are scanned using the Affymetrix GeneChip® Scanner, and the scan data are analyzed with the help of GeneChip® Targeted Genotyping Analysis Software. The intensity of fluorescence signal from a specific barcode feature reflects the specific SNP allele; it also provides a quantitative measure indicative of the copy number.

A much smaller quantity of the probe is needed than in other assay systems, which permits the inclusion of more probes in a single reaction leading to higher multiplexing. It also reduces the chances of interaction between two probes, which minimizes the background signal. In addition to highly accurate SNP genotyping at high levels of multiplexing, the MIP approach is highly quantitative and provides reliable information on allele copy number. The MIP technology can simultaneously analyze tens of

Table 13.1 A comparison among the common microarray-based high-throughput SNP genotyping systems

	Illumina		Affymetrix		Beckman Coulter
Feature	GoldenGate	Infinium	MIP[a]	GeneChip	SNPstream
Type of array	Capture oligos on beads (SAM)	Specific oligos on beads (BeadChip)	Capture oligos on glass	Oligos on glass	Capture oligos on glass
SNP allele discrimination based on	Allele-specific primer extension	Allele-specific primer extension	Single-base extension	Allele-specific hybridization	Single-base extension
Detection of primer extension by	Two-color fluorescence	Biotin–avidin interaction; single color	Two- or four-color fluorescence	Biotin–avidin interaction; single color	Two-color fluorescence
Number of SNPs assayed/samples processed	1,536 SNPs per miniature array; ~110,000 SNPs/SAM	Up to 300,000 SNPs	10,000 SNPs	Up to 690 K SNPs	Tens of SNPs from 100 s of samples per plate
Genomic regions examined	Specific regions with SNPs	Randomly amplified fragments	Specific regions with SNPs	Complexity reduced genomic fragments	Specific regions with SNPs
PCR amplification	Yes (of the ligation product)	Yes (of genomic fragments)	No[b]	Yes (of genomic DNA)	Yes

Based mainly on Tsuchihashi and Dracopoli (2002) and Gupta et al. (2008)
[a]*MIP* Molecular Inversion Probe, *SAM* Sentrix Array Matrix
[b]Circularized probe is amplified by rolling circle DNA replication catalyzed by DNA polymerase

thousands of loci, but efforts are being made to expand it to hundreds of thousands (Ji and Welch 2009). The MIP approach is seen as complementary to and not as a replacement for many of the available CNV analysis technologies. Some of the advantages of MIP technology include much less DNA requirement (37 ng DNA per sample), DNA quality is not that important, and sequence design is more flexible and can detect copy number up to 60, more accurate SNP allele information, and fewer individual MIPs needed for identification of CNV. Since the barcode array is universal, the array itself is inexpensive.

13.2.10 Whole-Genome-Based Microarray Platforms

There are several whole-genome-based microarray genotyping platforms; the main features of some of these systems are summarized in Table 13.1. The Affymetrix's GeneChip Axiom arrays and Beckman Coulter's SNPstream use chips prepared on glass. But the Illumina's GoldenGate and Infinium platforms use chips based on microbeads. These genotyping platforms also differ in terms of the nature and function of the DNA sequences immobilized on the chips and the basis of allele discrimination. For example, allele discrimination is based on hybridization in the case of GeneChip, on allele-specific primer extension in the cases of GoldenGate and Infinium, and on single-base primer extension in the cases of MIP and SNPstream. Some of these platforms have been developed for several important crop species. For example, the Universal Soybean Linkage Panel (USLP 1.0) allows simultaneous scoring of 1,536 SNP loci using GoldenGate assay, while an Infinium-based assay has been developed for 44,000 soybean SNPs. However, out of the total number of SNP loci included in an assay platform, one may expect about 33 % (in a population with low variability) to 66 % (in a population with large variability) of the loci to be polymorphic or informative. In addition, other

microarray-based high-throughput genotyping platforms like diversity array technology (DArT; Sect. 2.6) and single feature polymorphisms (SFP; Sect. 2.8) analyze DNA sequence polymorphisms irrespective of the type of sequence changes involved.

13.2.10.1 The Illumina Infinium HD Assay System

The Illumina Infinium® HD Assay Ultra protocol is based on microbeads, uses Infinium I and Infinium II probe designs, and allows genotyping for 3,000–300,000 SNP and CNV markers per sample. The 3′ end of Infinium II probes is placed immediately adjacent to the SNP site, while that of Infinium I probe is positioned over the SNP site. Therefore, Infinium I probe will generate signal only when there is a perfect match between the primer and the target DNA. When primer extension takes place, either a biotin-conjugated nucleotide or a dinitrophenyl-labeled nucleotide will be added to the primer. The dCTP and dGTP molecules are biotin labeled, while the dATP and dTTP molecules are labeled with dinitrophenyl. The overall signal-to-noise ratio of the assay is further improved by amplification of the signal from the incorporated label. The DNA sample is denatured, neutralized, and used for whole-genome isothermal amplification that increases the DNA amount several thousand-fold. The amplified DNA is enzymatically fragmented and hybridized with the 50 nt long locus-specific oligos attached to the beads of the BeadChip. The BeadChip is a silicon-based array device designed for running multiple samples simultaneously. Each bead type (there are up to 300,000 types of beads in a single BeadChip) has oligo specific for a unique locus. The beads are assembled into microwells of the BeadChip substrate, and a sample section has multiple copies of each bead type. The SNP locus discrimination and CNV determination are achieved by capture of the genomic fragments on the array beads on the basis of sequence-specific hybridization with the probe oligos on the capture beads. The detection of SNP allele is based on the single-base primer extension. This technology uses multi-sample BeadChip format, has high call rate and high accuracy, does not use PCR, and needs only 200 ng DNA per sample.

After single-base extension and probe hybridization steps, the BeadChip is scanned by Illumina HiScan System, iScan System, or BeadArray Reader, which are two-channel high-resolution laser imagers with automated BeadChip loading and unloading added on facility. The laser excites the fluorophore attached to the single base added to the oligos, and high-resolution images of the light emitted from them are recorded. The HiScan and iScan Systems have a much higher throughput than the BeadArray Reader and generate data of equally high quality. These systems automatically tilt and align the BeadChip to ensure their optimal positioning for scan. The image data from HiScan/iScan are used by the GenomeScan software platform to create data files for each channel of BeadChip. GenomeScan also analyzes the data file along with the data on individual bead types in a given channel to generate the genotype data (www.illumina.com).

13.2.10.2 Affymetrix Axiom® Genome-Wide Arrays

The main limitation of bead array technologies is that 5–20 % of the markers are randomly lost every time a bead pool is created or the array is prepared. The Affymetrix SNP genotyping arrays do not suffer from any such limitation, and they are optimized for specific populations and applications. The Affymetrix Mapping GeneChip® 500 K array was the first microarray with sufficient feature density for SNP genotyping in genome-wide association studies. Subsequently, a much higher density array, the Affymetrix SNP 6.0 (a 900 K human SNP array), was developed. These arrays used allele-specific hybridization for detection of the selected and validated polymorphic human SNPs. The Axiom™ Genome-Wide Arrays, like the GeneChip Arrays, have 30 nt long oligonucleotide probes directly synthesized on the array substrate. But each Axiom array has a total of ~1.38 million features available for experimental use. Further, the Axiom arrays use ligation reaction-based assay for detecting SNPs.

Therefore, the Axiom arrays use a set of two probes for each SNP locus. These probes represent the genomic region flanking, and including, the SNP locus on either the forward or the reverse DNA strand. One of the probes, the equivalent of the reporter oligo, is locus specific, does not include the SNP locus itself, and is fixed to the array substrate. The other probe is the equivalent of capture oligo, is allele-specific, and includes the SNP locus (Sect. 4.6.8); this probe is usually termed as solution probe. For an A/T (or a G/C) SNP locus, the solution probe is of two types having A or T (or G or C) as the terminal base corresponding to the SNP allele of the locus, and each of them has the attachment site for a different dye.

Each SNP locus is evaluated by two features, so that a total of ~690 K SNPs can be assayed by one Axiom array. But in the case of SNP loci having four alleles (A/T and C/G alleles), four features will be needed for each SNP locus. Since the attachments for the same two dyes are used for the two allele pairs, two distinct probe sequences located at different sites in the array would be required to distinguish them (Hoffman et al. 2011). A total of 200 or 300 ng of high purity genomic DNA is required for each sample. The DNA of large (>10 kb) molecular size is amplified, enzymatically fragmented, and precipitated. The fragments are resuspended in hybridization mastermix, denatured, and hybridized with the probes fixed on the array. The solution probe is now added and hybridized with the genomic DNA associated with the array probe. The two probes are ligated in an allele-specific manner, and the probes not involved in ligation are washed out. Now, staining reaction attaches the dye to the ligation product, and fluorescence signal from the features are recorded. The hybridization, staining, washing, and imaging steps are carried out in an automated manner in the Affymetrix GeneTitan® Multi-Channel Instrument. The image data are processed using the Axiom® Genotyping Algorithm version 1 (Axiom GT1). The Axiom genotyping platform allows automated parallel processing of 96 samples per plate. The Axiom arrays are available for humans and some other species like maize, wheat, rose, strawberry, etc., and they can also be custom made as per the researchers' requirements. For example, the Axiom 2.0 Assay genotypes for biallelic SNPs and simple InDels in humans using a fully automated workflow.

13.2.10.3 Agilent SurePrint Arrays

The Agilent SurePrint microarrays are prepared using the Hewlett Packard inkjet printing technology to precisely deposit the cDNAs onto the glass wafers coated with a substrate that enables strong binding of DNA to the wafers. Alternatively, the standard phosphoramidite chemistry is used to synthesize the oligonucleotide probes directly onto the glass wafers. The array printer uses components similar to those of inkjet printer to spot very precisely small volumes of liquids containing nucleic acids or phosphoramidite dNTPs onto the glass wafers. The printing technology enables sequential precision deposition of the dNTPs in the specified order at predetermined spots to enable the synthesis of the desired oligonucleotide probes. The array probes are long (60 nt) for high confidence polymorphism characterization. The probes are fixed permanently onto the wafers, and the array surface is deactivated to minimize nonspecific binding of sample DNA. Agilent offers SurePrint G3 Human Comparative Genomic Hybridization (CGH) arrays (for chromosomal aberration analysis) as well as SurePrint G3 Human CGH + SNP arrays both in catalogue and custom design formats. These arrays are in 60 K, 180 K, 400 K, and one million formats, which differ in their design scheme. For example, a 400 K array has 300,000 CGH probes and 120,000 SNP probes.

The assay procedure is simple and requires only 500 ng DNA per sample. The sample DNA is digested with the selected restriction enzyme, and the restriction fragments are labeled using the Klenow fragment. The labeled sample (target) DNA is hybridized with the array probes, and the arrays are scanned with the Agilent DNA Microarray Scanner. This scanner is not affected by the array surface characteristics, including roughness, curvature, etc. The Agilent Genomic

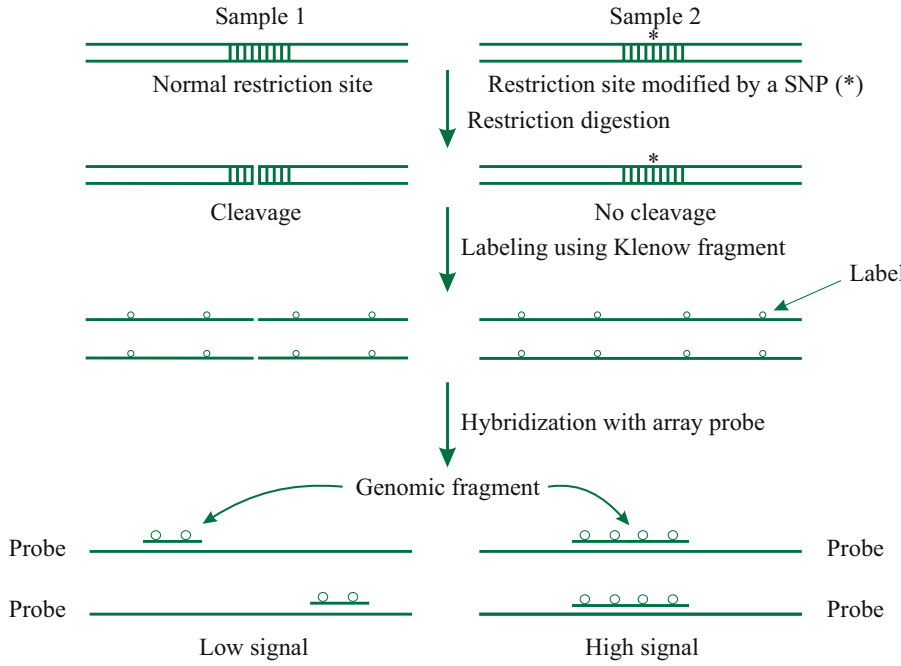


Fig. 13.9 The basis of SNP genotyping with Agilent's SurePrint arrays

Workbench software simultaneously analyzes both CGH and SNP data and allows evaluation of the quality of genotype data. When the sample DNA quality is high, the SNP call rate is over 95 % and the accuracy is >99 % (www.agilent.com/genomics). This technology is designed to detect and genotype SNPs located in the recognition sites of the restriction enzyme used to cut the sample DNA. In case a given restriction site in a sample DNA is intact, the sample DNA will be cleaved at this site, and only one of the two fragments is likely to hybridize with a probe oligo. As a result, the intensity of signal will be much less than when the restriction site is modified by a SNP, and the sample DNA is not cleaved at this site (Fig. 13.9).

13.2.10.4 GenomeLab SNPstream Genotyping System

The GenomeLab™ SNPstream® Genotyping System from Beckman Coulter, Inc. is a cost-effective high-throughput SNP genotyping platform based on single-base primer extension assay (Sect. 4.6.6). It can genotype 12 or 48 SNP loci in each well (multiplexing level), comes in a 384-well format, and generates 4,600 to over three million genotype data points per day. The SNPstream system is scalable and has a compact design that integrates hardware, reagents, and the necessary software. The cost per genotype or per sample remains the same irrespective of the throughput of the assay. It requires merely 2 ng genomic DNA per sample, which is subjected to multiplex (12 or 48 plex depending on the assay format) PCR amplification using pairs of primers specific to the individual SNP loci. The PCR products include the concerned SNP sites; they are purified and used as template for the single-base primer extension assay. This assay uses a single primer, called SNPware primer, for each SNP locus. This primer has a special design as follows. The 5′ half of a given primer is called tag and is complementary to a unique probe attached to the well surface. The 3′ half of this primer is complementary to the sequence on the 3′ side of a SNP locus (Fig. 13.10); the 3′ terminal base of the primer will be placed at the base next to the SNP locus. The SNPware primers specific for the SNP loci, amplified in the multiplex PCR, are used for

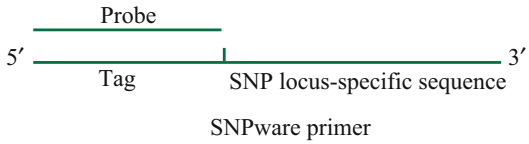


Fig. 13.10 The SNPstream SNP genotyping system. The probe oligo has unique sequence and is fixed to the well surface. The SNPware primer is used for single-base extension. Tag sequence is complementary to a probe sequence, while the 3′ half of the primer is specific to a given SNP locus

single-base primer extension reaction using fluorescence labeled chain terminating acyclonucleotides. Two different dyes are used for labeling the nucleotides corresponding to the two alleles of a given SNP locus; this permits direct determination of the SNP genotype. The extended SNPware primers are hybridized to the probes in the wells, and the fluorescence signal is captured by a CCD-based imager. The image data are analyzed using a comprehensive software suite that has advanced, easy-to-use tools for the complete workflow process, including image data analysis and genotype calling. The 48-plex SNPstream system yields genotype data with 99 % accuracy (https://www.beckmancoulter.com).

13.3 High-Throughput SNP Discovery and Genotyping

The NGS technologies afford extremely rapid sequencing at remarkably lower costs in comparison to the Sanger–Coulson technology. Therefore, the NGS technologies have allowed the development of several different methods that combine SNP discovery with SNP genotyping. These methods are based on sequencing of the whole genome or only a fraction of the genome. In these approaches, a genome complexity reduction technique needs to be employed to avoid repetitive and duplicated DNAs. Further, resequencing of the entire genome is still too costly for routine SNP genotyping. Therefore, researchers have been trying to develop strategies for reducing the total amount of sequencing required for SNP genotyping without compromising the SNP quality. These strategies can be grouped into two broad categories: (1) - genome-wide and (2) targeted marker discovery methods. The genome-wide methods are restriction enzyme based and identify markers distributed over the entire genome. These methods have been grouped into the following three classes: (a) reduced representation sequencing, (b) restriction site-associated DNA sequencing (RAD-Seq), and (c) low-coverage genotyping. The various features of some of the methods are summarized in Table 13.2. In contrast, the targeted marker discovery and genotyping approaches focus on specific regions of the genomes, e.g., RNA-Seq and sequence capture methods (Sects. 4.5.3 and 4.5.6) (Davey et al. 2011).

13.4 Reduced Representation Sequencing

In *reduced representation sequencing* approaches, only a subset of the genomic fragments is sequenced in each individual for marker discovery. It is expected that separate sets of genomic fragments will be sequenced in different individuals. Further, the sum total of sequenced fragments from all the individuals of the sample will represent most of the nonrepetitive genomic regions of the species. The RR-Seq strategies provide marker polymorphism information that is sufficient for many biological studies. These strategies are of the following two types: (1) reduced representation libraries (RRLs) and (2) complexity reduction of polymorphic sequences (CRoPS).

13.4.1 Reduced Representation Libraries

Reduced representation libraries are constructed by fully digesting the genomic DNA with a frequent cutting restriction enzyme and selecting fragments of ~300 bp or so for cloning/sequencing (Fig. 13.11). This approach reduces the fraction of genome represented in the library to

Table 13.2 A comparison among the various marker discovery methods based on NGS technologies. A frequent cutter restriction enzyme is used, except for RAD-Seq where a rare-cutter enzyme is used followed by random shearing

	Marker discovery method				
Feature	CRoPS	RRL	RAD-Seq	GBS	MGS
Amount of DNA needed	300 ng/sample	25 μg (all samples pooled)	300 ng/sample	100 ng/sample	10 ng/sample
PCR amplification	Required	Not known	Required	Required	Not known
Reference genome	Not required	Not required	Not required	Preferable	Preferable
Genomes with large repetitive fractions/higher ploidy levels	Suitable	Not suitable	Suitable	Suitable	Suitable
Detected polymorphism	High	High	Higher than RRL and CRoPS	Moderate	Moderate
Fragments produced by	Two restriction enzymes	One restriction enzyme	Restriction enzyme and mechanical shearing	One or two restriction enzymes	One restriction enzyme
Complexity reduction by	Selection nucleotides in PCR primers	Fragment size selection	Presence of restriction sites	Restriction enzyme and adapter ligation	Fragment size selection
Sequencing coverage	Moderate (5–10×)	Deep (20–30×)	Deep (~30×)	Low	Low (even ~1×)
Sequencing of	Complete fragment	Fragment ends	Fragment ends	Fragment ends	Complete fragment
Barcode	Used[a]	Not used[b]	Used	Used	Used
Suitable for studies with	Wild populations	Wild populations	Wild populations	Experimental populations	Experimental populations
QTL mapping and MAS	Low suitability	Low suitability	Moderate suitability	High suitability	High suitability

Based on Mir and Varshney (2013) and other sources
[a]Barcodes were first used for the CRoPS scheme
[b]Barcodes were not used in the original scheme, but they can be used

~1–10 %. RRLs were first used to develop a SNP map of human genome using the first-generation capillary sequencing (Altshuler et al. 2000). Later, this approach was adapted for the Illumina Genome Analyzer NGS platform (van Tassell et al. 2008), and the Applied Biosystem's SOLiD sequencing technology has also been used. The genomic DNAs from a large number of distantly related individuals/lines are restriction digested. The restriction enzyme may be selected on the basis of in silico analysis of a reference genome sequence, if available, for the concerned species. The restriction fragments from all the individuals/lines are pooled; fragments of 300–700 bp are selected and used for sequencing. This approach provides partial but genome-wide coverage at a fraction of the cost for whole-genome sequencing of all the individuals. RRLs have also been produced by selection of the target fragments hybridized to a microarray, but this approach would limit the SNPs to the coding sequences.

Generally, sequencing is limited to only the fragment ends, but the entire fragments may also be sequenced. The sequence reads can be mapped onto a high-quality reference genome, and SNPs can be identified in the same way as in the case of whole-genome resequencing. But when a reference genome is not available, long single reads from the 454 GS FLX+ system or paired-end reads, i.e., reads from both the ends of RRL fragments, from Illumina platform can be

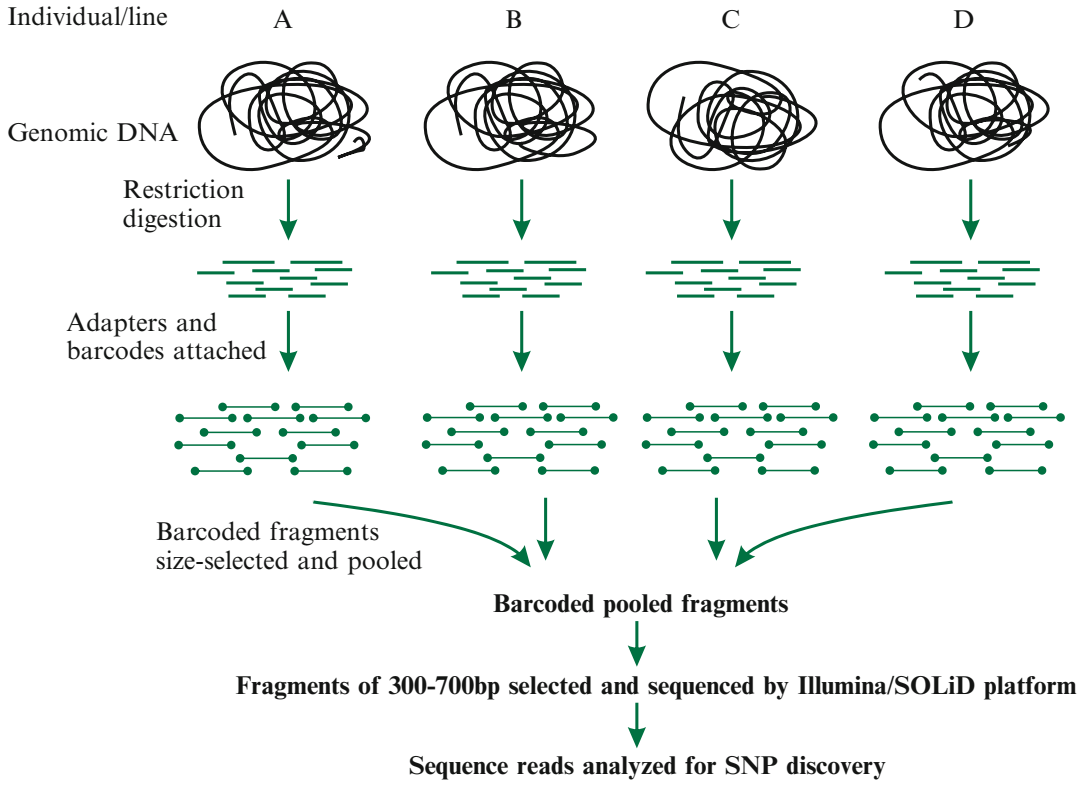


Fig. 13.11 A simplified schematic representation of reduced representation library (RRL) approach for SNP discovery and genotyping. The original RRL scheme did not use barcodes, but the use of barcodes would provide valuable additional information about genotypes of different individuals/lines (Based on van Tassell et al. 2008; Davey et al. 2011)

used to assemble the fragments. These assembled sequences are then used as reference sequence for calling SNPs. Paired-end reads also facilitate detection of structural variations. RRL method discovers high-quality putative SNPs; provides estimates of allele frequencies, including minor allele frequencies in the population; but is unable to provide any information about the individuals constituting the population. This problem can be overcome by using different barcodes for the fragments from different individuals. For this, the DNA fragments from each individual will have to be separately size selected and ligated to the barcode and adapter before they are pooled. The barcode sequences would enable clear-cut identification of the sequence reads from different individuals and thereby provide information on their SNP polymorphism.

RRL has great potential for SNP discovery, but is not suitable for large-scale genotyping since it requires between 10 and 50 μg of DNA, and the library preparation cost is much higher than that for normal whole-genome resequencing library due to the use of restriction enzymes. The RRL approach has been used to identify up to millions of candidate SNPs in maize, soybean, etc. For example, RRL was used to discover 14,550 and 25,047 SNPs in two independent studies with soybean (see Davey et al. 2011).

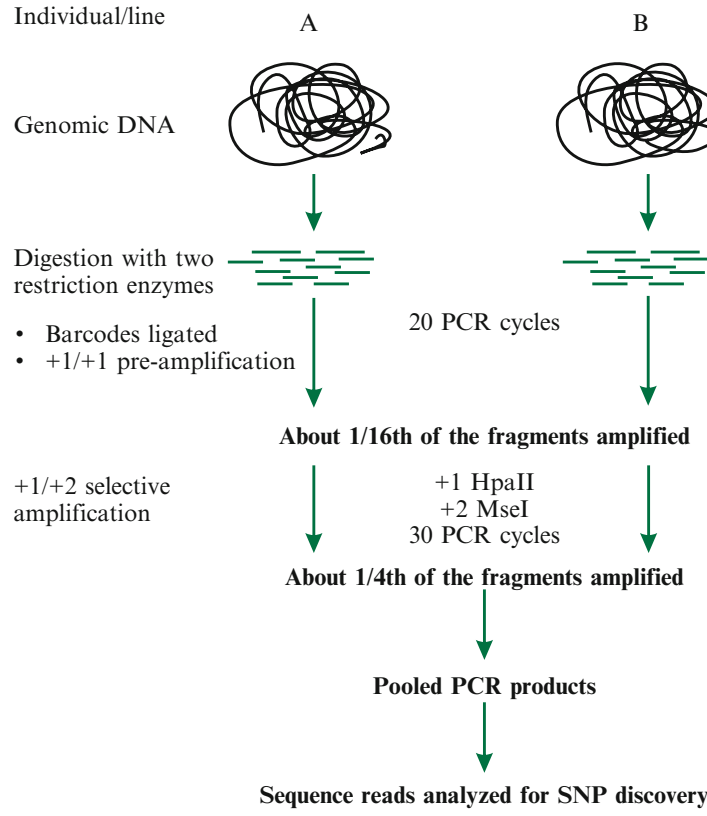


Fig. 13.12 A simplified schematic representation of complexity reduction of polymorphic sequences (CRoPS) scheme of SNP discovery and genotyping using a NGS platform. The complexity reduction is based on the AFLP approach (Based on vanOrsouw et al. 2007; Davey et al. 2011)

13.4.2 Complexity Reduction of Polymorphic Sequences

van Orsouw et al. (2007) combined complexity reduction procedure of AFLP with sequencing by the Genome Sequencer (GS) 20/GS FLX NGS technology to discover high-quality putative SNPs in two maize inbreds. This approach has been called *complexity reduction of polymorphic sequences* (*CRoPS*). They separately digested the genomic DNAs from the two inbreds with two restriction enzymes (*Hpa*II and *Mse*I), ligated appropriate adapters to the fragments, and pre-amplified them using AFLP primers with one selective nucleotide each (the *+1/+1 pre-amplification*) for 20 cycles. The pre-amplification product was suitably diluted and amplified using AFLP primers with one (*Hpa*II primer) and two (*Mse*I primer) selective nucleotides (the *+1/+2 selective amplification*) for 30 cycles (Fig. 13.12). The 5′ ends of amplification primers consisted of four-nucleotide tags (barcodes) to enable identification of the fragments from the two inbreds. The PCR products from the two inbreds were pooled and sequenced with 5–10× depth. The sequence reads were clustered, aligned, and mined for SNPs using custom-developed tools. They were able to discover over 1,200 high-quality putative SNPs; over 75 % of the SNPs in a random sample drawn from these SNPs were confirmed to be true SNPs.

Barcode identifier sequences or *barcodes* are unique short (4–8 nt long) nucleotide sequences that differ from each other for at least two bases. Distinct barcodes are attached to DNA fragments from different individuals for a clear-cut identification of the sequence reads using appropriate bioinformatics tools. Barcode sequences were first used for CRoPS analysis (vanOrsouw et al. 2007). SNP discovery is often hampered by the presence of highly repetitive sequences in species like maize (~80 % of the nuclear genome is highly repetitive), wheat, soybean, etc. The

CRoPS strategy can be used in species having high levels of repetitive DNA in their genomes and/or exhibiting low levels of polymorphism, e.g., in the elite (breeding) germplasm of crop plants. In addition, there is no need for a good quality reference genome.

13.5 Restriction Site-Associated DNA Sequencing

Restriction site-associated DNA sequencing approach is designed to sequence short regions surrounding all recognition sites present in the genome for the selected restriction enzyme. This method is derived from the RAD tag marker technique (Sect. 2.9) by adapting it to a NGS platform like Illumina Genome Analyzer (Baird et al. 2008; Davey and Blaxter 2010). In this method, DNA from each individual of the population is digested with a rare-cutter restriction enzyme (6–8 bp recognition site) that produces sticky ends. The fragments are ligated to the barcoded modified Illumina adapter P1 so that the fragments from different individuals can be identified (Fig. 13.13). The fragments from all the individuals are pooled and randomly sheared, and fragments of appropriate size (300–700 bp) are selected. These fragments will have the P1 adapter at both, one, or none of their ends. Now, the P2 adapter is ligated to these fragments. The P2 adapter is so designed that it is completed only when the given fragment has a P1 adapter at its other end. This is because replication primed by the P1 primer is needed for the completion of the P2 adapter. Since the P2 primer binds only to the completed P2 adapter, only the fragments having the P1 adapter at one or both their ends will be amplified. These fragments will have the barcode as well as the sequence located on one side of the recognition site for the selected restriction enzyme (see Davey et al. 2011).

The raw sequence reads can be aligned to a reference genome sequence of the species, and SNPs and InDels can be identified using a suitable software tool. In case a reference genome is not available, sequence reads representing the same genomic region are identified and classified into two or more groups on the basis of their sequence similarity, and each of these groups is treated as an allele. The sequences of these groups are analyzed for SNP/InDel discovery, and the sequencing errors are corrected by comparing the counts for each base at every position of the concerned sequence read. Since Illumina reads are of up to 150 bases, up to about 300 bases flanking each restriction site can be analyzed for SNP and InDel discovery. RAD-Seq can also detect polymorphism due to "presence"/"absence" of a restriction site revealed by the presence of a read sequence in some individuals/lines and its absence in the others. This polymorphism is the same as that detected by the RAD tag marker (Sect. 2.9).

RAD-Seq has been used to study population differentiation, identification of SNPs, and construction of linkage maps, e.g., in barley, ryegrass, etc. It can generate high-density genome-wide SNP and InDel markers. Paired-end sequencing of RAD-Seq libraries can be used to assemble contigs of ~500 bases for each SNP locus, which can be used to anchor the markers to existing genomic resources. The paired-end sequences can be used to design primers for high-throughput genotyping assays, particularly for such organisms for which well-assembled reference genomes are not available. The sample preparation is labor intensive and expensive: it requires mechanical shearing of DNA and end repair, two steps of adapter ligation, and two steps of gel purification. Finally, RAD-Seq requires 300 ng to 1 μg of genomic DNA per sample.

There are several modifications of the RAD-Seq approach (Table 13.3), including genotyping by sequencing (GBS), 2-enzyme GBS, double digest restriction-site-associated DNA sequencing (ddRAD-Seq), ezRAD-Seq, and 2b-RAD-Seq. The modification *ezRAD-Seq* uses the standard Illumina TruSeq library preparation kits, requires little laboratory expertise and equipment, and involves no additional costs than sequencing using the NGS technology. The TruSeq PCR-free Nano kits should be used when the genomic DNA is <1 μg per sample.

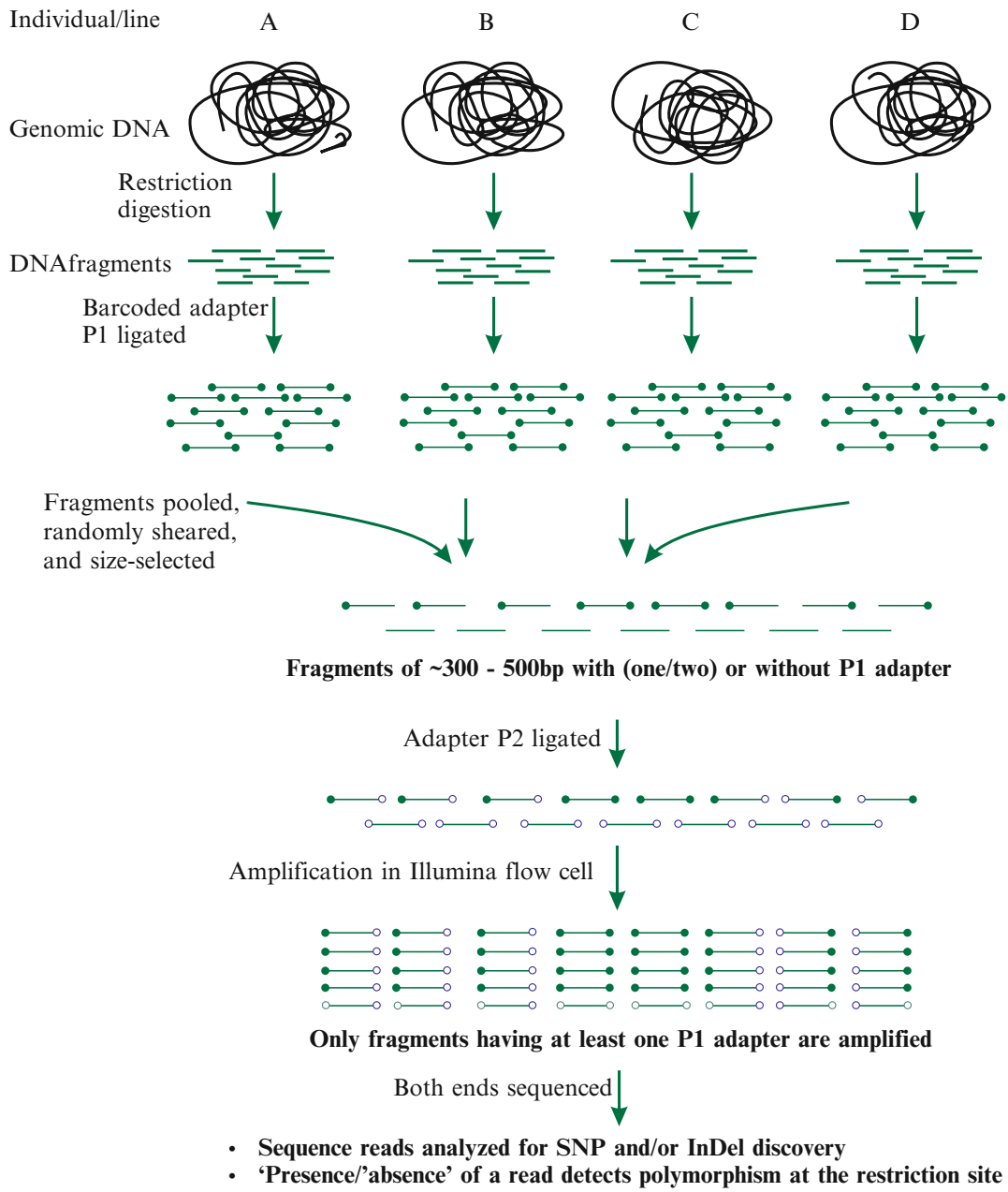


Fig. 13.13 A simplified representation of RAD-Seq approach for SNP discovery and genotyping. *Solid circles* represent P1 adapters and *open circles* denote P2 adapters. A fragment can be amplified only when it has at least one P1 adapter (having the barcode) (Based on Baird et al. 2008; Davey et al. 2011)

Any laboratory equipped for DNA extraction and restriction digestion of DNA can perform ezRAD analysis by sending the digests to an Illumina sequencing facility that includes library preparation as a part of the service. Any restriction enzyme or a combination thereof may be used to generate fragments of the desired size. This method generally targets a single restriction site; when two restriction enzymes are used, they are usually isoschizomers to eliminate the effects of

Table 13.3 Comparison among various RAD methods used for discovery and genotyping of DNA sequence polymorphisms

Feature	RAD tag	ezRAD	ddRAD	2b-RAD	GBS	2-enzyme GBS
Number of restriction enzymes used	One	One or two[a]	Two	One	One	Two
Targeted restriction sites	One	One	Two	One	One	Two
Type of enzyme(s)	Rare cutter	Frequent cutter	Frequent cutter	Frequent cutter	Rare or frequent cutter	Rare plus frequent cutter
Fragment shearing	Yes	No	No	No	No	No
Fragment size selection	Yes	Yes	Yes	No	No	No
Expertise and time needed for library preparation	High	Low	Moderate	Moderate	Moderate	Moderate
Barcode	Used	Not used	Used	Used	Used	Used
Library preparation cost per sample	Low	Moderate	Very low	Low	Moderate to very low	Moderate to very low
DNA/sample	300 ng	1 μg or less	100 ng	–	100 ng	–
Initial expenditure	High	Very low	High	High	High	High
Scalability[b]	Low	Low	Moderate	Moderate	Low	Low

Based on Toonen et al. (2013) and others
[a]When two enzymes are used, they are isoschizomers
[b]Scaling up of an operation reduces overall cost of the operation

restriction site methylation. The high molecular weight genomic DNA is digested with the selected enzyme(s), the fragments are end repaired, their 3′ ends are adenylated, and then the TruSeq adapters are ligated to them. Reagents can be saved by carrying out all the reactions in one-third of the recommended reaction volume for the TruSeq protocol. Then, size selection for 400–500 bp (including the ~120 bp adapters) fragments is implemented. There is no sonication step, and fragments from over 20 individuals can be pooled. The fragment size can be optimized by enzyme selection, and the number of fragments sampled from each individual can be manipulated by the size selection step. This method samples a set of the restriction sites present in the genome for the selected enzyme. It can be used with non-model organisms since a reference genome is not required for ezRAD. Thus, ezRAD is especially suited for SNP discovery and targeted amplicon sequencing in species with little genomic resources (Toonen et al. 2013).

A modification of RAD-Seq, called *double digest restriction-site-associated DNA sequencing*, is highly repeatable and randomly samples hundreds to hundreds of thousands of regions from a set of regions of the concerned genomes. The library preparation for ddRAD-Seq is considerably cheaper and much faster (<8 h hands-on time for dozens to hundreds of samples), requires less DNA (<100 ng) than RAD-Seq, and can be performed in microtiter plates. The library preparation is based on simultaneous digestion of genomic DNA with two restriction enzymes, and eliminates random shearing used in RAD-Seq. This reduces the cost of library preparation five-fold ($5 per sample compared to $25 for RAD-Seq). The size selection procedure is precise and repeatable; it selects only such fragments of the specified size that are produced due to one cut by each of the two enzymes. As a result, the fragments are randomly sampled from the same set of genomic regions of all the individuals, while they represent a random sample of genomic regions in the case of RAD-Seq. This sampling procedure selects for only a small proportion (<5 %) of the fragments and reduces the total number of sequence reads needed per sample to achieve a

given sequencing depth. As a result, a highly multiplexed library construction and sequencing becomes feasible. It also makes ddRAD-Seq robust to under-sampling of read counts, which is a common problem in pooled sequencing studies. It uses a combinatorial indexing scheme that allows dozens to hundreds of individuals to be pooled for sequencing in a single lane. The restriction fragments from different individuals are separately ligated to a small number of barcoded adapters. The fragments from different individuals are then pooled, size selected, and amplified using a primer with an index as per the Illumina multiplexed paired-end sequencing protocol. Several pools, each using uniquely indexed PCR primers, can now be pooled to achieve a high level of multiplexing without incurring the high cost associated with the use of a unique barcode for each individual. The open source computational pipeline developed for the ddRAD-Seq sequence data can simultaneously discover and genotype sequence polymorphisms with or without a reference genome. In case of de novo sequencing, the program incorporates 30–50 % more short reads than Stacks (Sect. 13.11.3); this is comparable to the efficiency reached with the use of a reference genome. Further, its graph clustering procedure can group haplotypes with any number of mismatches (Peterson et al. 2012).

13.6 Low-Coverage Genotyping

Typically, SNP discovery is based on deep sequencing (>30× sequencing depth) of a small number of individuals. This approach is adequate for discovery of markers for genes producing large effects. But it cannot be effectively used for detecting minor allelic variants and for association studies, where sequencing of up to thousands of individuals is required. It has been shown that genomic DNA from many individuals can be pooled and sequenced at low coverage. This sequence data can be analyzed to generate accurate genotype calls in those regions of the genome that have been sequenced in at least some of the individuals of a sample (Li et al. 2011b). In general, the accuracy of genotype calls increases with the number of individuals included in the pool provided the sequencing depth remains the same. For example, for a minor allele with frequency of >0.2 %, a 4× sequencing depth for a pool of >3,000 individuals provides detection power similar to that obtained from sequencing of 2,000 individuals at 30× depth. But the sequencing effort for the low-coverage genotyping would be only ~20 % of that for the deep sequencing approach.

The hidden Markov model (HMM) can be used for analysis of the sequence data across many individuals for imputation of missing genotype data (Li et al. 2011b). The HMM method identifies those stretches of chromosomes that are shared among individuals and uses this information to call genotypes for the genomic stretches missing in some of the individuals. The HMM method uses the following logic: if a pair of chromosomes from two different individuals shares a series of alleles flanking a site missing in one of these chromosomes, this site is likely to have the same alleles that are present in the other chromosome. Thus, with even a 2−4× sequencing depth, common as well as low-frequency SNP loci can be discovered and genotyped with high confidence. The low-coverage sequencing can be used to develop a reference panel, which would enable genotype imputation of additional individuals to further increase the detection power. The low-coverage sequencing of more individuals generates more power than that of high-coverage sequencing of fewer individuals. But comprehensive detection of very rare variants requires deep (20−30× depth) sequencing.

In low-coverage sequencing, only a subset of genomic regions will be sequenced in any individual, and these regions would differ among individuals. Therefore, markers sequenced at sufficient coverage and having known positions on a reference genome are used to impute genotypes of the markers not sequenced in an individual and to infer recombination breakpoints. This approach is suitable for genotyping of recombinant populations provided the genotypes of the parental lines/individuals

are either known or probabilities can be assigned to them. There are two main schemes for low-coverage genotyping: (1) genotyping by sequencing (GBS) and (2) multiplexed shotgun genotyping (MSG).

13.6.1 Genotyping by Sequencing

In *genotyping by sequencing* strategy, a frequent cutting restriction enzyme is used to digest the genomic DNA from each individual of the sample. One adapter with barcodes and a common adapter without barcodes are mixed and ligated to the fragments. As a result, a given fragment will have either the barcoded adapter or the common adapter at both the ends, or it will have a barcoded adapter at one end and the common adapter at the other end. The fragments from all the individuals are pooled and bridge amplified on an Illumina Genome Analyzer flow cell. For efficient amplification, a fragment has to be >1 kb in size and should have the common adapter at one end and the barcoded adapter at the other end (Fig. 13.14). Thus, a substantial proportion of the fragments will be filtered out, but still a large number of fragments will be retained. One end of all such fragments is sequenced, and the reads are mapped to a reference genome for calling SNPs. But a complete reference genome is not essential, and one may treat the consensus of the read clusters developed during the genotyping process itself as the reference genome (Elshire et al. 2011; Sonah et al. 2013).

The restriction enzyme may be selected for producing a large proportion of correct size fragments, or it may be a methylation-sensitive enzyme. Repetitive genomic regions can be avoided by using a methylation-sensitive enzyme like *Ape*KI; this also enriches the fragments for lower copy regions of the genome (Elshire et al. 2011). The use of a combination of two restriction enzymes further reduces the complexity (Poland et al. 2012) or a selection primer for the final amplification step. In the latter case, the primer extends one or two bases beyond the *Ape*KI site into the DNA fragment, and the additional nucleotides serve as selection nucleotides. This step reduces complexity, increases coverage depth, and permits greater multiplexing without reducing the coverage depth. The *depth of coverage of SNPs* denotes the number of reads containing a given SNP locus and depends on the number of copies of the given genomic sequence in the GBS library. The use of two restriction enzymes, one "rare cutter" and one "common cutter," e.g., *Pst*I and *Msp*I, leads to a greater degree of complexity reduction than when only *Ape*KI is used. In case of two enzymes, the barcode is attached to the adapter for the "rare cutter" (the forward adapter), while the adapter for the "common cutter" is the specially designed "Y adapter" (the reverse adapter). Only the fragments having the barcoded forward adapter at one end and the common reverse adapter at the other end will be amplified.

The GBS library construction is simple and suited for use with large numbers of individuals/lines. The genomic DNA is digested and the adapters are ligated in the same well, and 96–384 samples can be processed simultaneously. Further, only relatively small amount of DNA (100 ng) per individual/line is required. The GBS procedure leads to increase by several orders of magnitude in the sequencing throughput as well as permits multiplexing. This approach is good where a reference genome is available, but it can be used for de novo sequencing in case a reference genome is unavailable. In the latter case, the putative SNPs must be validated using a suitable assay. In case a reference genome is available, in silico mapping of the sequence reads onto the reference genome serves as validation. GBS can be used for SNP discovery in polyploid crops. One of the chief limitations of GBS is the large numbers of missing data; some imputation models, e.g., BEAGLE v3.0.2 and IMPUTE v2, have been developed to resolve this problem.

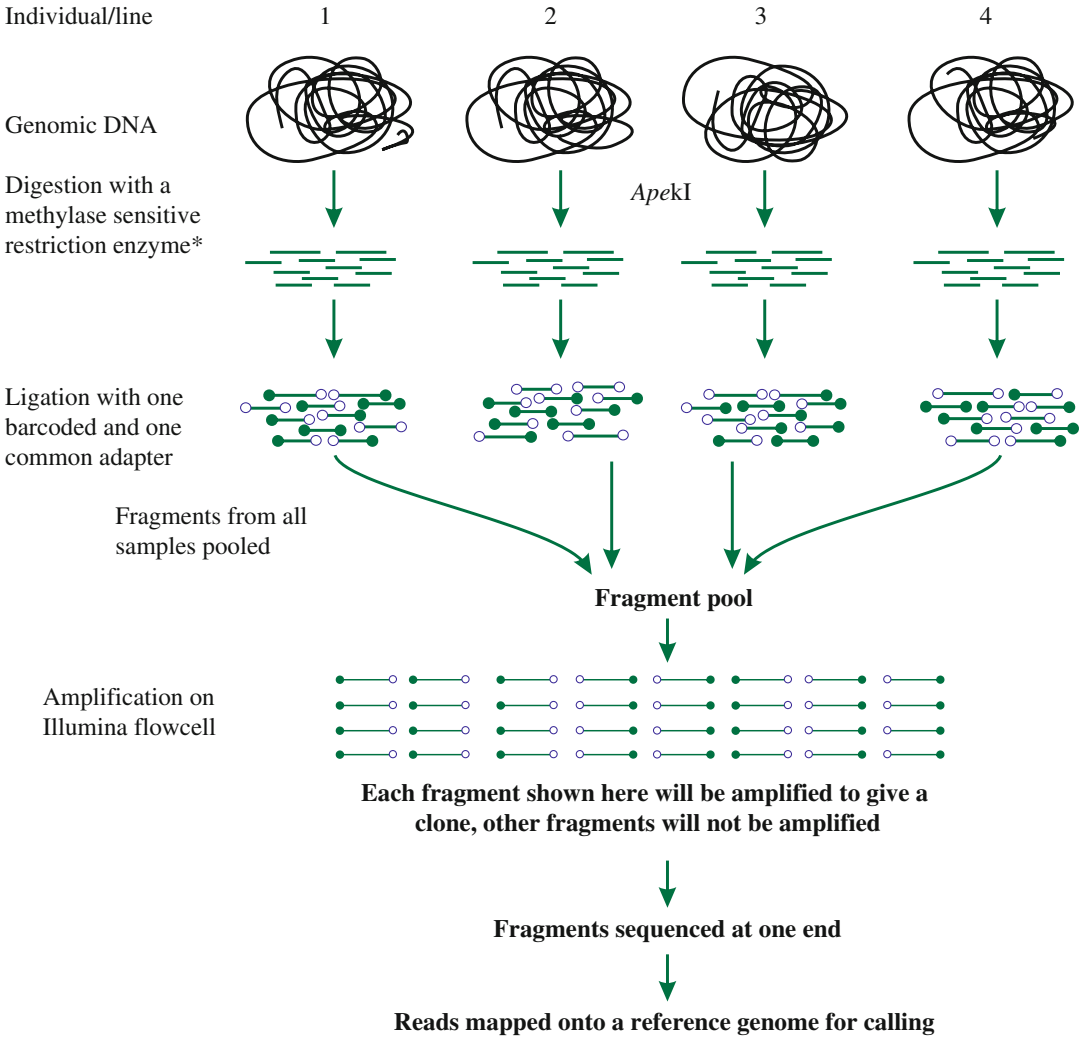


Fig. 13.14 A schematic representation of genotyping by sequencing (GBS) approach for SNP discovery and genotyping. *Solid circle*, barcoded adapter; *open circle*, common adapter. Only those fragments that are less than 1 kb and have the barcoded adapter at one end and the common adapter at the other will be amplified (Based on the procedure of Elshire et al. 2011). * A combination of two restriction enzymes can also be used for genomic DNA digestion

13.6.2 Multiplexed Shotgun Genotyping

In *multiplexed shotgun genotyping*, restriction fragments from several individuals are separately ligated to distinct barcodes, pooled, and sequenced using a NGS platform. The sequence reads are analyzed to discover SNPs as well as carry out genotyping of the individuals in the pool. Andolfatto et al. (2011) separately digested genomic DNAs from 96 F_1-backcross progeny from a *Drosophila* interspecific cross with the frequent cutter *Mse*I. The barcoded *Mse*I adapter was ligated to the fragments, and the fragments were then pooled and subjected to size selection (250–300 bp). Now, the standard Illumina adapters were ligated to the fragments. The fragments were sequenced; the sequence reads were grouped on the basis of barcodes and mapped onto the two parental genomes for

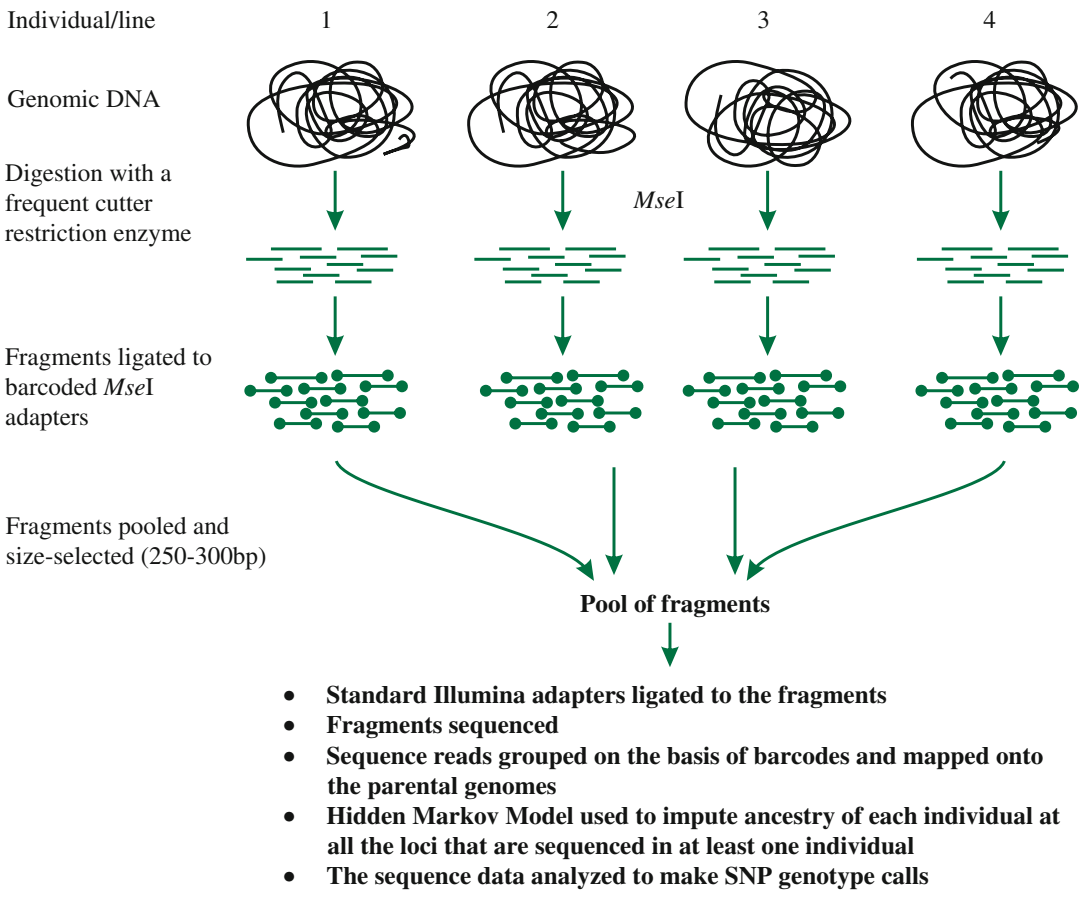


Fig. 13.15 A schematic representation of the multiplexed shotgun genotyping (MSG) for SNP discovery and genotyping (Based on the procedure of Andolfatto et al. 2011)

calling SNPs (Fig. 13.15). But the genome sequence of only one parent or even a reference genome can be used for this purpose. The sequence reads for any individual are sparsely distributed over the whole genome, and only few individuals of the population are sequenced for a given genomic region. Therefore, HMM was used to impute the ancestry for every individual at all such genomic locations that were sequenced in at least one individual of the pool. The logic for this approach is that the genomes of the parental strains are syntenic with the reference genome, and their sequences are sufficiently similar. This similarity permits mapping of the short reads onto the reference genome with high accuracy and imputation of ancestry at most of the genomic locations on the basis of the linkage relationships. The HMM was used to make "soft" probability-based ancestry assignments, which is well suited for the MSG sequence data. The MSG procedure has a much-simplified DNA sample preparation protocol, needs very small amount (10 ng) of DNA per sample, and is completed in ~2 days for 96 individuals/lines.

13.7 Applications of NGS-Based Marker Discovery and Genotyping Methods

The different NGS-based methods provide marker data with different density and depth of coverage and, consequently, are suited for different types of studies. For example, studies of wild populations are without a reference genome sequence. Therefore, it is desirable that a large

number of markers are scored accurately in most individuals to enable a precise estimation of population parameters. For such studies, RAD-Seq, ddRAD-Seq, RRLs, and CRoPS would be the most appropriate. But in case of experimental populations, usually a reference genome is available, polymorphism is relatively limited, and parental genotypes are known. In such populations, therefore, GBS and MSG would serve the purpose since the linkage relationships can be easily inferred. Therefore, GBS and MSG are well suited for marker-assisted selection and QTL mapping studies (Table 13.2; Davey et al. 2011).

13.8 A Comparison of NGS and Other SNP Genotyping Approaches

The array-based and other high-throughput SNP genotyping approaches are applicable to already identified SNP loci. In addition, the sequences of the flanking genomic regions of these loci must be known to permit the designing of suitable assays for them. In contrast, the NGS-based approaches (RRL, CRoPS, RAD-Seq and its variations, GBS, and MSG) allow simultaneous discovery and genotyping of up to hundreds of thousands of SNP loci. When a new population is genotyped for the same markers, the genomes of individuals of this population are sequenced afresh; this avoids the risk of bias in marker identification and genotyping. Davey et al. (2011) concluded that the cost of SNP genotyping using NGS technology was higher than that of genotyping of already identified SNPs using existing SNP arrays. Therefore, existing microarrays and other assays (Sect. 13.2) will continue to be used for most projects of individual laboratories. But where an array is not available, the development of new SNP arrays would be economical only for large consortia since the array would be used by several laboratories. But for individual laboratories and small groups of researchers, the cost of sequencing using one of the reduced representation approaches would be far lower than that of array development. Alternatively, a small panel of SNP or SSR markers may be developed by sequencing, and genotyping may be carried out using a suitable strategy (Chaps. 3 and 4).

13.9 Reduced Representation Versus Whole-Genome Sequencing

An experimenter has to decide whether he/she should use whole-genome resequencing or one of the reduced representation approaches for SNP discovery and genotyping. This decision will mainly depend on the available financial resources and, to some extent, on the density and accuracy of markers required for the study. The consideration of sequencing cost is markedly influenced by the genome size and the availability of a good quality reference genome. Whole-genome sequencing is much more expensive than the reduced representation approaches. Around the year 2011, a conservative estimate of the cost of complete sequencing of the human genome of ~3 Gb at 30× coverage was around UK£5,000. At this rate, the cost for sequencing of a population of 100 individuals will be about £500,000. The 30× coverage may be low for a species lacking a reference genome because de novo assembly of quality reference genome sequence from short reads of NGS platforms is not easy. But RAD-Seq avoids the genome assembly process and could sample 200,000 SNP markers in 100 humans at 30× coverage at a cost of merely about £14,000 (a 35-fold reduction in cost), while the cost for MSG and GBS would be merely around £1,000 (a 500-fold reduction in cost; Davey et al. 2011).

The sequencing throughput has been doubling every 5 months. If this trend continues, whole-genome sequencing was projected to cost, by the end of 2013, the same as the RAD-Seq approach in 2011, i.e., £14,000 for 100 humans at 30× sequencing depth. It may, therefore, be expected that whole-genome sequencing of populations for marker discovery will soon become affordable to most laboratories (Davey et al. 2011). However, reduced representation methods are likely to remain in use for following reasons: (1) much

lower cost, (2) easier analysis of larger populations than whole-genome sequencing, (3) generation of adequate marker data for many studies, and (4) the likelihood of enhanced quality of the marker genotypes with improved algorithms.

13.10 SNP Discovery in Polyploids

A polyploid species has two or more homoeologous or similar genomes. As a result, genes of such species have two or more similar copies located in the homoeologous chromosomes. Therefore, a proportion of SNPs detected in a polyploid species will be due to differences in the sequences of paralogous copies of genes present in the homoeologous chromosomes of the same individual. Such SNPs are called hemi-SNPs, and they generate artificial polymorphism. The biggest challenge is to separate the hemi-SNPs from true SNPs, which represent differences in the sequences of the alleles of the same genes usually present in different individuals. The following strategies have been used for SNP detection in polyploid species: (1) comparison of short EST sequences with the transcript assemblies from a genotype of the same species or from several related species, (2) candidate gene amplicon resequencing, and (3) identification of SNPs in a known diploid progenitor species. The reference sequence assemblies should have sufficient depth. Further, appropriate filtering algorithms are used to eliminate the hemi-SNPs. The above strategies have been used in such polyploid species as *Brassica napus*, *Avena sativa*, *Saccharum officinarum* (sugarcane), and *Triticum aestivum* (SNPs discovered in *Triticum tauschii*) (Deschamps and Campbell 2013).

13.11 Bioinformatics Tools for Marker Discovery from NGS Sequence Data

Several software packages have been developed for handling the NGS sequence data generated following various reduced representation schemes. These packages were originally designed for a particular reduced representation scheme, but they can be adapted for processing of data from other schemes as well.

13.11.1 PoPoolation

The *PoPoolation* software is designed for analysis of pooled NGS sequence data to discover nucleotide polymorphism. The sequence reads are analyzed in several steps, e.g., the reads are trimmed for low-quality bases, the Burrows–Wheeler Alignment (BWA) tool is used to map the reads against a reference genome, the aligned reads are converted into a pileup file using SAMtools, and then this file is used for genome-wide polymorphism analysis with the PoPoolation tool. This tool needs at least a partially assembled reference genome. The trimming step reduces sequencing errors to one-tenth. A 40-fold coverage seems to be sufficient for reliable polymorphism detection. PoPoolation is not recommended for use with cDNA or unigenes and pooled RNA samples. This tool provides several output options, e.g., display of a simple graphical view of the polymorphism present in the entire chromosome, a display in FlyBase and the UCSC Genome Browser, etc. Its source code can be downloaded from http://code.google.com/p/popoolation/ (Kofler et al. 2011).

13.11.2 RADtools 1.2.4

RADtools 1.2.4 software is designed for processing of RAD-Seq sequence data for discovering SNPs and structural variations. This software discovers candidate genetic markers from Illumina sequence reads. RADtools is compatible with the Linux and Mac OSX operating systems.

13.11.3 Stacks

An extension of the *Stacks* package carries out efficient analysis of GBS sequence data. The

Stacks (http://creskolab.uoregon.edu/stacks/) software analyzes RAD-Seq and GBS sequence data generated by using any restriction enzyme. It uses a sliding window algorithm for sequence data analysis against a reference genome. But Stacks can also de novo assemble the short reads. The analysis by Stacks is highly integrated: it begins with raw sequence reads, progresses through various steps to nucleotide variant allele and genotype calling, and finally generates the core population genetics statistics outputs. Stacks is robust, efficient and flexible and can handle sequence data from thousands of individuals. The MySQL database permits efficient data visualization, management, and modification. The input data can be in FASTA or FASTQ format, and there are several output format options. Stacks is a collection of several programs that can be run either individually or together with the help of a wrapper program (Catchen et al. 2013).

13.11.4 TASSEL

The *TASSEL* software package has multiple functions. It includes the *TASSEL-GBS* pipeline for discovery of SNPs from GBS sequence data. It can score millions of SNPs in up to 100,000 individuals. It can use even an unfinished reference genome sequence for SNP discovery. This tool comprises a discovery and a production pipeline. The discovery pipeline uses all the FASTQ files available till date for the given species to discover SNPs. These SNPs are ascertained, filtered, and stored in a physical map, called TOPM, in a "production-ready" state. The discovery step is performed periodically to keep the TOPM up to date. The Production pipeline uses the TOPM to quickly call SNPs and genotypes. TASSEL can be used with any operating system. Its use does not require expertise in statistics or computer science. Its command-line version, called the pipeline, permits users to program tasks using a script. It can be downloaded free from http://www.maizegenetics.net/tassel (Glaubitz et al. 2014).

13.11.5 SAMtools/BCFtools

SAMtools/BCFtools package is designed for discovery of SNPs, short InDels, and structural variations from NGS sequence data by using a reference genome sequence. The SAMtools collects summary information from the sequence data, computes the likelihood of obtaining the data for each possible given genotype, and stores the likelihood data in BCF (binary variant call) format. The BCFtools calls SNPs and InDels using the likelihood data. It is recommended for InDel discovery using sequence data from NGS technologies like Illumina that have low InDel error rate. This package allows the use of several functions like C50 (it reduces the effect of such reads that have excessive mismatches), D and S (it controls the read depth per sample as well as strand bias), etc. to increase the precision of variant calling. But this package is unable to properly handle data for variants having multiple alleles (http://samtools.sourceforge.net/mpileup/shtml).

13.12 Future Directions

It may be expected that improvements in the NGS marker discovery protocols will increase the quality and accuracy of marker sets. Further, some variant applications of these strategies would also be developed. For example, cDNA may be digested with restriction enzymes to identify a small set of markers from the transcriptome. The cDNA-based markers can be used for analysis of gene expression without transcriptome assembly. Improvement in the analysis of sequence data would be quite rewarding as this would increase the reliability of marker discovery and genotyping. Increased sequencing throughput would allow sequencing at greater depths, and this would increase the sequence accuracy and discovery of more markers per individual. Although whole-genome sequencing of populations is becoming increasingly feasible, the reduced representation protocols may remain useful, especially in the cases of non-model organisms (Davey et al. 2011).

Questions

1. Briefly describe any two of the whole-genome-based array platforms for SNP genotyping and discuss their merits and limitations.
2. Briefly describe one SNP genotyping platform each with low, moderate, and high multiplexing capability and discuss their merits and limitations.
3. "SNP marker system is amenable to high-throughput genotyping". Discuss this observation in the light of relevant information.
4. "NGS technologies have permitted SNP genotyping to be combined with SNP discovery". Discuss this statement in the light of available information.
5. Briefly describe the Illumina GoldenGate technology for SNP genotyping highlighting its features, merits, and limitations.
6. Briefly describe the reduced representation sequencing approaches for SNP genotyping and compare them with low-coverage genotyping approaches.
7. Briefly describe the RAD-Seq method for SNP genotyping, briefly outline some of its modifications, and highlight their distinguishing features, merits, and limitations.
8. Give a brief description of some of the software programs used for SNP discovery and genotyping using NGS sequence data.
9. "In near future, reduced representation sequencing for SNP genotyping may become redundant". Evaluate this observation with the help of available relevant information.

Bioinformatics Tools and Databases for Genomics Research

14

14.1 Introduction

Bioinformatics involves the development of statistical tools and techniques and computer software for acquisition, storage, analysis, and visualization of biological information. The term "bioinformatics" has been derived by the fusion of the terms "biology," "information technology," and "statistics." The discipline of bioinformatics has the following three main activities: (1) development of new algorithms and statistical techniques for the assessment of relationships among enormous biological datasets, (2) use of these tools and techniques for analyzing and interpreting the huge biological datasets, and (3) development of databases for an efficient storage and management of the huge amounts of information, and fast search, retrieval and/or analysis of the desired data. Bioinformatics evolved because new tools and techniques were necessary to handle the enormous amino acid and nucleotide sequence data being generated. During the early 1960s, the National Biomedical Research Foundation compiled the first comprehensive collection of amino acid sequences. The European Molecular Biology Laboratory (EMBL) organized their collection of data on nucleotide sequences in 1980; the European Bioinformatics Institute (EBI), Hinxton, UK, now maintains this nucleotide sequence database. The National Center for Biotechnology Information (NCBI), USA, was created during the early 1980s. Sometime later, the DNA Data Bank of Japan (DDBJ) was established. In 1984, the National Biomedical Research Foundation established the Protein Information Resource (PIR), which identifies and interprets the data on amino acid sequences.

14.2 Representation of Nucleotide and Amino Acid Sequences

The amino acid and nucleotide sequences are reduced to digital data. This is greatly facilitated by the use of single-letter codes for the amino acids and the organic bases (Tables 14.1 and 14.2). It may be noted that in RNA sequences the symbol U is used in the place of T. Even those amino acid/base positions that exhibit ambiguity can be adequately represented by single-letter codes. In case of DNA, the sequences of the two complementary strands of a DNA duplex are represented by the symbols for complementary bases, which can be deduced either manually (for short sequences) or by using a computer software. In databases, the nucleotide sequences are listed from the 5′ end (at the extreme left of the written sequence) to the 3′ end of a single strand. The representations of amino acid sequences of protein molecules begin at their N-termini and proceed to their C-termini.

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_14, © Author(s) 2015

Table 14.1 Single-letter codes for different bases found in nucleotide sequences

Symbol	Meaning	Logic for the symbol	Symbol for the complementary base
A	Adenine	***A***denine	T
C	Cytosine	***C***ytosine	G
G	Guanine	***G***uanine	C
T	Thiamine	***T***hiamine	A
R	G or A	Pu***r***ine	Y
Y	C or T	P***y***rimidine	R
M	A or C	A***m***ino group (bases having)	K
K	G or T	***K***eto group (bases having)	M
S	G or C	***S***trong base pairing	S[a]
W	A or T	***W***eak base pairing	W[a]
H	A, C, or T	Not G[b]	D
B	C, G, or T	Not A[b]	V
V	A, C, or G	Not U[b]	B
D	A, G, or T	Not C[b]	H
N	A, C, G, or T	***N***ucleotide	N

The codes are based on the recommendations of International Union of Pure and Applied Chemistry (IUPAC)
[a]The same symbol is used for the base on the complementary strand since G pairs with C (symbol S denotes both G and C), while A pairs with T (W denotes both)
[b]Not G, letter H comes immediately after letter G in the alphabet; not A, letter B is the next letter to A; not U, letter V follows letter U (denotes T in DNA); not C, letter D occurs just after C

Table 14.2 Single-letter symbols for different amino acids in protein sequences

Symbol	Amino acid	Three letter code
A	Alanine	Ala
B	Asparagine or aspartic acid	Asx
C	Cystine	Cys
D	Aspartic acid	Asp
E	Glutamic acid	Glu
F	Phenylalanine	Phe
G	Glycine	Gly
H	Histidine	His
I	Isoleucine	Ile
K	Lysine	Lys
L	Leucine	Leu
M	Methionine	Met
N	Asparagine	Asn
P	Proline	Pro
Q	Glutamine	Gln
R	Arginine	Arg
S	Serine	Ser
T	Threonine	Thr
V	Valine	Val
W	Tryptophan	Trp
Y	Tyrosine	Tyr
Z	Glutamine or glutamic acid	Glx
X	Any amino acid	Xaa

The codes are based on the recommendations of International Union of Pure and Applied Chemistry (IUPAC)

14.3 Bioinformatics Tools

The genome-sequencing projects triggered the development of such high-throughput technologies that generated sequence data at an unprecedented rapid pace. This necessitated the development of computer programs capable of acquiring, analyzing, classifying, and storing very large volumes of data and retrieving the desired data from the stored data. As a result, the computer hardware capabilities had to be greatly enhanced, new statistical techniques needed to be developed, appropriate computer programs were designed, and suitable data storage and management systems were also implemented. The various computer programs used for the acquisition and analysis of data and detection of associations and patterns as well as to achieve other specific objectives are often referred to as *bioinformatics tools* or simply as *tools*. A large number of different tools is available for achieving a variety of objectives. Some of the tools used for marker discovery and development, gene prediction, association analyses, data storage and management, etc. are briefly described in the following sections.

14.3.1 AutoSNP

Nucleotide sequences of expressed sequence tags (ESTs) can be analyzed to discover single-nucleotide polymorphisms (SNPs). Such SNPs are of great biological significance since they are based on the exons of expressed genes. The AutoSNP computer program carries out automated analysis of EST sequence data and identifies SNPs as well as insertion/deletion (InDel) variations present in them. It aligns the EST sequences and distinguishes between predicted SNPs and sequencing errors on the basis of the redundancy criterion. A putative SNP will be present in multiple reads, while a sequencing error would occur in one or two reads. For each candidate SNP, redundancy score and co-segregation score are estimated. The redundancy score of a predicted SNP locus is the frequency of polymorphism at this locus. The co-segregation score is the likelihood that the predicted SNP will be transmitted together with the other SNPs present in its vicinity in the EST sequence. The AutoSNP output includes the predicted SNPs and InDels along with their redundancy and co-segregation scores. Most of the SNPs and InDels predicted in maize using the AutoSNP tool were validated as true SNPs and InDels. The AutoSNP program is available to research workers free of cost on request to the authors of the program (email: dave.edwards@nre.vic.gov.au; Barker et al. 2003).

The SNPServer (http://hornbill.cspp.latrobe.edu.au/snpdiscovery.html) is a Web interface for using AutoSNP, BLAST, and CAP3 programs for SNP discovery in real time (Savage et al. 2005). BLAST identifies related EST sequences, CAP3 aligns and clusters these sequences, while AutoSNP analyzes the alignments to identify SNPs and InDels. The results from this SNP discovery pipeline, the source of the EST data, as well as their annotation are stored in autoSNPdb. This database can be accessed for free at http://autosnpdb.qfab.org.au/. The database has SNP data on rice, barley, and *Brassica* spp. AutoSNPdb allows identification of SNPs and InDels in specified genes or genes related to specific traits, and between genes of specified pairs/groups of plant varieties. A user-friendly GUI (graphical user interface) enables easy visualization of the SNPs in the database (Duran et al. 2009). Another tool, QualitySNPng, uses a haplotype-based strategy to enable the visualization and detection of SNPs from NGS data, and it does not require a fully sequenced reference genome (http://www.bioinformatics.nl/QualitySNPng/).

14.3.2 SNP2CAPS

The CAPS markers are valuable cost-effective tools for analysis of SNP and InDel polymorphisms in laboratories that are not highly equipped. This is particularly true when common restriction enzymes are used to analyze the CAPS markers. It is quite difficult to manually convert SNP markers into CAPS markers. The computer program dCAPS Finder 2.0 can be used to design PCR primers with such mismatches that either create a restriction site at the selected SNP locus or remove a site existing at the locus. This facilitates the conversion of SNPs to CAPS markers, but designing of such primers successfully is not a simple issue. The program SNP2CAPS (for SNP-to-CAPS) screens multiple aligned sequences for polymorphic restriction sites, analyzes these sites, and identifies such sites that are the most likely candidates for CAPS marker development. This generic program also evaluates the restriction enzymes for their suitability for CAPS analysis in the submitted sequences and selects those enzymes that show at least one restriction site polymorphism in each of the aligned sequences (Thiel et al. 2004).

When polymorphism at a restriction site (Fig. 14.1a) results from an authentic SNP, the restriction site will have an unambiguous sequence, i.e., it will consist of A, T, C, and G only (Fig. 14.1b). In some cases, however, there may be ambiguity (symbol N) at one or two positions in addition to the SNP polymorphism (Fig. 14.1c); however, this is unlikely to affect the CAPS development. Therefore, the above cases would be good candidates for CAPS development. But restriction site polymorphisms may arise merely due to an ambiguous sequence, in which case N would be present within the

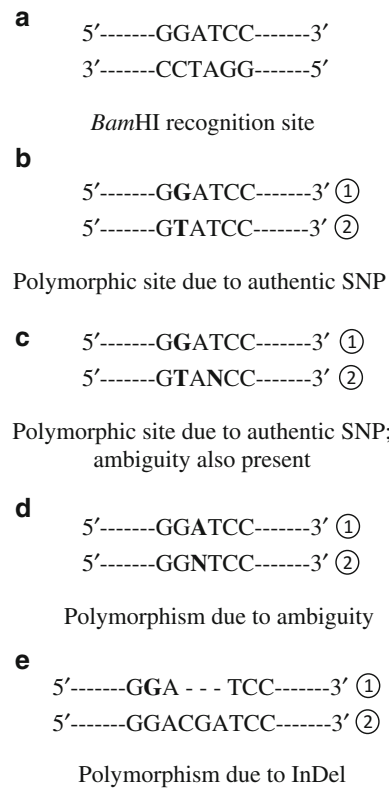


Fig. 14.1 Polymorphism in *Bam*HI recognition site. (**a**) A normal recognition site. (**b**) Polymorphism produced by SNP. (**c**) Polymorphism generated by SNP, but sequence ambiguity (N) is also present. (**d**) Polymorphism due to sequence ambiguity (N). (**e**) Polymorphism caused by InDel within the recognition site. The polymorphisms depicted in (**b**), (**c**), and (**e**) are good candidates for CAPS marker development. 1 and 2 are sequences of the same strand from two different individuals

restriction site sequence in some of the aligned sequences (Fig. 14.1d). Such restriction site polymorphisms are not suitable for CAPS development. In addition, insertion (or deletion) of one or more nucleotides into (or from) the restriction site will also generate restriction site polymorphism (Fig. 14.1e); these would be useful for CAPS development. Thus, the SNP2CAPS program analyzes the submitted sequences and identifies the recognition site polymorphisms suitable for CAPS marker development. The input for SNP2CAPS program is the multiple sequence alignment of the target sequences from different accessions. This input file may be in modified FASTA, ClustalW, MSF, MEME, etc. formats. In addition, it needs an input of restriction enzyme data, which can be downloaded from REBASE (the restriction enzyme database; http://rebase.neb.com/). A high proportion (90 %) of multiple aligned sequences of barley ESTs contained SNPs and InDels; they also had one or more restriction sites that were polymorphic. Further, over 30 % of these polymorphic restriction sites were for ten common restriction enzymes. SNP2CAPS offers a command line as well as a GUI. The SNP2CAPS is freely available from the website http://pgrc.ipk-gatersleben.de/snp2caps/.

14.3.3 TASSEL

The results from association analyses are often confounded by factors like selection, population structure, and family relationships, which may lead to incorrect marker–trait associations. The GLM and MLM approaches were developed to minimize the effects of population structure and/or family relationships on the findings from association studies. The GLM and MLM methods have been implemented in the software TASSEL (*T*rait *A*nalysis by a*SS*ociation, *E*volution and *L*inkage). The GLM method uses a structured association analysis based on a Q matrix to minimize the probability of false associations. The *Q matrix* reflects population structure and is computed by using the STRUCTURE program (Sect. 14.3.4) or by the principal components analysis (PCA) method. The MLM method uses in its model the kinship (K) matrix as well as the Q matrix in an effort to further reduce the risk of discovering false-positive associations. The estimates of K matrix representing the average relatedness between pairs of individuals/lines can be obtained from pedigree information or from genotype data for a large number of unlinked markers covering the whole genome of the organism. TASSEL carries out F-tests and permutation tests and estimates model effect means. When the trait in question does not have normally distributed residual error, some transformation function may be used to generate roughly normal error terms, or a permutation test may be used to generate distribution-independent p-values.

The TASSEL program can handle datasets from plant, animal, and human populations. It

enables the estimation of linkage disequilibrium (LD) as D' and as r^2 and allows graphical visualization of these estimates. Other features of this program include analysis of InDels, diversity analysis, execution of PCA, and imputation of missing data. This package includes several tools for extraction and visualization of data like sequence alignment viewer, neighbor-joining cladogram construction, and many data graphing functions. It has many data management functions and a data browser that provides an interface to relational databases. This software is in Java and is compatible with Windows, Mac, and Linux operating systems. The TASSEL executables, user manual, etc. are available for free from http://www.maizegenetics.net/tassel (Bradbury et al. 2007).

14.3.4 STRUCTURE

The *STRUCTURE* software (ver. 2.3.4 in 2012) is capable of detecting the presence of two or more homogeneous groups within a single population (Pritchard et al. 2000a). A *homogeneous group* is a group of individuals that is at Hardy–Weinberg equilibrium for all of the several random markers. This program implements a Bayesian (Markov chain Monte Carlo) algorithm for model-based clustering of individuals genotyped for several unlinked markers. It can use data from most genetic markers, including SSRs, SNPs, and AFLPs. It attempts to find out the number of homogeneous groups most likely to be present in the given population. The investigator should aim to find the smallest number of groups that accounts for the major structure in the population marker data. It also generates estimates of Q, which depict the likelihood that an individual belongs to a particular cluster. An individual may get assigned to two or more groups if its Q values indicate it to share the genetic properties of these groups. The accuracy of such assignments depends on several factors, including the numbers of individuals genotyped in the sample, groups present in the sample, the marker loci scored; the amount of admixture in the population; and the extent of allele frequency differences among the groups in the sample. This program has been used for detecting genetic structures in the sampled populations, assigning the individuals to different groups of the sample, population admixture, hybridization analysis, etc. Most studies show that STRUCTURE efficiently assigns different individuals to the populations of their origin, particularly when the population has two to four well-differentiated homogeneous groups. After starting the STRUCTURE program in a random configuration, it is run for, typically, 10,000–100,000 steps in simulation and then for another 10,000–100,000 or more steps to get accurate estimates of Q. The program is run several times, each time assuming a different number of groups, ranging from one to ten, in the dataset.

The executables of the program are compatible with Mac, Windows, Linux, or Sun. The computational part of the program, written in C, has a Java front end, which provides several helpful features. The data file should be a text file, and the missing data should be indicated by a number, often −9, which is not used anywhere else in the data file. The STRUCTURE software is free and is available at http://pritch.bsd.uchicago.edu/software/structure2_1.html.

14.3.5 Microarray Software

Microarray technology is a sophisticated precision experimental tool for studying genome-wide gene expression patterns and levels. It generates large quantities of data that require well-designed, user-friendly software for acquisition, analysis, storage, and management. The *TM4 software* is a suite of the following four tools: (1) A *M*icro*A*rray *DA*ta *M*anager (MADAM) tool guides the user through the microarray procedure beginning from RNA isolation to the analysis of data. It also facilitates the entry of data in the database and provides a platform for launching other data entry and management tools. (2) The TIGR Spotfinder tool rapidly and reproducibly analyzes the microarray images as well as quantifies the levels of gene expression. (3) The *MI*croarray *D*ata *A*nalysis *S*ystem (MIDAS) normalizes and filters the data generated by the Spotfinder tool. Finally, (4) the *M*ulti*E*xperiment

*V*iewer (MeV) tool analyzes the gene expression data files and displays the gene expression and annotation information obtained from the microarray experiments. Analysis modules implemented in MeV include the following: PCA, clustering (hierarchical and k-means), self-organizing maps and trees, etc. Bootstrapping and jackknifing procedures are used to generate consensus clusters. In addition, TM4 has a MySQL-based database for storage of the relevant data. This database conforms to the *M*inimal *I*nformation *A*bout a *M*icroarray *E*xperiment (MIAME) standards. TM4 was developed for spotted two-color microarrays, but it can be easily modified for single-color microarray formats. TM4 can be used for a wide variety of biological systems, including plant, animal, and microbial species. It is an extensible, open-source software suite available for free to research workers (http://www.tigr.org/software). The MADAM, MIDAS, and MeV tools can run on Windows, Mac OS X, Linux, and Unix platforms, but the TIGR Spotfinder runs only on Windows (Saeed et al. 2003).

14.3.6 A C. Elegans Database (AceDB)

The *AceDB* database management system was originally designed to handle the data generated from the *Caenorhabditis elegans* genome project. It has many powerful tools for handling genomic and bioinformatics data, which have now been made much more flexible. As a result, they are now used for management of genomic databases of many organisms, including plants. The AceDB system can handle diverse data types, including those pertaining to maps (both genetic and physical maps) and DNA sequences, and it can be easily modified to handle new types of data. AceDB has a full GUI and uses plain text input files, which greatly facilitates the management and distribution of genomic data. It is easy to modify and extend by simple text editing of a single file; this makes AceDB an ideal research tool. The AceDB system can be fully operated by a single biologist. The AceDB database management system is still being used for developing biological databases. Precompiled executables of AceDB for UNIX, Windows, and Macintosh environments along with the relevant documentation are available at the website http://www.acedb.org/.

14.3.7 MAPMAN

Whole-genome gene expression analyses using microarrays and metabolite profiling based on mass spectrometry generate huge amounts of data covering several parameters. The chief limitation in proper exploitation of this data is their proper analysis and interpretation. MAPMAN tool displays the large datasets in form of diagrams that depict the concerned metabolic pathways or other cellular functions and processes; this facilitates the interpretation of these datasets. This tool has two modules, viz., the SCAVENGER and the IMAGEANNOTATOR modules (Thimm et al. 2004). The *SCAVENGER module* collects data on gene expression and metabolite levels and classifies them into hierarchical groups termed as "Bins" and "subBins." A *Bin* corresponds to a specific area of metabolism, e.g., photosynthesis. A Bin can be further divided into *subBins*, e.g., "light reactions," "photorespiration," and "Calvin cycle" in the case of "photosynthesis." The different Bins and subBins are given specific numerical codes reflecting their hierarchical relationships. A specific SCAVENGER module is designed for each type of data: separate modules are used for gene expression and metabolite data. The TRANSCRIPT-SCAVENGER module handles data from gene expression arrays; it sorts the genes into Bins and subBins on the basis of their function deduced from gene annotation information. The assignment of the data to Bins and subBins involves automatic recruitment as well as manual correction. The guiding principles for the assignment are as follows: (1) as many genes as possible, including those with "supposed" annotation, should be assigned to specific Bins, (2) Bin structure should be modified, if needed, to accommodate the relevant data, and (3) as far as possible, each gene should be placed into a single Bin and

subBin. The METABOLITE-SCAVENGER classifies the metabolites into different groups on the basis of either their structures or the pathways in which they occur. The IMAGEANNOTATOR organizes the data groupings generated by the SCAVENGER and displays them as diagrams.

The modular structure of MAPMAN permits editing of the existing data categories, addition of new categories, and the development of SCAVENGER modules for new types of data. MAPMAN needs to be further developed for correction of deficiencies and inclusion of additional applications. For example, the SCAVENGER modules need to be developed for automatically updating the annotation and terminology by error-free acquisition of the GOC (*G*ene *O*ntology *C*onsortium) and other relevant releases. Modules are also needed for the removal of unnecessary redundancies, display of absolute levels of gene expression, or metabolite accumulation. In addition, modules capable of statistical analyses of the data also need to be developed. The IMAGEANNOTATOR module and the instructions for its use are freely available from the website http://gabi.rzpd.de/projects/MapMan/. The SCAVENGER modules can be obtained on request without any charge.

14.3.8 GenScan

The *GenScan* program (Burge and Karlin 1997) predicts complete gene structure, including introns, exons, and the exon–intron boundaries, promoter sites, and poly-A signals in genome sequences of many different types of organisms. The gene structure model used by GenScan is a "probabilistic model" developed for human genes. This model includes the description of the signals for transcription, translation, and splicing and the features related to the lengths and the base compositions of exons, introns, and the intergenic regions. It searches the query sequence for the features of this model, and the stretches of the sequence matching the descriptions of exons, promoters, etc. are identified, and a probability is assigned to each identified stretch. The identified "optimal exons" match the model with the highest probability ($P > 0.99$) and are considered to represent actual exons. GenScan also predicts "suboptimal exons" having acceptable probability levels ($P = 0.50–0.99$) of representing a true exon. Exons predicted with <0.50 probability are discarded as unreliable. This program is capable of predicting multiple genes as well as partial genes located in a given nucleotide sequence. The users can examine the "optimal" and the "suboptimal" sets of predictions to identify nonstandard gene structures like alternatively spliced genes. GenScan can accept and analyze nucleotide sequences of up to one million base pairs in length. It can analyze the sequence of either one or both the stands of a DNA duplex and make consistent prediction of groups of genes. GenScan has high accuracy but is sensitive to exon length. GenScan is by far the most comprehensive and sophisticated gene prediction tool available for free. The GenScan server can be accessed at http://genes.mit.edu/GENSCAN.html. Some other tools designed for gene prediction are FGENESH/FGENES, HMM Gene, GENE ID, GENE PARSER, etc. (Table 12.2).

14.3.9 ClustalW

A multiple sequence alignment is perhaps the most useful investigative procedure in bioinformatics; it is often said that it could be helpful in almost any situation. It helps in prediction of protein structure and function, and is the basis for phylogenetic analyses. The *Clustal* family of programs is perhaps the most extensively used for alignment of multiple sequences. There are two types of Clustal (ver. 2) programs: (1) ClustalW (has a command-line user interface) and (2) ClustalX (has a GUI) (Thompson et al. 1997; Chenna et al. 2003; Larkin et al. 2007). ClustalW is easy to use and is the most frequently used multiple sequence alignment tool. ClustalW uses a progressive method of alignment, in which the sequences are first compared in pairs for similarity. Each similar pair of sequences is then treated as a single

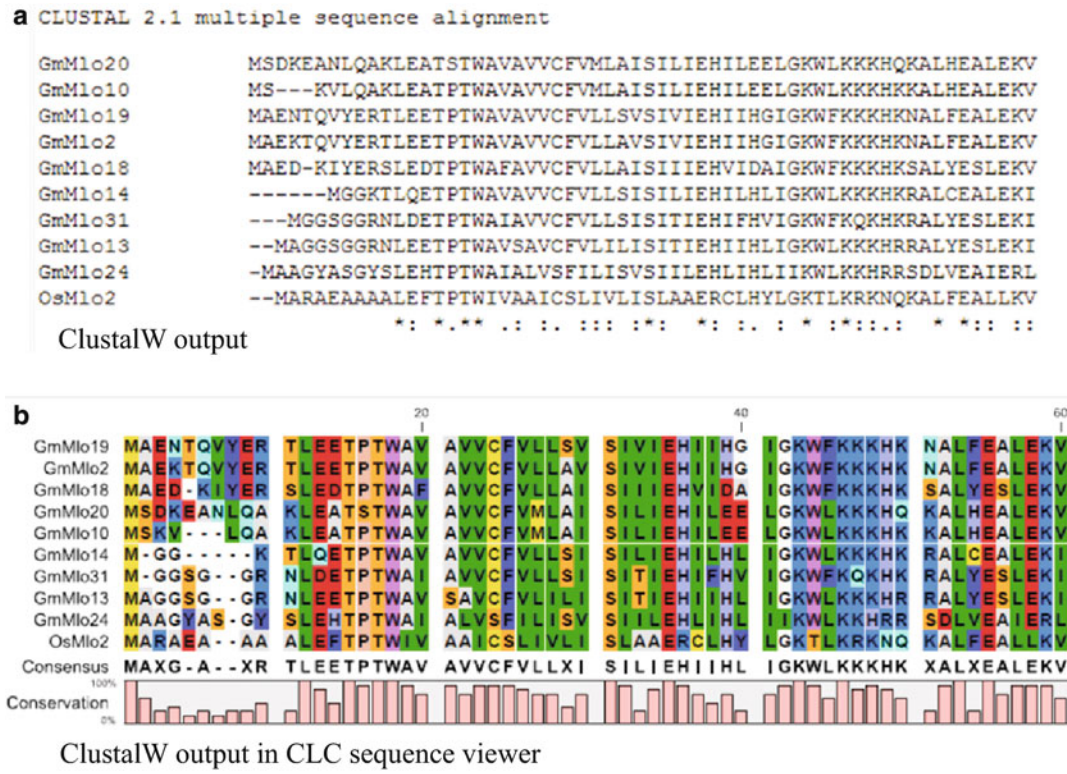


Fig. 14.2 Multiple sequence alignment of nine members of soybean *Mlo* (*GmMlo*) gene family and one rice *Mlo* (*OsMlo2*) gene. (**a**) The output from ClustalW program. (**b**) The ClustalW output visualized using CLC Sequence Viewer (Courtesy, Reena Deshmukh, Varanasi)

sequence, and the sequences so obtained are again compared two-by-two and aligned in pairs. This procedure is repeated till all the sequences are aligned together. A researcher can align the sequences using the default setting, but occasionally one may like to customize the setting to best suit one's needs. The main parameters that can be customized are the substitution matrix and the penalties for gap opening and gap extension.

Clustal programs offer several options for input/output formats, including Clustal, PHYLIP (output only), and FASTA (input only) formats. However, it is most convenient to use the FASTA format for ClustalW input sequences. However, judging the quality of a sequence alignment is essentially an educated guesswork. The bottom row of the ClustalW output of multiple sequence alignment contains stars (*), colons (:), and dots (.) (Fig. 14.2a). A star below a column indicates a fully conserved or an invariant amino acid residue, a colon (:) denotes that all the residues in the column have roughly the same size and hydrophobicity, a dot (.) signifies that the different amino acid residues in the column are either similar in size or hydrophobicity, while lack of a symbol indicates that the residues in the column differ both in size and hydrophobicity. A simple criterion of a sequence block with a good alignment is as follows: it is a gap-free continuous stretch of 10–30 amino acids having 1–3 stars, 5–7 colons, and a few dots scattered in the block. The aligned sequences can also be viewed through the CLC Sequence Viewer (6.8.1.); it uses a color code, in which the same color is used to depict amino acids having similar size and hydrophobicity. In addition, the level of conservation in each column is presented as a bar diagram; fully conserved columns are represented by a bar of full (100 %) height (Fig. 14.2b).

Clustal Omega is the latest addition to the Clustal family. This high-capacity program aligns hundreds of thousands of sequences in only a few hours. It can use multiple processors, and the quality of alignment is superior to those of the earlier versions. However, at present, Clustal Omega has a command-line interface and handles only protein sequences. It may be pointed out that it is preferable to work with protein sequences than nucleotide sequences. Precompiled executables and the source code of the programs (ver. 2.1) for Windows, Linux, and Mac OS X systems are available from www.clustal.org. EBI no longer maintains the ClustalW program, but it has Clustal Omega. ClustalW ver. 2.1 can be used at DDBJ (http://clustalw.ddbj.nig.ac.jp/) and the ClustalW servers (http://www.ch.embnet.org/software/ClustalW.html).

T-Coffee (igs-server.cnrs-mrs.fr/Tcoffee), MEME (http://meme.sdsc.edu/meme/website/intro.html; uses the expectation maximization method), HMMER (http://hmmer.janelia.org/; for protein sequence analysis), MUSCLE (http://www.drive5.com/muscle/; http://www.ebi.ac.uk/tools/muscle/index.html; aligns both DNA and protein sequences), MAP (http://genome.cs.mtu.edu/map.html), and COBALT tool of NCBI (for protein sequence alignments) are some of the other programs that can be used for multiple sequence alignment.

14.4 Bioinformatics Databases

A *database* is a systematized collection of vast amounts of information on a specific topic, e.g., nucleotide sequence, protein sequence, etc., in an electronic environment. The organization of databases is such that it allows regular updating of data and easy search and retrieval of the desired information. There are three types of databases, namely, flat-file, relational, and hierarchical databases. The *flat-file database* is the earliest and the simplest database type, is usually used for storing small amounts of data, and may consist of one or more files. It is easy to set up, but the storage methods are rather complex. In *relational databases*, the data are organized in form of tables. Further, the columns in these tables are indexed according to common features. These databases are constructed using the SQL (*S*tructured *Q*uery *L*anguage) programming language. This type of database has a well-defined design and architecture that minimizes redundancy of stored data, but its setting up is intellectually demanding. It supports very fast data search and can answer complex questions. In the *hierarchical databases*, the data are organized in a hierarchical (ordered tree) structure, and there are two or more levels of data organization (e.g., MAPMAN, Sect. 14.3.7). Construction, operation, and modification of these databases are simple, and data search and retrieval is fast. But they need more space, consume more time, and the same record may need to be stored at two or more places.

Special computer software programs, called database management systems (DBMS), are used to organize, search, and access the data. These programs not only contain raw data records, they also have operational instructions to help identify interconnections in the data records. The DBMS can be either relational or object oriented, the former being the most commonly used. For example, MySQL is a full-fledged, open-source relational DBMS, which has a three-tier architecture, viz., user interface, business logic, and data storage tiers. The biological databases are generally concerned with DNA and protein sequence data storage and management. The primary databases on nucleotide sequences are GenBank, EMBL, DDBJ, and GSDB (Genome Sequence Databases), while Swiss-Prot, TrEMBL, PIR, and MIPS (Martinsried Institute of Protein) are examples of primary databases on protein sequences. Further, there are secondary databases representing some specific sections of the primary databases, and in case of proteins there are composite and structure databases as well. In addition, there are several specialized databases devoted to specific organisms (Table 14.3).

Table 14.3 A selected list of the various "omics" resources and tools for major crop species/groups of crop species (whole-genome sequences, EST sequences, proteomics, transcriptomics, metabolomics, long-insert library, and HTG)

Crop species	Link to "omics" resource/tool
Apple	www.rosaceae.org
Arabidopsis[a]	https://www.arabidopsis.org/
Banana	http://www.musagenomics.org/index.php
Brassica spp. (*Brassica* ASTRA)	http://hornbill.cspp.latrobe.edu.au
Cassava	www.phytozome.net
Cotton	www.cottonmarker.org
GrainGenes[b]	http://wheat.pw.usda.gov/
Gramene[b]	http://www.gramene.org/
Grape	www.vitaceae.org
Legumes	http://www.comparative-legumes.org/
Lotus	http://www.kazusa.or.jp/lotus/
Maize	www.maizegdb.org; http://www.maizegenome.org/
Medicago	http://www.medicago.org/genome/)
Poplar	http://genome.jgi-psf.org/Poptr1/
Potato	www.potatogenome.net
Rape seed	www.brassica.info
RGP[b]	http://rgp.dna.affrc.go.jp/
Rice	http://rice.plantbilogy.msu.edu
Rosaceae	(GDR; http://www.genome.clemson.edu/gdr/)
Sorghum	www.phytozome.net/sorghum
Soybean	http://soybase.org
Tobacco	http://www.intl-pag.org/13/abstracts/PAG13_P027.html)
Tomato	http://solgenomics.net/;
	http://sgn.cornell.edu/help/about/tomatosequencing.html
Wheat	http://www.wheatgenome.org
Metabolic Network[c]	http://www.plantcyc.org/
Protein annotations[c]	http://salad.dna.affrc.go.jp/salad/en/

A comprehensive list of bioinformatics tools/software is available at http://seqanswers.com/wiki/Special:BrowseData/Bioinformatics-application

[a]The model dicot plant species

[b]These databases incorporate analytical, visualization, and interrogation tools

[c]Databases of plant metabolic etwork (PMN) and Surveyed conserved motif ALignment diagram and the Associating Dendrogram (SALAD)

14.4.1 GenBank

GenBank (http://www.ncbi.nih.gov/Genbank) is the main primary genomic DNA sequence database held at NCBI, USA, since 1992. The NCBI staff builds this database from sequences obtained through submissions from various laboratories, exchanges with EMBL and DDBJ nucleotide sequence databases, and patented sequence data from the US Patent and Trademark Office. The sequences stored in GenBank are divided into organismal and functional categories. The nucleotide sequences from specific organisms are stored in the organismal category, while those representing specific functions, such as high-throughput genome (HTG), expressed sequence tags (ESTs), etc. are stored in the functional category irrespective of their source organism. The translations of nucleotide sequences stored in GenBank are stored in the PIR database. The GenBank participates in the International Nucleotide Sequence Database Collaboration (INSDC) along with the EMBL and the DDBJ nucleotide sequence databases. GenBank is interconnected with the EMBL and the DDBJ databases, and these three databases exchange sequence information every day.

14.4.2 Phytozome

There are three different comparative genomic databases on green plants, namely, GreenPhylDB, Plaza, and Phytozome. These databases aim to support studies on genomics-based plant evolution and to facilitate the application of the information on functional genomics obtained from model plants for the improvement of crop plants. The Phytozome database is accessible at the website http://www.phytozome.net. It provides comparative data on genomes and gene families and the tools for their analysis. This database provides complete information on the pattern of changes in the nucleotide sequence and structure of each gene during evolution as well as the evolutionary history of plant gene families and their genomic organization. The large-scale gene family phylogenetic trees are constructed using distance-based methods or, sometimes, distance-plus-character-based methods. It also contains sequences and functional annotations of complete genomes of several (25 in 2012) plant species. The Phytozome Web portal has several widely used tools for search, identification, and evaluation of gene families. It provides information on genomic context of plant genes, gene homologues, and paralogues, RNA transcripts from the given genes, alternatively spliced RNA transcripts and the resulting peptide sequences, and functions of gene families. It also permits putative functions to be assigned to new DNA sequences. It allows access to complete genome sequences available in the database. Retrieval of genes and gene families can be done by either using keyword or through a search based on sequence similarity. Genome-centric views of the genomes present in Phytozome are available through Gbrowse and can be accessed from the Phytozome homepage. The BioMart tool of the database allows customized construction of sequences and annotations of genes and/or gene families (Goodstein et al. 2012).

14.4.3 European Molecular Biology Laboratory Nucleotide Sequence Database

The *EMBL Nucleotide Sequence Database* is maintained by EBI, UK, which is an outstation of the EMBL, Germany. This database is a part of the INSD Collaboration and can be accessed at http://www.ebi.ac.uk/embl. The nucleotide sequence collection of the database is comprehensive and includes all such annotation that is available from public sources. This database serves as the primary source of nucleotide sequences for Europe. The nucleotide sequence data and third party annotation (TPA) are generally submitted by the Webin tool, while sequence alignment data are submitted through Webin-Align. The nucleotide sequence data generated by large-scale genome-sequencing projects and those available from the European Patent Office can be submitted using automated procedures. The INSD Collaboration assigns an accession number to each sequence submitted to the database. Only the accession number of the nucleotide sequence submitted to the database needs to be cited in the publication, and the sequence is considered as part of this publication. The above is a mandatory requirement for publication of sequence information in most journals. The TPA entries carry the prefix "TPA" to distinguish them from primary data. The data are grouped into divisions on the basis of either the methodology used for their generation or the taxonomic status of the source organism; in addition, there are some specialized divisions as well. The other genomic databases held at EBI are Ensembl (a database of genome annotation) and Genome Reviews (has the curated complete genome sequences stored in the EMBL Nucleotide Sequence Database). The daily releases of the database contain new submissions and updated sequence data, while every 3 months the entire database is released. Access to the sequence data is available via FTP, several WWW interfaces, and e-mail. The FTP Server provides access to the daily releases, periodic updates, and the collective files having all types of data. Dbfetch (database fetch) tool is used for the retrieval of sequence data of up to 50 entries via http. The SOAP (*S*imple *O*bject *A*ccess *P*rotocol) tool, on the other hand, enables communication with other systems. The Web-based Sequence Retrieval System (SRS) can access nucleotide sequence data available in other databases at EBI. In addition, tools like FASTA and BLAST are also available (Kanz et al. 2005).

14.4.4 Swiss-Prot

Swiss-Prot (established in 1986; http://www.ebi.ac.uk/) database contains curated and fully annotated protein sequences. Each protein sequence entry consists of core data and the annotation information. The core data comprise sequence and taxonomic data and citation information. The annotation information includes function, domains and sites, secondary and quaternary structures, posttranslational modifications, etc. Effort is made to keep the level of redundancy to the minimum and to afford a high level of cross-referencing to other databases. The Swiss-Prot format closely follows the format of the EMBL nucleotide sequence database. The database *TrEMBL* (*Tr*anslation from *EMBL*) is a supplement to Swiss-Prot and contains unannotated protein sequences. The protein sequences included in TrEMBL are obtained by translating each coding sequence (CDS) available in the EMBL nucleotide sequence database, but the CDSs contained in Swiss-Prot are excluded from this database. The TrEMBL entries are recorded in a Swiss-Prot-like format (Bairoch and Apweiler 1996). The Swiss-Prot and TrEMBL databases are now maintained as two sections of the UniProt Knowledgebase (Sect. 14.4.5).

14.4.5 UniProt Knowledgebase (UniProtKB)

The *UniProt Knowledgebase (UniProtKB*, http://www.uniprot.org) is an expertly curated central database of protein knowledge. This database was developed by the UniProt Consortium and is updated by them every 4 months. The chief objective of UniProtKB is to unify the available information on protein sequence and function and to provide the same to the users. The Consortium comprises workers from Swiss Institute of Bioinformatics (SIB), EBI, UK, and PIR. The UniProtKB is extensively cross-referenced, and all data are freely available to scientists. This database has two main sections: (1) the UniProtKB/Swiss-Prot section (it comprises manually curated records) and (2) UniProtKB/TrEMBL section (it contains automatically curated, annotated, and classified records). The UniProtKB has the following databases. (1) The UniProt Archive (UniParc) database serves as an all-inclusive protein sequence repository. (2) In the UniProt Reference Clusters (UniRef) database, the protein sequences are grouped into different clusters on the basis of sequence identity; this facilitates the search for closely related proteins. (3) Finally, the UniProt Metagenomic and Environmental Sequences (UniMES) database has been designed to serve the emerging area of metagenomics. In addition, UniProtKB contributes gene ontology annotations to the GOC. *Ontology* is a controlled terminology that is used by researchers working with databases of different taxa. For example, gene ontology (GO) pertains to the description of gene product features from several species. Plant ontology (PO) relates to the terminology needed for describing the anatomical and developmental features of different plant species. Similarly, trait ontology (TO) lists the details of evaluation procedure and the environments, in which a specific trait of a given species was assayed. The use of ontologies allows databases to share information with other databases (Magrane and UniProt Consortium 2011; UniProt Consortium 2013).

The main access point to the UniProtKB is the website. The homepage of this website provides a quick introduction to the UniProtKB via a website tour. It also provides tools for a variety of purposes, including querying, data analysis, documentation, Google-like full-text searches, database identifier (ID) mapping, etc. The ID mapping tool converts UniProt IDs to corresponding IDs of several other databases. It also has the BLAST tool for searching similar sequences and ClustalW for aligning multiple sequences. The full-text search does not require information about the data being searched; in addition it is quick and easy. But more complex enquiries can be processed using a field-based text search. The BioMart data management system enables processing of multiple biological queries by accessing the UniProtKB, InterPro, Ensembl, and PRIDE resources. The search results are presented as a table, which can be reorganized by the users.

14.4.6 Gramene

Gramene (http://www.gramene.org) grew out of the RiceGenes project to encourage analysis of functions that are conserved across species as well as those that are specific to individual species. This database began in 2002 as a resource for the rice community and as a collection of comparative mapping studies in grasses. Gramene was organized around the rice genetic, physical, and sequence-based maps. Further, a set of corresponding anchor genetic markers was used to develop the comparative maps of several grass species such as rice, wheat, barley, maize, etc. Gramene has now become a comparative genomic resource for important plant species like *Arabidopsis*, *Brachypodium*, poplar, etc. (Ware et al. 2002; Youens-Clark et al. 2011). It contains data on a variety of relevant topics, including metabolic pathways, QTLs, genes, proteins, genetic diversity, etc. The data stored in the database are integrated with genetic, physical, bin, etc. maps as well as genome browsers. Gramene carries out alignments on the whole genome as well as gene-to-gene basis. It also predicts orthologous and paralogous relationships by constructing gene trees, and implements a synteny analysis to confirm homology.

The main navigation bar on each page of Gramene provides links to various databases, search tools, submission forms, etc. The genetic map can be accessed for a linkage group of a species or for maps having specific features like molecular markers. With the help of the integrated map of rice, one can find the location in the rice genome sequence that corresponds to the given position in the maize, wheat or barley genetic map. Gramene has a curated database containing all publicly available mutants of rice, molecular markers as well as proteins; it also includes descriptions of the variants for physiological or morphological characters, etc. Gene symbol, gene name or Gramene accession number can be used to search for the gene of interest. The search result presents an all-inclusive summary of all the data related to the specified phenotypic variant. The rice genome browser of Gramene displays a variety of information, including gene predictions, genetic markers, coding sequences, etc., and information about the protein encoded by the predicted gene.

Gramene uses ontologies for describing proteins, genes, alleles, and phenotypes, which allows information sharing with other databases. This also permits a gene affecting a developmental stage or an organ in one species to be used for predicting the gene involved in a similar function in another species. Gramene provides several Web services like a Distributed Annotation Server and useful tools like BLAST. Updates of the database are released regularly, while major additions are released twice every year. All Gramene databases and software are free; the downloads can be made via ftp://ftp.gramene.org.

14.4.7 GrainGenes

GrainGenes database is devoted to genetic and genomic information on the following cereal crops and their wild relatives: wheat, barley, oat, and rye (http://wheat.pw.usda.gov). This international database serves as data storehouse as well as information center. It contains curated data on genetic as well as physical maps, probes utilized for constructing the maps, ESTs, EST-derived simple sequence repeats, oligogenes, and QTLs. It contains the contact details of the researchers working with these crop species. It also stores data on sequences, genetic resources, pathology, literature references (on genetics and genomics), and provides links to other databases. The genetic and physical maps available at http://wheat.pw.usda.gov/ggpages/maps are summarized and interactive, and include links to the information on the concerned mapping study. It permits comparison among maps from different populations. QTL data for the relevant species are gathered from various sources, and they are referenced to similar QTLs for rice (in Gramene) and maize (in MaizeGDB). GrainGenes contributes to the development of trait ontology in collaboration

with other databases. It has a database of genotypes (alleles of molecular markers) and phenotypes for wheat and barley varieties (Carollo et al. 2005).

14.4.8 MaizeGDB

The *Maize Genetics and Genomics Database* (*MaizeGDB*) serves the maize research community as the chief source of data on genetics and genomics of maize (http://www.maizegdb.org). MaizeGDB stores data on DNA sequences, genetic studies, QTL experiments, gene products, relevant literature references, and the list of persons/organizations involved in maize research. The genetic data include allelic diversity, maps, primers/probes used for mapping, metabolic pathways, and phenotypic image data. MaizeGDB serves as a portal for the maize genome-sequencing information. Maize sequences available at GenBank are downloaded, curated, analyzed, and assembled into contigs at PlantGDB (Plant Genome Database; http://www.plantgdb.org). However, they are stored at MaizeGDB, which makes them accessible to the users. Cytological map images have been added to the database, and plant ontology terms are being associated with the database records. A researcher can enter the database by typing the term of his/her interest, e.g., *adh1*, into the search field. Appropriate links can be followed to obtain other information related to the search item. The genome browser enables data visualization and displays the data within their chromosomal context. In addition, the MaizeGDB Web service permits the use of data analysis tools like BLAST and GeneSeqer. MaizeGDB homepage offers a set of video tutorials to facilitate effective use of the database (Lawrence et al. 2005).

14.4.9 RiceGeneThresher

RiceGeneThresher (http://rice.kps.ku.ac.th) is a public domain rice genomics resource. The MySQL-based database has integrated information on genetics, genetic markers, genome annotation, ESTs, metabolic pathways etc. It also contains information on protein–protein interaction predictions, and genes that respond to various stresses. RiceGeneThresher is fast and flexible, and has tools for whole-genome mining for QTLs governing the specified traits. The data from studies on inheritance, molecular biology, and various "omics" approaches are analyzed to find the most promising candidate genes within a genomic/QTL region, and to infer their functions. The search for candidate genes may use as query either the relevant DNA marker or base sequence of the target region of the genome. Alternatively, the gene/locus name or gene annotation keywords like gene locus ID may be used as query. The retrieved information is displayed both as graphical and standard Web pages. It includes physical locations of the candidate genes, and nucleotide sequences of the complete genes (including their upstream regions) and their coding regions as well as the amino acid sequences of the proteins produced by their translation (Thongjuea et al. 2009).

14.4.10 Microarray Databases (ArrayExpress and Gene Expression Omnibus)

In addition to in-house databases, there are centralized databases like ArrayExpress and Gene Expression Omnibus (GEO) for gene expression data obtained from microarray experiments. *ArrayExpress* (http://www.ebi.ac.uk/arrayexpress) database is maintained at EBI, UK, and is accessible to the public (Brazma et al. 2003; Parkinson et al. 2005). This database is capable of structured storage of well-annotated data generated by any microarray platform. The annotation used by ArrayExpress conforms to MIAME standard. It is also able to exchange data in Microarray Gene Expression Markup Language (MAGE-ML) format. The other available online facilities include MIAMExpress (the data submission tool), the interface for searching the database, and the analysis tool Expression

Profiler. The query may relate to the experimenter, the laboratory where the work was done, the organism studied, the type of gene expression experiment and/or the type of microarray used in the study. The submissions can be of the following three types: arrays, experiments, and protocols.

The *Gene Expression Omnibus (*http://www.ncbi.nlm.nih.gov/geo) is a public domain storehouse for data on gene expression as well as genomic hybridization generated by high-throughput platforms (Edgar et al. 2002). Heterogeneous datasets can be easily deposited into this database where they are safely stored and can be readily retrieved when required. The objective of GEO is to complement in-house databases for gene expression and to serve as a tertiary central data distribution center. The Platforms, Samples, and Series modules of the GEO function as its central data entities. The *Platform* comprises a list of probes, which determine the molecules detected in the study. The *Sample* describes the group of molecules that was investigated in the study. This description is related to a single platform that was employed to produce the data on abundance of various molecules. The *Series* organizes the samples into meaningful datasets that constitute the experiment.

14.4.11 HarvEST

HarvEST software was initially developed for visualization of EST databases. At present, it supports several activities, including identification of SNPs, designing of genotyping platforms, comparative genomics, association of physical maps with the concerned genetic maps, and designing of microarrays. This software is available for banana/plantain, barley, *Brachypodium*, cassava, citrus, coffee, cowpea, soybean, rice, and wheat; the programs for barley and cowpea are the most complete. HarvEST has databases of crop species-specific EST sequences that have been generally trimmed free of vector and assembled using the CAP3 program, except in the cases of ESTs of rice, soybean, and wheat. The ACE file viewer allows the examination of sequence alignment and identification of SNPs. HarvEST offers a variety of assemblies, synteny, and sequence alignment analyses, archived information on BLAST hits, Boolean and other searches, etc.; these applications do not require Internet connectivity. The query may specify the genes involved in the trait of interest like tolerance to a stress, a developmental stage, or a specific tissue. The search output is either displayed on the screen, or it can be exported as a summary table/sequence file. HarvEST provides links to other sequence databases and supports online searches. The HarvEST software operates in the Windows environment and can be downloaded free for academic use from www.harvest-web.org.

14.5 Sources of Multiple Databases and Tools

There are several bioinformatics resources that provide multiple databases and database search and data analysis tools. It may be pointed out that bioinformatics is developing very rapidly, and software programs that are the best in their respective fields today may become less preferred or even obsolete tomorrow. Therefore, it may often be quite helpful to consult one of the bioinformatics resource locators listed in Table 14.4 to find out the Web locations of the tools that are currently the most appropriate for the needs of a research worker.

14.5.1 National Center for Biotechnology Information

The *National Center for Biotechnology Information* is located in Maryland, USA (www.ncbi.nlm.nih.gov/). The NCBI is an organization of the United States National Library of Medicine (NLM), which itself is a part of the National Institutes of Health (NIH). The NCBI serves as the primary provider of information relevant to biotechnology and biomedicine. The NCBI

Table 14.4 A list of selected bioinformatics resource locators

Resource locator	Web address	Resources related to
ArrayExpress	www.ebi.ac.uk/microarray/	DNA chips
ExPASy	www.expasy.ch	Servers dedicated to proteins; the home of Swiss-Prot
Pasteur	http://bioweb.pasteur.fr/intro-uk.html	Many online tools; miscellaneous links
Phylip	http://evolution.genetics.washington/edu/	Everything related to phylogeny
RNA World	www.imb-jena.de/RNA.html	Links related to RNA
Swbic	www.swbic.org/	Miscellaneous links
NCBI primers	www.ncbi.nlm.nih.gov/education	Very good introductory material on many subjects
Coffee Corner	www.ncbi.nlm.nih.gov	Online protein news
Bio-informer	www.ebi.ac.uk/Information/News	The EBI online news

maintains a series of databases (total number, 66) covering relevant literature, health, organisms, genomes, nucleotide sequences, genes, proteins, chemicals, and pathways (Table 14.5). The major databases maintained at NCBI are GenBank (Sect. 14.4.1) and PubMed (bibliographic database for biomedical literature). Other databases include the Gene, Genome, Epigenomics, Gene Expression Omnibus (Sect. 14.4.10), RefSeq, Structure, Database of Short Genetic Variation (dbSNP), TAXONOMY, etc. The NCBI also provides a variety of tools (58 tools; Table 14.6) for database search (some of the tools) and/or data analysis (most of the tools). The Entrez search engine of NCBI is its main system for text search and retrieval. The other tools include 1,000 Genomes Browser, BLAST, CDTree, Cn3D, Genetic Codes, Open Reading Frame Finder (ORF Finder), SNP Database Specialized Search Tools, TAXONOMY BROWSER, etc. The NCBI Handbook provides a description of the databases and some of the tools like Entrez and BLAST. It also contains information on the manner in which the databases work and the approaches for their utilization. In addition, each NCBI resource (databases and tools) has online help documentation to assist the user in their proper utilization. The NCBI research group conducts studies on the relevance of sequencing errors for database search, develops new algorithms for database search and multiple sequence alignment, constructs nonredundant sequence databases, builds mathematical models for the estimation of statistical significance of sequence similarity, and develops vector models for text retrieval.

14.5.2 Kyoto Encyclopedia of Genes and Genomes

The *Kyoto Encyclopedia of Genes and Genomes* (*KEGG*) is an integrated bioinformatics storehouse in the public domain (http://www.genome.ad.jp/kegg/). It aims to deduce from the genome information an understanding of the higher-order biological functions and their relevance to cells/organisms (Kanehisa et al. 2004). It integrates the current information on genes, proteins, enzymes, reactions, biochemical compounds, and molecular interaction networks. KEGG maintains a suite of databases, namely, PATHWAY, GENES, SSDB, KO, COMPOUND, GLYCAN, REACTION, and enzyme databases. These databases signify graph objects belonging to the following three categories: (1) protein network, (2) gene universe, and (3) chemical universe (Table 14.7). Subsequently, a resource on glycome informatics was developed; this resource integrates protein network, genomic, as well as chemical information (http://www.genome.jp/kegg/glycan/). The GLYCAN database (for carbohydrate structures) along with two useful tools is a part of this resource. In addition, this resource has glycan-related pathways and a map depicting all the possible structural variations in the carbohydrates of the biological world (Hashimoto et al. 2006).

The KEGG databases are organized in a hierarchical manner. These databases interact with numerous external databases. For example, the GENES database selects its entries semiautomatically from different sources, which include

Table 14.5 A list of the genomic resource related databases (except those devoted exclusively to animals) available at the NCBI website (www.ncbi.nlm.nih.gov/)

Database	Type of data stored
BioProject (former Genome Project)	Studies on genomics, functional genomics and genetics, and links to the concerned datasets
BioSample	Description of the biological materials used for experimental assays
Bookshelf	Selected freely downloadable biomedical books covering molecular biology, biochemistry, cell biology, genetics, etc.
CloneDB (former Clone Registry)	Clones and libraries, including sequence data, map positions, etc.
Computational Resources from NCBI's Structure Group	Databases and tools used for the study of structures of macromolecules, conserved domains, classification of protein, etc.
Conserved Domain Database (CDD)	Protein sequence alignments and profiles of the domains that have been conserved during evolution
Database of Expressed Sequence Tags (dbEST)	A division of GenBank comprising short single-pass reads of cDNA sequences
Database of Genome Survey Sequences (dbGSS)	Short single-pass reads of genomic DNA; a division of GenBank
Database of Genomic Structure Variation (dbVar)	Large changes in genome structure, including large InDels, translocations, and inversions
Database of Genotypes and Phenotypes (dbGaP)	Results from studies designed to elucidate the relationship between genotype and phenotype, includes findings from genome-wide association studies (GWAS)
Database of Short Genetic Variation (dbSNP)	Information on single-nucleotide variations, microsatellites, and small-scale InDels
Epigenomics	Richly annotated epigenomics datasets
GenBank	Annotated collection of all publicly available DNA sequences
Gene	Database of genes, especially from completely sequenced genomes
Gene Expression Omnibus (GEO) Database	A repository of MIAME-compliant gene expression data
Gene Expression Omnibus (GEO) Datasets	Curated datasets on gene expression and molecular abundance; derived from the GEO database
Gene Expression Omnibus (GEO) Profiles	Expression and molecular abundance profiles of individual genes; derived from the GEO database
Genome	Sequence and map data for the whole genomes of over 1,000 organisms
HomoloGene	A gene homology tool; identifies putative orthologs; curated orthologs from a variety of sources included via the Gene database
Journals in NCBI Databases	Journals referenced in NCBI database records, including PubMed abstracts
NCBI Glossary	Description of NCBI tools, acronyms, data representation formats, and bioinformatics terms
NCBI Handbook	Description of the various features of NCBI databases and software
NCBI Help Manual	Details of many NCBI resources, including the Entrez system, Gene, SNP, LinkOut, etc.
NCBI Website Search	Static NCBI Web pages, documentation, and online tools
Nucleotide Database	Nucleotide sequences from several sources, including GenBank, RefSeq, TPA, and PDB (Protein Data Bank)
PopSet	Related DNA sequences originating from comparative studies
Protein Clusters	Related protein sequences (clusters) consisting of Reference Sequence proteins
Protein Database	Protein sequences from a variety of sources, including GenPept, RefSeq, Swiss-Prot, PIR, and PDB
Reference Sequence (RefSeq)	Curated, nonredundant DNA, RNA, and protein sequences
Retrovirus Resources	Designed to support research on retroviruses; includes a genotyping tool
Sequence Read Archive (SRA)	A collection of sequence data from the next-generation sequencing platforms

(continued)

Table 14.5 (continued)

Database	Type of data stored
Structure (Molecular Modeling Database)	Macromolecular 3-D structures derived from PDB; also has tools for their visualization and comparative analysis
TAXONOMY	Names and phylogenetic lineages of more than 160,000 organisms
Third Party Annotation (TPA) Database	Sequences and annotations built from the existing primary sequence data in GenBank
Trace Archive	DNA sequence chromatograms (traces), base calls, and quality estimates for single-pass reads
Transcriptome Shotgun Assembly (TSA) Database Computationally assembled sequence database derived from NGS technologies Unigene	Sets of RNA transcripts computationally identified to be encoded by the same gene/pseudogene; also has information on protein similarities and gene expression
Unigene Library browser	Expressed sequence tag (EST) libraries arranged on the basis of organism, tissue type, and developmental stage
UniSTS	Sequence-tagged sites (STSs) derived from STS-based maps and other experiments
Viral Genomes	Biology of viruses, links to viral genome sequences in Entrez genome, etc.
Virus Variation	Sequence sets of selected viruses; also tools for their analyses

Table 14.6 A list of selected tools available at the NCBI website (total tools 58; www.ncbi.nlm.nih.gov/)

Tool	Application
1,000 Genome Browser	An interactive graphical viewer; this tool allows variant calls, genotype calls, aligned sequence reads, and other applications
Amino Acid Explorer	It uses characteristics of amino acids to predict changes in protein sequences due to mutations and functions of amino acid residues in conserved domains
Assembly Archive	This tool links raw sequence with the sequences available in GenBank, EMBL, and DDBJ; it allows viewing of the multiple sequence alignments
Blast Link (Blink)	Blink displays the results of BLAST search of a protein sequence against the protein sequence database at NCBI
Blast Microbial Genomes	BLAST search for similar sequences present in the selected complete eukaryotic/prokaryotic genomes
Blast RefSeqGene	BLAST search of the genomic sequences in the RefSeqGene/LRG set
Basic Local Alignment Search Tool (BLAST)	Searches for regions of local similarity between new nucleotide or protein sequences and those present in the sequence databases
Batch Entrez	Retrieves records from many Entrez[a] databases; results can be saved in various formats
CDTree	Classification of protein sequences; analysis of their evolutionary relationships; creation and updating of protein domain alignments
COBALT	A protein multiple sequence alignment tool; it uses RPS-BLAST, BLASTP, and PHI-BLAST
Cn3D	Visualization of 3-D structures of proteins from NCBI's Entrez retrieval service; displays structure, sequence, and alignment
Concise Microbial Protein BLAST	A specialized BLAST search of database consisting of all proteins from complete microbial genomes
Conserved Domain Architecture Retrieval Tool (CDART)	Displays the functional domains of a protein sequence and lists proteins with similar domain architectures
Conserved Domain Search Service	Identifies conserved domains present in a protein sequence
Digital Differential Display (DDD)	Compares EST profiles to identify genes with significantly different expression levels
Electronic PCR (e-PCR)	Identifies sequence-tagged sites (STSs) present in DNA sequences
Gene Expression Omnibus (GEO) BLAST	Aligns a nucleotide or protein sequence with GenBank sequences included in the GEO database

(continued)

Table 14.6 (continued)

Tool	Application
Gene Plot	Pair-wise comparison of two prokaryotic genomes; displays pairs of protein homologues
Genetic Codes	Provides genetic codes for organisms in the TAXONOMY database
Genome BLAST	Similarity search for query nucleotide or protein sequences in the genomic sequence databases using BLAST tool
Genome ProtMap	Maps each protein from a COG or a VOG back to the genome
Remap Tool	Allows projection of annotation data and other features from one genomic assembly to another or to RefSeqGene sequences
Genome Workbench	Permits viewing and analysis of sequence data
Map Viewer	For special browsing of maps and assembled sequences for a subset of organisms
OSIRIS	Assessment of multiplex short tandem repeat (STR) DNA profiles
Open Reading Frame Finder (ORF Finder)	Detection of all ORFs in the submitted sequence or in a sequence present in a database
Primer-BLAST	Designs PCR primers for a template sequence
ProSplign	Aligns proteins to genomic nucleotide sequences
Related Structures	Finds 3D structures from the Molecular Modeling Database (MMDB) for sequences similar to the query protein sequence
SNP database Specialized Search Tools	Search of the SNP database by genotype, method, population, etc.
Sequence Viewer	Generates a graphical display of a nucleotide or protein sequence and its annotated features; this display is configurable
Splign	Computes cDNA-to-genomic sequence alignments
Tax Plot	Compares genomes on the basis of the encoded protein sequences
TAXONOMY BROWSER	Searches taxonomy tree using partial taxonomic names, common names, etc.
Taxonomic Common Tree	Generation of a taxonomic tree for the selected group of organisms
VecScreen	Quick identification of sequences of vector origin
Vector Alignment Search Tools (VAST)	Identifies similar 3-D structures of proteins
Viral Genotyping Tools	Identifies the genotype of a viral sequence

[a]Entrez: NCBI's primary text search and retrieval system; integrates the PubMed database with 39 other databases

the sequences submitted to the GenBank, RefSeq, and EMBL nucleotide sequence databases, as well as organism-specific databases accessible to the public. These entries are re-annotated by assigning K numbers after the genes are classified in the KO groupings. It is intended that the KEGG would be self-reliant in connecting genome information to cellular functions. It is hoped that KEGG will eventually be able to enable in silico analysis of various biological systems and to create computer representations of cells and organisms. The GenomeNet website (http://www.genome.ad.jp/kegg/) enables easy access to KEGG. The table of contents page of KEGG can be accessed at http://www.genome.ad.jp/kegg/kegg2.html. This page permits access to the databases of KEGG. Academic users can utilize the SOAP server (at http://www.genome.ad.jp/kegg/soap/) to gain computerized access to KEGG.

14.5.3 Molecular Biology Database Collection

The Nucleic Acids Research launched the *Molecular Biology Database Collection (MBDC)* in the year 2000. The MBDC is a centralized online compilation of molecular biology databases (http://www.oup.co.uk/nar/Volume_28/Issue_

Table 14.7 The KEGG databases and their subject matter

Graph object (subject domain)	Major databases	Information stored	Information source
Gene universe	GENES	Information about individual genes	Submissions to GenBank, RefSeq, EMBL databases, and other publicly available databases
	SSDB	Sequence similarity database with ortholog and paralog clusters	GENES database
	KO	Classification of functions of genes in the SSDB database	SSDB ortholog clusters
Chemical universe	COMPOUND	Chemical structures of metabolic and some pharmaceutical and environmental compounds	Manually entered and verified
	GLYCAN	Carbohydrate structures; links to complex lipid and carbohydrate metabolism pathways	CarbBank database; direct entries
	REACTION	Formulas for enzyme catalyzed reactions	
	ENZYME	Enzyme nomenclature; links to the various KEGG databases	The enzyme nomenclature website[a]
Protein network (most unique data object)	PATHWAY	Network of gene products, including protein–protein interactions, metabolic networks, and gene regulatory networks	A collection of manually drawn diagrams called the KEGG Reference Pathway Diagrams (maps)

Based on Kanehisa et al. (2004)
[a]http://www.chem.qmul.ac.uk/iubmb/enzyme/

01/html/gkd115_gml.html). The aim of this compilation is to make these databases more accessible to the scientific community by helping them select the databases best suited to their needs and access them using the links provided. The emphasis of MBDC is to include databases with new value added by data curation, annotation, connections to new data, or inclusion of some other innovative features. The database list, classified on the basis of the information content of the databases, contains minimum redundancy, and the links to the databases are regularly updated (Baxevanis 2000).

14.5.4 Architecture for Metabolomics (ArMet)

ArMet (*Architecture for Metabolomics*; http://www.armet.org) is a data model designed to provide a standard format for describing metabolomics experiments and representing the data obtained from them (Jenkins et al. 2004). It covers the description of the biological source materials, the experimental details (sample collection, preparation, and analysis), and the results obtained. ArMet has the following nine packages: (1) Admin, (2) Biological source, (3) Growth, (4) Collection, (5) Sample handling, (6) Sample preparation, (7) Analysis-specific sample preparation, (8) Instrumental analysis, and (9) Metabolome estimate. These packages are applicable to a wide range of experiments. The core dataset that provides the lowest common denominator for comparison of different datasets is also described. ArMet is not a publicly available data repository; it merely serves as a basis for designing data storage facilities and appropriate data handling tools. ArMet uses controlled vocabularies to allow correct interconnection with other databases. It enables uniform recording of experimental details, ensures completeness and internal consistency of datasets, affords dependable exchange of data, allows a meaningful comparison of data from a range of techniques, supports designing of new experiments, and promotes standard operating procedures. It is designed to handle greenhouse-/phytotron-grown *A. thaliana*, field-grown potatoes, harvest and storage of these

materials, preparation of the materials for analysis, and the analysis itself. In addition, it can handle preparation and storage of peak lists as well. The results of analyses are represented in the Universal Modeling Language (UML) 1, which allows clear data specification and dissemination. The ArMet is built using Oracle with a Microsoft Office front end. It allows designing of sub-packages for new analytical and experimental techniques.

14.5.5 Database Search and Analysis Tools

The amount of data stored in each database is so enormous that it is a daunting task to find the information of interest from a given database. In addition, the data extracted from a database have to be suitably analyzed for the detection of hidden patterns, associations, and other features of interest and for their proper interpretation. These tasks are facilitated by computer programs that have been designed for either searching and retrieving the data from various databases or for searching, retrieving, as well as analyzing the retrieved data. The programs designed for search and, often, analyses of the retrieved data are generally provided by the websites hosting the databases.

14.5.5.1 Entrez

Entrez (introduced in 1991) is one of the most popular search engines at the NCBI, USA, and is accessible at www.ncbi.nlm.nih.gov/entrez/. It is a highly versatile and adaptable text-based search and retrieval system. It can search all major NCBI databases, including PubMed, databases of nucleotide and amino acid sequences, TAXONOMY, Swiss-Prot, etc. Entrez collects data from several sources and retrieves indexes and properly organizes the biomedical information. The data from different sources may have different formats and designs, but they all are integrated into a uniform information model and retrieval system. Entrez has nine nodes [published articles, nucleotide sequences, protein sequences, taxonomy, structure, genomes, Online Mendelian Inheritance in Man (OMIM), PopSet, and books]. Each of these nodes is, in fact, an assemblage of all such data that have been grouped and indexed together; this collection of data is referred to as an *Entrez database*. Each object in an Entrez database has a unique ID number and represents, as far as possible, a stable, objective observation of data. Entrez offers a variety of search criteria for a large number of information types, e.g., all possible citations from a given author that deal with a given subject, standard names for given genes, a given nucleotide/protein sequence in the databases, etc. Entrez also helps deduce relationships between different types of data by linking with the selected nodes and carrying out the necessary computations. The associations detected in this way may be helpful in planning of future experiments as well as facilitate the interpretation of existing information.

14.5.5.2 EBI Search

The *EBI Search*, also known as EB-eye, is a text-based search engine accessible at the website http://www.ebi.ac.uk/ebisearch/advancedsearch.ebi. EB-eye enables easy and consistent access, via a network of cross-references, to the databases hosted at EMBL-EBI. These databases cover nucleotide and protein sequences, structures, gene expression data, reaction maps and pathway models, literature pertaining to the biomedical sciences, as well as the intellectual property relevant to these disciplines. The EBI Search indexes the molecular data and other information contained in these databases and organizes the resources in a hierarchical manner to facilitate search. One can access the EBI Search through the Web or through an interface of the SOAP Web service. The Web page showing the search results gives a summary of hits, i.e., matches, for each query category/domain, the actual list of hits, and other related data, including alternative views. The summary page contains information about the gene, its expression pattern, the encoded protein, the protein structure, and the relevant literature. The summary page can be exported or printed as a

report. The user can look for orthologues of a given gene in another species.

14.5.5.3 BLAST

The *BLAST* (*Basic Local Alignment Search Tool*; Altschul et al. 1990) is the most popular data-mining tool developed ever. The BLAST is a family of user-friendly sequence similarity search tools for the identification of database sequences homologous to the query or submitted nucleotide or amino acid (protein) sequences. This allows prediction of the functions of the submitted sequences and helps in the modeling of 3-D structures of the concerned protein sequences. The BLAST algorithm does not simultaneously use the entire query sequence for similarity search. Instead, it divides the query sequence into several pieces of 11 nucleotides or three amino acids each and uses one piece at a time for similarity search. This is why BLAST is called *local alignment search tool*. The above strategy facilitates a much faster search of database sequences homologous to the query sequence. It may be pointed out that BLAST is much faster and more accurate in using protein sequences than nucleotide sequences. BLAST can be used to find genes in a genome, predict function and/or 3-D structure of a protein, and find members of a gene family. *It is often said that BLAST tool can do almost anything*.

There are five main types of the BLAST (ver. 2.0) tool: BLASTp, BLASTn, BLASTx, tBLASTn, and tBLASTx. The *BLASTp* program is used for comparing the submitted protein sequence against a protein database. The two most popular BLASTp online services are available at the NCBI server (www.ncbi.nlm.nih.gov/BLAST) and the Swiss EMBnet-ExPASy server, the latter offering more options to the users. *BLASTn* compares the query nucleotide sequence with a nucleotide sequence database. The *BLASTx* tool translates the submitted nucleotide sequence into a sequence of amino acids and compares this sequence with the sequences listed in a protein database. This tool is helpful in the correction of sequencing errors and may find a better sequenced corresponding DNA segment deposited in the database. The tool *tBLASTn* uses a protein sequence as query to search a nucleotide sequence database by translating the latter into protein sequences. Finally, *tBLASTx* translates the submitted nucleotide sequence as well as the nucleotide database sequence into protein sequences and searches for homology between the two. In addition to the above, there are several specialized BLAST programs like PSI-BLAST, PHI-BLAST, etc.

Thus, similarity search for a query protein sequence can be done using either BLASTp or tBLASTn programs. BLASTp is the most suitable for indicating the function of a protein, while tBLASTn is the best for searching new genes encoding similar proteins. Similarly, a nucleotide sequence can be used by BLASTn, BLASTx, and tBLASTx tools. BLASTn identifies similar DNA sequences irrespective of the query sequence being a coding or noncoding DNA. But BLASTx analyzes a coding query sequence and identifies similar proteins in the database. tBLASTx, on the other hand, discovers proteins and ESTs encoded by nucleotide sequences comparable to that submitted as the query sequence. When nucleotide sequences are used as query, it is advisable to restrict the search to a subset of the database since *BLAST search using DNA sequences is much slower than that based on protein sequences*.

PSI-BLAST (*P*osition-*S*pecific *I*terated *BLAST*) is used to identify all the members of a very large gene family, which cannot be accomplished by the simple BLAST programs. The first round of PSI-BLAST search is a simple BLASTp search using BLOSUM62 substitution matrix. After this search, the PSI-BLAST program develops a revised substitution matrix on the basis of alignments of the search results with the query sequence. The revised matrix is used for the next round of BLAST search, after which the matrix is again updated; this process is repeated several times. In each repeat of the search (iteration), PSI-BLAST identifies genes that are more distantly related to the query sequence than those detected in the previous rounds of the search. In this way, all such genes are identified that show conservation of amino

acid residues at some positions in the amino acid sequences of the proteins encoded by them. Obviously these positions may be expected to be involved in the cellular functions performed by these proteins so that the amino acid residues at these positions have been conserved in all the members of the gene family.

Suppose a researcher has generated a nucleotide (DNA/RNA) or protein sequence and wishes to identify homologous sequences present in the relevant database. The selected BLAST tool carries out the homology search as follows:

1. The submitted sequence is broken down into small pieces or "words" of 11 consecutive nucleotides or three amino acids each. These "words" are the units for base-per-base (or amino acid-per-amino acid) comparison with the database sequences. It should be noted that *repetitive sequences are masked by the default setting of the BLAST program.*
2. Each match is awarded a positive score, while each mismatch is penalized by a negative score. These scores are awarded according to a substitution-scoring matrix. The gaps introduced in either sequence for making the alignment are also suitably penalized. The amino acid substitution matrices are more reliable than nucleotide substitution matrices. There are 100 or so different amino acid substitution matrices, of which the matrices of the PAM (*P*oint *A*ccepted *M*utation) and BLOSUM (*BLO*cks *SU*bstitution *M*atrices) series are the most popular. *BLOSUM62 is the default substitution matrix for the BLAST program.*
3. The sequence alignment is assigned an overall score, which is the summation of the scores for each of its amino acids/nucleotides. The top-scoring alignments are ranked according to set criteria, which distinguish between a similarity due to ancestral relationship (homologous sequences) and that due to random chance (similar sequences).
4. Discovered homologies or matches are further examined using information accessible through ENTREZ and other search tools.

The general steps followed for BLAST search are as follows (Fig. 14.3). (1) The first step is to specify the parameters of BLAST search. The default parameters are optimal and well tested. But the user may modify them if he/she has some specific reason for doing so. (2) The sequence, for which homology search is to be made, is submitted into the "input sequence" box of BLAST interface. This sequence has to be in the FASTA format. This format is used for representation of nucleotide and amino acid sequences used in alignment/database scanning programs. The first line of FASTA format starts with >; it is the "definition line" that includes a unique identifier. The sequence to be submitted is in single-letter code and begins in the next line. FASTA is the default format for most of the sequence analysis software. FASTA is also a similarity search tool that is much slower than BLAST, but may work better than BLAST with satellite sequences. In contrast, the sequence part of the FASTA format is called *RAW format*; it is used in some sequence analysis software. (3) The appropriate BLAST program is selected depending on the type of similarity search to be

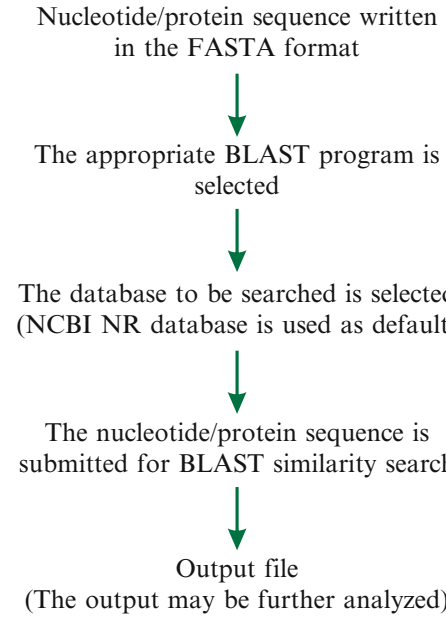


Fig. 14.3 A simplified schematic representation of the steps during sequence similarity search by BLAST. The default settings for BLAST are usually the optimum, but they can be specified by the user. The BLAST program works best with protein sequences

made. In general, it is preferable to work with a protein sequence than with a nucleotide sequence unless it is noncoding DNA/RNA. (4) The appropriate database, from which homologous sequences are to be searched, is specified. The default database used by BLAST is the NR database [nonredundant entries from GenBank, EMBL, DDBJ, and PDB (Protein Data Bank) databases] of NCBI. (5) The query sequence is now submitted to the BLAST server. (6) The results of BLAST search can be obtained either by e-mail or seen at the BLAST interface.

The so-called "traditional" BLAST report has been designed for easy readability. It has three major sections: (1) the header (it lists the BLAST program used, the sequence submitted as query, and the name of the database searched by BLAST), (2) description of each database sequence matching the query sequence, and (3) alignments with the query sequence for each matched database sequence. In addition, the bit score and *E*-values (expectation values) are also provided for each match. This report displays, by default, up to 500 sequences that matched the query sequence. But one can change this number in the case of advanced BLAST page. *As a general rule, when 25 % of the amino acid residues of any two proteins are identical, they are considered as homologous. Similarly, two DNA sequences are regarded as homologous if 70 % of their nucleotides are identical*. It may be pointed out that the above criteria do not work well when the query sequence is less than 100 nucleotides or amino acids in length. If an alignment has a small number of gaps and a few segments with high similarity, it is considered to be a good alignment. But an alignment having some identical amino acid/nucleotide residues distributed here and there over the entire sequence is not regarded as a good alignment. This criterion is useful particularly when the frequency of identical amino acids is around 25 %.

The bit score indicates the degree of similarity and depends on the alignment of similar or identical residues and the gaps, if any, needed for aligning the similar/identical residues in the query and the identified sequences. Therefore, as the bit score increases, the quality of alignment of the submitted sequence to the identified sequence also improves. The *E*-value indicates the likelihood that the detected similarity of the query to the sequence identified from the database is merely due to chance. Therefore, the match between the two sequences increases as the *E*-value decreases. For example, when an identified sequence is the same as the submitted sequence, the *E*-value will be zero. Therefore, to be certain of the homology between the query and the identified sequences, the *E*-value should be lower than 0.0001. The *E*-value depends on the size of the searched database and the system of scoring used for the search. The database sequences identified by BLAST can be considered as homologous to the query sequence only when the two are similar in the same region or, at least, in overlapping regions.

14.5.5.4 Entrez Gene

Entrez Gene, the successor to *Locus Link* program, handles queries concerning various loci. It differs from Locus Link with respect to the following two important features: (1) it has greater scope and (2) it is integrated with the Entrez search and retrieval system. Entrez Gene provides greater access than Locus Link to the genomes that are represented by the Reference Sequences of NCBI. It provides information about genes, including their official names, and allows search for genes homologous to a given gene and to obtain information about them. For example, one can easily obtain information about mouse genes, or genes of several other organisms, that are homologous to a given human gene.

14.5.5.5 Open Reading Frame Finder (ORF Finder)

The *ORF Finder* (*O*pen *R*eading *F*rame *Finder*) tool carries out graphical analysis to identify ORFs present in the submitted nucleotide sequence. Alternatively, it can analyze the nucleotide sequences retrieved from a database to identify the ORFs contained in them. The ORF Finder can detect all the ORFs that equal or exceed specified minimum size. It identifies ORFs using either the standard or an alternative

set of genetic codes from 16 different sets of genetic codes. Several different formats can be used for saving the amino acid sequences deduced from the identified ORFs. Further, these sequences can be used for BLAST searches of the sequence databases. The completeness and the accuracy of the sequence submissions are likely to be improved by the use of ORF Finder tool. The Sequin software package for sequence submission has the ORF Finder tool as a part of the package.

14.5.5.6 Search Tools for SNP Database (dbSNP)

A variety of tools are available on the left side bar of the homepage of SNP database (dbSNP) for searching the dbSNP. These programs allow SNP search by SNP genotype, SNP discovery method, the population in which the SNP was discovered, the researcher who submitted the SNP data, and marker and sequence similarity criteria. Entrez SNP is the main tool for searching dbSNP since it is a part of the Entrez search system. The search for SNPs can be based on qualifiers (aliases) or a specific search field. The Entrez SNP site lists combinations of 25 distinct search fields that can be used for this purpose. The "between markers positional query" is used when a researcher wishes to find SNPs located within any genomic region that is delineated by two STS markers. In addition, the NCBI Map Viewer tool offers other map-based queries.

14.5.5.7 Genome Remapping Service

The *Remap* tool of NCBI is used to project annotation data from one genome sequence assembly to another genomic assembly or to sequence assemblies of the RefSeqGene. It can also be used to transfer the locations of various genomic features across different genomic assemblies. The remapping is made possible through a base-by-base analysis of the concerned nucleotide sequences. The users have the option to either specify the stringency of remapping or to use the default settings. This tool displays the summary of remapping results on the Web page. However, the complete results, including the annotation and the remapped features, have to be downloaded and viewed by using the Genome Workbench graphical viewer of NCBI.

14.5.5.8 Primer-BLAST and Electronic PCR (e-PCR)

The tool *Primer-BLAST* designs pairs of PCR primers with the help of the Primer3 program. These primer pairs are designed for the amplification of the given template nucleotide sequences. The designed primers are used in an in silico PCR reaction using the given sequence as template, and the potential products are determined. These products are automatically used for a BLAST search against the databases specified by the user to assess the specificity of the designed PCR primers to the intended target sequences. The *Electronic PCR (e-PCR)*, on the other hand, is a computational program, which is able to identify STSs within the given nucleotide sequences. This tool searches the nucleotide sequences for potential STSs by using the PCR primers for similarity search. It then assesses the identified sequences matching the PCR primers for their correct order, orientation, and spacing to determine if they can serve as primers for generating known STSs in the given nucleotide sequence.

14.5.5.9 COBALT

The *COBALT* tool of NCBI carries out multiple alignments of protein sequences. The BLAST tools, viz., RPS-BLAST, BLASTp, and PHI-BLAST, are used for sequence similarity searches of the Conserved Domains Database (CDD) and the protein motif database. The protein domains are searched in the CDD by the RPS-BLAST. The tool PHI-BLAST (*P*attern *H*it *I*terated *BLAST*), on the other hand, carries out iterative searches for such sequences that have the pattern stipulated by the user. Prior to each new round of search, the PHI-BLAST program revises the substitution score matrix via PSI-BLAST. The COBALT tool uses the search results to discover a group of pair-wise constraints that are used for the alignment of the multiple protein sequences.

14.5.5.10 Splign and ProSplign

The *Splign* tool is a computer program that carries out cDNA-to-genomic sequence alignments. Similarly, *ProSplign* program aligns protein sequences to genomic nucleotide sequences. Both Splign and ProSplign programs use a variant of the Needleman–Wunsch global alignment algorithm. This algorithm enables these programs to specifically take into account introns and the splice site sequences. As a result, they are able to accurately determine the splice sites in genomic sequences and are tolerant to sequencing errors.

14.5.5.11 PredictProtein

The *PredictProtein server* provides, perhaps, the most comprehensive analysis of protein structure. This server carries out multiple sequence alignments, predicts secondary structures of proteins, detects functional motifs listed in PROSITE, predicts composition-bias regions, finds putative domain structures, and achieves fold recognition by prediction-based threading. It predicts transmembrane helix location and topology, globular as well as coiled-coil regions of proteins, the regions that switch structures, and the sites having cysteine bonds. In addition, it makes use of a collection of methods and databases to predict the presence of signal peptides and the locations of their cleavage sites, glycosylation sites, and cleavage sites for certain proteases, the presence of N-terminal chloroplast transit peptide and its probable cleavage site, and the three-dimensional structures (3-D structures) of protein molecules. It is capable of evaluation of secondary structure prediction accuracy and an automatic evaluation of prediction methods. It is also able to detect remote homologues of the submitted protein sequences.

The query protein sequence may be submitted to the PredictProtein server at cubic.bioc.columbia.edu/predictprotein/. But it may often be easier and faster to access the server at one of the mirror sites, e.g., www.sdsc.edu/predictprotein/, www.embl-heidelberg.de/predictprotein/, and www.cmbi.kun.nl/bioinf/predictprotein/. Alternatively, a request may be submitted to the META-PP server of the PredictProtein website; this server allows the query sequence to be submitted to many servers at once. The META-PP server automatically collects the results from these analyses and provides it to the user. However, it may often be faster and less confusing to access the relevant servers directly rather than using the META-PP server link to them. One may find the most suitable server by checking out EVA (*EV*aluation of *A*utomatic protein structure prediction), the secondary structure server monitoring system, at the website http://cubic.bioc.columbia.edu/eva/.

14.5.5.12 Cn3D and CDTree

Cn3D is a stand-alone tool for viewing 3-D structures of protein sequences obtained from the Entrez search and retrieval service of NCBI. In addition to the 3-D structure, it displays the protein sequence as well as the alignment; it also has powerful annotation and alignment editing features. This program runs on Windows, Macintosh, and UNIX systems. In addition, its configuration can be altered to make it capable of receiving data from the popular Web browsers. *CDTree* is a stand-alone software program that analyzes the amino acid sequences of proteins to determine their evolutionary relationships and also classifies them. It is capable of importing the existing records and hierarchies from the CDD as well as analyzing and updating them. The users can utilize this program for the construction of their own CDTrees and for creating and updating protein domain alignments. This tool is integrated with Entrez-CDD and the Cn3D application.

14.5.5.13 ScanProsite

ScanProsite (www.expasy.ch/tools/scanprosite/) compares the submitted protein sequence with the patterns and profiles listed in the PROSITE database. *PROSITE* database maintained at the ExPASy site (www.expasy.ch) has a collection of functional sites and short sequence patterns or motifs found in many proteins and shown to be associated with some biological property of the proteins. It also contains domain profiles, which

describe every position of an entire protein family. The entries in PROSITE are generally linked to Swiss-Prot and other relevant databases. The PROSITE file includes the sequence entries of the relevant databases that share the matched sequence motif of interest. The characterized motifs are well documented to minimize redundancy. The ScanProsite search result contains information on the sequence and the name of the detected pattern, its likely biological function, name of and hyperlink to the 3-D structure of the pattern (if available), and the list of segments of the submitted protein sequence having this pattern.

14.5.5.14 TAXONOMY BROWSER

The *TAXONOMY BROWSER* search tool provides taxonomic information about various species. The TAXONOMY database of NCBI has information (including scientific and common names) about all organisms, for which some sequence information is known; it includes over 79,000 species. The TAXONOMY server provides genetic information and the taxonomic relationships of the species in question. TAXONOMY has links with other servers of NCBI, e.g., Structure and PubMed.

14.5.6 Genamics SoftwareSeek

The *Genamics SoftwareSeek* is a database of both free and commercial software programs used in molecular biology and biochemistry. This database can be accessed at the website http://genamics.com/software/. This website serves as repository of the listed tools as well. This database has over 1,300 entries, which are growing rapidly. The various tools in the database are classified into 24 different categories, including Biochemistry (101 tools), Chemistry (99 tools), DNA sequence analysis (242 tools), Genetics (206 tools), Genome analysis (95 tools), Image analysis (49 tools), Molecular modeling (177 tools), PCR (15 tools), Analysis of phylogenetic relationships (90 tools), Identification of proteins (48 tools), Analyses of protein sequences (182 tools) and protein structures (96 tools), Prediction of protein (72 tools), and RNA structures (19 tools), and Alignment (165 tools) and presentation of sequences (62 tools). The numbers listed within parenthesis indicate the numbers of tools available under the respective categories. Many of the listed tools can be downloaded from this website. These tools can operate on Windows, MS-DOS, Mac, UNIX, and Linux platforms. This website also has online tools that run through an Internet browser.

14.5.7 Sequence Manipulation Suite

The *Sequence Manipulation Suite* (www.bioinformatics.org/sms2/) is a collection of Web-based computer programs designed for analysis, formatting, and preparation of textual representations of both DNA and protein sequences. The tools available at this Web portal can be used free of charge. The ver. 2 of this portal has a total of 62 tools, which are grouped into the following four categories: (1) format conversion, (2) sequence analysis, (3) sequence figures, and (4) random sequences (Table 14.8). The "format conversion tools" are the second largest in number, and they convert DNA and protein sequences written in one format into another format; they also allow some other types of sequence manipulations. For example, the "EMBL to FASTA" tool converts an EMBL DNA sequence file into the FASTA format. The "sequence analysis" tools form the largest category; these tools analyze the submitted sequences and extract the desired information. For example, the "ORF Finder" tool identifies ORFs in DNA sequences, while the "Reverse Translate" tool converts the submitted protein sequence into the most likely nondegenerate coding DNA sequence on the basis of a codon usage table. The tools in the "sequence figures" category prepare textual representation of sequences. For example, the "Restriction Map" tool identifies and depicts the positions of restriction enzyme cut sites in the submitted DNA sequence. Similarly, the "Translation" tool uses the submitted DNA sequence to prepare a textual map of its translated protein sequence. The tools in the

Table 14.8 A list of the web-based tools available at the sequence manipulation suite ver. 2

Application	Tools
Format conversion	Combine FASTA, EMBL to FASTA, EMBL Feature Extractor, EMBL Trans Extractor, Filter DNA, Filter Protein, GenBank to FASTA, GenBank Feature Extractor, GenBank Trans Extractor, One to Three, Range Extractor DNA, Range Extractor Protein, Reverse Complement, Split Codons, Split FASTA, Three to One, Window Extractor DNA, Window Extractor Protein
Sequence analysis	Codon Plot, Codon Usage, CpG Islands, DNA Molecular Weight, DNA Pattern Find, DNA Stats, Fuzzy Search DNA, Fuzzy Search Protein, Ident and Sim, Multi Rev Trans, Mutate for Digest, ORF Finder, Pairwise Align Codons, Pairwise Align DNA, Pairwise Align Protein, PCR Primer Stats, PCR Products, Protein GRAVY, Protein Isoelectric Point, Protein Molecular Weight, Protein Pattern Find, Protein Stats, Restriction Digest, Restriction Summary, Reverse Translate, Translate
Sequence figures	Color Align Conservation, Color Align Properties, Group DNA, Group Protein, Primer Map, Restriction Map, Translation Map
Random sequences	Mutate DNA, Mutate Protein, Random Coding DNA, Random DNA Sequence, Random DNA Regions, Random Protein Sequence, Random Protein Regions, Sample DNA, Sample Protein, Shuffle DNA, Shuffle Protein
Miscellaneous	IUPAC codes, Genetic codes, Browser compatibility, Mirror this site, Use this site off-line, About this site, Acknowledgments, Reference

www.bioinformatics.org/sms2/

"random sequence" category either generate entirely random sequences or random sequences from a given sample sequence or introduce mutations in the submitted sequence. Each tool has a window for submission of the DNA or protein sequence, and the results are returned as a new page. The output of each program is in the form of HTML commands, which is converted into a standard Web page by the Web browser. One can print, save, or edit the output with the help of either an HTML or a text editor.

14.5.8 PHYLIP

The *PHYLIP* (*PHYL*ogeny *I*nference *P*ackage) is free and comprises a collection of programs designed to construct evolutionary or phylogenetic trees from several types of data. The tree construction methods implemented by PHYLIP include the distance matrix, the parsimony, and the maximum likelihood methods. PHYLIP also carries out bootstrapping. It can use data on discrete characters, distance matrices, nucleotide and protein sequences, frequencies of genes, restriction sites, and DNA fragments. The PHYLIP package comprises 25 different programs for phylogenetic analyses. For example, *protpars* and *dnapars* determine phylogenies of protein and DNA sequences, respectively, by the parsimony method. The tools *proml*, *dnaml*, and *restml* use the maximum likelihood method; they are designed for phylogenetic analyses using data on protein and DNA sequences and the presence/absence of restriction sites, respectively. The *neighbor* package implements Neighbor-Joining and UPGMA methods for phylogenetic tree construction. The package *contml* draws phylogenetic inferences from data on quantitative traits and gene frequencies by the maximum likelihood method. The *drawgram* and the *drawtree* tools are used for drawing rooted and unrooted trees, respectively. The *consense* tool uses the majority rule for drawing consensus trees, while *retree* package rearranges trees, including conversion between rooted and unrooted trees.

The programs in the PHYLIP package are menu driven. The users can select the parameters for the phylogenetic analyses from among the available options. The input data is in the form of a text file prepared in flat ASCII or Text Only format by a word processor or a text editor. In case of nucleotide and protein sequence data, the input should be a high quality multiple sequence alignment of the concerned sequences. The multiple alignment may be done by a suitable program like ClustalW, which can write the output data files in the PHYLIP format. PHYLIP output files have names like outfile and outtree; the outtree files have trees written in a format used

by a number of major phylogeny packages. The source code and the precompiled executables for all the 25 programs of the PHYLIP package for Windows, Mac OS, Mac OS X, and Linux operating systems are available at bioweb.pasteur.fr/intro-uk.html.

Questions

1. "BLAST is the most popular data-mining tool developed ever." Discuss this statement giving appropriate justification for your arguments.
2. "NCBI provides a variety of useful bioinformatics tools." Examine this observation in the light of existing evidence.
3. What is a database? Briefly describe the various types of databases, and summarize the salient features of one nucleotide sequence and one protein sequence database in the public domain.
4. Briefly describe couple of tools used for molecular marker development.
5. Discuss the various applications of Clustal and PHYLIP software packages.
6. "The TASSEL and STRUCTURE programs are primarily relevant for association studies, but TASSEL software has some other applications as well." Evaluate this statement in the light of available relevant information.
7. Briefly describe one bioinformatics resource for each of the following: understanding of the higher-order biological functions, the various molecular biology databases, and tools for sequence manipulation.

Phenomics

15

15.1 Introduction

The first step in the analysis of a trait of an organism is the determination of the type and/or the level of expression of the concerned trait; this is referred to as *phenotyping*. Many biological investigations, including mutant isolation, genomic selection, genome-wide association studies, and selection in plant breeding populations, require evaluation of thousands of lines/plants within a short period of time. In general, accurate phenotyping is far more difficult than accurate genotyping for a variety of reasons, including the vast number of phenotypic traits and their sensitivity to the environmental factors (Table 15.1). Acquisition of sufficient relevant phenotype data on plot/plant basis is still challenging, especially for quantitative traits like tolerance to abiotic stress, polygenic disease resistance, and yield potential. The phenotyping of plant populations is one of the most demanding activities for the following reasons: (1) replicated trials conducted over multiple environments (locations and seasons), (2) assays at fixed times/developmental stages, (3) slow and expensive assay procedures, and (4) the need for precision phenotyping in some of the cases. Therefore, considerable research effort is being directed at the development of high-throughput phenotyping methods. These methods are generally based on a suitable combination of novel technologies like spectroscopy, imaging and image analysis, as well as high-performance computing.

15.2 Phenomics

Phenomics is a twenty-first century discipline that means different things to different groups. Gerlai (2002) coined this term for the discipline devoted to collection of large amounts of data on behavioral and other phenotypic effects of gene mutations. Later, phenomics was defined as a systematic genome-wide study of phenotypes of an organism. *Plant phenomics* is described, in simple terms, as the study of plant growth, architecture, performance, and composition using high-throughput methods of data acquisition and analysis. *Forward phenomics* uses these methods to screen germplasm collections for valuable traits. *Reverse phenomics*, on the other hand, consists of detailed analysis of a trait to unravel the various physiological, biochemical, and biophysical processes, and the genes involved in control of the trait. The term *phenome* describes the sum total of phenotypes at various levels ranging from molecules to organs and the whole organism (Table 15.2). Thus, phenomics is a transdisciplinary field requiring expertise in several disciplines, including systems biology, cell biology, genetics, molecular biology, mathematical modeling, statistics, and information sciences.

The phenomics approaches, in general, evaluate many or all of the variable traits of relatively large populations that are grown in, preferably, multiple environments. Generally, different developmental stages and several cell/tissue/

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_15, © Author(s) 2015

Table 15.1 A comparison between various features of genotype and phenotype

Feature	Genotype	Phenotype
Level of description	Single: DNA sequence	Many: from molecules to organism and even groups of individuals
Effect of the environment	Virtually no effect	Marked effects
Effect of stage of development	No effect	Tremendous effect; some phenotypes depend on specific stages of development
Effect of tissue or cell type	No effect; same in all tissues	Varies with tissue and cell type
Effect of epigenetic changes	Very little[a]	Considerable; may be important in phenomena like heterosis
Unambiguous description	Always	Extremely difficult
The list of complete set of components	Relatively simple	Extremely complex
High-throughput determination	Techniques available; improving rapidly	Techniques being developed; need considerable improvement
Complete set of components termed as	*Genome*	*Phenome*
Discipline devoted to genome-wide study	*Genomics*	*Phenomics*

[a]Methylation of DNA bases, especially, cytosine

Table 15.2 The various levels of phenome and the status of their high-throughput determination

Phenome level	Genome-wide estimation	Remarks
Transcriptome	High-throughput techniques available	Databases exist for model organisms
Proteome	Identification and quantification highly sophisticated	Expensive equipment, diverse concentrations, and properties of the components
Metabolome		
Physiological traits	Imaging techniques; high-throughput challenging	Heterogeneous response due to variation in time and space
Plant growth and development	High-throughput platforms based on image analysis	Available for laboratory and greenhouse settings; study of underground structures problematic
Field-based phenomics	High-throughput methods used for certain traits	Canopy spectral reflectance in routine use for measuring nitrogen or water-use efficiency
Chemical composition	High-throughput techniques being developed	Analysis methods mostly based on classical biochemical analyses

organ types are analyzed. The key features of growth conditions are specified and closely monitored, and the data, including metadata, are collected in formats that facilitate analysis, storage, and sharing among researchers. *Metadata* relate to the details of experimental conditions and the procedures followed for the concerned studies. Metadata are usually in digital form, and they allow comparison among the datasets obtained from the same experiment conducted by different researchers/laboratories. The various tools/procedures used in phenomics are either classical assays that depend on visual observations, various measurements and/or analyses of biochemical constituents, or the more recent target-specific, highly automated evaluation procedures. Many methods of the second category are, in fact, improved versions of the techniques that were originally developed for measuring total leaf area and were applicable to small rosette plants like *Arabidopsis*. The other methods of this group are refinements of the low-resolution imaging technology developed for remote-sensing applications. The chief advantage of high-

throughput phenomics approaches is the speed of data collection: field data that may take several days for acquisition by traditional approaches can be gathered in few hours using multiple sensors mounted onto a couple of vehicles (a field-based phenotyping platform). This would save time and allow multiple observations of a given plant/plot in a single day.

Phenomics is an area of intense ongoing research. The existing phenomics tools and techniques are being refined, their capabilities are being enhanced, and new approaches are being developed. There have been several recent reviews on the subject, including those by Araus and Cairns (2014), Cobb et al. (2013), Fiorani and Schurr (2013), Walter et al. (2012), Ollinger (2011), Berger et al. (2010), Munns et al. (2010), Chaerle et al. (2009), and Chaerle and Van Der Straeten (2001). In addition, two books dedicated to phenotyping protocols have recently been published (Normanly 2012; Panguluri and Kumar 2013). This discussion is based on the above publications and the other references listed in the References section of the book.

15.3 The Imaging Technology

The images used for phenotyping are acquired by one or more of the following high-resolution video cameras: (1) RGB (red, green, and blue; the visible range), (2) near infrared (NIR), (3) infrared (IR), (4) fluorescence, and (5) hyperspectral cameras (Table 15.3). The images are acquired either manually or by an automated high-throughput phenotyping system. In manual imaging, the camera is mounted on a tripod, and the plant to be imaged is placed on a table in front of, preferably, a light blue color wall. A white mulch/gravel may be used to cover the pot soil for easy separation of the plants in the images. Plants like wheat, barley, etc. may need physical support when grown in pots; this support should not be of green color. A light source is installed on the top and on either side of the plant. The illumination should be of optimum intensity and homogeneous, and shadows and reflections should be avoided; a cloudy day illumination is preferable. Further, underexposure is preferable to overexposure. The plants are imaged from two or three different views: one image is taken from the top view and one or two images (Sect. 15.11.2) are acquired from the side view. In case of plants like tomato, imaging from all the four side views is desirable since each view may provide some additional information. The images may be acquired by either keeping the plants fixed and moving the camera (for plants like *Arabidopsis* that have simple architecture) or keeping the camera fixed and moving the plants (for plants like wheat, barley, maize, etc. that have complex architecture). Generally, a suitable software (often provided by the camera manufacturer) is used to regulate image acquisition. A color card and a ruler should be included in the images to enable conversion of pixels into millimeters and for normalization of image colors and magnification. This simple step would allow comparison among images acquired by different imaging setups. The images should be stored in a file format like PNG or TIFF (both are in common use) that does not lead to information loss due to compression (Hartmann et al. 2011). The acquired images can be processed using a freely available image analysis tool (Sect. 15.17).

The automatic imaging systems may use either moving cameras or moving plants. The PHENOPSIS and GROWSCREEN systems use cameras that move over the plants. The LemnaTec system (used at Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben), on the other hand, uses fixed cameras, and the plants are moved for imaging. The LemnaTec system is designed for 312 - pot-grown barley plants that are kept under a controlled environment. Each pot containing a single plant is placed in a carrier tagged with a chip for a reliable tracking of the plant during the entire experimental period. The carriers are mounted onto a conveyor belt system designed for automatic retrieval of each plant and passing it through the imaging units. Images in visible, NIR, and ultraviolet (UV) spectra are captured in three separate imaging units. Every imaging unit uses one camera to acquire the top view and one

Table 15.3 Main features of the video cameras used for imaging-based high-throughput phenotyping

Camera type	Spectral sensitivity	Exposures per second	Remarks/applications
RGB camera	Visible, 400–950 nm (with filters: 400–700 nm)	17	Morphological and growth phenotyping: node number, leaf length, morphology, growth phases, nutrient deficiency, disease, and senescence analyses
Fluorescence camera	Same as RGB camera	17[a]	Low light conditions; excitation, blue light (<500 nm); stress identification and quantification, photosynthesis and chlorophyll content
NIR camera	900–1,700 nm	30	Root imaging, water distribution and dynamics, especially in response to drought
IR camera	8,000–14,000 nm	40	Temperature (within leaves and between plants): stomatal conductance; drought, heat, etc. stress studies
Hyperspectral camera	Reflectance in visible to NIR range	–	Pigment composition, nitrogen use efficiency, other biochemical features

The camera function is regulated by suitable software
[a]Recent improved models, up to 50 exposures per second

Table 15.4 Automated facilities for plant phenomics experiments accessible to the phenomics research workers (based on Fiorani and Schurr 2013)

Facility	Address
European Plant Phenotyping Network	http://www.plant-phenotyping-network.eu
Australian Plant Phenomics Facility	http://www.plantphenomics.org.au
International Plant Phenomics Network	http://www.plantphenomics.com

camera to capture the side views. The conveyor belt system is designed to enable the lifting of each plant and rotating it in the horizontal plane as per need. After imaging, the plants are taken to watering and weighing units. The huge amounts of data generated by the automated phenotyping platforms are stored in database systems. The images are analyzed by image-processing software dedicated to the specific platforms (Hartmann et al. 2011). The establishment of fully automated phenotyping platforms is quite expensive, and their running costs are high. The automated systems are designed for plants like *Arabidopsis* and the major cereal crops, but they are not capable of simultaneous phenotyping of multiple species. These systems have to be adapted to individual plant species or groups of few similar species, which requires considerable optimization of procedures. The construction of databases and their maintenance, particularly in the public domain, is a challenge. Therefore, the phenomics research community has established a network of phenomics facilities (Table 15.4), which make these facilities accessible to workers with limited financial resources (Fiorani and Schurr 2013).

15.4 Advantages of Image-Based Phenotyping

Image-based phenotyping offers several important advantages as follows. (i) These approaches are noninvasive and nondestructive. (2) As a result, the same plants can be imaged in a sequence throughout the life cycle/experiment to measure dynamic traits like growth. (3) The image-based assays are sensitive, especially when a high-resolution camera (>1 megapixel) is used. (4) They are relatively easy to perform. (5) The whole plant, plot, or even the entire field can be included in a single image. (6) Therefore, analysis of a single image would allow quantification of several traits. (7) The digital images can be stored in databases and reanalyzed later using an improved image-processing algorithm or to evaluate some new questions/hypotheses. This capability would eliminate the need for fresh experimentation, which will greatly reduce

costs and save valuable time. (8) The morphological features and the symptoms like necrosis, senescence, nutrient deficiency, etc. are quantified in place of being assayed in arbitrary units by humans. Finally, (9) imaging exploits fluorescence, NIR, and IR spectra and permits the analysis of those traits that cannot be evaluated by human eye.

However, (1) all the phenotypic traits are not measurable by imaging and image analysis. Further, (2) for the estimation of some traits like leaf angle, leaf temperature, etc., imaging may have to be done within a relatively narrow time window. In addition, (3) imaging generates huge amounts of data that have to be stored, managed, and analyzed; this requires appropriate computer hardware and software. Finally, (4) the use of sensor-/image-based data for estimating phenotypes depends on calibration of the relationship between a specific sensor-/image-based dataset and a given phenotypic feature. The calibration involves evaluation of a sample representing, preferably, the entire population with respect to variation in the relevant sensor-/image-based data. The phenotype data for the target trait are acquired from this sample using a well-established conventional method. The sensor-/image-based data and the phenotype data are used to build a robust model for prediction of the phenotype on the basis of sensor-/image-based data. This prediction model should be validated before it is used for phenotype prediction.

15.5 Reflectance Imaging

Crop leaves strongly absorb incident light in the range of 400–700 nm and strongly reflect light in the NIR region (700–1,200 nm). In addition, there is some reflectance in the visible region with a peak in the green region (~550 nm). Canopy reflectance in the visible to NIR regions is measured either as multispectral data or hyperspectral data. The *multispectral reflectance data* are acquired at few selected wavelengths. In contrast, the *hyperspectral reflectance data* are collected at narrow (1–2 nm) bandwidths ranging from 270 to 1,100 nm. The spectral indices that indicate stress levels in plants were identified from analysis of hyperspectral measurements. However, these indices are not amenable to high-throughput phenotyping. Multispectral data/imaging reveals alterations in internal and surface structures, leaf chemical composition as well as leaf water contents. The multispectral approach may be supplemented by reflectance data in the visible region. Illumination of plants with physiologically inactive NIR radiation (700–1,100 nm) allows reflectance imaging both under light and dark conditions. But under light conditions, leaves are visualized with higher contrast than in the dark. Leaf cover is efficiently discriminated from field soil due to the higher reflectance by leaves.

15.5.1 Visual Imaging

Digital imaging in the visible wavelength (400–700 nm) is called *visual imaging*. It is one of the simplest and rather useful methods of imaging, which provides information on plant size as well as color. This information allows quantitative measurement of growth, senescence, nutrient deficiencies, pathogen infections, and the consequences of stress-response mechanisms. In addition, it allows identification of the type of stress likely to be responsible for the observed changes. For example, visual imaging permits the separation of the various responses of plants to salinity stress into the following two categories: responses due to the salts themselves (salt-dependent responses) and those independent of the concerned salts (salt-independent responses). The stomata close shortly after exposure to high salt concentration, and plant growth is rapidly inhibited; this inhibition is independent of salt accumulation in plant tissues. Later, leaf senescence begins in response to salt accumulation. The separation of leaf areas into yellow and green areas allows quantitative assessment of senescence. Visual imaging also allows phenotypic analysis of large plant populations for mutant isolation or linkage mapping. Time-lapse visual imaging permits

the estimation of growth rate and visualization of effectiveness of the strategies to limit insect damage. Visual images are generally used as reference images in conjunction with other imaging techniques. The visual images to be used as reference are acquired either at the same time or just before the acquisition of the other types of images (Chaerle and Van Der Straeten 2001).

15.5.2 Near Infrared Imaging

The reflectance is high in the NIR region, especially between 800 and 1,300 nm. The reflectance declines beyond 1,300 nm due to absorption by tissue water: there are three characteristic water absorption bands at 1,450, 1,930, and 2,500 nm. The NIR region reflectance can be measured/imaged and used for various analyses, including calculation of some useful indices like water index (Sect. 15.14.6), normalized difference vegetative index (NDVI), etc. These indices make the reflectance data nearly independent of sunlight intensity. Sequences of NIR images can be analyzed to determine leaf growth and leaf growth rate. Multispectral imaging enables the detection of alterations in leaf angles of plants generated by various factors like drought stress since leaf angle alterations also change the sunlight reflection pattern. Further, information about the contents of various pigments like chlorophylls, carotenoids and xanthophylls, photosynthetic activity, and water content can be obtained from multispectral imaging (Ollinger 2011; Chaerle and Van Der Straeten 2001). The changes in chlorophyll and/or nitrogen content(s) of the leaf, and in water status and health of the tissue lead to characteristic alterations in leaf color. Even slight differences in leaf color can be detected by measuring reflectance at different wavelengths. The issues relevant to the use of NIR reflectance for routine phenotyping concern the cost, management, analysis, and interpretation of the huge amounts of data; researchers are beginning to address these issues (Fiorani and Schurr 2013).

The use of hyperspectral reflectance spectroscopy in plant breeding has been rather limited due to various reasons, including its dependence on solar radiation. But LED-based easy-to-use portable spectrometers for measuring NDVI are in common use. NDVI is perhaps the most extensively used vegetation index for assessing the responses of plants to drought, salinity, and nutrient deficiency stresses and to predict yield in the field. NDVI is estimated as follows (Ollinger 2011):

$$\mathrm{NDVI} = (R_{\mathrm{NIR}} - R_{\mathrm{red}})/(R_{\mathrm{NIR}} + R_{\mathrm{red}}) \quad (15.1)$$

where R_{NIR} and R_{red} denote reflectance in the NIR (at 800 nm) and red (at 680 nm) regions of the spectrum, respectively. Theoretically, NDVI values can range between −1.0 and +1.0. In general, healthy plants show higher NDVI values than unhealthy plants. NDVI is often directly related to photosynthetic activity, chlorophyll content, leaf area index, biomass, and yield. Sometimes, the ratio vegetation index ($R_{\mathrm{NIR}}/R_{\mathrm{red}}$) is also used, but NDVI is considered more desirable.

15.6 Infrared Imaging

Objects maintained at ambient temperature give off IR radiation of ~10,000 nm. The IR radiation between 8,000 and 14,000 nm is detected by handheld IR thermometers or canopy temperature "guns" and converted into a temperature reading. IR thermometers are relatively cheap, but are not in routine use in breeding programs. This is mainly because walking through the hundreds of plots of an experiment takes a long time, during which the environmental and the physiological conditions of the crop are bound to change. One way to avoid this problem is to mount several IR sensors onto a boom or pole attached to, say, a tractor and use them to scan the plots. A network of distributed sensors capable of wireless communication can be used to continuously monitor canopy temperature throughout the growing season. In any case, this method can be used only after canopy closure, i.e., after the soil within a row is covered by canopy, because these sensors cannot

differentiate between radiation from plants and that from the soil.

An imaging IR sensor or IR/thermal camera based on IR thermometers enables a much faster data collection than IR thermometers. These imaging systems convert IR radiation patterns into pseudo-color images; this constitutes *thermal imaging* or *thermography*. Current thermal cameras have temperature resolution of 0.1 °C. Analysis of thermal images permits the estimation of leaf/canopy temperature. The canopy temperature estimates, however, are affected by several factors, including canopy closure, air temperature, leaf size and angle, incident sunlight, wind speed, etc. Therefore, thermal sensing generally involves comparison of known healthy plants (used as controls) with the unknown test plants. The IR cameras can be mounted on an aerial device like a tethered balloon to cover large experimental area, but this would reduce spatial resolution of the images. For over 30 years, canopy temperature has been extensively used to assay crop water use, photosynthesis, and even for predicting yield. Thermography can also be used to estimate stomatal conductance and to observe the progress of freezing in plants. Further, it can reveal infections and spontaneous cell death before the symptoms become visible. Finally, an IR time-of-flight distance camera can be used to resolve individual leaves, allowing automated determination of leaf orientation.

15.7 Fluorescence Imaging

When a molecule absorbs light of a given wavelength and emits light of a longer wavelength, the phenomenon is called *fluorescence*, and such molecules are termed as *fluorophores* or *fluorescent molecules*. When fluorescence is due to some endogenous molecule, e.g., chlorophyll in plants, it is called *autofluorescence*. Wherever autofluorescence is not available or usable, either an exogenous fluorophore has to be supplied or a transgene expressing a fluorophore needs to be introduced into the plants. Illumination with the light of appropriate wavelength will excite the fluorophores, and the resulting fluorescence can be monitored at the cell (using a fluorescence microscope), organ, whole plant, or canopy level; this is known as *fluorometer*. Fluorometer is of two types, viz., imaging and non-imaging fluorometer. In *imaging fluorometer*, images of the fluorescing objects are acquired with the help of a fluorescence imaging system. The imaging system includes a light source for homogeneous illumination of the target surface, a fluorescence detector, and a computer to control data acquisition and analysis. The blue or short wavelength red light is the most often used for excitation of chlorophyll fluorescence. UV illumination allows the detection of both blue-green fluorescence and chlorophyll fluorescence. But the excitation of chlorophyll fluorescence by UV radiation will be problematic if UV-absorbing substances were present in the epidermis of the test plants. Chlorophyll fluorescence measurement in the field involves either complete protection of the plants from sunlight using a shield or box or the use of strong lasers to induce the fluorescence. Generally, the light sources function in the pulsed mode, i.e., they are modulated; this eliminates the interference by the ambient as well as the reflected light in the background. Generally, illumination is provided by either LEDs (light emitting diodes) or Xenon/halogen lamps with band-pass filters.

Monochrome CCD cameras are used to capture the fluorescence images; these cameras are set to operate in synchrony with the light pulses, and have appropriate filters for proper imaging. For example, a red filter is used to block all light of <650 nm for detection of chlorophyll fluorescence. A cooled CCD camera is used for detection of low fluorescence signals, e.g., for the imaging of F_0. The newly developed modulated imaging systems can measure fluorescence at sufficiently fast rates, e.g., FluorCams (www.psi.cz) can capture up to 50 images/s. The images reveal the fluorescence characteristics of whole leaves/plants. Each pixel in the image can serve as a distinct measurement. These images allow estimation of expected area of leaf and the rate of growth when successive images are taken over a

period of time. Fluorescence imaging can be used for analysis of various physiological processes including photosynthesis, gene expression, signaling pathways, and plant-microbe interactions. It has also been used to determine the early effects of biotic and abiotic stresses like water stress, insect attack, etc.

The *non-imaging fluorometer* uses portable handheld fluorometers, e.g., plant efficiency analyzers (PEA), to measure fluorescence from few square millimeter leaf areas. The leaf area to be monitored is first dark adapted by covering it with a specially designed clip for few minutes. The clip is then removed to expose the covered area to a saturating flash of light and the resulting fluorescence is recorded over a period of few seconds. Most modern fluorometers are modulated, and they are tuned to detect the fluorescence excited only by their own light sources. As a result, these devices can be used to measure chlorophyll fluorescence even in full sunlight. Non-imaging fluorescence measurements are extensively used to characterize leaf tissues and can be used to determine chlorophyll contents of leaves, seeds, etc. Seed chlorophyll content shows significant negative correlation with germination potential in case of cabbage. Therefore, cabbage seed lots can be classified into high and low germination potential groups on the basis of chlorophyll fluorescence.

15.7.1 Chlorophyll Fluorescence

Chlorophyll is the most abundant endogenous fluorescent molecule in plants. The light energy absorbed by green plants meets one of the following three fates: (1) One part of this energy is used for electron transport and carbon assimilation (*photochemical quenching*). (2) Another fraction of the incident light energy is dissipated as heat via the xanthophyll cycle (*non-photochemical quenching*). (3) Finally, the remainder of light energy is emitted as fluorescence (Maxwell and Johnson 2000). The proportion of light energy used for each of the above three reactions depends mainly on the quantity of the incident light energy and the physiological condition of the test plant. When a dark-adapted plant (kept in the dark for 10–15 min) is exposed to light, chlorophyll fluorescence begins at the minimal level (F_o) as all the photosystem II (PS II) reaction centers are open. The fluorescence increases very rapidly to reach the maximum level (F_m) in ~1 s. F_m is the level of chlorophyll fluorescence in the absence of electron transport, but with heat dissipation. Thereafter, the level of fluorescence slowly declines as photosynthesis is activated till it reaches an equilibrium level, F_s, in a few minutes time (Fig. 15.1). The chief limitations of chlorophyll fluorescence assays are as follows: dark adaptation would require a few minutes for each plant, measurements have to be done before

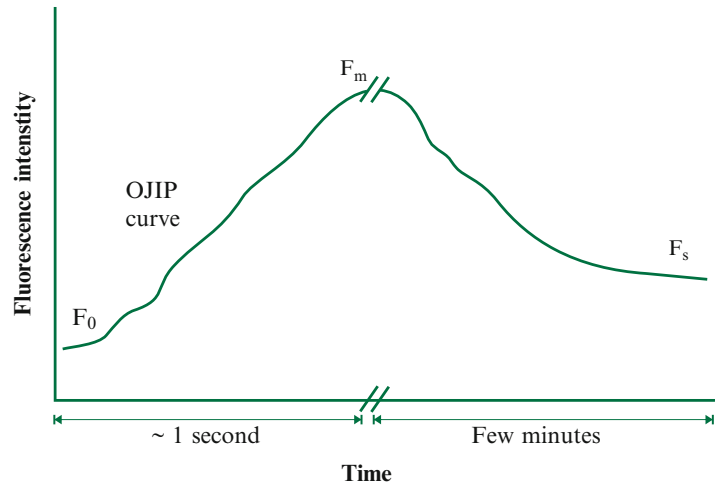


Fig. 15.1 A simplified representation of the chlorophyll fluorescence curve for a dark-adapted (kept in the dark for 10–15 min) plant exposed to high intensity illumination. The rising portion of the curve is generally termed as OJIP curve. F_o minimal fluorescence, F_m maximum fluorescence, F_s steady-state fluorescence

dawn or after dark, and application of 2-D fluorescence cameras is generally limited to plants like *Arabidopsis* (Fiorani and Schurr 2013).

Chlorophyll fluorescence primarily assesses the photosynthetic function: in general, it shows negative correlation with the level of photosynthesis. Therefore, chlorophyll fluorescence has been widely used to study photosynthesis and to detect and quantify different stresses that affect photosynthesis. The most common technique for measuring photosynthetic efficiency from chlorophyll fluorescence is the pulse-amplitude-modulated (PAM) approach. The PAM approach uses saturating light flashes of ~1 s at high intensity (~8,000 μmol m^{-2} s^{-1}) to transiently close all PSII reaction centers leading to the maximum fluorescence. The PAM approach is being used by the automated phenotyping platforms to screen genotypes for their photosynthetic performance under different environments. But measurements with PAM are feasible only at a short distance from the photosynthetic tissue so that only single leaves or small plants can be monitored. Further, the repeated use of saturating light pulses may photo-inactivate PS II (Furbank and Tester 2011).

The laser-induced fluorescence transient (LIFT) approach is another approach for measuring chlorophyll fluorescence. In this approach, a low-intensity laser light is used to alter the level of photosynthesis and to measure the resulting fluorescence transient (Normanly 2012). A fluorescence model is then used to estimate the maximum fluorescence level. This maximum fluorescence level is related to the minimum fluorescence level in the same way as it is in the case of the PAM method. Further, the findings from the LIFT approach show good correlation with those from the PAM approach. The LIFT method has been used to monitor photosynthetic efficiency of plants under controlled environments, to measure photosynthesis at canopy level, and to detect the impact of cold stress on photosynthesis. A fluorescence imaging system that uses UV-A laser for excitation has been widely used to characterize leaf tissues. This system can also be used to assess postharvest fruit quality of apples much before visible surface damage. A flash lamp-based chlorophyll fluorescence imaging system provides high-resolution imaging; this system is considered as a cost-effective alternative to the system based on UV-A laser excitation.

15.7.2 Green Fluorescence Protein

The *green fluorescent protein* (*GFP*) encoding gene is derived from jelly fish (*Aequorea victoria*). The GFP emits green (511 nm) fluorescence when excited by blue (498 nm) light or UV rays. Usually, modified GFP proteins with enhanced properties and different [red (RFP) or yellow (YFP)] fluorescence colors are used. GFP fluorescence can be assayed by true color imaging using high-resolution cooled CCD chip cameras. Suitable software, e.g., LemnaTec Bonit, is used to analyze the images to detect and quantify even slight differences in fluorescence intensity, which corresponds to the level of GFP expression in the sample. GFP has been expressed either alone or as a fusion protein. GFP can be fused with a protein that is to be tracked within cells using, especially, confocal laser-scanning microscopy. Appropriate signal sequences can be fused with GFP to guide it to specific subcellular locations. The GFP gene may be fused with different cDNAs for identification of the subcellular locations of the proteins encoded by the cDNAs and to follow their relocation during various cellular functions. GFP has also been used to develop chameleon indicators that permit monitoring of various cellular functions, including the alterations in signal transduction generated by specific stresses (Chaerle and Van Der Straeten 2001; Chaerle et al. 2009).

Pathogens like tobacco mosaic virus (TMV) may be modified to express GFP and so that their movement in the host tissues can be monitored during pathogenesis. They can also be used to visualize the resistant and susceptible host responses using confocal microscopy. Root colonization has been monitored using GFP expressing bacterial strains (Chaerle and Van Der Straeten 2001). The GFP gene functions as a marker for the detection of linked transgenes in

transgenic plant populations by using handheld long-wavelength UV lamps. Control plants lacking GFP will fluoresce red (due to chlorophyll), while those expressing GFP will fluoresce yellow due to the overlapping of green and red signals from GFP and chlorophyll, respectively. The difference in GFP fluorescence intensity permits the identification of homozygous from heterozygous transgenic seeds/seedlings. In several recent studies, any adverse effect of GFP on plant development was not observed. However, GFP expression should, preferably, be limited in time by using an inducible promoter. But autofluorescence of certain tissues might interfere with GFP detection. Further, continuous observation at high light intensities (needed for GFP fluorescence) often leads to photo-bleaching of GFP.

15.8 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) enables nondestructive, high-resolution visualization of the distribution of bound and free water, which indicates the spatial organization of organs and tissues. For example, proton-MRI microscopy has been used for nondestructive imaging of fruit development. But the findings from in vivo MRI imaging need to be verified by studies using other microscopic techniques. There is a rapid increase in proton-MRI signal as a result of tissue freezing. Therefore, MRI permits localization of frozen and unfrozen water in tissues and identification of healthy and frost-damaged tissues. Since MRI data acquisition is slow, thermography is preferable for monitoring more rapid events during freezing. MRI can be used to determine the effect of environmental conditions on water distribution within plants, and analytical tools for determining water content from MRI data have been developed. But MRI cannot be used to screen plant populations. Efforts are being made to combine MRI with positron emission tomography (PET) for investigating the structures of plants and their transport processes. However, MRI–PET data analysis remains a challenging task (Jahnke et al. 2009).

15.9 Multi-sensor Monitoring Approaches

In a *multi-sensor approach*, images of the same object are captured using different sensors, e.g., visible and thermal sensors, or a combination of visible, thermal, and fluorescence sensors. During image analysis, the images from different sensors are laid over each other with the help of predetermined reference points within the concerned images. This greatly facilitates separation of the imaged object from the background materials. A combination of two or more imaging methods might generate more phenotypic information of greater reliability (Chaerle et al. 2009). The individual imaging techniques can reveal the symptoms of a wide range of stresses at an early stage, but the use of images from multiple sensors may permit the identification of the stress responsible for the observed symptoms. For example, leaf chlorophyll content decreases in response to both water stress and nitrogen deficiency; this change is readily detected by fluorescence imaging. However, water stress also leads to stomatal closure; this in turn leads to increased leaf/canopy temperature, which is easily detected by thermography. In contrast, nitrogen deficiency does not affect stomatal closure so that there is no change in leaf/canopy temperature. Therefore, a combination of fluorescence and thermal imaging would allow the determination of whether water stress or nitrogen deficiency is the real cause of the observed decrease in chlorophyll contents of the test plants. Clearly, the optimal combination of different sensors would depend on the physiological effects of the different stresses that are to be distinguished.

15.10 Field-Based Phenomics

Initially, the phenomics techniques were developed for phenotypic evaluation of individual plants grown under controlled environments in phenotyping systems that combine robotics with automatic image acquisition and analysis. They

Table 15.5 A summary of the salient features of phenomics studies carried out in environmentally controlled phenotyping facilities and of those conducted under field conditions

Feature	Controlled conditions	Field conditions
Environmental variation	Minimized; a defined set of conditions maintained	Range of environmental conditions occurring at the target sites
Trait heritability	Higher than that in the field	Relatively lower
Measurement precision	Increased	Affected by the environment
Number of replicates for a given precision level	Smaller	Larger
Automation	High	Low
Standardization	High	Low
Genotype × environment interaction	Minimum or absent	Prevalent
Relevance to breeders and farmers	Low	High
Chief application	Development of hypotheses for evaluation in the field	Evaluation of the hypotheses developed from the controlled environment studies

are well suited for certain targeted applications, e.g., rapid screening for specific traits like shoot/root architecture. The controlled environments differ from the field conditions with respect to a number of factors, including soil, nutrients, water, and radiation, and in terms of supporting the production of flowers and seeds. Therefore, the phenomics studies under the controlled environments differ from those conducted under field conditions with respect to several salient features (Table 15.5). As a result, the phenotype data for plant breeding applications should be collected from field studies, for which field-based phenomics approaches (Table 15.6) are being developed (White et al. 2012; Araus and Cairns 2014). However, the data from controlled environment systems may complement the findings from the field studies. A good field-based phenotyping system should be rapid, flexible, reliable, applicable to small plots, and capable of evaluation of multiple traits in a single passage through the plots. It should provide high resolution, allow multiple view angles, control illumination, and permit regulation of the distance between the target plant/plot and the sensors. The field-based phenotyping (FBP) system should permit repeated measurements during the season and even in a single day.

A FBP platform requires the following: (1) instruments for data collection from the field; (2) systems for providing power, protection from weather, etc.; (3) systems for integrating the different instruments; (4) vehicles for rapid and accurate positioning of the instruments in the field; (5) high-throughput analysis systems for leaf, seed, etc. samples collected from the field; (6) computer programs for analysis of the huge datasets; and (7) integrated management procedures to maximize the reliability and efficiency of the FBP. In general, multiple instruments would have to be simultaneously used for evaluation of the various traits that need to be assayed. Vehicle options for a FBP include high-clearance tractors, manned/unmanned aircraft, and secured balloons; each of these options has its advantages and limitations. For instance, tractors would image the plants from close range and would give high resolution. But imaging of the entire field would take time, during which the intensity of light, the speed of wind, etc. might change. This would be particularly relevant for thermal imaging. Further, tractors would compact the soil, damage plants, and spread diseases and pests. Aircraft can cover large areas rapidly and would permit detection of disease epidemic initiation. However, aircraft cannot be used in rough weather, would incur high operating costs, and may not provide sufficient resolution. The use of secured balloons would allow simultaneous imaging of many plots, and the images can be acquired automatically at regular intervals (White

Table 15.6 A selection of promising phenomics approaches (based on White et al. 2012 and other sources)

Trait measured	Surrogate measurement	Measurement technique	Remarks
Drought/salinity/ nutrient deficiency response	NDVI[a]	LED-based spectrometers (NI)[a]	Also used for prediction of yield
Stomatal conductance; photosynthesis	Canopy temperature	IR thermometers (NI); thermal imaging[b]	Used after canopy closure; affected by environmental conditions
Chlorophyll content	Chlorophyll fluorescence	Fluorescence imaging[b]/ measurement (NI)	The F_{735}/F_{700} is linearly proportional to chlorophyll content
	NDVI	Measurement of red and NIR reflectance[b]	Correlated with chlorophyll content
	Red edge NDVI $[(R_{750} - R_{705})/(R_{750} + R_{705})]$	NIR measurement (NI)[b]	This index indicates chlorophyll content of the leaves
	Transmittance in the red and NIR regions	SPAD chlorophyll meters (NI)	These hand-held meters are being used by many large breeding programs
Leaf/shoot area/ biomass	Predicted leaf/shoot area/biomass	Visual imaging[b]; NDVI[b] (NI)	Usually, one top and two side view images analyzed
Growth rate	Predicted leaf/shoot area/biomass	Analysis of a sequence of visual images[b]	Visual images of the same plant taken at several points of time
Senescence, chlorosis or necrosis	Estimation of the affected leaf area	Visual imaging and fluorescence imaging[b]	The color scheme for analysis specified by the user; reduced fluorescence in affected leaf areas
Root traits	Root trait measurements based on image analysis	Visual image analysis	Seedlings grown in aeroponics/hydroponics or on agar medium; used by phenotyping platforms
		Neutron radiography	Root systems of plants grown in soil; in experimental stages
Seed/fruit size, color, etc.	Traits determined from image analysis	Visual image analysis	Images acquired with cloudy day illumination
Photosynthesis	Chlorophyll fluorescence	Fluorescence imaging[b]/ measurement	Fluorescence negatively correlated with photosynthesis; F_v/F_m ratio multiplied by light intensity
	Reflectance at 531 and 570 nm	Photochemical reflectance index (NI)[b]	Indicates diurnal radiation use efficiency
Water-use efficiency	SPAD chlorophyll meter readings	SPAD chlorophyll meters	Close direct relationship with transpiration efficiency
	Leaf/canopy temperature	IR thermometers (NI); thermal imaging[b]	Precise calibration is necessary
Stomatal conductance	CID value	Carbon isotope composition of plant dry matter	CID value is a reliable indicator of transpiration efficiency
Pathogen infection	Chlorophyll fluorescence	Fluorescence imaging[b]	Detection of infection before symptoms are visible; quantification of resistant and susceptible responses
Plant water status	Stomatal conductance	Porometers (often, hand-held)	Not a good indicator of water status in most crop species
Drought stress tolerance	Stay-green phenotype	Visual imaging[b]	Indicates ability for photosynthesis under water stress; a combination of visual, fluorescence and thermal imaging is more informative

(continued)

Table 15.6 (continued)

Trait measured	Surrogate measurement	Measurement technique	Remarks
	Water index, normalized difference water index	NIR reflectance measurement (NI)[b]	Permits the detection of onset of water stress
Heat and drought stress tolerance	Canopy temperature depression	IR thermometers (NI); thermal imaging[b]	Temperature should be measured during mornings and the afternoons

A *surrogate measurement* quantifies a trait that has dependable, significant and high correlation with the target trait
[a]NDVI, normalized difference vegetation index; NI, non-imaging
[b]Techniques having potential for field-based phenomics

et al. 2012; Araus and Cairns 2014). Imaging is increasingly replacing point measurement devices as it offers several advantages (Sect. 15.4), including being amenable to multi-sensor approaches (Sect. 15.9).

A FBP vehicle prototype has been developed; this vehicle has sensors for measuring plant height, canopy temperature, and reflectance at three selected wavelengths. The observations by this platform are geo-referenced using GPS (geographical positioning system) to minimize inadvertent errors by the driver(s) of the vehicle(s). An automated FBP system uses the "light-curtain system" for monitoring of morphological traits like shoot height, total leaf area, etc. of maize plants and the spectral reflectance of canopy. This platform comprises a tractor that carries a set of light barriers deployed on suitable vertical poles. These light barriers are moved along the rows of young maize plants growing in the field. Combine-mounted NIR spectroscopy instrument can measure protein or oil content of seed. After acquisition, the data are transferred to a computing facility, where they are analyzed to deduce canopy structure, spectral indices, etc. It may be pointed out that quantitative imaging in the field is affected by variable illumination and reflectance from plant canopies, altered quality of sunlight due to weather conditions, plant movements caused by wind or rain, etc. The experimental design should be able to account for soil and microclimate heterogeneity and to minimize their effects on the findings. In order to be able to estimate genotype × environment interactions, a set of environmental factors should be recorded during the experimental period (White et al. 2012; Araus and Cairns 2014). The FBP approaches are still in the developmental stages.

15.11 Morphological and Growth Analyses

The morphological features of plants, e.g., plant height, number of nodes, internode length, number of leaves, etc., provide valuable information about their structure and function. For example, the leaf area or volume of a plant allows reliable prediction of its biomass. Further, monitoring of leaf area/volume of a plant over a period of time permits the estimation of its growth rate. In addition, leaf color serves as a good indicator of plant health, including the effects of diseases, nutrient deficiencies, and senescence. Imaging technologies allow a nondestructive qualitative as well as quantitative assessment of the morphological features and growth rates of plants in a high-throughput manner.

15.11.1 Dynamic Measurement of Leaf Area

Growth may be defined as an increase in the total dry mass, volume, height, and/or total area of a plant, which may result from cell division, cell expansion, and/or cellular differentiation.

Several mechanical methods, e.g., linear voltage differential transducers, rotary resistance transducers, etc., have been used to measure growth both in the field and indoors. These methods measure only elongation, are labor intensive, and/or have low resolution. Recent methods for estimation of leaf growth are based on analysis of a sequence of visual images of the target plants. These methods are classified as morphometric, optical flow, or particle-/marker-tracking approaches (Mielewczik et al. 2013). The *morphometric methods* use segmentation algorithms to create leaf/rosette outlines and then calculate the projected leaf area from these outlines. These methods are generally used by high-throughput phenotyping platforms. These methods have been used to assess growths of single leaves, but are the most effective in the study of whole shoots. The *optical flow methods* provide high resolution in both time and space for leaf, root, and hypocotyl growth analyses. They analyze the movement of structural features like vein intersections, trichomes, etc. or artificially created marks like ink dots applied onto the leaf surface. But these methods are sensitive to brightness fluctuations since it is often not feasible to ensure constant illumination. In case of *particle-/marker-tracking* methods, a set of landmarks is created within the target root/leaf/hypocotyl. A pattern-matching algorithm is used to follow the displacement of these landmarks in the consecutive images. This procedure allows the estimation of root, leaf, or hypocotyl area (Mielewczik et al. 2013). In addition, relative growth rate (RGR) can be estimated with high temporal resolution (Sect. 15.11.3). The image-based estimates of leaf area show high positive correlation with those obtained by actual measurements. The accuracy of this technique will be high during seedling stage and will decline in the later stages, e.g., after tillering in cereals, due to overlapping of leaves.

15.11.2 Plant Biomass Estimation

Digital analysis of visual images provides estimates of plant height, plant volume, and biomass. Such analyses require three different images: one image of the top view and two images representing the side view of each plant. The plants are rotated by 90° in the horizontal plane before taking their second side view image. The three digital images of a plant are processed using suitable software to estimate its leaf area and plant volume, which is used to predict its biomass. The biomass estimation requires calibration of the relationship between plant area/volume and biomass at different stages of growth of the given plant species. The calibration step involves growing the plants under controlled conditions to the desired developmental stage, taking their images and then harvesting them. The leaf area of these plants is measured by a conventional method and their fresh weights are determined. Then, the leaves, stems, etc. of the plants may be separated, if desired, and their dry weights (biomass) recorded after drying them in an oven. A calibration curve is then plotted for the shoot area estimated from the images and the actual shoot area and/or biomass determined as above. Based on this curve, a suitable prediction model may be developed and used for biomass prediction. The main limitations of this approach are the overlapping of leaves in older plants and the background noise caused by soil. In addition, the calibration curve will change in response to stress, the developmental stage, the plant species, and even the camera setups; this necessitates recalibration whenever one of the above factors is changed (Fiorani and Schurr 2013).

This approach was first developed for *A. thaliana*, but is now used for monitoring growth of major grain crops and ornamentals. The available evidence shows that plant area/volume estimated from images shows a good correlation with biomass in species like wheat, barley, sorghum, and tomato. Commercial facilities for leaf area and quasi-3-D growth analyses under controlled environmental conditions are available (www.plantphenomics.com and www.plantphenomics.org.au). A platform for high-throughput estimation of biomass of field-grown cereals uses multiple sensors fixed on a tractor trailer having two light curtains (Busemeyer et al. 2013). This platform has three laser distance sensors, two 3-D-time-of-flight cameras, and a hyperspectral imaging

system. The raw data from these sensors are processed using a software package to extract the volume and the density of plants in each plot. The hyperspectral imaging allows estimation of dry matter content, but appropriate steps are needed to minimize the noise in data. The data from time-of-flight cameras, laser distance sensors, and light curtains is used to calculate plant height, penetration depth, and coverage density and also to get an estimate of the number of plants. Multiple linear regression was used to develop a biomass prediction model, which was calibrated with actual biomass data of the previous year. The predicted biomass showed high positive association with the actual biomass (the coefficient of determination, R^2, being 0.92).

In addition to the imaging techniques, there are some nonoptical approaches for direct determination of plant biomass, e.g., electrical determination of the water content of a plant, positron emission tomography, and portable nuclear magnetic resonance (NMR) imaging devices. These approaches assess biomass nondestructively and can resolve 3-D plant architecture. It may be expected that technological advancements in the future may permit their field application.

15.11.3 Basic Plant Growth Analysis

Analysis of a sequence of visual images of the same plant that were taken at different points of time allows estimation of several growth parameters like growth rate, stay-green period, flowering period, ripening dynamics, etc. The increase in shoot area of a plant over a given time period can be estimated by analyzing the images of the plant taken at the specified points of time. When images are taken at a series of time points, the shoot area may be plotted against time to yield a growth curve. This curve is sigmoidal for most plant species when plotted from the seedling stage to the early reproductive stage. However, leaf senescence during seed ripening leads to a decrease in the projected leaf area. Color information of the leaves may be used to determine green leaf area. Further, the data on shoot area would allow estimation of shoot biomass. They would also permit the development of a growth model through curve fitting. This model can be used to estimate absolute growth rate (AGR) as follows (Tessmer et al. 2013):

$$\mathrm{AGR} = (W_2 - W_1)/(t_2 - t_1) \qquad (15.2)$$

where W_2 and W_1 are estimates of shoot dry weights (derived from the growth model) at time points t_2 and t_1, respectively. The AGR will indicate the increase in shoot weight of the plant per unit time. The time taken by a genotype to attain the maximum absolute growth can be treated as a trait, which is affected by factors like drought, salinity, and other stresses.

Relative growth rate (RGR) can be estimated by the following formula (Tessmer et al. 2013):

$$\mathrm{RGR} = (\ln W_2 - \ln W_1)/(t_2 - t_1) \qquad (15.3)$$

where, ln is natural log. In case W_1 and W_2 represent sample means, the natural logs of individual plant data (and not the actual data themselves) should be averaged to determine the values of ln W_1 and ln W_2. This procedure is necessary since averaging of the actual individual plant data introduces bias in the estimate of RGR (Hoffmann and Poorter 2002). The RGR is generally the highest in the seedling stage, after which it declines gradually. RGR is independent of plant size; therefore, it allows comparison between genotypes having different growth habits. Leaf area duration can be estimated from the growth model; this estimate would provide a measure of the total leaf area present over the given period of time. In addition, morphological parameters like height, width, compactness, and leaf area-to-plant height ratio of the plant can also be estimated. Compactness of a plant describes the area that completely encloses the plant; this is a useful feature in plants like *Arabidopsis*. The leaf area-to-plant height ratio, on the other hand, is relevant for plants like wheat and barley. It is important that all seedlings used in such a study are of the same size at the start of the study. A large memory for data storage would be needed even for relatively small size experiments on growth analysis.

15.11.4 Assessment of Structure/ Development

Analysis of plant architecture is based on three (one top and two side view) visual images of each plant acquired with cloudy day illumination. These images are segmented and skeletonized. Skeletonization involves determination of the midlines of the stems and leaves and drawing these lines on the image. The skeletonized images are analyzed to gain information about plant architecture, including maximum height and width of the plant, internode length, leaf angle, number of leaves, etc. A graphic image of the plant is then made to permit reliable identification of nodes on the basis of the minimum length of branching structure specified by the user. The image analysis software tracks the leaf from node to the leaf tip and estimates leaf area. The estimates of leaf angle, stem area, internode length, etc. provide a good idea of plant architecture, morphology, and growth. The active male inflorescence (tassel) of maize plants is detected by the presence of yellow or violet anthers at the top of the shoot. Similarly, other plant organs, e.g., panicles in rice, ears in barley and wheat, etc., are identified on the basis of color definition of these organs provided by the user. Once identified, the length, width, and areas of these organs can be estimated from the visual images. Further, specific organs, e.g., panicles of rice plants, may be imaged after they are separated from the plants and are properly spread out. Such images would greatly facilitate the estimation of branch orientation, spikelet number, panicle area, etc. Finally, imaging of densely grown plants would allow estimation of such universal parameters as ground coverage, color and orientation of plants, total panicle area, etc.

15.11.5 Measurement of Senescence/ Necrosis

The measurement of traits like senescence, necrosis, etc. that lead to color changes in leaves and other plant organs is based on color analysis of visual images of the plants. In general, three (one top and two side view) images of each plant are taken. A cloudy day illumination is used so that small changes in color are clearly discernible. The images are segmented, and the part of image representing the plant is divided into different regions based on a color classification scheme specified by the user. The color scheme should have two or three shades of green plus the color classes for chlorotic, necrotic, and bleached regions. The image areas corresponding to the different regions defined by the color scheme are quantified; these estimates provide a measure of tissue damage, necrosis, and/or senescence. The progression of senescence, chlorosis, etc. over time within a plant can be monitored by imaging the plant at several points of time. Further, different plants/genotypes can be compared for these traits. A control group is included to provide a relative measure of the various changes. Fluorescence images can particularly facilitate the determination of senescing areas of leaves since such areas would show reduced chlorophyll content as well as chlorophyll fluorescence.

15.11.6 Analysis of Root Systems

The analysis of root traits of the field-grown plants generally involves visualization of excavated root systems. Root systems are also analyzed using camera systems that are introduced into the soil through small tubes made of Plexiglas. Finally, root biomass may be indirectly quantified, for example, by analysis of changes in the electrical properties of the soil due to water uptake by roots. Imaging techniques for 2-D and 3-D analyses of soil-grown root systems include X-ray computed tomography, neutron radiography (NR), and magnetic resonance imaging. NR permits analysis of complete root systems growing in soil. There is strong contrast between the soil and the roots, and the image processing is relatively easy. But most automated phenotyping platforms use aeroponic or hydroponic culture systems for direct visualization and imaging of roots. The root images are analyzed to

determine total root length, branching angles, and other root traits. A high-throughput root-monitoring platform (PlaRoM) has been developed for analysis of root systems of up to 50 seedlings grown on agar medium in rectangular plates. This platform is supported by automatic image-processing software that yields results of very high accuracy (Gregory et al. 2009; Yazdanbakhsh and Fisahn 2009). A prototype of an automated shoot and root-imaging platform for plants grown in soil-filled rhizoboxes or rhizotrons has been recently described. This system is able to simultaneously acquire growth rates for roots, shoots, and shoot-to-root ratios. In such a system, the proportion of roots observable at the transparent side of a rhizotron depends mainly on the plant species, and this proportion is loosely correlated with average root diameter (Nagel et al. 2012). The root system image analysis tools (Sects. 15.17.6, 15.17.7, 15.17.8 and 15.17.9) generally analyze grayscale (dark roots in a bright background) images.

15.11.7 Seed and Fruit Phenotyping

Image analysis can reveal several important features like size (estimated from seed area), volume, length, diameter, shape, and color of seeds. The seeds are separated from each other and spread on a flat surface, placed individually in wells of multiwell plates, or held in specially designed seed holders to keep them in the desired position. Whenever possible, backlight images of the seeds are taken. However, top light images should be taken whenever color is to be scored. A cloudy day illumination is preferable as it minimizes reflections. Generally, visual images are taken, but fluorescence and IR images may also be acquired where desired. In specific cases, 3-D laser-scanning technology (Sect. 15.11.8) can be used to assay seed volume. Analysis of seed color would permit detection of infections and the identification of specific traits like vitreousness of maize kernels. But calibration would be required for the estimation of traits like seed volume, seed weight, etc. Similarly, visual imaging allows phenotyping of fruits for color, size, and shape. In addition, flatbed scanners and transmitted light can be used for rapid analysis of seed size and shape. Finally, NIR spectroscopy is used for large-scale determination of seed water, protein, oil, and starch contents, but this application requires extensive calibration (Fiorani and Schurr 2013).

15.11.8 Laser Scanning: 3-D Plant Morphology

Digitized 3-D structures of plants have been obtained by using ultrasonic or electromagnetic devices, but this is labor intensive and time-consuming. The laser-scanning technology uses 3-D scanners to generate up to millions of accurate measurements of an object relatively rapidly and, to a large extent, automatically. The 3-D scan data are processed using appropriate software to construct 3-D images of the object and to derive precise information about its various morphological features. A 3-D scanner is a device that uses light, lasers, or X-rays to capture the geometry of physical objects with a very large number of measurements called dense point clouds. The 3-D scanners may be designed for short-range (<1 m) or mid- and long-range (>2 m) scanning. These scanners and the 3-D scan processing software are fast, accurate, and affordable. A portable LIDAR (light detection and ranging) scanner is a 3-D scanner that measures the distance from the sensor to the target object by either time-of-flight, optical-probe or light-section method. In the case of time-of-flight method, the distance is estimated from the time a laser pulse takes to travel from the sensor to the target object and return back to the sensor. A LIDAR instrument was used to scan tomato canopy from three positions around the canopy spaced at 120° from each other, and the scan data from these positions were registered together. The points representing leaves were taken out and used to create leaf images, and leaf areas were accurately estimated from these images. Further, 3-D images of the canopy were generated, and vertical leaf area density, leaf area index, and leaf inclination angle were estimated from the images (Hosoi et al. 2011).

Field-grown barley, oat, and wheat crops were scanned six times during the growing season using a laser 3-D scanner. The scanner was mounted on a movable rack of ~3 m height so that it scanned the ground beneath the scanner. Analysis of the scanning data allowed estimation of plant height and grain yield; these estimates correlated well with the actual measurements of the traits (Lumme et al. 2008). However, leaf shape can be captured accurately only if scanning were done under windless conditions so that the leaves do not move due to wind.

15.12 Analyses of Chemical and Physiological Parameters

Physiological processes are important determinants of the performance of crop species. These processes, in turn, are affected by several factors like chlorophyll content, plant water status, soil water content, etc. Further, the chemical composition of plant produce determines its quality and, ultimately, its end use. The traditional methods for estimation of physiological processes and chemical composition are generally slow and not amenable to high throughput. Efforts are being made to develop high-throughput surrogate measurements for rapid and cost-effective phenotyping for these traits. A *surrogate measurement* determines the level of some other trait that shows reliable and predictable relationship with the target trait. For example, the SPAD chlorophyll meter measures transmittance of red and IR light through plant leaves. These readings are readily converted into chlorophyll contents of the leaves (Ling et al. 2011). Thus, *the red and IR transmittance readings serve as surrogate measurements for leaf chlorophyll content*.

15.12.1 Estimation of Relative Chlorophyll Content

Leaf chlorophyll content is affected by abiotic and biotic stresses and senescence. Therefore, this trait is a valuable indicator of plant health. Conventional chlorophyll content estimation is based on spectrophotometry of plant extracts, which is not amenable to high throughput. But analysis of fluorescence images permits high-throughput estimation of chlorophyll content and photochemical yield of PS II. The ratio of chlorophyll fluorescence at 735 nm (F_{735}) to that in the range of 700 nm and 710 nm (F_{700}) is reported to show linear relationship with the chlorophyll content, the R^2 being >0.95 (Gitelson et al. 1999). Therefore, the F_{735}/F_{700} ratio seems to be a precise surrogate measurement for the estimation of chlorophyll content of leaves. In addition, reflectance measurements in the NIR region are used to estimate chlorophyll content (and the contents of other pigments), and several indices have been proposed for this purpose (Ollinger 2011). One such index, the *Red Edge NDVI*, is estimated as the ratio $(R_{750} - R_{705})/(R_{750} + R_{705})$, where R_{750} and R_{705} represent reflectance at 750 and 705 nm, respectively.

The commercially available SPAD chlorophyll meters, e.g., SPAD-502, nondestructively estimate relative chlorophyll contents of leaves of a range of plant species (Ling et al. 2011). These hand-held devices measure transmittance of red (650 nm) and NIR (940 nm) regions of light through the target leaf. The red and NIR lights are generated by LEDs in the illumination system of the meter. The target leaf is inserted into the sample slot of the meter for measuring transmittance through an area of 2 mm × 3 mm, but the leaf should not be more than 1.2 mm thick. If fine veins are present in the target area of the leaf, the average of several measurements should be used to minimize error. The meter should be protected from sunlight during the measurement to avoid interference from sunlight. The transmittance values are converted into SPAD readings, which are proportional to chlorophyll contents of the leaves. The absolute chlorophyll contents of leaves can be estimated with the help of a calibration curve. The calibration curve is constructed by determining the SPAD readings of a sample of leaves with different chlorophyll contents. The chlorophyll contents of these leaves are then determined by

spectrophotometry of leaf extracts. Finally, the relationship between the SPAD readings and the chlorophyll contents is used to develop a prediction model. The SPAD meters give reliable estimates of chlorophyll content, and they are being routinely used in many breeding programs.

15.12.2 Monitoring Photosynthesis

Photosynthesis is one of the principal contributors to the yields of crop plants. The traditional measurements of photosynthesis have low throughput and are labor intensive. The high-throughput estimations of photosynthesis are based on either measurement of oxygen evolution at a supersaturating CO_2 concentration or quantification of chlorophyll fluorescence. *The amount of chlorophyll fluorescence shows negative correlation with photosynthetic activity*. The maximum quantum yield of PS II, i.e., the efficiency when all the PS II reaction centers are open, is given by the ratio F_v/F_m, where F_v equals $F_m - F_o$ (Maxwell and Johnson 2000). The F_v/F_m ratio generally ranges between 0.78 and 0.84 for healthy plants. The efficiency of PS II is multiplied by the light intensity to estimate the rate of linear electron transport, which is correlated with CO_2 fixation. But this relationship may not hold good in the field due to factors like photorespiration, nitrogen metabolism, etc. The fluorescence measurements can be combined with determination of gas exchange to estimate the quantitative relationship between carbon assimilation and PS II efficiency. This estimate would allow the determination of actual carbon assimilation from the amount of incident light energy used for photosynthesis; the latter can be estimated from fluorescence images. It is problematic to subject crop plants to high-throughput screening for chlorophyll fluorescence after they passed the seedling stage. This is because the vegetative structure of older plants is complex. It is difficult to obtain accurate images of such plants without constructing full 3-D models. The 3-D model construction, however, requires images acquired from many angles. Another approach uses the Hue parameter estimates for monitoring photosynthesis. The Hue parameters are extracted from RGB images of leaf disks (http://www.fact-archive.com/encyclopedia/HSV_color_space); these values may provide a rough estimate of photosynthesis. But this would require establishment of a linear relationship between the Hue parameters and photosynthesis in the concerned species under the conditions prevailing in the experiment.

15.12.3 Assessment of Water Use

Water-use efficiency (WUE) of a plant is the ratio of the amount of water used for metabolism by the plant to that lost through transpiration. In crop production, generally the WUE of productivity or the integrated WUE is relevant. The integrated WUE, in simple terms, is the ratio of biomass produced by a plant to its transpiration rate. Transpiration rate can be measured by gravimetric monitoring of water consumption over time by the plants. The plant growth can be estimated nondestructively by digital imaging. Another approach to estimate WUE is based on the phenomenon of carbon isotope discrimination (CID) by plants. It is well known that plants preferentially avoid using $^{13}CO_2$ that comprises the heavy isotope of carbon. The $^{13}CO_2$ is present in the atmospheric CO_2, and this discrimination is exercised during both CO_2 diffusion into leaves and carbon assimilation. This discrimination is reflected in the carbon isotope composition or CID value of the plant dry matter. *The CID value of a crop genotype is strongly associated with its stomatal conductance and is a reliable indicator of its transpiration efficiency, especially in C_3 crops.* CID values have been used in wheat to identify and characterize the genetic variation in transpiration efficiency. It has also been utilized to develop wheat varieties having higher WUE and improved yield, and they have been used for commercial cultivation. The CID values can be determined from plant samples collected at the end of the crop season, which is a great advantage of this approach. These values reflect the integrated effect of WUE of the given genotypes

for the entire growing season. The chief limitations of this method are high cost (up to US $ 30/sample) and the need for normalization of CID values for photosynthetic capacity or yield potential. Further, sometimes CID values are not correlated with yield potential. For these reasons, CID value is not in common use in breeding programs (Furbank and Tester 2011; Walter et al. 2012; Chaerle et al. 2009).

Another surrogate measure of water-use efficiency is based on SPAD chlorophyll meter readings. In case of groundnut, these readings are recorded on the second or third completely expanded leaf, and they show a close direct relationship with transpiration efficiency (see Panguluri and Kumar 2013). Finally, thermography is a powerful tool for monitoring plant-water relations, including transpiration and canopy water use, through analysis of leaf or canopy temperature. In general, leaf or canopy temperature increases as transpiration rate decreases primarily due to stomata closure. But this application requires a precise calibration of the leaf/canopy temperature data obtained from thermography with the actual canopy temperature and with the real transpiration rate data collected by traditional methods. The calibration is necessary because canopy temperature readings are likely to have a variable contribution from soil temperature due to differences in canopy closure. Further, the genotypes being compared may be in different developmental stages, which may affect their transpiration rates and, consequently, leaf/canopy temperature. Finally, differences in velocities of wind and/or sunlight/shade status between different points of the canopy would also affect the temperature data extracted from the thermal images (Furbank and Tester 2011).

15.12.4 Estimation of Soil Water Content

Soil water content affects chemical, physical, and biological properties of soils, including water uptake by plants. Many methods directly measure soil water content, but several others use a surrogate measure. The standard direct method for determining soil water content involves collection of soil sample from the field and weighing the sample before any water is lost. The sample is then dried in an oven, and its weight is determined. The loss in weight due to drying gives a direct measure of the amount of water present in the soil sample. This amount of water is generally converted into a normalized standard unit of soil water content, viz., mg mg^{-1} or m^3 m^{-3} of soil. There are several indirect methods for estimation of soil water content, including neutron probes, time domain reflectometry (TDR) and remote sensing. The indirect methods rely on calibration of the relationship between the data obtained from the given method with the actual soil water content measured by the direct method (Anonymous 2008). The direct method of soil water content determination is slow, labor intensive, and time-consuming. The neutron probes do not provide accurate estimates of water content of the soil near the ground surface. The results from TDR may be affected by soil salinity and temperature. The remote sensing is widely used to estimate soil water content for ecological and hydrological studies involving large areas, but it has limited usefulness at the field level.

Mobile NIR and visual spectrophotometers can be used for estimation of soil water content on plot and field scales. But a method of more general applicability uses visual images of soil for prediction of soil water content. Soil surfaces reflect visible light and generate variations in color and brightness levels in the visual images. The visual images can be processed using computer software to obtain unique gray levels corresponding to the different colors and brightness levels of the images. The relationship between gray level data and water contents of soil samples was used to develop and validate a prediction model. The soil surface image gray level for a given soil showed a negative linear relationship with square of the soil water content ($r = > -0.91$). This model was able to predict the soil water content from gray level data with a high level of accuracy ($r = 0.99$). The chief advantages of this method are the ease, the speed, and the low costs of image acquisition and image analysis (Zhu et al. 2011).

15.12.5 Analysis of Chemical Composition

High-throughput platforms for several types of chemical analyses have been developed/are being developed. For example, highly sensitive and high-throughput quantitative analyses of six phytohormones (auxins, cytokinins, abscisic acid, gibberellins, salicylic acid, and jasmonic acid) have been developed. These analyses involve solid-phase extraction using 96-well column plates, followed by liquid chromatography coupled with mass spectrometry. A method using gas chromatography and mass spectrometry for analysis of amino acids present in sub-milligram quantities in the fresh plant materials is easily adapted to high-throughput screening approaches. Methods for profiling of plant membrane lipids by electrospray ionization mass spectrometry have been developed and used successfully in *A. thaliana*. Flow cytometry is easy to use, highly accurate, highly reproducible, and cheap; it is used for measuring plant genome sizes (nuclear DNA contents). Methods are being standardized for transient RNAi (RNA interference) assays (based on PEG-mediated genetic transformation) in 96-well plates for high-throughput gene function analysis in plants. Carbohydrate microarrays can be used to screen large numbers of samples for antibodies/proteins and carbohydrate-binding modules and to investigate enzyme activities. NIR spectroscopy is routinely used in laboratory for analyses of grain and fodder quality traits, e.g., contents of nitrogen, moisture, fiber, carbohydrates, amino acids, etc., in maize, wheat, sorghum, etc. (Normanly 2012).

15.13 Biotic Stress Detection

Any environmental factor that limits the performance of a crop genotype is termed as *stress*. Infections by viruses, fungi, nematodes, and bacteria and infestations by insects constitute *biotic stresses*. Most of the biotic stresses markedly affect host metabolism, including photosynthesis. Fluorescence imaging would reveal the changes in photosynthetic activities of leaves infected by pathogens. For example, when resistant tobacco plants are infected by TMV, there is increase in chlorophyll fluorescence, blue-green fluorescence, and leaf temperature during the early stages of infection. But with the onset of cell death, chlorophyll fluorescence and leaf temperature decrease sharply, while the blue-green fluorescence remains at a high level. Wounding of leaves by insects can reduce chlorophyll fluorescence, particularly around the sites of damage (Chaerle and Van Der Straeten 2001). Digital chlorophyll fluorescence imaging and visual imaging have been used for some years for monitoring the progression of symptoms of diseases in leaves. Visual imaging is as sensitive as visual scoring of the disease symptoms. In addition, it permits high-throughput quantification of lesions/chlorotic areas of leaves and tracking the progression of lesions over time. Chlorophyll fluorescence images of leaves allow the detection of fungal pathogen-affected areas before the symptoms become visible and also quantification of the susceptible and resistant responses (Chaerle and Van Der Straeten 2001; Chaerle et al. 2009). A mobile fluorescence imaging system has been developed to screen plant populations in the field for viral infections before the appearance of disease symptoms. Many biotic stresses have been detected on the basis of leaf temperature. Thermography allows the detection of infections at both plant and field levels even before symptoms become visible. Thermal imaging can permit detection of root diseases like rots and wilts as these reduce water uptake and, ultimately, transpiration. In general, leaf temperature measured at a single time point is not diagnostic, and some additional information is required to determine the cause of temperature change (Chaerle and Van Der Straeten 2001; Chaerle et al. 2009). Phenomics approaches would be more useful where resistance response is quantitative, and its scoring requires examination of plants at several time points during the disease progression.

15.14 Monitoring Drought Stress

An abiotic environmental factor that produces an adverse effect on crop performance is called *abiotic stress*. Plant breeding approaches usually consider growth and photosynthesis parameters while assessing tolerance to various abiotic stresses. Drought is the most important abiotic stress, but creation of a regulated water deficit condition is a challenging task. Soil drying (withholding water from soil-grown plants) is regarded as the most practical approach for approximating field drought conditions in the laboratory. But soil drying is difficult to control, is not homogeneous, and may affect nutrient uptake. In field situations, rain-out shelters in combination with regulated irrigation can be used to exercise some control on the water stress. Different genotypes of a crop species may take up water at different rates and from different depths of soil, which may lead to somewhat different levels of water stress in them. Therefore, the level of soil moisture and its distribution in the soil should be determined at the beginning of the field trial, and it should be monitored throughout the experiment. In case of green house, the pots with the plants can be regularly weighed to monitor their water use. In automated screening facilities, water supply can be regulated by the classical water withdrawal approach, maintenance of a constant soil water status, and a procedure that mimics the drought condition of the target environment (Munns et al. 2010).

In case of pots, there may be problems of drainage and variable water potential in the drying soil. Therefore, pots should be tall and "inorganic soils" like fritted or calcined clay should be used to facilitate drainage. Hydroponics avoids these problems and creates water stress by using an osmoticum. The osmoticum may be nonionic (mannitol, melibiose, sorbitol, or high-molecular-weight polyethylene glycol, PEG) or ionic (NaCl or a mixture of salts). The osmoticum molecules enter plant roots, PEG limits O_2 diffusion to roots, and carbohydrates support bacterial growth. Most crop species tolerate salt (NaCl) up to the concentration of 100 mM. But a mixture of salts, such as the macronutrients of Hoagland's solution, is preferable to the other osmotica because the uptake of these salts is tightly regulated. Drought and salinity stresses produce several similar phenotypic effects, and their screening methods show considerable overlap. Stomatal closure is one of the first effects of salinity stress and is caused mainly by osmotic effect (chemical drought). Stomatal closure reduces photosynthesis and transpiration and increases canopy temperature. One or more of the following assays can be used to monitor the effects of drought: (1) stomatal conductance, (2) leaf/canopy temperature, (3) visible imaging, (4) IR thermography, (5) chlorophyll fluorescence, and (6) estimation of tissue water content (Munns et al. 2010).

15.14.1 Stomatal Conductance

Water loss from plants depends on stomatal conductance and, over a longer period, the total leaf area of the plant. Stomatal conductance itself depends on the degree of stomatal closure, which is quickly affected by soil water potential and is more sensitive to water deficit than photosynthesis. Therefore, it can be used to quantify the response of different plants to low soil water potential, but its assessment in the field by porometer is labor intensive (Munns et al. 2010). Further, stomatal conductance is highly variable as it is affected by factors like vapor pressure deficit and CO_2 concentration. This is particularly so when hand-held porometers are used in enclosed spaces like growth chambers. In case of barley plants subjected to salt stress, stomatal conductance was positively associated with growth rate, number of tillers, and shoot biomass and was helpful in selection for yield. In general, drought leads to stomata closure and rise in leaf temperature. However, stomatal conductance shows variable relationship with leaf water status depending on the plant species. In case of most crop plants, e.g., barley, wheat, soybean, etc., stomatal

closure is not a dependable surrogate measure of water deficit. *Drought tolerant lines with small change in stomatal conductance might be suitable for cultivation under arid conditions provided irrigation is available. But lines showing large changes in stomatal conductance would be suited for cultivation under conditions of long-term drought.* However, drought tolerant lines should be selected only after their performance under controlled environmental conditions have been verified in experiments conducted under the target soil and environmental conditions (Munns et al. 2010; Furbank and Tester 2011).

15.14.2 Leaf/Canopy Temperature

Canopy temperature largely depends on stomatal conductance. In breeding programs, it is usually measured by IR thermometers. Genotypes with lower canopy temperature are reported to take out greater quantities of water from deeper layers of soil. Canopy temperature depression (CTD) is the difference between air and canopy temperatures so that a positive CTD value indicates a canopy that is cooler than the air. In case of wheat, CTD during specific growth stages, e.g., the grain filling period, has been used as a basis for selection for tolerance to heat and drought stresses. Under drought conditions, CTD seems to be more useful when measured in the mornings and afternoons. The wheat-breeding program for rain-fed environments at CIMMYT, Mexico, routinely uses CTD under drought conditions for evaluation of all F_3 and F_4 bulks. Canopy temperature data have been used to identify stress tolerant genotypes of rice and wheat (Furbank and Tester 2011; Fiorani and Schurr 2013; Panguluri and Kumar 2013).

15.14.3 Visible Imaging

Leaf growth decreases in response to drought even before a decline in stomatal conductance or photosynthesis, but this decrease can vary greatly even within a single species. Digital visual images of plants grown under controlled environments allow estimation of biomass and RGR, which can be used for screening genotypes for drought response. Further, the color information would allow estimation of the degree of senescence. Senescence of older leaves during drought may suggest an escape or avoidance process. In addition, genotypes with stay-green phenotype can be identified; these genotypes would be useful as they would be able to continue photosynthesis under water stress (Berger et al. 2010; Furbank and Tester 2011).

15.14.4 IR Thermography

Leaf stomata regulate transpiration, which affects plant temperature. As a result, a direct relationship is seen among transpiration rate, stomatal conductance, and leaf temperature. Therefore, thermal imaging can be used to screen large numbers of lines for stomatal conductance; it is being used for this purpose under controlled environmental conditions. But its use in the field presents problems, including separation of canopy from soil, variable shading, and variation in illumination due to clouds, time of the day, plant height, etc. New image analysis algorithms allow the separation of canopy from soil. The variation in leaf temperature can be used to detect stomatal closure since the temperature variance in case of stressed canopies is expected to be higher than that of well-watered canopies. IR thermography at the young seedling stage has been used to select such genotypes of wheat and barley that can maintain stomatal conductance when subjected to osmotic stress. This technique permits rapid, low-cost, high-throughput screening of pot-grown seedlings to identify lines tolerant to drought during the vegetative phase (Berger et al. 2010; Munns et al. 2010; Furbank and Tester 2011). However, considerable improvements in thermography protocols are still required. IR thermometers are less expensive and easier to handle than IR cameras. But data collection with IR thermometers would take a long time, and the environmental conditions in the field may change during this period. In

contrast, several plots can be imaged at the same time using IR cameras. Thermal images of untreated control plants and of plants subjected to an abiotic stress can be compared to deduce the magnitude of difference in canopy temperatures. This difference can be used to assess the relative sensitivity of different genotypes to the applied stress provided the environmental conditions are carefully controlled. In general, the smaller the difference between the control and the stressed plants, the lower the sensitivity of the genotype to the applied stress. It may be pointed out that visual, fluorescence, and thermal imaging may be used in combination to obtain a more precise temperature measurement than thermal imaging alone. This could also allow diagnosis of specific stresses and quantification of genotype responses (Chaerle et al. 2009; Munns et al. 2010; Furbank and Tester 2011).

15.14.5 Chlorophyll Fluorescence

The ratio F_v/F_m is the easiest to measure and the most widely employed chlorophyll fluorescence parameter in stress studies. Fluorescence imaging permits high-throughput estimation of average F_v/F_m for whole plants/target leaves, particularly for rosette species like tobacco, cotton, etc. *F_v/F_m parameter is not much sensitive to the level of drought stress, but it can serve as a screen for survival under drought conditions*. In *Arabidopsis*, the threshold value of F_v/F_m for plant survival seems to be about 33 % of the mean value for the watered control plants (Woo et al. 2008). However, this screen has limited utility for most annual crops. Further, large-scale screening of crop plants for chlorophyll fluorescence is feasible in seedling stage only. Finally, fluorescence imaging alone does not permit detection of early stages of water stress, although it can complement other imaging techniques (Chaerle et al. 2009; Munns et al. 2010; Furbank and Tester 2011).

15.14.6 Estimation of Tissue Water Content

The water content or the water potential of plants determines their water status. The plant water content, in turn, depends on the amounts of water absorbed by roots and the water lost through transpiration. The water content of plant tissues may serve as an indicator of the stress due to water deficit. The relative water content (RWC) of a plant is traditionally measured by detaching the plant leaves and hydrating them for 3–4 h in distilled water, after which the increase in their water contents in relation to their dry weights are determined. Thus, RWC indicates the degree of leaf dehydration and leaf water status. But the stressed plants may respond to water stress by osmotic adjustment. In such cases, leaves of the stressed plants will have higher solute concentration than those of the control plants. As a result, the leaves of stressed plants would take up more water than those of the non-stressed control plants, and RWC will be a poor indicator of water status. In such cases, the leaves of drought-stressed plants should be rehydrated by preventing transpiration; this can be achieved by keeping the plants still established in soils of low water potential in complete darkness (Munns et al. 2010). Thus, the traditional measurements of leaf water potential or leaf water content are destructive, they physically disturb the plants, and their findings may not be representative of the plant as a whole.

Thermal/IR imaging measures leaf/canopy temperature and is the most commonly used approach to monitor water status of plants. In the laboratory and greenhouse, the water absorption bands (bands absorbed by water; Sect. 15.5.2) between 1,450 and 1,600 nm are used for imaging, while the minor water bands at 970 nm and 1,200 nm are used in the field. The NIR reflectance indices, Water Index (R_{900}/R_{970}) and Normalized Difference Water Index $[(R_{860} - R_{1200})/(R_{860} + R_{1200})]$ relate to the reflectance between 850 nm and 1,200 nm of

the NIR spectrum (Ollinger 2011). The water index shows significant correlation with plant water content, especially when samples with a wide range of water content are included. The above two water indices are the most widely used indices for monitoring crop responses to different water stress regimes. The sensitivity of these indices can be enhanced by recording NIR reflectance over a period of time. NIR reflectance might not permit an absolute measurement of relative water content, especially under mild drought stress conditions, but it allows the detection of the onset of water stress in plants after water withdrawal (Munns et al. 2010).

15.15 Molecular Biomarkers

Molecular biomarkers are those dynamically expressed molecules that can be measured and can be used as indicators of specific phenotypic features. These biomarkers often allow the prediction of phenotypes even before they become observable. The molecular biomarkers include RNA transcripts, proteins, metabolites, and lipids. For example, the biomass and freezing tolerance of *Arabidopsis* accessions could be predicted on the basis of combinations of different metabolites. Similarly, the inclusion of metabolite levels along with DNA marker genotype data of the parental lines significantly improved the prediction of heterosis in their progeny. The use of molecular biomarkers is not dependent on the availability of genetic/genomic information about the crop species. Further, their use in breeding programs may save much time and cost since selection would involve estimation of only the targeted biomarker molecules. Gene expression microarrays and the next-generation sequencing technologies permit high-throughput characterization of the entire transcriptomes. The high-throughput analyses for the other molecules (the proteome, the metabolome, etc.) may soon become practicable and affordable. These developments may be expected to enable the identification of specific sets of molecular biomarkers indicative of the various phenotypic states (Normanly 2012).

15.16 Image Analysis

The general steps common to digital image processing for measuring plant size and to perform subsequent growth analysis are as follows. *Image retrieval* is the first step in image processing and involves loading of the images from the database/storage folder into the image analysis software. It may be necessary to crop the images or to specify the region of interest in order to reduce computing time and/or to minimize noise. The different images present in each image stack are aligned using reference points located outside the target object; this can be done using an ImageJ plugin like MultiStackReg. An *image stack* consists of all the images of a single object taken at different time points. *Image preprocessing* is the second step and comprises the use of filters to minimize noise or increase sharpness. It may be pointed out that this step may lead to a loss of information. If thresholding is to be used as the basis of segmentation, the color images have to be converted to grayscale images (Normanly 2012).

In the *image segmentation* step, the image is divided into objects of interest and the objects to be excluded from analysis. A segmented image is known as *binary image*, in which pixels that belong to the object of interest are scored as 1, while all other pixels are set at 0. There are several methods for segmentation, including histogram thresholding, feature space clustering, region-based approaches, fuzzy approaches, and neural networks; these methods are applied to grayscale images. Of these, the thresholding segmentation methods are perhaps the most commonly used (Hartmann et al. 2011). Morphological operations are then used to correct the imperfections like holes present in the binary images; this constitutes *noise reduction*. Objects like leaves often become fragmented during the earlier steps of image processing. In the next step, the *image composition* step, individual fragments of an object are merged to create the object. *Image description* involves quantification of such features as area, height, width, etc. of the object. *Color classification*

involves extraction of the object, i.e., the plant, from the original RGB image. The leaves can be subdivided into different areas on the basis of color information of the original image, and these areas can be measured to reveal the extent of necrotic or senescent leaf area (Normanly 2012).

15.17 Image Analysis Software

High-throughput phenotyping began with automated greenhouses, in which images are automatically acquired at regular intervals and analyzed using computer software to detect and quantify trait phenotypes. There are several fully automated high-throughput platforms for growing and phenotyping of plants, e.g., PHENOPSIS (used by French National Institute for Agricultural Research, INRA) for *A. thaliana,* GROWSCREEN (used at Research Center Jülich) for different plant species, TraitMill™ for rice (developed by CropDesign), and LemnaTec. The GROWSCREEN FLUORO platform simultaneously monitors leaf growth and chlorophyll fluorescence to detect stress tolerance in rosette plants. These platforms use dedicated image analysis software, which cannot be easily modified (Hartmann et al. 2011). However, several free, open-source image analysis software are also available, some of which are described below.

15.17.1 ImageJ

ImageJ is a flexible, free, open-source, Java-based image analysis software developed at the NIH, USA. It has several useful plugins and tools and a graphical user interface (GUI). It can display, edit, process, analyze, save, and print the following types of images: 8-bit color and grayscale images, 16-bit integer images, and 32-bit floating-point images. It is able to read images in many formats and raw formats, and support handling of image stacks. It can perform time-consuming operations in parallel on multi-CPU hardware. It calculates area of the selected portion of an image or of the entire image, measures lengths (in units like millimeters) and angles, and calibrates using density or grayscale standards. It generates histograms and profile plots; cuts, copies, or pastes images or selections; and adds text and various shapes to the images. It performs a variety of operations, including contrast manipulation, sharpening, smoothing, scaling, cropping, resizing, rotation, etc. It can zoom an image up and down by a factor of 32 and analyze the images so derived. It can analyze any number of images simultaneously if adequate memory were available. ImageJ is extensible via Java plugins. It has an editor as well as a Java compiler that can be used to develop the desired plugins. It can be run either online or on a computer that has a Java 5 or later version virtual machine. The compiled versions for Windows, Mac OS/OS X, Linux, etc. are available at http://rsbweb.nih.gov/ij/index.html.

15.17.2 HTPheno

HTPheno is a freely available (http://htpheno.ipk-gatersleben.de/), open-source, automated image analysis software package designed for high-throughput phenotyping of plants. HTPheno is a flexible plugin for ImageJ. It estimates plant height, plant width, projected shoot area, etc. by analyzing top and side view color images of the test plants (Hartmann et al. 2011). It can be adapted for analysis of images acquired either manually or by an automated phenotyping system. HTPheno retrieves from the image file both single images as well as series of images and analyses them automatically. It has two functions, viz., (1) calibration and (2) image processing. The calibration function specifies different parameters for segmentation and translates pixels into millimeters. The automatic image-processing function is used after calibration, each processing step being displayed as an image. First, the original image is loaded, and the regions occupied by various objects in the image are clearly demarcated as distinct rectangles. The original image is then segmented using a pixel-based multidimensional

histogram thresholding procedure, which generates a color-coded segmented image. Now, an opening operation removes small objects from the image. Finally, calculations are performed, the analysis results are placed onto the original image, and a scale bar (100 mm) is drawn on this image. The results are also presented in a tabular form, which can be exported. HTPheno is several-fold faster than manual image analysis and yields good results. But HTPheno cannot detect yellow and brown parts, e.g., leaf tips, of plants.

15.17.3 Rosette Tracker

Rosette Tracker is a free, open-source tool for image analysis that requires minimal user input. It can be used as a plugin for ImageJ and can be adapted for use with various low-budget imaging systems. This image analysis tool is the first image-processing package that is capable of analyzing visual, chlorophyll fluorescence and/or thermal/IR time-lapse images. It performs (1) calibration, (2) image segmentation, (3) rosette detection, and (4) plant growth analysis. The user may enter the pixels to millimeter ratio, if known, or the graphical tool of Rosette Tracker allows easy setting of the scale. It uses the Gaussian probability distribution functions to arrive at the correct segmentation of visual images. This segmentation algorithm is able to exclude green pixels not belonging to the rosette. This method is easily adapted for segmentation of chlorophyll fluorescence images. But the segmentation of an IR image involves projection of the segmentation results from visual/fluorescence image of the object onto its IR image. Plant growth can be measured as area, diameter, stockiness, RGR, average intensity, and/or compactness of the rosette. This tool can detect many rosettes in an image provided the individual rosettes are clearly separated from one another and the image has high resolution. A simple GUI permits the tuning of parameters for handling of time-lapse images captured with different imaging setups. The compiled version and source code of Rosette Tracker are available at http://telin.ugent.be/~jdvylder/RosetteTracker/ (De Vylder et al. 2012).

15.17.4 Martrack Leaf

Martrack Leaf implements a marker-tracking method of monitoring growth of individual leaves over a period of few days (Mielewczik et al. 2013). A single growing leaf from each target plant of soybean was fixed using a setup that supports both the plant and the selected leaf. The leaf was kept fixed in the focal plane of a camera that was positioned to acquire top-view image. Artificial landmarks were created by fixing with glue 5 mm diameter black plastic beads at selected positions on the leaf border. Successive images of the leaf were taken after every 90 s using a CCD camera. The user has to define, through a simple GUI, the template size (an area around each black bead) and the search length (the region of the image in which a bead is to be tracked) for image analysis. The center of every black bead is selected by a mouse click in the first frame of the image stack for each leaf. This tool then tracks each bead in the entire image sequence and displays the paths of their displacement with red lines in the first image of the stack. A normalized cross-correlation is now used to determine the position of each bead in the latest image of the sequence. The positions of all the beads in an image describe a polynomial. The area of this polynomial is estimated; this area corresponds to leaf area and can be used to compute RGR. Martrack Leaf performs better than optical flow approaches under both indoor and outdoor conditions and requires fewer images. It provides more experimental flexibility than morphometric and mechanical methods of leaf growth analysis. But it is sensitive to changes in the reflectivity of the background. Martrack Leaf is implemented in Matlab 7.12. The compiled package for Linux, Mac, and Windows environments are provided online as additional files in Mielewczik et al. (2013).

15.17.5 HPGA (High-throughput Plant Growth Analysis)

High-throughput plant growth analysis (*HPGA*) is a high-throughput image analysis tool that

estimates plant area, AGR, and RGR of rosette plants like *A. thaliana* (Tessmer et al. 2013). Area of a plant is estimated from top-view image by the following four-step procedure: identification of plant center, identification of leaf tip, estimation of leaf area, and measurement of plant area. It avoids segmenting of all the leaves in the image. It takes care of the leaf overlap by detecting all leaf tips and then measuring leaf lengths. The conversion of leaf length into leaf area uses a species-specific model, which needs to be trained for each new species. But sometimes tips of some inner leaves may be missed, for which the plant area estimate has to be suitably corrected. HPGA uses a three-parameter logistic nonlinear growth model to generate precise and robust growth curves. The growth model allows calculation of AGR and RGR from the biomass/area of the plants estimated at successive points of time. The image processing by HPGA is better than that by the global thresholding approach. HGPA is available at http://www.msu.edu/~jinchen/HPGA.

15.17.6 Root System Analyzer

Root System Analyzer is the first fully automatic tool for analysis of 2-D images of root systems grown in soil (Leitner et al. 2014). It uses a segmented image for skeletonization on the basis of morphological features of the roots and creates a graph representation of the root system. The user initiates tracking of the primary root in this graph by a mouse click. The tool then automatically tracks the primary and the lateral roots and recovers root architecture parameters. If needed, the user may manually correct the tracking of individual roots. This tool uses a dynamic root architecture model that allows detection of developmentally valid roots and permits root development analysis from a sequence of images. It simulates the growth of each root on the optimal path (the shortest path between two points) and is capable of analyzing complex root systems. The output for each root includes branching order, time of emergence, and root length and root area at observation time. These data allow the estimation of the total number of roots, total root length, root elongation rate, etc. The user has to be familiar with Matlab for using this tool. The package is available free from http://www.csc.univie.ac.at/rootbox/rsa.html.

15.17.7 SmartRoot

SmartRoot is designed for analysis of primarily 2-D grayscale images of root systems (Lobet et al. 2011). It carries out the following operations: (1) segmentation (by thresholding), (2) skeletonization, (3) root tracking, and (4) data analysis. The user has to select each root by a mouse click for skeletonization. The root tracking step works directly on the image source and generates data about root structure. The data are processed to obtain individual root data and the parameters needed for root system modeling. The analysis of image sequences of a root system requires the user to determine the anchor points in each image to avoid problems due to small shifts in root positions in the different images. The data processing is user-friendly, and the output is in the form of text files and images.

15.17.8 RootReader2D

RootReader2D is a tool for analysis of mostly 2-D grayscale images of root systems (Clark et al. 2013). It carries out segmentation by thresholding, and skeletonization is based on morphological features of roots. It generates a graph of the root system from the skeleton. This graph is used for root tracking using a search algorithm, which discovers the shortest route connecting two points. The user has to manually select each root, by a mouse click, for tracking. Tracking yields data on root structure in terms of connectivity between roots and the root positions in the space. Analysis of this data yields information that is used for modeling the root system.

15.17.9 RootReader3D

RootReader3D (www.plantmineralnutrition.net) is written in Java. It reconstructs 3-D root systems from 2-D images acquired using a 3-D imaging platform (Clark et al. 2011). The steps for reconstruction of the 3-D models can be visualized by using viewing interfaces. Further, the user can interact with the tool through mouse and keyboard commands to carry out either semi-automatic or automatic reconstruction. The seedlings to be imaged are grown in growth cylinders containing gellan gum. One 2-D image of the root system is taken at every 9° rotation of each seedling. Thus, a total of 40 images of each seedling are taken at a given point of time. The 2-D images are cropped, converted to grayscale, and segmented by thresholding. The segmented images are analyzed using RootReader3D tool to reconstruct a 3-D model of the root system. This tool can quantify 27 different root system traits, including primary root length, total root system length, and maximum root system width. The estimates for the different root traits obtained from this tool were significantly positively correlated with those generated by the 2-D measurement methods.

15.18 Applications of Phenomics

The phenomics approaches have the following potential applications: (1) rapid identification of stresses in a plant population, (2) rapid and efficient screening for mutants, (3) evaluation of phenotypic effects of uncharacterized transgenes, (4) detection and monitoring of disease epidemics in fields, (5) detection of root attacks by fungi, insects and other pathogens, and (6) modeling of biomass production. (7) These techniques would be helpful in the study of various physiological processes underlying specific plant functions. Further, (8) they would facilitate screening of germplasm collections for identification of accessions with the genes of interest. (9) The phenomics approaches would also facilitate the selection of superior/desired genotypes from breeding populations. (10) They would, ultimately, allow the huge genomic information to be reliably related to specific phenotypes and (11) permit a systematic study of the pleiotropic effects of the genes. When all these happen, (12) it might become feasible to precisely predict the phenotypic effects of the changes at DNA sequence level. In short, (13) the above developments would enable a planned and precise use of the available genetic diversity for crop improvement to achieve increased agricultural productivity.

15.19 Achievements

High-throughput phenotyping is being extensively used by the private sector laboratories. The information about the technologies used by the private sector is generally not available in the public domain. In any case, where and when this information does become available, it happens only after considerable delay and the information is more likely than not to be incomplete. Rapid advances are being made in imaging technologies developed for high-throughput phenotyping of model plants like *A. thaliana*. The image analysis provides information on growth dynamics, morphological traits, and photosynthetic parameters. For rosette plants like *A. thaliana*, estimation of growth rates from digital image data is relatively simple and accurate. A number of noninvasive high-throughput assays for various physiological and morphological traits have been/are being developed (Table 15.6). Chlorophyll fluorescence imaging has been used for evaluation of the effects of drought stress on photosynthesis and identification of stress tolerant genotypes. IR thermometers are being used in many programs for the identification of heat/drought tolerant genotypes. LED-based instruments for measuring NDVI are also being used in many breeding programs. SPAD chlorophyll meters are used to estimate chlorophyll content of leaves and to assay water-use efficiency. Efforts are being made to develop FBP platforms to facilitate basic studies and plant breeding efforts.

15.20 Future Directions

Phenomics involves both extensive and intensive phenotyping. In *extensive phenotyping*, a large number of phenotypes are evaluated in a limited number of contexts, e.g., analysis of the expression of all genes in a single tissue at one stage of development. But in *intensive phenotyping*, one or few phenotypes are characterized in a great detail, e.g., monitoring of the expression of one gene in all the tissues over the various developmental stages. Further, the phenomics data should be collected from a population exposed to a large number of different environments. Therefore, phenomics studies will always involve prioritization of the phenotypes to be investigated and the environments in which they are to be evaluated (Houle et al. 2010). The phenotype data can be used to build a database that may ultimately allow linking of gene sequences to plant structure and function. This development will allow the linking of information in genomics with plant function and, ultimately, with agricultural traits. The data for the above database will have to be measured under clearly defined environmental conditions, described objectively in mathematical terms, and stored in a digitized and easily searchable format. Further, information in a standardized format on experimental details like the plant material used in the experiments, the conditions for growth of the plants, and the techniques used for phenotyping (the metadata) need to be reported so that data from different experiments can be compared when required. It may be pointed out that, at present, it may not be possible to standardize the experimental conditions and the phenotyping techniques on a large scale. In view of this, Poorter et al. (2012) have proposed a practical guide for growing plants for experimental purposes and have attempted to develop the logic for determining the minimum set of data to be reported in cases of plant phenotyping.

The current phenomic studies are largely extensive; it is important that these studies also adopt the intensive approach. This change would necessitate considerable enhancement in the capabilities, including the throughput, of the phenotyping techniques coupled with a large reduction in their costs. Such phenomic data, of necessity, will comprise a very large number of different phenotypes (p) scored on a large number of individuals (N), but the p will always be much larger than N. There is an urgent need to develop suitable statistical models for analysis of such data since the available methods are not entirely satisfactory. It will be highly desirable that the software developed for the statistical analysis of the phenomic data permit automated data analysis. Further, models need to be developed that would allow prediction of phenotypes from the available information on genetic variation and other molecular data. So far, the attempts to develop such models have produced encouraging results in plants and, especially, in animals. Finally, the phenomics studies by different research groups should be integrated in a manner similar to that for the genomics efforts; the phenomics investigations, however, would have to be on a much larger scale than the genomics initiatives. It is also important to undertake a systematic comparison of the accuracy and the strength of different analysis tools and software packages. In addition, the phenomics teams have to be truly transdisciplinary involving plant biologists, physicists, mathematicians, and engineers (Bilder et al. 2009; Houle 2010; Fiorani and Schurr 2013).

The priority research needs in the field-based phenomics are (1) effective management of the data (from their acquisition in the field to the genetic analysis); (2) combining the data from different sensors and from a GPS receiver; (3) development of protocols for testing of promising instruments for different functions, including calibration, stability over temperatures, and ease with which they can be integrated in the overall setup; and (4) the development of better algorithms for data analysis. There is an increasing tendency for obtaining broad patents for phenomics techniques/procedures. Such patents would inhibit innovation in the instruments and the software required for FBP. It has been suggested that patenting may be limited to sufficiently novel innovations in designs of specific instruments. In addition, some instruments generate outputs in proprietary formats; this situation complicates

their integration with the other components of the setup. Therefore, it is highly desirable that the controls of the instruments and their outputs are readily accessible without additional proprietary hardware/software (White et al. 2012).

Questions

1. Discuss the various applications and limitations of, and future prospects for field-based phenomics.
2. Briefly describe the general procedures for image acquisition and image analysis for phenomics. Discuss the advantages and limitations of imaging technology.
3. Briefly describe the various imaging technologies used in phenomics studies.
4. Discuss the various applications of reflectance imaging in phenomics.
5. Explain the use of imaging technologies for plant structure and growth analysis.
6. Discuss the various applications of fluorescence imaging and monitoring technologies in phenomics studies.
7. Briefly describe the infrared imaging technology and discuss its applications and limitations in plant phenomics studies.
8. Discuss the application of imaging technologies in root architecture analysis.
9. Explain the usefulness of various imaging technologies in evaluation of biotic stresses in plant populations.
10. Briefly describe the applications of various high-throughput phenotyping technologies in monitoring of drought responses of plants.
11. Briefly describe the salient features of some of the software for shoot or root image analysis.
12. Explain the meaning of plant phenomics. Discuss the relevance of plant phenomics studies and the main issues for future studies.

Glossary

***A* matrix** It contains the proportion of alleles that are identical by descent for each pair of individuals in a sample.

***Ab initio* gene prediction** Gene prediction by using specialized software for searching genome sequences for the presence of genes.

Abiotic stress An abiotic environmental factor that produces an adverse effect on crop performance.

Acceptable cluster A group of two or more entities, for which within cluster GD is less than the overall mean GD.

Accuracy of mapping The closeness of the QTL position and effect size estimates obtained from a study to their "true" values.

Ad hoc index A selection index based on marker genotype and trait phenotype data from the same population in which it will be used.

Address sequence In SNP genotyping, the tag oligo for a given bead type.

Admixture Gene flow between genetically distinct populations of the same species.

Advanced intercross line population A population developed by intermating the individuals of F_2 and subsequent generations from a suitable cross.

AFLP primer A primer having the adapter sequence plus 1–3 arbitrary nucleotides at its 3′ end.

Allele-specific oligo In case of Illumina GoldenGate assay, two ASOs are used for each SNP locus; the 3′ region of an ASO is complementary to the sequence on the 3′ side of the SNP locus, and its 3′ terminal base is complementary to one of the two alleles at the SNP locus.

Allele-specific PCR Selective amplification of only one of the alleles at a SNP locus.

Allelotyping Estimation of the relative abundance of alleles of a SNP locus in a pool of DNA samples.

Allozymes Variants of an enzyme encoded by different alleles of the same gene.

Amplicon The DNA segment amplified by PCR.

Amplified fragment length polymorphism It involves digestion of genomic DNA with two restriction enzymes, ligation of appropriate adapters to the fragments, selective amplification of a much smaller set of these fragments by using AFLP primers, and the separation of the PCR products by denaturing polyacrylamide gel electrophoresis.

Anchored ISSR primer In case of ISSR, the primer has a microsatellite sequence plus a short (usually, two nucleotides long) arbitrary sequence either at its 3′ or 5′ end.

Anchored microsatellite-primed PCR See *inter-SSR PCR*.

Anchored simple sequence repeats See *inter-SSR PCR*.

Arbitrary primed PCR A single arbitrary sequence primer of 18–32 nt is used for amplification; the first two PCR cycles are carried out at low stringency.

Ascertainment bias A systematic bias generated in a dataset by the manner in which the data were collected.

Association mapping See *population mapping*.

Association mapping panel See *association mapping population*.

Association mapping population A large random sample from a natural population,

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0, © Author(s) 2015

a germplasm core collection, a collection of breeding lines, or a population derived from a set of multiparent crosses and used for AM.

Association panel See *association mapping population*.

Associative transcriptomics The analysis of SNP genotype data generated from RNA-Seq data to detect marker–trait associations.

Autofluorescence Fluorescence due to some endogenous molecule, e.g., chlorophyll in plants.

Backcross Cross between the F_1 generation (or a later generation) of a cross and one of its parents.

Backcross breeding A useful trait is transferred from a donor parent (DP) into a recurrent parent (RP) by repeated backcrossing to the RP.

Backcross inbred lines Homozygous lines developed by backcrossing the F_1 from a cross between two homozygous lines to one of the parents, followed by continued selfing of the BC_1F_1 progeny.

Background selection Marker-assisted selection for the RP genomic regions, except for the target gene/QTL.

Barcode identifier sequences Unique short (4–8 nt long) nucleotide sequences that differ from each other for at least two bases.

Barcodes See *barcode identifier sequences*.

Base calling In DNA/RNA sequencing, deduction of bases on the basis of light color and intensity signals.

Bayes factor The ratio of the probability of getting the observed data when H_1 is correct to that when H_o is true.

Beavis effect The smaller is the size of a mapping population, the smaller is the number of detected QTLs for a trait and the larger are the estimates of their effects.

Bin mapping In this approach, the linkage map is divided into several relatively small segments called bins, and the markers are mapped within individual bins. Also see *selective mapping*.

Bin In case of linkage maps, a relatively small (typically 10–20 cM long) segment of a linkage group that is flanked by fixed core, anchor, or framework marker loci. In case of selective mapping, an interval in a linkage group within which a breakpoint has not occurred in any individual included in the sample. In case of MAPMAN tool, a specific area of metabolism, e.g., photosynthesis.

Binary image A segmented image.

Bioinformatics tools Computer programs for acquisition and analysis of data, for detection of associations and patterns, or for achieving other specific objectives.

Bioinformatics Derived from the terms "biology," "information technology," and "statistics"; it involves development of statistical tools and techniques and computer software for acquisition, storage, analysis, and visualization of biological information.

Biotic stress Infection by a virus, fungus, nematode, or bacterium or infestation by an insect.

BLAST (Basic Local Alignment Search Tool) A family of user-friendly sequence similarity search tools for identification of database sequences homologous to the query or submitted nucleotide or protein sequence. *The most popular data-mining tool developed ever*.

BLASTn It compares the query nucleotide sequence with a nucleotide sequence database.

BLASTp It compares the submitted protein sequence against a protein database.

BLASTx It translates the submitted nucleotide sequence into an amino acid sequence and compares the latter with a protein database.

Blotting Transfer of molecules from a gel onto a solid support. In Southern hybridization, transfer of DNA fragments from the gel onto a solid support like nitrocellulose filter membrane.

Bottleneck A marked reduction in the size of a population for one or more generations.

Breeder's exemption The protected variety can be freely used for scientific purposes and for creation of genetic variability in plant breeding programs.

Breeding population In case of genomic selection, the population subjected to GS.

Breeding value Of an individual/line, the expected phenotypic value of its progeny.

Bulked segregant analysis Equal amounts of DNA from 10 plants from each of the two

most extreme phenotypic groups for the target trait are bulked to create two bulks; DNAs from the two parents and the two bulk DNAs are screened with a large number of markers to detect markers putatively linked to the genes governing the target trait.

Bulked segregant RNA-Seq A modification of BSA; it uses RNA sequence data from the two phenotypic extreme bulks to identify markers tightly linked to the gene responsible for the target trait.

Candidate gene approach of association mapping Association analysis is restricted to the genomic regions having the candidate genes/QTLs for the trait(s) of interest.

Candidate gene A gene that is expected, on the basis of previous knowledge, to be involved in the control of a trait of interest.

Capture oligo The oligo that is complementary to the 3′ side of the SNP locus and includes the polymorphic nucleotide. See also *tag oligonucleotide*.

Cases In association mapping, individuals carrying the allele of a gene responsible for a disease.

cDNA library A population of bacterial transformants or phage lysates containing recombinant DNA molecules, in which all the mRNA species isolated from an organism or tissue are represented as cDNA inserts.

Centimorgan The distance between two genes/loci that is expected to lead to one percent crossing over between them.

Chromosome landing Direct identification, in a single step, of the clone with DNA insert having the target gene using one or more markers located very close to the gene, made possible by high-resolution mapping, as a result of which the physical distance between the markers and the target gene is less than the average insert size of the genomic library.

Chromosome segment substitution lines A series of homozygous lines, each having a single distinct chromosome segment from a DP in the chromosome background of RP.

Chromosome walking A technique combining restriction mapping, subcloning, and nucleic acid hybridization to analyze DNA inserts of a genomic library in an effort to locate the gene of interest.

Cleaved amplified polymorphic sequence Detection of length polymorphism following restriction digestion of specifically amplified PCR products; syn., *PCR–RFLP*.

Clone Asexual progeny of a single asexually propagated plant.

Clustal A group of multiple sequence alignment tools, e.g., ClustalW, ClustalX, and Clustal Omega.

Cluster analysis The entities with similar features are grouped into the same cluster.

Coefficient of co-ancestry See *kinship coefficient*.

Colinear markers Markers located in the same linear order in two different chromosomes of the same species or in chromosomes of two different species.

Color classification In image analysis, extraction of the object, i.e., the plant, from the original RGB image.

Combined marker-assisted selection MAS is used in combination with phenotypic screening/selection.

Comparative mapping A comparative study of linkage maps of different species.

Complementation hypothesis See *dominance hypothesis*.

Complete line conversion Markers used for foreground as well as background selections in a backcross program.

Complete linkage map A linkage map containing sufficiently large number of genetic markers so that every point in the genome of the species is genetically linked to at least one marker.

Complexity reduction of polymorphic sequences Genomic DNA is digested with two restriction enzymes and the complexity is reduced by selective amplification procedure of AFLP.

Complexity Of a genome or DNA preparation, the total number of different sequences present in the genome/DNA.

Composite interval mapping It first carries out single marker analysis, then builds multiple QTL model, and uses QTLs present in the

other marker intervals as cofactors in the model.

Concatenated cDNA sequencing It involves isolation and pooling of multiple cDNA clones, their enzymatic concatenation, followed by shotgun sequencing.

Confirmation of marker–trait linkage Evaluation of the observed marker–trait linkage in another mapping population developed from the same cross or in the same mapping population by another worker.

Confirmation of QTL Replication studies conducted to ensure that the detected QTL is real and to verify its position and effect size.

Consensus accuracy (of reads) Accuracy of the sequence of a fragment obtained as consensus sequence of all the reads of the fragment.

Consensus linkage map A linkage map created by merging two or more linkage maps for a given species. Syn., *merged linkage map, integrated linkage map*.

Conserved DNA-derived polymorphism markers Markers based on conserved DNA regions of a selected set of well-characterized plant genes.

Conserved orthologous sequences Orthologous sequences with almost similar sequence.

Conserved orthologous set of genes A group of genes conserved in sequence and copy number during evolution.

Conserved orthologous set of markers Markers based on conserved orthologous set of genes.

Conserved region amplification polymorphism Markers based on one primer derived from an exon of a gene and the other primer targeting introns, similar to TRAP markers.

Contig A series of clones containing overlapping DNA inserts covering a specific genomic region.

Control In association mapping, individuals lacking the disease and unrelated to the cases.

Copy number variation Variation in the number of copies of a given genomic region in the genomes of different individuals/lines of a species.

Core marker A highly polymorphic marker that is expected to be polymorphic in most, if not all, mapping populations of the given species. Syn., *anchor* or *framework marker*.

Correct reading frame The longest reading frame in a given DNA sequence that is uninterrupted by a stop or termination codon.

Coverage of sequencing The sequencing depth for the whole-genome, e.g., 10x, 20x, 30x, etc., coverage or depth.

Crossing over A physical exchange of ordinarily strictly homologous segments between homologous chromosomes.

Cytogenetic map A genetic map depicting the locations of various genes in the chromosomes of a species relative to specific microscopically observable landmarks in the chromosomes.

Dark-adapted plant A plant kept in the dark for 10–15 min.

Database A systematized collection of vast amounts of information on a specific topic, e.g., nucleotide sequence, protein sequence, etc., in an electronic environment.

Demographic history Changes over time in population size, development of subgroups within a population, etc., in natural populations of species.

Denaturing/temperature gradient gel electrophoresis It reveals differences in the movement of double-stranded DNA molecules from the same genomic regions of different individuals when they are subjected to electrophoresis in a gel in which the denaturing conditions increase with the distance from the loading well.

Depth of coverage of a SNP The number of sequence reads containing a given SNP locus.

Derived cleaved amplified polymorphic sequence A restriction enzyme recognition site is introduced into the PCR product by one of the primers so that the product yields a CAPS marker. Syn., *mismatch PCR–RFLP*.

Designer crops Varieties developed by any methodology, including MAS, that exhibit a specified phenotype.

Digested RAMP Markers developed by digesting the RAMP products with a restriction enzyme.

Discovery array A microarray having all the fragments amplified following the complexity reduction procedure.

Distinctness In PBR, the new variety must be distinguishable from other varieties of the same crop for one or more identifiable morphological, physiological, or other characteristics.

Diversity array technology A high-throughput, low-cost modification of AFLP procedure that uses microarray-based nucleic acid hybridization for detection of polymorphism.

Diversity panel A sample representing as much genetic diversity of the parent population as is practically feasible.

DNA amplification fingerprinting A dominant marker system; typically, a single 4–6 nt long single primer is used for amplification of genomic DNAs.

DNA barcode A standardized genomic DNA sequence of over 400 bp length used for a reliable identification of organisms.

DNA chip See *microarray*.

DNA fingerprinting Multilocus/single locus assays based on minisatellite/microsatellite DNAs for unequivocal identification of individuals on the basis of gel patterns of the derived fragments.

DNA sequencing Determination of base sequence of a DNA fragment.

Dominance hypothesis Heterosis is the result of favorable dominant alleles masking the deleterious effects of concerned recessive alleles in the heterozygotes.

Donor parent A homozygous line from which the gene of interest is transferred into another line.

Double digest restriction-site-associated DNA sequencing Genomic DNA is simultaneously digested with two restriction enzymes, and a precise size selection procedure is used to select only such fragments of specified size that are produced due to one cut by each of the two enzymes.

Dynamic allele-specific hybridization Specific probes are hybridized with the target PCR products, and melting temperatures of the duplexes so produced are used to deduce the SNP allele.

Dynamic traits Quantitative traits, e.g., plant height, which show different patterns of development in different genotypes that may, in the end, show comparable/different values for the concerned trait.

Economic heterosis Superiority of F_1 hybrids over the best commercial variety of the crop.

Effective number of markers It is that number of markers with which the standard deviation of the estimates is not significantly affected by reducing or increasing the number of loci/bands analyzed.

Efficient mixed-model association EMMA corrects AM for population structure and kinship; it uses an algorithm for deducing the phylogenetic kinship matrix from genome-wide markers and applies it to the linear mixed model.

Endonucleases Enzymes that produce internal cuts or cleavages in DNA molecules.

Epigenetic changes Changes involving DNA methylation, RNA interference, and histone modification (acetylation, methylation, phosphorylation, and ubiquitination).

Epigenetic marks Epigenetic changes.

Epigenetic recombinant inbred lines RILs with the same genotype, but differing from each other in terms of epigenetic modifications.

Epigenetics Study of a change in gene function without any change in the gene base sequence.

Epigenomics A genome-wide study of epigenetic marks.

Epimutation Heritable genetic variation generated by epigenetic changes.

Epistasis Interaction between two and more different genes; the expression of a gene is modified due to the influence of other interacting genes.

Epistatic QTLs These QTLs interact with the main effect QTLs to influence the trait phenotype.

Epistatic selection Selection for a phenotype produced by interaction between two and more genes.

Epistasis hypothesis Heterosis results from epistatic gene interactions.

eQTL hotspot A genomic region containing eQTLs affecting the expression of many different genes located in the same genomic region.

Essentially derived variety A variety predominantly derived from the initial variety so that it retains the expression of essential characteristics from the genotype or combination of genotypes of the initial variety.

Euheterosis See *heterosis*.

Ex situ germplasm conservation Maintenance of germplasm accessions as seed samples, in the field as growing plants, as slow-growth shoot cultures, or as tissues/organs frozen in liquid nitrogen in gene banks located away from their natural habitats.

Expressed SSRs SSR markers derived from ESTs.

Expression level polymorphism Variation in the expression levels of genes.

Expression proteomics A comparative quantitative analysis of the patterns of protein expression between samples differing for some variable.

Expression QTLs QTLs concerned with regulation of expression levels of the genes.

ezRAD-Seq A simplified RAD-Seq technique; it uses the standard Illumina TruSeq library preparation kits and sequencing by the NGS technology. Complexity is reduced by size selection (400–500 bp).

Factored spectrally transformed linear mixed model It reduces computation time by using a low-rank relatedness matrix based on a few thousand SNPs in place of all the SNPs used for AM.

False discovery rate The expected ratio of the wrongly rejected null hypotheses to the total number of H_0 rejected in the experiment multiplied by the probability of making at least one rejection of H_0.

Family mapping Linkage mapping using populations created by crossing usually two homozygous lines.

Farmer's privilege Farmers can save a portion of their harvest of a protected variety and use it as seed for planting their next crop, but they can neither sell nor exchange this produce.

Feature In case of SFP, the 25 nt long oligonucleotides used to construct the microarrays.

Fine mapping Analysis of very large populations using a sufficiently large number of markers to identify markers located very close to a gene known to be linked to a marker. Syn., *high-resolution mapping*.

First-generation DNA sequencing Sequencing methods based on in vivo cloning and chemical or enzymatic sequencing procedures.

Flanking markers One marker located on either side of the target gene/QTL.

Flat-file database The earliest and the simplest database type suitable for storing small amounts of data.

Fluorescence A molecule absorbs light of a given wavelength and emits light of a longer wavelength.

Fluorometer Monitoring the level of fluorescence following illumination with light of the appropriate wavelength.

Fluorophore A molecule that generates fluorescence.

Foreground selection Marker-assisted selection for the target gene/QTL.

Forward phenomics Use of high-throughput methods for screening germplasm collections for valuable traits.

Founder parents In case of NAM, several individuals/lines that are crossed with each of one to few nested parents.

Founder population The group of individuals that initiated the population under consideration.

Framework map A map constructed using a large mapping population for precisely mapping a set of framework markers selected on the basis of their even distribution throughout the genome.

Full MAS See *complete line conversion*.

Functional AM The phenotype of a dynamic trait may be measured at several different time points during the development and used for AM either independently or jointly to enable

the identification of different genes/QTLs that are expressed at specific developmental stages of such traits.

Functional genomics The study of gene expression patterns and the functioning of metabolic pathways.

Functional GWAS See *functional AM*.

Functional map A special category of linkage maps depicts locations of different genes of the given species prepared by using gene-based markers, or the gene sequences themselves are used as markers.

Functional markers Markers derived from such polymorphic sites within genes that have a causal relationship with specific phenotypes of the concerned trait.

Functional proteomics Analysis of the characteristics of molecular protein networks involved in a living cell.

GC clamp A stretch of ~30 bp containing only GC bases; used at the 5′ end of a primer in case of D/TGGE.

GD-based clustering methods Clustering based on a pairwise distance matrix.

Gene conversion A process in which a small segment of one chromosome is copied in the place of its homologous segment of the homologous chromosome during meiosis.

Gene finding See *gene prediction.*

Gene hunting See *gene prediction.*

Gene ontology It describes features of gene products in multiple species.

Gene prediction Identification, by genome sequence analysis, of genomic regions that function as genes.

Gene pyramiding In general terms, bringing together two or more genes controlling a single trait in a single line/variety.

Gene space study See *gene space-based association mapping*.

Gene space-based association mapping Association analysis using SNP data generated from sequences of transcribed genes.

Gene space The fraction of genome that corresponds to the protein-coding genes and, also, the distribution pattern of these genes.

Gene-based markers Markers based on polymorphic sites within genes, but their relationships with the relevant trait phenotypes are not known. Syn., *gene-targeted markers, gene-specific markers*.

Gene-specific markers See *gene-based markers*.

Gene-targeted markers See *gene-based markers*.

Gene/QTL introgression Transfer of the target gene(s)/QTL(s) into the RP using backcross procedure.

Genetic bit analysis See *single-base extension*.

Genetic distance In genetic diversity analysis, a quantitative measure of the genetic difference between two entities in terms of differences in their DNA sequences, gene frequencies, etc. In linkage mapping, recombination frequency corrected for the occurrence of multiple crossovers between the concerned gene/marker pair.

Genetic diversity analysis Estimation of genetic similarity (or dissimilarity) between pairs of entities and use of these estimates for grouping of the entities.

Genetic diversity The sum total of genetic differences present among different individuals, genotypes, strains, clones, or populations of a species.

Genetic drift Random change in gene frequency of a population due to random sampling of gametes that unite to produce the finite number of individuals in each generation.

Genetic map A schematic representation of genetic markers in the same order, in which they are located in a chromosome along with the genetic distances between them.

Genetic marker locus The specific location in a genome identified by a genetic marker.

Genetic marker A trait that is polymorphic, easily and reliably identified, and readily followed in segregating generations and indicates the genotype of the individuals that exhibit the trait.

Genetic resources See *germplasm.*

Genetical genomics Genetic analyses of the expression values of genes in the same way as any other phenotype.

Genome The complete set of nuclear and cytoplasmic genes present in an organism.

Genome annotation Identification of genes, their 5′- and 3′-regulatory sequences, as well as their functions.

Genome-wide association studies The markers used for association studies are distributed, preferably evenly and densely, over the whole genome.

Genome-wide transcription map A map depicting the transcriptional status as well as the expression levels of all the genes present in the genome.

Genomic estimated breeding value Of an individual, the sum total of effects (on trait phenotype) associated with all the marker alleles included in the GS model applied to the population under selection.

Genomic library A collection of plasmid clones or phage lysates containing recombinant DNA molecules that together, ideally, represent the entire genome of the concerned organism.

Genomic resources The sum total of information about the structural and functional aspects of the genome of the concerned species.

Genomic selection A specialized form of MAS; information from genotype data on all the markers covering the entire genome form the basis of the selection.

Genomic tiling microarray A microarray with a set of overlapping oligonucleotide probes that together represent a part of the genome of a species at very high resolution.

Genomics The field of study of whole genomes in terms of their organization and function.

Genotype calling Assigning of SNP alleles to different individuals in the sample.

Genotypic value Of an individual/line, the phenotype expected from its genotype.

Genotyping array In case of DArT, it contains such genomic DNA segments of a species, which are known to be polymorphic in a range of germplasm of interest.

Genotyping by sequencing The genomic DNA from each individual is digested separately with a frequent cutting restriction enzyme, the fragments are ligated to a barcoded adapter and a common adapter, and only the fragments of >1 kb in size and having the common adapter at one end and the barcoded adapter at the other end are sequenced.

Genotyping Screening of the individuals of a mapping population with polymorphic markers.

Germplasm In theory, the sum total of the genetic information, i.e., all the alleles of various genes, present in a crop species and its wild relatives. In practice, a large collection of different accessions of the concerned species and its wild relatives.

Group I transposons See *retrotransposons.*

Group II transposons These sequences transpose as DNA molecules.

Growth An increase in dry mass, volume, length, or area of a plant as a result of division, expansion, and differentiation of its cells.

GS model training Estimation of the GS model parameters from the phenotype and marker genotype data of the training population.

Haldane mapping function It corrects recombination frequencies for multiple crossing over events assuming lack of interference.

Half-length of D′ The physical/genetic distance at which the value of D' between two loci declines to 0.5.

Haplotype association mapping See in silico *association mapping.*

Haplotype tagging SNPs See *tagSNPs (tSNPs).*

Haplotype The group of alleles of different genes that are located in the same chromosome and tend to be inherited together as a block.

Hemi-SNP A SNP that detects variation at homoeologous/paralogous loci in the two or more genomes of a polyploid species.

Heritability The proportion of genetic variance for a trait to its phenotypic variance.

Heterogeneous stock In outcrossing species with a short-generation time, a population generated by crossing several inbred strains/

lines and maintained by random mating or mating in pairs a suitably large number of individuals.

Heterologous probes Probes prepared from one species and used in another species. Syn., *heterospecific probes*.

Heterosis Superiority of an F_1 hybrid over both its parents in terms of yield or some other trait.

Heterosis QTLs QTLs involved in the expression of heterosis.

Heterotic group Crosses between members of the same heterotic group show little or no heterosis, but those between members of different heterotic groups show moderate to high heterosis.

Heterotic pattern The pattern of variation in the extent of heterosis observed in crosses between members of different heterotic groups.

Heterotic pool See *heterotic group*.

Hierarchical databases The data are organized in a hierarchical (ordered tree) structure, and there are two or more levels of data organization.

High-throughput genotyping Simultaneous genotyping for few to several hundreds or thousands of markers in hundreds to thousands of individuals.

Hitchhiking An increase in the frequencies of alleles at essentially neutral loci located on either side of a locus subjected to selection.

Homoeo-SNPs See *hemi-SNPs*.

Homogeneous group A group of individuals at Hardy–Weinberg equilibrium for all of several random markers/loci.

Homogeneous reaction All the steps of such a reaction are carried out in a single vessel.

Homoplasy The situation of two SSR alleles of identical size being different by descent.

Horizontal disease resistance This type of resistance is governed by polygenes that reduce disease development and, particularly, the pathogen reproduction rate. It is effective against all the races/pathotypes of the pathogen and is ordinarily durable.

Hybrid varieties F_1 generations from crosses between two and more purelines, inbreds, clones, or other genetically dissimilar populations/lines used for commercial cultivation.

Hybrid vigor See *heterosis*.

Hyperspectral reflectance data Data collected at narrow (1–2 nm) bandwidths between 270 and 1,100 nm.

Hypervariable DNA See *variable number of tandem repeats (VNTRs)*.

Identical by descent The copies of an allele of a gene present in two individuals/lines originated by replication of the same ancestral copy of the allele.

Image analysis Processing of an image for deriving the desired information; it involves retrieval, preprocessing, segmentation, reduction, composition, description, and classification steps.

Image composition Individual fragments of an object are merged to create the object.

Image description Quantification of such features as area, height, width, etc., of the object.

Image preprocessing The use of filters to minimize noise or increase sharpness.

Image retrieval Loading of the images from the database/storage folder into the image analysis software.

Image segmentation The image is divided into objects to be analyzed and those to be excluded from analysis.

Imaging fluorometer A fluorescence imaging system is used to acquire images of the fluorescing objects.

Immortalized F_2 population The population of single cross F_1s produced by intercrossing a set of RILs in pairs or as per some other scheme.

In silico association mapping AM based on phenotype and genotype data on inbreds and breeding lines routinely collected in the breeding programs.

In situ conservation of germplasm Establishment of gene sanctuaries or biosphere reserves in areas of high variability within the centers of origin of the concerned crop species by protecting the demarcated areas from human interference.

Inbreeding Mating between individuals related by descent.

Inbreeding depression Reduction in vigor and fertility as a consequence of inbreeding.

Inclusive composite interval mapping It uses a modification of the CIM algorithm: the markers having significant regression coefficient estimates are selected as background markers or cofactors and the regression coefficients for the remaining markers are set at zero; this is done only once during the entire analysis.

Inducer pollinator strain In maize, a strain that induces high-frequency haploid development when it is used a pollinator.

Infrared imaging Imaging using IR sensor or IR/thermal camera based on IR thermometers, and the IR radiation pattern is converted into pseudo-color images.

Initial variety A variety used for the development of a new variety.

Insertional mutagenesis Induction of mutations by insertion of either a transposable element or *Agrobacterium* T-DNA sequence within the concerned gene.

Intellectual property rights Legal recognition of the right of an inventor or his assignee to derive exclusive economic benefits from his intellectual property.

Intellectual property An idea, a design, an invention, a manuscript, etc., which can ultimately generate a useful product/application.

Inter-simple sequence repeat marker A marker based on a single primer composed of a microsatellite sequence. Syn., *inter-SSR PCR marker*.

Inter-SSR amplification See *inter-SSR PCR*.

Inter-SSR PCR Markers generated by anchored ISSR primers.

Inter-SSR PCR marker See *inter-simple sequence repeat (ISSR) marker*.

Interconnected mapping populations Populations produced by crossing a set of homozygous parental lines in such a way that two or more crosses have at least one parent in common.

Intercross recombinant inbred lines RILs produced following few to several generations of random mating or intermating among the individual plants beginning in the F_2 generation.

Intervarietal substitution lines See *chromosome segment substitution lines*.

Interference Occurrence of crossing over at a chromosomal site interferes with the occurrence of another crossing over in its surrounding regions.

Introgression line library A NIL set; the sum total of DP genome segments present in these NILs, ideally, represents the entire DP genome.

Introgression lines See *chromosome segment substitution lines*.

Intron-targeting polymorphism markers Each ITP marker is based on a pair of primers specific to the conserved regions of exons flanking an intron.

Invader technology It exploits the ability of certain enzymes to specifically recognize the invasive nucleic acid structure and cleave at a specific site one of the strands forming this structure.

Invasive nucleic acid structure A replication forklike structure formed when the single-stranded 3′ end of a DNA molecule invades a homologous DNA duplex and displaces the strand having the same sequence as the invading strand.

Ion semiconductor sequencing DNA sequencing method uses a semiconductor-sensing device or ion chip for detection and quantification of the H^+ ions liberated during DNA synthesis by DNA polymerase.

Isogenic lines Lines having identical genotype, except for the alleles of a single gene.

Isoschizomers Two restriction enzymes that recognize the same target sequence, but one of them is methylation sensitive and the other is methylation insensitive.

Isozymes Different forms of an enzyme present in the same individual and having the same catalytic function. In strict sense, each variant is encoded by a different gene and, in practice, includes allozymes.

Joint inclusive composite interval mapping An extension of the ICIM algorithm for analysis of

data from multiple cross populations sharing one common parent, e.g., NAM population.

Joint linkage and association mapping AM analysis of a sample drawn from a natural population and linkage analysis of the open-pollinated progeny from this sample. The multiparent populations like NAM and MAGIC allow creation of populations suitable for JALM with relative ease.

Jumping library The DNA inserts represent sequences on one side each of two neighboring cutting sites of the restriction enzyme used to create the library.

Kinship coefficient The degree of relatedness between pairs of individuals/lines of the sample. Alternatively, it is the probability that the alleles of a randomly chosen gene in a pair of individuals/lines are identical by descent.

LD decay plot Pairwise values of LD are plotted against the genetic distances (in cM) or physical distances (in bp) between pairs of markers.

LD decay Decline in the magnitude of LD with each generation due to recombination between the linked genes.

Ligation In case of DNA, joining of two oligonucleotides/DNA fragments by phosphodiester linkage.

Linkage The tendency of two or more genes or loci being inherited together because they are located close to each other in the same chromosome.

Linkage disequilibrium A specific allele at one locus occurs with a specific allele at the second locus more often than expected on the basis of random assortment of the two loci.

Linkage drag The negative effect of genes linked to the target gene/QTL on the performance of lines produced by gene transfers.

Linkage map A schematic representation of the relative locations of various genetic markers present in the chromosomes of an organism as determined from the frequency of recombination between pairs of markers.

Linking library The DNA inserts represent sequences located on both the sides of a single restriction site for the enzyme used to construct the library.

Locus-specific oligo In case of Illumina GoldenGate assay, its 3′ region is specific to the sequence on the 5′ side of the SNP locus, and the middle sequence is complementary to one of the capture oligos.

LOD score (z) The log to the base 10 of the ratio of probability of obtaining the given data assuming linkage between the two genes with a specified frequency of recombination to the probability of getting the same data with independent segregation.

LOD score threshold The lowest value of LOD score that is accepted as evidence for linkage.

Low-coverage genotyping The genomic DNA from many individuals is pooled and sequenced at low (~2x–4x) coverage.

Main effect QTL A QTL that produces direct effect on expression of the concerned trait.

Major QTL A main effect QTL that explains 10 % or more of the phenotypic variance for the concerned trait.

Map-based cloning See *positional cloning*.

Mapping functions Formulas used for converting recombination frequency into genetic distance.

Mapping population A population that is suitable for linkage mapping of genetic markers, genes, and/or QTLs.

Marker index The product of multiplex ratio and the average PIC score for a marker system in a given population.

Marker-assisted backcrossing A backcross program based on molecular markers.

Marker-assisted plant breeding The use of molecular marker data for enhancing the effectiveness of various breeding activities, including planning and execution of breeding programs, and improving selection efficiency.

Marker-assisted selection Selection for the desirable allele of a gene/quantitative trait locus (QTL) on the basis of molecular marker(s) linked to it in place of the phenotype generated by this allele.

Matrix substance An organic molecule that has the same energy absorption spectrum as the

selected laser wavelength and does not interact chemically with the target biomolecule.

Melting temperature (T_m) The temperature at which 50 % of the DNA duplexes would dissociate into separate strands.

Meta-analysis Combining the results from many different studies concerning a single research issue to identify common patterns, sources of disagreements, and any other relationships among their findings.

Meta-QTLs QTLs identified by QTL meta-analysis. Also called *true QTLs*.

Metabolic QTLs These QTLs control the rates of various metabolic reactions and metabolite levels.

Metabolome All the metabolites, representing the end products of cellular processes, present in a cell, tissue, organ, or organism.

Metabolomics A systematic study of the characteristic small-molecule metabolite profiles generated by the various cellular metabolic processes.

Metadata The details, usually in digital form, of experimental conditions and the procedures followed for the phenomics studies (or any other study).

Microarray A small plaque/wafer of silicon, glass, or metal, onto which one end of multiple copies of each of a large number of different single-stranded DNA molecules is covalently linked and the different molecules are arranged in separate spots.

Microsatellites Usually, <100 bp long sequences comprising tandem repeats of 2–7 bp.

Microsatellite markers See *simple sequence repeat (SSR) markers*.

Microsatellite-primed PCR See *single primer amplification reaction*.

Microsynteny Synteny based on DNA sequence.

Migration See *admixture*.

Mini-sequencing See *single-base extension*.

Minisatellites Sequences, typically 0.2–2 kb long, made up of 11–60 bp long tandem repeat units having identical or almost identical sequences.

Minor QTL A main effect QTL that explains less than 10 % of the phenotypic variance for the concerned trait.

Missing heritability The part of phenotypic variation in a quantitative trait that is not explainedb y the QTLs identified by various AM studies.

Mixed linear model The markers and the population structure (Q) are treated as fixed linear effects and the additive effects of the multiple background QTLs are considered as linear random effects.

Model-based clustering methods Cluster membership is based on some parametric evolutionary model.

Molecular beacons Specially designed oligonucleotide hybridization probes used for identification of SNP alleles.

Molecular biomarkers Those dynamically expressed molecules that can be measured and used as indicators of specific phenotypic features.

Molecular inversion probe A single 120 nucleotide (nt) long oligonucleotide that hybridizes to a specific sequence of the genome and forms a circle that has a single base pair gap at the SNP site. The assay involves primer extension and ligation producing a closed circular molecule.

Molecular Plant Breeding. See *marker-assisted plant breeding*.

Morgan A measure of genetic distance; one Morgan (M) equals 100 centimorgans.

Morphological markers Simply inherited and easily scored morphological traits; the earliest genetic markers.

Multitrait mixed model It extends the linear mixed-model approach of AM to analysis of pairs of correlated traits.

Multilocus mixed model It includes multiple loci as cofactors in the AM model and employs a simple stepwise mixed-model regression analysis combined with forward inclusion and backward elimination of loci in the model.

Multiparent advanced generation intercross populations A collection of RILs produced

from a complex crossbred/outbred population involving several parental lines.

Multiple interval mapping An approach for simultaneous QTL mapping in multiple marker intervals.

Multiple QTL mapping It combines simple interval mapping with multiple regression analysis and includes all the significant QTLs in the genetic model used for mapping.

Multiplex PCR Two or more primer pairs used for amplification of two or more loci in a single PCR reaction tube.

Multiplex ratio The average number of markers scored per assay of a marker system.

Multiplex-endonuclease genotyping approach AFLP A modification of AFLP; four or more endonucleases are used for digestion of the sample DNA; only one pair of adapters used for amplification.

Multiplexed shotgun genotyping Size-selected (250–300 bp) restriction fragments from several individuals are separately ligated to distinct barcodes, pooled, and sequenced using a NGS platform.

Multiplexing Carrying out two or more different reactions, e.g., PCR amplification, in a single tube or separating the products of two or more PCR reactions in a single gel lane.

Multispectral reflectance data Reflectance data acquired at few selected wavelengths.

Multitrait introgression Introgression of genes governing two or more different traits into a single RP.

Multivariate linear mixed model It allows testing of associations between markers and multiple correlated phenotypes and is able to control population structure.

Multivariate methods Methods for analysis of data on multiple traits for each entity.

MutMap scheme A quick, reliable, and cost-effective method for mapping of causal SNPs in induced mutations. It uses a single bulk of the mutant plants from the F_2 generation of mutant $\times$ parent cross and a reference genome for alignment of the NGS sequence data from this bulk.

MutMap-Gap scheme It identifies causal mutations located in the genomic regions missing from the parental/reference genome.

Near-isogenic lines Pairs of homozygous lines that are identical in genotype, except for a single gene/locus; in practice, they also differ for a variable length of the genomic region flanking this locus and some random genomic segments located elsewhere in the genome.

Nebulization Mechanical shearing of DNA.

Negative selection The use of molecular markers closely linked to undesirable alleles of known genes/QTLs to select against these alleles.

Nested parents In case of NAM, one or few parents crossed with all the founder parents.

Next-generation DNA sequencing methods These methods use PCR for in vitro cloning and sequence multiple copies of a very large number of relatively small DNA fragments.

Noise reduction In image analysis. morphological operations to correct the imperfections like holes present in the binary images.

Non-anchored ISSR primer In case of ISSR, the primer consists solely of a microsatellite sequence.

Non-imaging fluorometer Use of portable handheld fluorometers to measure fluorescence from few square millimeter leaf area.

Non-photochemical quenching The part of light energy lost as heat via the xanthophyll cycle.

Normalized difference vegetative index The ratio of difference between reflectance in the NIR (at 800 nm) and red (at 680 nm) regions to the total of the two.

Normalized difference water index The ratio $[(R_{860}R_{1200})/(R_{860}+R_{1200})]$, where R_{860} and R_{1200} are reflectance at 850 and 1,200 nm, respectively.

Novelty A variety should not have been commercially exploited for more than one year before the grant of PBR protection.

Null allele The specific primers for a SSR locus consistently fail to amplify a detectable product.

Oligonucleotide ligation assay Assay for SNP genotyping based on hybridization of a pair of oligos with the target PCR products, followed by ligation of the two oligos by DNA ligase.

Oligonucleotide microarrays Oligonucleotides synthesized at a very high density (up to one million oligonucleotides/cm^2) directly on thin wafers of silicon glass. Syn., *DNA chips*.

Ontologies Controlled vocabularies shared by database communities working on different taxa.

Open reading frame A correct reading frame that begins with an initiation codon and ends with a termination codon.

Orthologous genes Genes of different species performing the same function.

Orthologous sequences Sequences from different species that originated from the same ancestral sequence.

Overdominance hypothesis Heterozygotes at certain loci are superior to the two homozygotes for the locus leading to heterosis.

Overgo probes Gene-specific oligonucleotide-based probes designed from ESTs.

***p*-value** The probability of Type I error or the probability of null hypothesis, e.g., a lack of marker–trait association, being wrongly rejected.

Paired-end sequencing In NGS, sequencing of both the ends of each DNA fragment.

Paralogous genes Two or more genes present in the genome of the same species that originated from the same ancestral gene and have the same function.

PCR-based markers Markers based on DNA sequence polymorphisms detected by PCR amplification of sample DNAs. Often called *second-generation molecular markers*.

Phenome The sum total of phenotypes at various levels ranging from molecules to organs and the whole organism.

Phenotypic selection Selection based on phenotypes of the target traits.

Phenotyping Evaluation of the individuals of a mapping population for phenotypic expression of the target trait.

Photochemical quenching The part of light energy used for electron transport and carbon assimilation.

Phylogenetic analysis Grouping the various entities included in a study on the basis of their genetic relationships indicating the degrees of genetic similarities–dissimilarities among them.

Phylogenetic trees A graphic or textual representation of the evolutionary relationships among various entities based on similarities and differences in their physical and genetic characteristics.

Physical distance The distance in terms of base pairs.

Physical map The genes/molecular markers are depicted in the same order as they occur in the chromosomes, but the distances between adjacent genes/markers are depicted in terms of base pairs.

Plant breeder's rights The right granted to a plant breeder, originator, or owner of a plant variety/hybrid to exclude others from producing or commercializing the propagating material of that variety/hybrid; the protection period 15–20 years.

Plant breeding The discipline that aims to change the genetic constitution of crop plants so as to make them more useful to humans.

Plant ontology It relates to anatomical features and developmental stages in different plant species.

Plant phenomics In simple terms, the study of plant growth, architecture, performance, and composition using high-throughput methods of data acquisition and analysis.

Pleiotropy A single gene influences the phenotypic expression of more than one trait.

Polygenes Genes producing small individual effects on the trait phenotype, but the effects of all the polygenes affecting a given trait are cumulative.

Polygenic effect term See *a matrix*.

Polymorphic information content The probability of a marker locus being polymorphic between two random individuals/lines selected from a given population, often referred to as *expected heterozygosity*.

Pooled mapping Plants homozygous for the recessive phenotype of the target trait are

selected from a large segregating population of a suitable cross and are divided into several random pools, each pool is analyzed with many markers, and markers closely linked to the target gene are identified.

Population mapping Mapping based on estimates of linkage disequilibrium (LD) obtained from populations consisting of a diverse set of individuals/lines drawn from either natural or breeding populations.

Population structure The level of genetic differentiation among different homogeneous groups present in a population.

Positional cloning Isolation and cloning of a gene on the basis of its close linkage with a DNA marker involves identification of a pair of markers flanking the mutant allele, identification and isolation of the DNA fragment containing the mutant allele, and determination of function of the concerned gene.

Positive false discovery rate The expected ratio of the wrongly rejected null hypotheses (H_0) to the total number of H_0 rejected in the experiment when positive findings have occurred.

Power of association mapping The probability of detecting "true" marker–trait associations in a sample using AM.

Power of QTL detection The probability of detecting a QTL with a given effect size and the stated level of Type I error.

Preamplification step Amplification of fragments using two AFLP primers, each having one selection nucleotide at its 3′ end.

Precision of QTL mapping The dispersion of repeated independent estimates of the QTL position or that of the genetic effects of the QTL alleles.

Primary mapping populations Populations created by hybridization between two homozygous lines usually having contrasting forms for the traits of interest.

Primary study A study that enables a discovery, e.g., detection and mapping of a QTL governing a trait of interest.

Primer extension In case of SNP genotyping, a specially designed primer is annealed to the target PCR product, extended by one to few nucleotides using DNA polymerase, and the products of the extension are analyzed to deduce the SNP allele.

Prior index An index based on marker genotype and trait phenotype data from materials other than the population being subjected to selection.

Probes Small DNA or RNA fragments of, usually, 500–3,000 bp; used for hybridization to detect specific fragments from among mixture of many different fragments.

Protein quantity QTLs They govern variation in the cellular contents of specific proteins.

Protein-based markers Electrophoretic variants of proteins, including enzymes.

Proteome The complete set of proteins expressed in a cell during a specific developmental stage and under the given environmental conditions.

Proteomics The study of proteome using a diverse array of techniques.

Pseudo-overdominance Heterozygote superiority is due to repulsion phase linkage.

PSI-BLAST It is used to identify all the members of a very large gene family, which is not possible by using the simple BLAST programs.

Pureline Self-pollinated progeny of a single homozygous plant of a self-pollinated species.

Pyrosequencing A method of DNA sequencing based on the use of pyrophosphate released on addition of a nucleotide to a growing chain for generation of, ultimately, light by luciferase that is detected.

***Q* matrix** A Q value indicates the likelihood that an individual belongs to a given putative homogeneous cluster/group present in a population.

QTL See *quantitative trait locus*.

QTL × environment interaction The effect size estimates for the same QTL in the same mapping population vary from one environment to the other, and some QTLs may not even be detected in some of the environments.

QTL analysis QTL detection, mapping, and fine mapping.

QTL confidence interval The genomic region in which the QTL is likely to be located. Syn., *QTL support interval*.

QTL fine mapping Identification of markers located very close to (<1 cM) the concerned QTL.

QTL hotspot A genomic region having QTLs involved in the control of different correlated traits.

QTL meta-analysis Integration of findings from different QTL studies to determine the "actual" number of QTLs affecting a trait, estimate their "actual" positions in the genome, and reduce the QTL confidence intervals.

QTL support interval See *QTL confidence interval*.

QTL validation Confirmation of the marker–QTL association and the QTL position in unrelated germplasm and assessment of the effect of genetic background on QTL expression.

QTL-NIL An NIL that contains a single segment of DP genome having, ideally, a single favorable QTL.

QTL-Seq An extension of BSA for QTL mapping by whole-genome resequencing of the two DNA bulks, i.e., the "high" and "low" trait phenotype bulks.

Qualitative trait A trait governed by one or few genes with large effects, the phenotypic expression of which is relatively little affected by the environment so that individuals can be readily classified into two or more distinct classes on the basis of their phenotype.

Quantitative trait A trait governed by several genes each having a small individual effect on the trait phenotype, which are, usually, cumulative. The phenotypic expression of such traits is markedly affected by the environment so that the individuals cannot be classified into distinct classes on the basis of trait phenotype.

Quantitative trait locus The genomic region associated with the expression of a quantitative trait; a QTL may contain one or more genes affecting the given trait**.**

Quantitative trait nucleotides The nucleotide changes that produce different alleles of a QTL.

RAD tag Short genomic sequences flanking the restriction site for the concerned restriction enzyme.

Randomly amplified microsatellite polymorphism Markers generated by using a 5′ anchored SSR primer in combination with a RAPD primer.

Random DNA markers DNA markers derived from polymorphic sequences located at random sites in the genome.

Randomly amplified polymorphic DNAs A dominant marker system; uses a single, typically, 10 nt long primers with an arbitrary base sequence for amplification of sequences from genomic DNAs of test individuals.

Read accuracy Accuracy of the sequence of a single read.

Read alignment. Aligning the short sequence reads onto a reference genome sequence.

Read length The length of the sequence generated from one DNA fragment.

Read mapping See *read alignment*.

Reading frame The arrangement of sets of three bases, each representing a codon, beginning at a specific nucleotide in a DNA sequence.

Recipient parent See *recurrent parent*.

Recognition sequences The sequences recognized by restriction enzymes. Syn.*, recognition sites, restriction sites.*

Recognition sites See *recognition sequences*.

Recombinant The individual having a new combination of linked genes.

Recombinant selection Marker-assisted selection against the DP genomic regions flanking the target gene/QTL.

Recombination Production of new combinations of linked genes.

Recurrent parent A homozygous line to which the F_1 plants and subsequent generations are backcrossed.

Recurrent selection backcross population A population developed by backcrossing the F_1 from a cross between lines having high

(DP) and low (RP) values for a quantitative trait and the subsequent generations to the RP; in each backcross generation, a predetermined number of individuals with the top phenotypic values, i.e., values close to the DP phenotype, for the trait are selected for backcrossing.

Red Edge NDVI It is the ratio of $(R_{750}-R_{705})$ to $(R_{750} + R_{705})$, where R_{750} and R_{705} represent reflectance at 750 and 705 nm, respectively.

Reduced representation libraries Libraries constructed by fully digesting the genomic DNA with a frequent cutting restriction enzyme and selecting fragments of ~300 bp or so for cloning/sequencing; ~1–10 % of the genome is represented in a RRL.

Reduced representation sequencing Only a subset of the genomic fragments is sequenced in each individual for marker discovery.

Reference parents See *nested parents*.

Regulatory QTLs See *expression QTLs*.

Relational database A database constructed using the SQL and organized in tables; the columns of tables are indexed according to common features.

Replication studies Subsequent studies conducted to confirm and validate the results from primary studies.

Reporter oligo The oligo complementary to the 5′ side of the SNP locus and lacking the polymorphic nucleotide.

Resistance gene analog markers Markers based on primers derived from conserved regions of disease resistance genes of plants.

Restriction enzyme An endonucleases that cleaves DNA only within or near a site that has a specific base sequence.

Restriction fragment length polymorphism A single restriction enzyme produces fragments of different lengths from the same genomic regions of different individuals/strains/lines.

Restriction-site-associated DNA markers Polymorphisms in the recognition sites for the restriction enzyme used for preparation of the assay sample.

Restriction-site-associated DNA sequencing Short regions surrounding all recognition sites present in the genome for the selected restriction enzyme are sequenced, derived from the RAD tag marker technique.

Restriction sites See *recognition sequences*.

Retrotransposon-based insertion polymorphism Detects retrotransposon insertions at specific sites using PCR amplification; one primer derived from the concerned retrotransposon and a pair of primers derived from the sequence flanking the insertion site.

Retrotransposons These sequences transpose via RNA intermediates; Syn., *Group I transposons*.

Reverse phenomics Detailed analysis of a trait to unravel the various physiological, biochemical, and biophysical processes and the genes involved in control of the trait.

Ridge regression In this model, all the marker effects are considered to belong to a normal distribution with mean zero and variance σ_g^2; it shrinks all marker effects toward zero.

RNA-Seq Sequencing of the complete transcriptome using a NGS technology.

Round robin mating scheme Each member of a set of lines is mated as male to a defined number of lines and as female to an equal number of other lines.

Second-generation DNA sequencing methods See *next-generation DNA sequencing methods*.

Secondary mapping populations Populations developed by crossing two lines/individuals selected from a mapping population and created mainly for fine mapping of the genomic region of interest.

Segregation distortion A significant deviation of the observed segregation ratio for a marker locus from the expected ratio.

Selection bias In case of SIM, the tendency to detect only large effect QTLs, leading to an upward bias in the effect size estimates for the detected QTLs.

Selection index A numerical score that combines information on all the traits associated with the dependent variable, usually, yield.

Selection nucleotide An arbitrary nucleotide added to the 3′ end of the primers to reduce

the number of different fragments amplified by PCR.

Selection Differential reproduction rates for different genotypes.

Selective genotyping An extension of the BSA approach; the population is phenotyped for the trait of interest, 30–50 plants/lines with extreme high and a similar number of plants with extreme low phenotypic values for the trait are selected, plants/lines are subjected to precision phenotyping and genotyping for a large number of markers, and the data from the two groups are analyzed to identify the markers linked to the target trait.

Selective mapping A small sample is drawn from the population on the basis of chromosomal breakpoints, and the individuals are genotyped with a large number of new markers, which are assigned to appropriate bins.

Sequence capture A targeted SNP discovery strategy applied to specific genomic regions.

Sequence characterized amplified regions A marker derived from an RAPD marker; a pair of forward and reverse primers, usually, of 20–24 nt long, specific for the two terminal sequences of the RAPD marker, are used for PCR amplification.

Sequence-related amplified polymorphism A marker based on amplification of open reading frames (ORFs); one primer targets exons and the other targets introns and promoters.

Sequence-specific amplification polymorphism A modification of AFLP; the restriction fragments are amplified using one AFLP primer and one primer based on a conserved sequence of a transposable element (TE). Syn., *transposon display procedure*.

Sequence-tagged microsatellite profiling A modification of AFLP; one AFLP primer and one primer based on a SSR sequence (anchored at its 3′ end) are used for amplification of the restriction fragments.

Sequence-tagged microsatellite site markers See *simple sequence repeat markers*.

Sequence-tagged site A locus that can be unambiguously defined by a pair of primer sequences that are used for its amplification.

Sequencing by synthesis DNA sequencing based on DNA synthesis.

Sequencing depth For a specific nucleotide, the total number of all reads, in which the given genomic position or base pair is represented. For the whole genome, the average number of times each base of the entire genome of an individual has been sequenced.

Shotgun sequencing In NGS, sequencing of one end of each DNA fragment.

Simple interval mapping A systematic one-dimensional search for one QTL at a time treating each marker interval independent of other intervals.

Simple sequence length polymorphisms See *simple sequence repeat markers*.

Simple sequence repeat markers A special version of STS markers; a microsatellite locus is amplified using a pair of specific primers based on the unique sequences flanking the locus. Syn., *sequence-tagged microsatellite site markers, simple sequence length polymorphisms, microsatellite markers*.

Single-base extension (SBE) A method of SNP genotyping; the primer is extended by a single nucleotide only.

Single feature polymorphism Allelic variation detected between pairs of lines of a species by using gene sequence-based high-density oligonucleotide microarrays for hybridization with their genomic fragments/cDNAs.

Single marker analysis Each marker is separately tested for its association with the target trait.

Single nucleotide polymorphism Variation in single base pairs of DNA.

Single point analysis See *single marker analysis*.

Single primer amplification reaction Markers generated by non-anchored ISSR primers.

Single-strand conformation profile/polymorphism A marker system, in which detection is based on differential movement of single-stranded DNA molecules representing identical genomic regions from different individuals.

Smart plant breeding See *marker-assisted plant breeding*.

SNP calling Determination of the genomic positions at which single nucleotide polymorphisms occur.

SNP index The ratio of the number of short reads with the mutant allele at a SNP locus to the total number of short reads covering the SNP locus.

SNP locus A specific position in the genome, at which different nucleotides occur in the same DNA strand of different individuals of the species.

SNP mining SNP discovery by analysis of genomic and/or EST sequences of different individuals of a species available in the databases.

SNP validation Evaluation that a discovered SNP is a true SNP and not a product of sequencing error, faulty read alignment, etc.; that its alleles represent homologous genomic regions; and that it segregates in a typical Mendelian fashion.

Somaclonal variation Heritable variation generated in cells and tissues grown in vitro.

Southern hybridization A DNA, DNA hybridization procedure named after E.M. Southern, involves restriction digestion of DNA followed by gel electrophoresis, blotting, hybridization with a labeled probe, and detection of the bands hybridized with the probe.

Spotted microarrays DNA fragments representing different genes of an organism spotted onto a suitable solid support.

SSD procedure One seed is harvested from each plant of the F_2, and the later generations and seeds from all the plants are composited and planted to raise the next generation.

Stability In case of PBR, the new variety must be stable in appearance and its clonal characteristics over successive generations under the specified environment.

Stable QTL The phenotypic effect of such a QTL is little affected by the environment so that it is detected across environments.

Standard heterosis See *economic heterosis*.

Start codon-targeted marker Markers based on single primers designed on the basis of the sequence of the short conserved region around the start codon, ATG, of plant genes.

Stepwise regression It treats marker effects as fixed and fits the markers into the model either singly or in small groups.

Strength of a QTL The proportion of total phenotypic variance for the target trait explained by the QTL.

Stress An environmental factor that limits the performance of a crop genotype.

Stringency of distinctness criterion In case of PBR, the number of bands required to differ between a pair of varieties/lines for them to be accepted as distinct varieties/lines.

Structural genomics Determination of the complete genome sequence and the complete set of proteins produced by an organism.

Structural proteomics Mapping of the 3-D structure and nature of protein complexes present specifically in a particular cell/organelle.

Structured association model An AM model designed to tackle the problems due to population structure.

SubBin In case of MAPMAN tool, a subdivision of a Bin.

Subcloning In chromosome walking, cloning of a small segment representing one end of the DNA fragment being analyzed.

***SupF* marker** A tRNA gene with a mutated anticodon that recognizes a polypeptide chain termination codon generated by a suppressor-sensitive mutation within a gene.

Surrogate measurement It determines the level of some other trait that shows a reliable and predictable relationship with the target trait.

Syntenic markers Markers located in the same chromosome.

Tag oligonucleotide In SNP genotyping, a unique oligonucleotide unrelated to the sequences of the locus to be genotyped; each tag oligo is attached to a known position on the microarray/chip.

tagSNPs (tSNPs) The subset of SNP loci that together enable a reliable deduction of genotypes at the remaining SNP loci of the haplotype block.

Target region amplification polymorphism A marker that detects polymorphism around the desired candidate genes; it uses a PCR primer based on the target EST, while the other primer targets either an intron or an exon.

tBLASTn It converts the submitted protein sequence into a nucleotide sequence and compares it with a nucleotide sequence database.

tBLASTx It translates the submitted nucleotide sequence as well as the nucleotide database sequence into protein sequences and searches for homology between the two.

Template switching During production of cDNA, reverse transcriptase uses one RNA molecule as template for some distance and then uses another RNA molecule as template; as a result, the cDNA molecule is made up of the 3′ region of the first RNA template and the 5′ region of the second RNA template.

Tentative ESTs Contigs obtained by de novo assembly of RNA-Seq sequence data.

Tentative unique sequences See *tentative ESTs*.

Thermal imaging See *infrared imaging*.

Thermography See *infrared imaging*.

Third-generation DNA sequencing methods These methods sequence single large DNA molecules.

Three-endonuclease AFLP A modification of AFLP, in which three restriction enzymes are used to digest the sample DNA.

Threshold characters Such characters require a specific environment, i.e., a threshold environment, for their expression.

Throughput The number of assays, e.g., SNP genotyping, carried out by an assay system in a unit time.

Tools See *bioinformatics tools*.

Training population In case of GS, the population used for training the GS model and for obtaining estimates of the marker-associated effects.

Trait ontology It lists the details of evaluation procedure and the environments, in which a specific trait of a given species was assayed.

Transcriptome The full complement of RNA molecules, including their quantities, produced by a cell during a specific developmental stage and under a given environment.

Transcriptomics Cataloguing of all the species of RNA transcripts expressed in a tissue/organ, their expression levels, splicing patterns, etc. and the effects of developmental stages and environmental conditions on their expression.

Transferability of SSR markers Primers for SSR markers developed in one species can be used in some other species.

Transposable elements Mobile DNA sequences.

Transposon display procedure See *sequence-specific amplification polymorphism*.

Transposons See *transposable elements*.

True heterosis See *heterosis*.

"True" QTL See *meta-QTL*.

True SNP In case of polyploid species, allelic variation between homologous loci of the same genome present in the same or different polyploid species.

Type I error The probability of null hypothesis, e.g., lack of marker–trait association, being wrongly rejected, i.e., being rejected even when it is correct.

Uniformity One of the DUS criteria for PBR protection; the new variety must be sufficiently uniform in appearance, particularly for the traits used to establish its distinctness, on plant-by-plant basis under the specified environment of its adaptation.

Unigenes Unique gene sequences; a set of non-redundant EST sequences for a species, each of which has a unique identity and genomic position.

Unstable QTL The phenotypic effect of such a QTL is markedly affected by the environment so that it is detected in only some of the environments.

Useful heterosis See *economic heterosis*.

Useful LD The level of LD that is useful for association mapping.

Validation of marker–trait linkage Evaluation of the observed marker–trait linkage in a fairly large number of unrelated germplasm showing variation for the concerned trait.

Variable number of tandem repeats Stretches of DNA composed of variable numbers of tandemly repeated sequences of, usually, 2–60 bp.; syn., *hypervariable DNA*.

Vegetation index The ratio of reflectance in the NIR region (at 800 nm) to that in the red region (at 680 nm).

Vertical disease resistance Disease resistance governed by major genes that exhibit gene-for-gene relationship, generate hypersensitive response to specific races/pathotypes of the concerned pathogen, and usually block disease development soon after the infection stage so that plants are virtually disease-free.

Visual imaging Digital imaging in the visible range (400–700 nm).

Water index The ratio of R_{900} to R_{970}, where R_{900} and R_{970} are reflectance at 900 and 970 nm, respectively.

Water-use efficiency The ratio of the amount of water used for metabolism by a plant to that lost through transpiration.

Wild form Wild species from which a crop species is considered to have directly evolved.

Wild relatives All the wild species that are phylogenetically related to a crop species.

References

Abdurakhmonov IY, Abdukarimov A (2008) Application of association mapping to understanding the genetic diversity of plant germplasm resources. Int J Plant Genomics 2008:1–18

Abe A, Kosugi S, Yoshida K et al (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. Nat Biotechnol 30:174–179

Agarwal M, Shrivastava N, Padh H (2008) Advances in molecular marker techniques and their applications in plant sciences. Plant Cell Rep 27:617–631

Ahmadi N, Albar L, Pressoir G et al (2001) Genetic basis and mapping of the resistance to rice yellow mottle virus. III. Analysis of QTL efficiency in introgressed progenies confirmed the hypothesis of complementary epistasis between two resistance QTLs. Theor Appl Genet 103:1084–1092

Akbari M, Wenzl P, Caig V et al (2006) Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor Appl Genet 113:1409–1420

Akey JM, Zhang K, Xiong M (2003) The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. Mol Biol Evol 20:232–242

Albini G, Joets FM (2003) ActionMap: a web-based software that automates loci assignments to framework maps. Nucleic Acids Res 31:3815–3818

Albrecht T, Wimmwe V, Auinger H-J et al (2011) Genome-based prediction of test-cross values in maize. Theor Appl Genet 123:339–350

Allard RW (1960) Principles of plant breeding. Wiley, New York

Alpert KB, Tanksley SD (1996) High-resolution mapping and isolation of a yeast artificial chromosome contig containing *fw2.2*: a major fruit weight quantitative trait locus in tomato. Proc Natl Acad Sci U S A 93:15503–15507

Altschul SFW, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altshuler D, Pollara VJ, Cowles CR et al (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513–516

Anand D, Prabhu KV, Singh AK (2012) Analysis of molecular diversity and of commercially grown Indian rice hybrids. J Plant Biochem Biotechnol 21:173–179

Anderson JR, Lubberstedt T (2003) Functional markers in plants. Trends Plant Sci 8:554–560

Andolfatto P, Davison D, Erezyilmaz D et al (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res 21:610–617

Anonymous (2008) Field estimation of soil water content: a practical guide to methods, instrumentation and sensor technology, Training course series no. 30. International Atomic Energy Agency, Vienna

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19:52–61

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Genetics 3:299–309

Arends D, Prins P, Jansen RC et al (2010) R/qtl: high-throughput multiple QTL mapping. Bioinformatics 26:2990–2992

Arens P, Mansilla C, Deinum D et al (2010) Development and evaluation of robust molecular markers linked to disease resistance in tomato for distinctness, uniformity and stability testing. Theor Appl Genet 120:655–664

Areshchenkova T, Ganal MW (2002) Comparative analysis of polymorphism and chromosomal location of tomato microsatellite markers isolated from different sources. Theor Appl Genet 104:229–235

Ashikari M, Sakakibara H, Lin S et al (2005) Cytokinin oxidase regulates rice grain production. Science 309:741–745

Austin RS, Vidaurre D, Stamatiou G et al (2011) Next-generation mapping of Arabidopsis genes. Plant J 67:715–725

Ayliffe MA, Lawrence GJ, Ellis JG (1994) Heteroduplex molecules formed between allelic sequences cause non-parental RAPD bands. Nucleic Acids Res 22:1632–1636

Babu R, Nair SK, Prasanna BM et al (2004) Integrating marker-assisted selection in crop breeding – prospects and challenges. Curr Sci 87:607–619

Babu R, Nair SK, Kumar A et al (2005) Two-generation marker-aided backcrossing for rapid conversion of normal maize lines to quality protein maize (QPM). Theor Appl Genet 111:888–897

Babu R, Prasanna BM (2014) Molecular Breeding for Quality Protein Maize (QPM). In: Tuberosa R, Graner A, Frison E (eds) Genomics of plant genetic resources,

DOI 10.1007/978-81-322-2316-0, © Author(s) 2015

vol 2, Crop productivity, food security and nutritional quality. Springer Science + Business Media, Dordrecht, pp 489–505

Babu KN, Rajesh MK, Samsudeen K et al (2014) Randomly amplified polymorphic DNA (RAPD) and derived techniques. Methods Mol Biol 1115:191–209

Bachem CW, van der Hoeven RS, de Bruijn SM et al (1996) Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. Plant J 9:745–753

Bagge M, Lubberstedt T (2008) Functional markers in wheat: technical and economic aspects. Mol Breed 22:319–328

Baird N, Etter P, Atwood T et al (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376

Bairoch A, Apweiler R (1996) The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. Nucleic Acids Res 24:21–25

Ball RD (2007) Statistical analysis and experimental design. In: Oraguzie NC, Rikkerink EHA, Gardiner SE et al (eds) Association mapping in plants. Springer Science + Business Media, LLC, New York, pp 133–196

Bandillo N, Raghavan C, Muyco PA et al (2013) Multiparent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice 6:11, http://www.thericejournal.com/content/6/1/11

Banerjee S, Yandell BS, Yi N (2008) Bayesian quantitative trait loci mapping for multiple traits. Genetics 179:2275–2289

Baranwal VK, Mikkilineni V, Zehr UB (2012) Heterosis: emerging ideas about hybrid vigour. J Exp Bot 63:6309–6314

Barker G, Batley J, O'Sullivan H et al (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics Appl Note 19:421–422

Basten CJ, Weir BS, Zeng ZB (1997) QTL Cartographer: a reference manual and tutorial for QTL mapping. North Carolina State University, Raleigh

Baxevanis D (2000) The Molecular Biology Database Collection: an online compilation of relevant database resources. Nucleic Acids Res 28:1–7

Beavis WD (1994) The power and deceit of QTL experiments: lessons from comparative QTL studies. In: Proceedings of the forty-ninth Annual Corn and Sorghum Research Conference, Chicago, IL, 7–8 Dec 1994. Am Seed Trade Assoc, Washington, DC, pp 250–266

Beavis WD (1998) QTL analysis: power, precision and accuracy. In: Paterson AH (ed) Molecular dissection of complex traits. CRC Press, Boca Raton, pp 145–162

Becker J, Heun M (1995) Mapping of digested and undigested random amplified microsatellite polymorphisms in barley. Genome 38:991–998

Beckmann JS, Soller M (1983) Restriction fragment length polymorphism in genetic improvement: methodologies, mapping and cost. Theor Appl Genet 67:35–43

Beckmann JS, Soller M (1986) Restriction fragment length polymorphism in genetic improvement. Oxford Surv Plant Mol Biol 3:197–250

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Bennetzen JL, Ma J (2003) The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. Curr Opin Plant Biol 6:128–133

Bentley DR, Smith AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59

Berger B, Parent B, Tester M (2010) High-throughput shoot imaging to study drought responses. J Exp Bot 61:3519–3528

Bernacchi D, Beck BT, Eshed Y et al (1998) Advanced backcross QTL analysis in tomato. I. Identification of QTLs for traits of agronomic importance from *Lycopersicon hirsutum*. Theor Appl Genet 97:381–397

Bernardo R (2001) What if we knew all the genes for a quantitative trait in hybrid crops? Crop Sci 41:1–4

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last twenty years. Crop Sci 48:1649–1664

Bernardo R (2009) Genome-wide selection for rapid introgression of exotic germplasm in maize. Crop Sci 49:419–425

Bernardo R (2010) Genome-wide selection with minimal crossing in self-pollinated crops. Crop Sci 50:624–627

Bernardo R (2013) Genome-wide markers for controlling background variation in association mapping. Plant Genome 6:1–9

Bernardo R, Charcosset A (2006) Usefulness of gene information in marker-assisted recurrent selection: a simulation appraisal. Crop Sci 46:614–621

Bernardo R, Kahler AL (2001) North American study on essential derivation in maize: inbreds developed without and with selection from F2 populations. Theor Appl Genet 102:986–992

Bernardo R, Yu J (2007) Prospects for genome-wide selection for quantitative traits in maize. Crop Sci 41:1–4

Bilder RM, Sabb FW, Cannon TD et al (2009) Phenomics: the systematic study of phenotypes on a genome-wide scale. Neuroscience 164:30–42

Bink MCAM, Boer MP, ter Braak CJF et al (2008) Bayesian analysis of complex traits in pedigreed plant populations. Euphytica 161:85–96

Bink MCAM, Totir LR, ter Braak CJF et al (2012) QTL linkage analysis of connected populations using ancestral marker and pedigree information. Theor Appl Genet 124:1097–1113

Bink MCAM, Jansen J, Madduri M et al (2014) Bayesian QTL analyses using pedigreed families of an outcrossing species, with application to fruit firmness in apple. Theor Appl Genet. doi:10.1007/s00122-014-2281-3

Boitard S, Abdallah J, de Rochambeau H et al (2006) Linkage disequilibrium interval mapping of quantitative trait loci. 8th World Congress on Genetics Applied to Livestock Production, 13–18 August 2006, Belo Horizonte

Bonnet DG, Rebetzke GJ, Spielmeyer W (2005) Strategies for efficient implementation of molecular markers in wheat breeding. Mol Breed 15:75–85

Borevitz JO, Liang D, Plouffe D et al (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res 13:513–523

Bortiri E, Jackson D, Hake S (2006) Advances in maize genomics: the emergence of positional cloning. Curr Opin Biol 9:1–8

Botstein D, White R, Skolnick M et al (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Bouchez A, Hospital F, Causse M et al (2002) Marker assisted introgression of favorable alleles at quantitative trait loci between maize elite lines. Genetics 162:1945–1959

Bradbury PJ, Zhang Z, Kroon DE et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics Appl Note 23:2633–2635

Braun A, Little DP, Koster H (1997) Detecting CFTR gene mutations by using primer oligo base extension and mass spectrometry. Clin Chem 43:1151–1158

Brazma A, Parkinson H, Sarkans U et al (2003) ArrayExpress - a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31:68–71

Brennan JP, Martin PJ (2007) Returns to investment in new breeding technologies. Euphytica 157:337–349

Breseghello F, Sorrells ME (2006) Association analysis as a strategy for improvement of quantitative traits in plants. Crop Sci 46:1323–1330

Broman KW, Wu H, Sen S et al (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890

Brown SM, Kresovich S (1996) Molecular characterization for plant genetic resources conservation. In: Paterson AH (ed) Genome mapping in plants. RG Landes Company, Austin, pp 85–93

Buckler ES, Holland JB, Bradbury PJ et al (2009) The genetic architecture of maize flowering time. Science 325:714–718

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Burns MJ, Barnes SR, Bowman JG et al (2003) QTL analysis of an intervarietal set of substitution lines of *Brassica napus*: (i) Seed oil content and fatty acid composition. Heredity 90:39–48

Burr B, Burr FA (1991) Recombinant inbreds for molecular mapping in maize: theoretical and practical considerations. Trends Genet 7:55–60

Burr B, Burr FA, Thompson KH et al (1988) Gene mapping with recombinant inbreds in maize. Genetics 118:519–526

Busemeyer L, Ruckelshausen A, Moller K et al (2013) Precision phenotyping of biomass accumulation in triticale reveals temporal genetic patterns of regulation. Sci Rep 3:2442. doi:10.1038/srep02442

Button P (2006) New developments in the International Union for the protection of new varieties of plants. Acta Hortic 714:195–210

Caetano-Anolles G, Bassam BJ, Greshoff PM (1991) DNA amplification fingerprinting using very short arbitrary oligonucleotide primers. BioTechnology 9:553–557

Caldeira RL, Carvalho OS, Lage RCG et al (2002) Sequencing of simple sequence repeat anchored polymerase chain reaction amplification products of *Biomphalaria glabrata*. Mem Inst Oswaldo Cruz, Rio de Janeiro 97:23–26

Carollo V, Matthews DE, Lazo GR et al (2005) GrainGenes 2.0. An improved resource for the small-grains community. Plant Physiol 139:643–651

Castiglioni P, Pozzi C, Heun M et al (1998) An AFLP-based procedure for the efficient mapping of mutations and DNA probes in barley. Genetics 149:2039–2056

Castro AJ, Capettini F, Corey AE et al (2003) Mapping and pyramiding of qualitative and quantitative resistance to stripe rust in barley. Theor Appl Genet 107:922–930

Catchen J, Hohenlohe PA, Bassham S et al (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140. doi:10.1111/mec.12354

Cavanagh C, Morell M, Mackay I et al (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol 11:215–221

Chaerle L, Van Der Straeten D (2001) Seeing is believing: imaging techniques to monitor plant health. Biochim Biophys Acta 1519:153–166

Chaerle L, Lenk S, Leinonen L et al (2009) Multi-sensor plant imaging: towards the development of a stress-catalogue. Biotechnol J 4:1152–1167

Charmet G, Robert N, Perretant et al (1999) Marker-assisted recurrent selection for cumulating additive and interactive QTLs in recombinant inbred lines. Theor Appl Genet 99:1143–1148

Chen XM, Line RF, Leung H (1998) Genome scanning for resistance gene analogs in rice, barley and wheat by high-resolution electrophoresis. Theor Appl Genet 97:345–355

Chenna R, Sugawara H, Koike T et al (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31:3497–3500. doi:10.1093/nar/gkg500

Chepelev I, Wei G, Tang Q et al (2009) Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Res 37:e106

Choi HK, Kim D, Uhm T et al (2004) A sequence-based genetic map of *Medicago truncatula* and comparison of marker colinearity with *M. sativa*. Genetics 166:1463–1502

Choumane W, Winter P, Weigand F et al (2000) Conservation and variability of sequence tagged microsatellite

sites (STMSs) from chickpea (*Cicer arietinum* L.) within the genus *Cicer*. Theor Appl Genet 101:269–278
Chudyk JP, Rusch TL, Fieweger KM et al (2006) Automating microsatellite genotyping with array tape. J Assoc Lab Autom 11:260–267
Churchill GA, Deorge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971
Churchill GA, Giovannoni JJ, Tanksley SD (1993) Pooled-sampling makes high-resolution mapping practical with DNA markers. Proc Natl Acad Sci U S A 90:16–20
Clark RT, MacCurdy RB, Jung JK et al (2011) Three-dimensional root phenotyping with a novel imaging and software platform. Plant Physiol 156:455–465
Clark RT, Famoso AN, Zhao K et al (2013) High-throughput two-dimensional root system phenotyping platform facilitates genetic analysis of root growth and development. Plant Cell Environ 36:454–466
Cobb JN, DeClerck G, Greenberg A et al (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–-phenotype relationships and its relevance to crop improvement. Theor Appl Genet 126:867–887
Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans R Soc Lond B Biol Sci 363:557–572
Collard BCY, Mackill DJ (2009a) Start codon targeted (SCoT) polymorphism: a simple, novel DNA marker technique for generating gene-targeted markers in plants. Plant Mol Biol Rep 27:86–93
Collard BCY, Mackill DJ (2009b) Conserved DNA-derived polymorphism (CDDP): a simple and novel method for generating DNA markers in plants. Plant Mol Biol Rep 27:558–562
Collard BCY, Jahufer MZZ, Brouwer JB et al (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. Euphytica 142:169–196
Concibido VC, Young ND, Lang DA et al (1996) Targeted comparative genome analysis and qualitative mapping of a major partial-resistance gene to the soybean cyst nematode. Theor Appl Genet 93:234–241
Crossa J, Campos G, Perez P et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724
Crossa J, Perez P, Campos G et al (2011) Genomic selection and prediction in plant breeding. J Crop Improv 25:239–261
Darvasi A, Soller M (1995) Advanced intercross lines, an experimental population for fine gene mapping. Genetics 141:1199–1207
Datta K, Baisakh N, Maung Thet K et al (2002) Pyramiding transgenes for multiple resistance in rice against bacterial blight, yellow stem borer and sheath blight. Theor Appl Genet 106:1–8
Davey JW, Blaxter ML (2010) RADSeq: next-generation population genetics. Brief Funct Genom 9:416–423
Davey JW, Hohenlohe PA, Etter PD et al (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510
de los Campos G, Perez P (2010) BLR: Bayesian linear regression. R Package Version 1.2. http://cran.r-project.org/web/packages/BLR/index.html
de Vienne D (ed) (2003) Molecular markers in plant genetics and biotechnology. Science Publishers, Enfield
de Vienne D, Causse M (2003) Mapping and characterising quantitative trait loci. In: de Vienne D (ed) Molecular markers in plant genetics and biotechnology. Science Publishers, Enfield, pp 89–124
de Vienne D, Santoni S, Falque M (2003) Principal sources of molecular markers. In: de Vienne D (ed) Molecular markers in plant genetics and biotechnology. Science Publishers, Enfield, pp 3–46
De Vylder J, Vandenbussche F, Hu Y et al (2012) Rosette Tracker: an open source image analysis tool for automatic quantification of genotype effects. Plant Physiol 160:1149–1159
Dekkers JCM (2007) Prediction of response to marker-assisted genomic selection using selection index theory. J Anim Breed Genet 124:331–341
Deschamps S, Campbell MA (2010) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. Mol Breed 25:553–570
Deschamps S, Campbell MA (2013) Genetic variant discovery and its use in genome characterization of ergonomically important crop species. In: Henry RJ (ed) Molecular markers in plants. Wiley, Hoboken, pp 138–167
Deshmukh R, Singh A, Jain N et al (2010) Identification of candidate genes for grain number in rice (*Oryza sativa L.*). Funct Integr Genomics 10:339–347
Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322
Duran C, Appleby N, Clark T et al (2009) AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. Nucleic Acids Res 37(Database issue):D951–D953. doi:10.1093/nar/gkn650
Eathington SR, Crosbie TM, Edwards MD et al (2007) Molecular markers in a commercial breeding programme. Crop Sci 47:S154–S163
Echt C, Knapp S, Liu B-H (1992) Genome mapping with non-inbred crosses using Gmendel 2.0. Maize Genet Coop Newslett 66:27–29
Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210
Edwards KJ (1998) Randomly amplified polymorphic DNAs (RAPDs). In: Karp A, Isaac PG, Ingram DS (eds) Molecular tools for screening biodiversity. Chapman and Hall, London, pp 171–175
Edwards M (2013) Whole-genome sequencing for marker discovery. In: Henry RJ (ed) Molecular markers in plants. John Wiley & Sons, Inc., Ames, Iowa, USA, pp 21–34
Edwards MD, Gifford DK (2012) High-resolution genetic mapping with pooled sequencing. BMC

Bioinformatics 13(Suppl 6):S8. http://www.biomedcentral.com/1471-2105/13/S6/S8
Edwards MD, Stuber CW, Wendel JF (1987) Molecular-marker-facilitated investigations of quantitative trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics 116:113–125
Ellis THN, Poyser SJ, Knox MR et al (1998) Polymorphism of insertion sites of *Ty1-copia* class retrotransposons and its use for linkage and diversity analysis in pea. Mol Gen Genet 260:9–19
Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379
Eshed Y, Zamir D (1994) A genomic library of *Lycopersicon pennellii* in *L. esculentum*: a tool for fine mapping of genes. Euphytica 79:175–179
Eshed Y, Zamir D (1995) An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics 141:1147–1162
Fan J-B, Gunderson KL, Bibikova M et al (2006) Illumina universal bead arrays. Methods Enzymol 410:57–73
Fang M (2012) A fast expectation-maximum algorithm for fine-scale QTL mapping. Theor Appl Genet 125:1727–1734
Fiorani F, Schurr U (2013) Future scenarios for plant phenotyping. Annu Rev Plant Biol 64:267–291
Fischer RA, Byerlee D, Edmeades GO (2014) Crop yields and global food security: will yield increase continue to feed the world? ACIAR Monograph No. 158. Australian Centre for International Agricultural Research, Canberra
Flint-Garcia SA (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374
Frisch M, Melchinger AE (2005) Selection theory for marker-assisted backcrossing. Genetics 170:909–917
Frisch M, Bohn M, Melchinger AE (1999) Comparison of selection strategies for marker-assisted backcrossing of a gene. Crop Sci 39:1295–1301
Fulton TM, Van der Hoeven R, Eannetta NT et al (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14:1457–1467
Furbank RT, Tester M (2011) Phenomics – technologies to relieve the phenotyping bottleneck. Trends Plant Sci 16:1–10
Gardiner JM, Coe EH, Melia-Hancock S et al (1993) Development of a core RFLP map in maize using an immortalized F_2 population. Genetics 154:917–930
Gaut BS, Long AD (2003) The lowdown on linkage disequilibrium. Plant Cell 15:1502–1506
Gebhardt C, Ritter E, Debener T et al (1989) RFLP mapping and linkage analysis in *Solanum tuberosum*. Theor Appl Genet 78:65–75
Gebhardt C, Ritter E, Barone A et al (1991) RFLP maps of potato and their alignment with the homoeologous tomato genome. Theor Appl Genet 83:49–57
Geering ADW (2013) Molecular markers for plant biosecurity. In: Henry RJ (ed) Molecular markers in plants. Wiley, Hoboken, pp 99–117
Geiger HH, Gordillo GA (2009) Doubled haploids in hybrid maize breeding. Maydica 54:485–499
Geiringer H (1944) On the probability theory of linkage in Mendelian heredity. Ann Math Stat 15:25–57
Gerlai R (2002) Phenomics: fiction or the future? Trends Neurosci 25:506–509
Gianola D, Fernando R, Stella A (2006) Genomic-assisted prediction of genetic values with semiparametric procedures. Genetics 173:1761–1776
Gilbert DG (1985a) Estimating single gene effects on quantitative traits. 1. A diallel method applied to *Est6* in *D. melanogaster*. Theor Appl Genet 69:625–629
Gilbert DG (1985b) Estimating single gene effects on quantitative traits. 2. Statistical properties of five experimental methods. Theor Appl Genet 69:631–636
Gitelson AA, Buschmann C, Lichtenthaler HK (1999) The chlorophyll fluorescence ratio F735/F700 as an accurate measure of chlorophyll content in plants. Remote Sens Environ 69:296–302
Glaubitz JC, Casstevens TM, Lu F et al (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9:e90346. doi:10.1371/journal.pone.0090346
Goddard M (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245–257
Goff SA (2011) A unifying theory for general multigenic heterosis: energy efficiency, protein metabolism, and implications for molecular breeding. New Phytol 189:923–937
Goffinet B, Gerber S (2000) Quantitative trait loci: a meta-analysis. Genetics 155:463–473
Gonzalez-Camacho JM, de los Campos G, Perez P et al (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theor Appl Genet 125:759–771
Goodstein DM, Shu S, Howson R et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40(Database issue):D1178–D1186
Gopalakrishnan S, Sharma RK, Anand Rajkumar K et al (2008) Integrating marker assisted background analysis with foreground selection for identification of superior bacterial blight resistant recombinants in Basmati rice. Plant Breed 127:131–139
Gorelick R, Laubichler MD (2004) Decomposing multilocus linkage disequilibrium. Genetics 166:1581–1583
Gowen JW (ed) (1952) Heterosis. Iowa State College, Ames
Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–874
Gregory PJ, Bengough AG, Grinev D et al (2009) Root phenomics of crops: opportunities and challenges. Funct Plant Biol 36:922–929
Gresshoff PM (2005) Positional cloning of plant developmental genes. In: Meksem K, Kahl G (eds) The handbook of plant genome mapping. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, pp 233–256
Grodzicker T, Williams J, Sharp P et al (1974) Physical mapping of temperature-sensitive mutations of

adenoviruses. Cold Spring Harb Symp Quant Biol 39:439–446
Guo Z, Tucker DM, Lu J et al (2012) Evaluation of genome-wide selection efficiency in maize nested association mapping populations. Theor Appl Genet 124:261–275
Gupta PK, Varshney RK (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. Euphytica 113:163–185
Gupta HS, Babu R, Agrawal PK et al (2013) Accelerated development of quality protein maize hybrid through marker-assisted introgression of opaque-2 allele. Plant Breed 132:77–82
Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr Sci 80:524–535
Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol 57:461–485
Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. Heredity 101:5–18
Gupta PK, Kumar J, Mir RR (2010) Marker-assisted selection as a component of conventional plant breeding. Plant Breed Rev 33:145–217
Gupta PK, Kulwal PL, Jaiswal V (2014) Association mapping in crop plants: opportunities and challenges. In: Friedmann T, Dunlap J, Goodwin S (eds) Advances in genetics, vol 85. Academic, Elsevier, Waltham, MA, USA, pp 109–148
Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome assisted breeding values. Genetics 177:2389–2397
Hackett CA, Milne I, Bradshaw JE et al (2007) TetraploidMap for Windows: linkage map construction and QTL mapping in autotetraploid species. J Hered 98:727–729
Haldane JBS (1919) The combination of linkage values and the calculation of the distance between the loci of linked factors. J Genet 8:299–309
Haldane JBS, Waddington C (1931) Inbreeding and linkage. Genetics 16:357–374
Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324
Hanson WD (1959) Early generation analysis of lengths of heterozygous chromosome segments around a locus held heterozygous with backcrossing or selfing. Genetics 44:833–837
Hartmann A, Czauderna T, Hoffmann R et al (2011) HTPheno: an image analysis pipeline for high-throughput plant phenotyping. BMC Bioinformatics 12:148
Hashimoto K, Goto S, Kawano S et al (2006) KEGG as a glycome informatics resource. Glycobiology 16:63R–70R
Hass-Jacobus B, Jackson SA (2005) Physical mapping of plant chromosomes. In: Meksem K, Kahl G (eds) The handbook of plant genome mapping, genetic and physical mapping. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, pp 133–149
Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. Genetics 117:331–341
Heffner EL, Sorrels ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci 49:1–12
Heffner EL, Lorenz AJ, Jannink J-L et al (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop Sci 50:1681–1690
Helentjaris TG (1992) RFLP analysis for manipulating agronomic traits in plants. In: Stalker HT, Murphy JP (eds) Plant breeding in the 1990s. CAB International, Wallingford, pp 357–372
Hill JT, Demarest BL, Bisgrove BW et al (2013a) MMAPPR: mutation mapping analysis pipeline for pooled RNA-seq. Genome Res 23:687–697
Hill TA, Ashrafi H, Reyes-Chin-Wo S et al (2013b) Characterization of *Capsicum annuum* genetic diversity and population structure based on parallel polymorphism discovery with a 30K Unigene pepper GeneChip. PLoS One 8:e56200. doi:10.1371/journal.pone.0056200
Hill-Ambroz KL, Brown-Guedira GL, Fellers JP (2002) Modified rapid DNA extraction protocol for high throughput microsatellite analysis in wheat. Crop Sci 42:2088–2091
Hirosawa M, Ishikawa K, Nagase T et al (2000) Detection of spurious interruptions of protein-coding regions in cloned cDNA sequences by GeneMark analysis. Genome Res 10:1333–1341
Hoffmann WA, Poorter H (2002) Avoiding bias in calculations of relative growth rate. Ann Bot 90:37–42
Hoffmann TJ, Kvale MN, Hesselson SE et al (2011) Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics 98:79–89
Hosoi F, Nakabayashi K, Omasa K (2011) 3-D modelling of tomato canopies using a high-resolution portable scanning Lidar for extracting structural information. Sensors 11:2166–2174
Hospital F (2005) Selection in backcross programmes. Philos Trans R Soc Lond Biol Sci 360:1503–1511
Hospital F, Charcosset A (1997) Marker-assisted introgression of quantitative trait loci. Genetics 147:1469–1485
Houle D (2010) Numbering the hairs on our heads: the shared challenge and promise of phenomics. Proc Natl Acad Sci U S A 107(Suppl 1):1793–1799
Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. Nat Rev Genet 11:855–866
Howes NK, Woods SM, Townley-Smith TF (1998) Simulations and practical problems of applying multiple marker assisted selection and doubled haploids to wheat breeding programs. Euphytica 100:225–230
Hu J, Vick BA (2003) Target region amplification polymorphism: a novel marker technique for plant genotyping. Plant Mol Biol Rep 21:289–294

Hua J, Xing Y, Xu C (2002) Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance. Genetics 162:1885–1895

Hua J, Xing Y, Wu W et al (2003) Single locus heterotic effects and dominance-by-dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. Proc Natl Acad Sci U S A 100:2574–2579

Huang N, Angeles ER, Domingo J et al (1997) Pyramiding of bacterial blight resistance genes in rice: marker-assisted selection using RFLP and PCR. Theor Appl Genet 95:313–320

Huang BE, Verbyla KL, Verbyla AP et al (2015) MAGIC populations in crops: current status and future prospects. Theor Appl Genet DOI 10.1007/s00122-015-2506-0

Ingvarsson PK, Street NR (2011) Association genetics of complex traits in plants. New Phytol 189:909–922

Ishii T, Yonezawa K (2007) Optimization of the marker-based procedures for pyramiding genes from multiple donor lines: I. Schedule of crossing between the donor lines. Crop Sci 47:537–546

Iwata H, Ninomiya S (2006) AntMap: constructing genetic linkage maps using an ant colony optimization algorithm. Breed Sci 56:371–377

Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bull Soc Vaudoise Sci Natl 44:223–270

Jaccoud D, Peng K, Feinstein D et al (2001) Diversity arrays: a solid state technology for sequence independent genotyping. Nucleic Acids Res 29:e25

Jackson SA, Hass-Jacobus B, Pagel J (2004) The gene space of the soybean genome. In: Wilson RF, Stalker HT, Brummer EC (eds) Legume crop genomics. AOCS Press, Champaign, pp 187–193

Jahnke S, Menzel MI, van Dusschoten D et al (2009) Combined MRI-PET dissects dynamic changes in plant structures and functions. Plant J 59:634–644

Jannink J-L, Jansen RC (2001) Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics 157:445–454

Jannink J-L, Wu X-L (2003) Estimating allelic number and identity in state of QTL in interconnected families. Genet Res 81:133–144

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9:166–177

Jansen RC (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447–1455

Jansen RC, Jannink J-L, Beavis WD (2003) Mapping quantitative trait loci in plant breeding populations: use of parental haplotype sharing. Crop Sci 43:829–834

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314:67–73

Jena KK, Mackill DJ (2008) Molecular markers and their use in marker-assisted selection in rice. Crop Sci 48:1266–1276

Jenkins H, Hardy N, Beckmann M et al (2004) A proposed framework for the description of plant metabolomics experiments and their results. Nat Biotechnol 22:1601–1606

Ji H, Welch K (2009) Molecular inversion probe assay for allelic quantitation. Methods Mol Biol 556:67–87

Jiang G-L (2013) Molecular markers and marker-assisted breeding in plants. InTech http://dx.doi. org.10.5772/52583

Jiang C, Zeng Z-B (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 140:1111–1127

Joehanes R, Nelson JC (2008) QGene 4.0, an extensible Java QTL-analysis platform. Bioinformatics 24:2788–2789

Johannes F, Porcher E, Teixeira FK et al (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. PLoS Genet 5:e1000530

Jones CJ, Edwards KJ, Castaglione S et al (1997) Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Mol Breed 3:381–390

Joseph M, Gopalakrishnan S, Sharma RK et al (2004) Combining bacterial blight resistance and Basmati quality characteristics by phenotypic and marker-assisted selection in rice. Mol Breed 13:377–387

Jourjon M-F, Jasson S, Marcel J et al (2005) MCQTL: multi-allelic QTL mapping in multi-cross design. Bioinformatics 21:128–130

Kaeppler S (2012) Heterosis: many genes, many mechanisms—end the search for an undiscovered unifying theory. ISRN Botany 12 pages. doi:10.5402/2012/682824

Kaeppler SM, Philips RL, Kim TS (1993) Use of near-isogenic lines derived by backcrossing or selfing to map quantitative traits. Theor Appl Genet 87:233–237

Kahl G, Mast A, Tooke N et al (2005) Single nucleotide polymorphisms: detection techniques and their potential for genotyping and genome mapping. In: Meksem K, Kahl G (eds) The handbook of plant genome mapping, genetic and physical mapping. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, pp 75–107

Kanehisa M, Goto S, Kawashima S et al (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32(Database issue):D277–D280

Kang HM, Zaitlen NA, Wade CM et al (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723

Kanz C, Aldebert P, Althorpe N et al (2005) The EMBL nucleotide sequence database. Nucleic Acids Res 33 (Database issue):D29–D33

Kao CH, Zeng Z-B, Teasdale RD (1999) Multiple interval mapping. Genetics 152:1203–1216

Khanna A, Sharma V, Ellur RK et al (2015) Development and evaluation of near-isogenic lines for major blast resistance gene(s) in Basmati rice. Theor Appl Genet. doi:10.1007/s00122-015-2502-4

Kofler R, Orozco-ter Wengel P, De Maio N et al (2011) PoPoolation: a toolbox for population genetic analysis

of next generation sequencing data from pooled individuals. PLoS One 6:e15925. doi:10.1371/journal.pone.0015925
Komori T, Ohta S, Murai N et al (2004) Map-based cloning of a fertility restorer gene, *Rf-1*, in rice (*Oryza sativa* L). Plant J 37:315–325
Konieczny A, Ausubel FM (1993) A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. Plant J 4:403–410
Koornneef M, Vanderveen HJ (1983) A marker line that allows the detection of linkage on all Arabidopsis chromosomes. Genetica 61:41–46
Korol A, Frenkel Z, Cohen L et al (2007) Fractioned DNA pooling: a new cost-effective strategy for fine mapping of quantitative trait loci. Genetics 176:2611–2623
Korte A, Vilhjalmsson BJ, Segura V et al (2012) A mixed model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 44:1066–1071
Kosambi DD (1944) The estimation of map distances from recombination values. Ann Eugen 12:172–175
Kover PX, Valdar W, Trakalo J et al (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. PLoS Genet 5: e1000551
Krishnan GS, Singh AK, Waters DLE et al (2013) Molecular markers for harnessing heterosis. In: Henry RJ (ed) Molecular markers in plants. Wiley, Hoboken, pp 119–136
Kuchel H, Ye G, Fox R et al (2005) Genetic and economic analysis of targeted marker-assisted wheat breeding strategy. Mol Breed 16:67–78
Kuchel H, Fox R, Reinheimer J et al (2007) The successful application of marker-assisted wheat breeding strategy. Mol Breed 20:295–308
Kumar M, Sharma CM, Rajwar GS (2004) A study on the community structure and diversity of a sub-tropical forest of Garhwal Himalayas. Indian Forester 130:207–214
Kump KL, Bradbury PJ, Wisser RJ et al (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat Genet 43:163–168
Kuroshu RM, Watanabe J, Sugano S et al (2010) Cost-effective sequencing of full-length cDNA clones powered by a de novo-reference hybrid assembly. PLoS One 5:e10517. doi:10.1371/journal.pone.0010517
Labate JA, Lamkey KR, Lee M et al (2000) Hardy-Weinberg and linkage equilibrium estimates in the BSSS and BSCB1 random mated populations. Maydica 45:243–255
Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756
Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199
Lander ES, Green P, Abrahamson J et al (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174–181
Landergren U, Kaiser R, Sanders J et al (1988) A ligase-mediated gene detection technique. Science 241:1077–1080
Larkin MA, Blackshields G, Brown NP et al (2007) ClustalW and ClustalX version 2. Bioinformatics 23:2947–2948. doi:10.1093/bioinformatics/btm404
Law JR, Donini P, Koebner RMD et al (1998) DNA profiling and plant variety registration. III: the statistical assessment of distinctness in wheat using amplified fragment length polymorphisms. Euphytica 102:335–342
Lawrence CJ, Seigfried TE, Brendel V (2005) The Maize Genetics and Genomics Database. The community resource for access to diverse maize data. Plant Physiol 138:55–58
Lawson DM, Lunde CF, Mutschler MA (1997) Marker-assisted transfer of acylsugar-mediated resistance from the wild tomato, *Lycopersicon pennellii*, to the cultivated tomato, *Lycopersicon esculentum*. Mol Breed 3:307–317
Lecomte L, Duffe P, Buret M et al (2004) Marker-assisted introgression of five QTLs controlling fruit quality traits into three tomato lines revealed interactions between QTLs and genetic backgrounds. Theor Appl Genet 109:658–668
Lee SH, van der Werf JHJ (2007) Fine mapping of multiple interacting quantitative trait loci using combined linkage disequilibrium and linkage information. J Zhejiang Univ Sci B 8:787–791
Lee M, Godshalk EB, Lamke KR et al (1989) Association of restriction fragment length polymorphism among maize inbreds with agronomic performance of their crosses. Crop Sci 29:1067–1071
Lee EA, Ash MJ, Good B (2006) Re-examining the relationship between degree of relatedness, genetic effects, and heterosis in maize. Crop Sci 47:629–635
Lefebvre V, Palloix A, Caranata C et al (1995) Construction of an intraspecific linkage map of pepper using molecular markers and doubled-haploid progenies. Genome 38:112–121
Leitner D, Felderer B, Vontobel P et al (2014) Recovering root system traits using image analysis exemplified by two-dimensional neutron radiography images of lupine. Plant Physiol 164:24–35
Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67
Lewontin RC (1974) The genetic basis of evolutionary change. Columbia University, New York
Li G, Quiros CF (2001) Sequence-related amplified polymorphism (SRAP), a new marker system based on a simple PCR reaction: its application to mapping and gene tagging in *Brassica*. Theor Appl Genet 103:455–461
Li H, Ye G, Wang J (2007a) A modified algorithm for the improvement of composite interval mapping. Genetics 175:361–374
Li Y, Li Y, Wu S et al (2007b) Estimation of multilocus linkage disequilibria in diploid populations with dominant markers. Genetics 176:1811–1821

Li H, Bradbury P, Ersoz E et al (2011a) Joint QTL linkage mapping for multiple-cross mating design sharing one common parent. PLoS One 6:e17573. doi:10.1371/journal.pone.0017573

Li Y, Sidore C, Kang HM et al (2011b) Low-coverage sequencing: implications for design of complex trait association studies. Genome Res 21:94–951

Lin CH, Yeakley JM, McDaniel TK et al (2009) Medium- to high-throughput SNP genotyping using veracode microbeads. Methods Mol Biol 496:129–142

Lincoln SE, Daly MJ, Lander ES (1993) Constructing genetic linkage maps with MAPMAKER EXP V3.0: a tutorial and reference manual. A Whitehead Institute for Biomedical research Technical Report. http://www.mapmaker@genome.wi.mit.edu

Ling Q, Huang W, Jarvis P (2011) Use of a SPAD-502 meter to measure leaf chlorophyll concentration in *Arabidopsis thaliana*. Photosynth Res 107:209–214

Lippert C, Listgarten J, Liu Y et al (2011) FaST linear mixed models for genome-wide association studies. Nat Methods 8:833–835

Lippman ZB, Zamir D (2007) Heterosis: revisiting the magic. Trends Genet 23:60–66

Lister C, Dean C (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant J 4:745–750

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397–401

Liu B-H (1998) Statistical genomics: linkage mapping and QTL analysis. CRC Press, Boca Raton

Liu P, Zhu J, Lu Y (2004) Marker-assisted selection in segregating generations of self-fertilizing crops. Theor Appl Genet 109:370–376

Liu S, Yeh C-T, Tang HM et al (2012) Gene mapping via bulked segregant RNA-Seq (BSR-Seq). PLoS One 7: e36406. doi:10.1371/ journal.pone.0036406

Livak KJ (1999) Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. Genet Anal 14:143–149

Lobet G, Pagès L, Draye X (2011) A novel image-analysis toolbox enabling quantitative analysis of root system architecture. Plant Physiol 157:29–39

Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120:151–161

Lumme J, Karjalainen M, Kaartinen H et al (2008) Terrestrial laser scanning of agricultural crops. Int Arch Photogram Remote Sens Spat Inform Sci (Beijing) 37B:563–566

Luo ZW, Wu C-I, Kearsey MJ (2002) Precision and high-resolution mapping of quantitative trait loci by use of recurrent selection, backcross or intercross schemes. Genetics 161:915–929

Luo ZW, Zhang RM, Kearsey MJ (2004) Theoretical basis for genetic linkage analysis in autotetraploid species. Proc Natl Acad Sci U S A 101:7040–7045

Luo X, Wu S, Tian et al (2011) Identification of heterotic loci associated with yield-related traits in Chinese common wild rice (*Oryza rufipogon* Griff.). Plant Sci 181:14–22

Lusser M, Raney T, Tillie P et al (2012a) International workshop on socio-economic impacts of genetically modified crops co-organised by JRC-IPTS and FAO, Workshop proceedings. Food and Agriculture Organization of the United Nations (FAO), Rome

Lusser M, Parisi C, Plan D et al (2012b) Deployment of new biotechnologies in plant breeding. Nat Biotechnol 30:231–239

Lyamichev V, Mast AL, Hall JG et al (1999) Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. Nat Biotechnol 17:292–296

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland

Lyttle TW (1991) Segregation distorters. Annu Rev Genet 25:511–557

Mackay L, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends Plant Sci 12:57–63

Mackill DJ, Junjian N (2001) Molecular mapping and marker-assisted selection for major gene traits in rice. In: Khush GS, Brar DS, Hardy B (eds) Rice genetics IV. Proc Fourth Rice Genetics Symposium, 22–27 Oct 2000. Science Publishers, Enfield and International Rice Research Institute, Los Banos, pp 137–151

Magrane M, UniProt Consortium (2011) UniProt knowledgebase: a hub of integrated protein data. Database vol 2011, Article ID bar009. doi:10.1093/database/bar009

Mammadov J, Aggarwal R, Buyyarapu R et al (2012) SNP markers and their impact on plant breeding. Int J Plant Genomics vol 2012. doi.org/10.1155/2012/728398, 11 pages

Maniatis N, Collins A, Xu C-F et al (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci U S A 99:2228–2233

Manly KF, Olson JM (1999) Overview of QTL mapping software and introduction to map manager QT. Mamm Genome 10:327–334

Manly KF, Cudmore JRH, Meer JM (2001) Map Manager QTX, cross-platform software for genetic mapping. Mamm Genome 12:930–932

Mather KA, Caicedo AL, Polato NR et al (2007) The extent of linkage disequilibrium in rice (*Oryza sativa* L.). Genetics 177:2223–2232

Matise TC, Perlin M, Chakravarti A (1993) MultiMap: an expert system for automated genetic linkage mapping. Proc Int Conf Intell Syst Mol Biol 1:260–265

Maurer HP, Melchinger AE, Frisch M (2004) Plabsoft: software for simulation and data analysis in plant breeding. Proc. 17th EUCARPIA General Congress, 8–11 Sept 2004, Tullu, pp 280–284

Maxwell K, Johnson GN (2000) Chlorophyll fluorescence – a practical guide. J Exp Bot 51:659–668

McGregor CE, Lambert CA, Greyling MM et al (2000) A comparative assessment of DNA fingerprinting

techniques (RAPD, ISSR, AFLP and SSR) in tetraploid potato (*Solanum tuberosum* L.) germplasm. Euphytica 113:135–144

Meksem K, Kahl G (2005) The handbook of plant genome mapping, genetic and physical mapping. WILEY-VCH Verlag GmbH & Co, KGaA, Weinheim

Melchinger AE (1990) Use of molecular markers in breeding for oligogenic disease resistance. Plant Breed 104:1–19

Melchinger AE (1999) Genetic diversity and heterosis. In: Coors JG, Pandey S (eds) The genetics and exploitation of heterosis in crop plants. ASA-CSSA, Madison, pp 99–118

Melchinger AE, Lee M, Lamkey KR et al (1990) Genetic diversity for restriction fragment length polymorphism and heterosis for two diallel sets of maize inbreds. Theor Appl Genet 80:488–496

Melchinger AE, Utz HF, Schon CC (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. Genetics 149:383–403

Meuwissen THE (2009) Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Select Evol 41:35. doi:10.1186/1297-9686-41-35

Meuwissen THE, Goddard ME (2001) Prediction of identity by descent probabilities from marker haplotypes. Genet Sel Evol 33:605–634

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Meyer K, Benning G, Grill E (1996) Cloning plant genes based on genetic map location. In: Paterson AH (ed) Genome mapping in plants. R.G. Landes Co., Austin, pp 137–154

Michaels SD, Amasino RM (1998) A robust method for detecting single-nucleotide changes as polymorphic markers by PCR. Plant J 14:381–385

Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to disease resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci U S A 88:9829–9832

Mielewczik M, Friedli M, Kirchgessner N et al (2013) Diel leaf growth of soybean: a novel method to analyze two-dimensional leaf expansion in high temporal resolution based on a marker tracking approach (Martrack Leaf). Plant Methods 9:30–42

Miller JC, Tanksley SD (1990) RFLP analysis of phylogenetic relationships and genetic variation in the genus Lycopersicon. Theor Appl Genet 80:437–448

Miller MR, Dunham JP, Amores A et al (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17:240–248

Min J, Chunyu Z, Khalid H et al (2012) Pyramiding resistance genes to northern leaf blight and head smut in maize. Int J Agric Biol 14:430–434

Mir RR, Varshney RK (2013) Future prospects of molecular markers in plants. In: Henry RJ (ed) Molecular markers in plants. Wiley, Hoboken, pp 169–190

Mitchelle SE, Kresovich S, Jester CE et al (1997) Application of multiplex PCR and fluorescence-based, semiautomated allele sizing technology for genotyping plant genetic resources. Crop Sci 37:617–624

Mohammadi SA, Prasanna BM (2003) Analysis of genetic diversity in crop plants – salient statistical tools and considerations. Crop Sci 43:1235–1248

Moll RH, Lonnquist JH, Fortuna JV et al (1965) The relation of heterosis and genetic divergence in maize. Genetics 52:139–144

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318

Muehlbauer GJ, Specht JE, Thomas-Compton MA et al (1988) Near isogenic lines – a potential resource in the integration of conventional and marker linkage maps. Crop Sci 28:729–735

Muller-Starck G (1998) Isozymes. In: Karp A, Isaac PG, Ingram DS (eds) Molecular tools for screening biodiversity. Chapman and Hall, London, pp 75–81

Mullis KB (1990) The unusual origin of the polymerase chain reaction. Sci Am 262:36–43

Munns R, James RA, Sirault XRR et al (2010) New phenotyping methods for screening wheat and barley for beneficial responses to water deficit. J Exp Bot 61:3499–3507. doi:10.1093/jxb/erq199

Murray HG, Thompson WF (1980) Rapid isolation of high molecular weight DNA. Nucleic Acids Res 8:4321–4325

Muthusamy V, Hossain F, Thirunavukkarasu N et al (2014) Development of β-carotene rich maize hybrids through marker-assisted introgression of β-carotene hydroxylase allele. PLOS One 9:e113583. doi:10.1371/journal.pone.0113583

Myles S, Peiffer J, Brown PJ et al (2009) Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 21:2194–2202

Nagel KA, Putz A, Gilmer F et al (2012) GROWSCREEN-Rhizo is a novel phenotyping robot enabling simultaneous measurements of root and shoot growth for plants grown in soil-filled rhizotrons. Funct Plant Biol 39:891–904

Nakaya A, Isobe SN (2012) Will genomic selection be a practical method for plant breeding? Ann Bot. doi:10.1093/aob/mcs109

Neeraja C, Maghirang-Rodriguez R, Pamplona A et al (2007) A marker-assisted backcross approach for developing submergence-tolerance rice cultivars. Theor Appl Genet 115:767–776

Nei M, Li W (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A 76:5269–5273

Nelson RR (1978) Genetics of horizontal resistance to plant diseases. Annu Rev Phytopathol 16:359–378

Nielsen R, Paul JS, Albrechtsen A et al (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443–451

Normanly J (ed) (2012) High-throughput phenotyping in plants, methods and protocols. Humana Press/ Springer, New York

Okayama H, Curiel DT, Brantly ML et al (1989) Rapid, nonradioactive detection of mutations in the human genome by allele-specific amplification. J Lab Clin Med 114:105–113

Ollinger SV (2011) Sources of variability in canopy reflectance and the convergent properties of plants. New Phytol 189:375–394

Olson ML, Hood L, Cantor C et al (1989) A common language for physical mapping of the human genome. Science 245:1434–1435

Oraguzie NC, Rikkerink EHA, Gardiner SE et al (2007) Association mapping in plants. Springer Science +Business Media, LLC, New York

Orita M, Iwahana H, Kanazawa H et al (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc Natl Acad Sci U S A 86:2766–2770

Ozsolak F, Milos PM (2011) Transcriptome profiling using single-molecule direct RNA sequencing. Methods Mol Biol 733:51–61

Palaisa KA, Morgante M, Williams M et al (2003) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell 15:1795–1806

Panaud O, Chen X, McCouch SR (1996) Development of microsatellite markers and characterization of simple sequence length polymorphism (SSLP) in rice (*Oryza sativa* L.). Mol Gen Genet 252:597–607

Pandey V, Nutter RC, Prediger E (2008) Applied Biosystems SOLiD™ system: ligation-based sequencing. In: Janitz M (ed) Next generation genome sequencing: towards personalized medicine. Wiley-VCH, Weinheim, pp 29–42

Pandit A, Rai V, Bal S et al (2010) Combining QTL mapping and transcriptome profiling of bulked RILs for identification of functional polymorphism for salt tolerance genes in rice (*Oryza sativa* L.). Mol Genet Genomics 284:121–136

Panguluri SK, Kumar AA (eds) (2013) Phenotyping for plant breeding: applications of phenotyping methods for crop improvement. Springer, New York

Paran I, Michelmore RW (1993) Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. Theor Appl Genet 85:985–993

Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc 103:681–686

Parkinson H, Sarkans U, Shojatalab M et al (2005) ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 33(Database issue):D553–D555

Paterson AH, Lander ES, Hewitt JD et al (1988) Resolution of quantitative traits into Mendelian factors using a complete linkage map of restriction fragment length polymorphism. Nature 335:721–726

Paterson AH, Damon S, Hewitt JD (1991) Mendelian factors underlying quantitative traits in tomato: comparison across species, generations and environments. Genetics 127:181–197

Peleman JD, van der Voort JR (2003) Breeding by design. Trends Plant Sci 8:330–334

Peleman JD, Sorensen AP, van der Voort JR (2005) Breeding by design: exploiting genetic maps and molecular markers through marker-assisted selection. In: Meksem K, Kahl G (eds) The handbook of plant genome mapping. Wiley-VCH Verlag GmbH & Co KGaA, Weinheim, pp 109–129

Peterson BK, Weber JN, Kay EH et al (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7:e37135. doi:10.1371/journal.pone.0037135

Piepho HP (2009) Ridge regression and extensions for genome-wide selection in maize. Crop Sci 49:1165–1176

Poczai P, Varga I, Laos M et al (2013) Advances in plant gene-targeted and functional markers: a review. Plant Methods 9:6. doi:10.1186/1746-4811-9-6

Podlich DW, Winkler CR, Cooper M (2004) Mapping as you go: an effective approach for marker-assisted selection of complex traits. Crop Sci 44:1560–1571

Poland JA, Brown PJ, Sorrells ME et al (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7:e32253. doi:10.1371/journal.pone.0032253

Poorter H, Fiorani F, Stitt M et al (2012) The art of growing plants for experimental purposes: a practical guide for the plant biologist. Funct Plant Biol 39:821–838

Powell W, Morgante M, Andre C et al (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breed 2:225–238

Price AL, Patterson NJ, Plenge RM et al (2006) Principal component analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Pritchard JK, Stephens M, Rosenberg NA et al (2000a) Association mapping in structured populations. Am J Hum Genet 67:170–181

Pritchard JK, Stephens M, Donnelly P (2000b) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Pumphrey MO, Bernardo R, Anderson JA (2007) Validating the *Fhb1* QTL for Fusarium head blight resistance in near-isogenic wheat lines developed from breeding populations. Crop Sci 47:200–206

Qin P, Wang Y, Li Y et al (2013) Analysis of cytoplasmic effects and fine-mapping of a genic male sterile line in rice. PLoS One 8:e61719. doi:10.1371/journal.pone.0061719

Radoev M, Becker HC, Ecke W (2008) Genetic analysis of heterosis for yield and yield components in rapeseed (*Brassica napus* L.) by quantitative trait locus mapping. Genetics 179:1547–1558

Rafalski JA, Tingey SV (1993) Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. Trends Genet 9:275–280

Rajendrakumar P, Biswal AK, Balachandran SM et al (2007) A mitochondrial repeat specific marker for distinguishing wild abortive type cytoplasmic male sterile rice lines from their cognate isogenic maintainer lines. Crop Sci 47:207–211

Rakitsch B, Lippert C, Stegle O et al (2013) A lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics 29:206–214

Randhawa HS, Mutti JS, Kidwell K et al (2009) Rapid and targeted introgression of genes into popular wheat cultivars using marker-assisted background selection. PLoS One 4:e5752. doi:10.1371/journal.pone.0005752

Reif JC, Melchinger AE, Xia XC et al (2003) Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. Crop Sci 43:1275–1282

Remington DL, Thornsberry JM, Matsuoka Y et al (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci U S A 98:11479–11484

Reynolds M, Foulkes MJ, Slafer GA et al (2009) Raising yield potential in wheat. J Exp Bot 60:1899–1918

Ribaut J-M, Betran J (1999) Single large-scale marker-assisted selection (SLS-MAS). Mol Breed 5:531–541

Ribaut J-M, Ragot M (2007) Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations and alternatives. J Exp Bot 58:351–360

Ribaut J-M, Hu X, Hoisington D et al (1997) Use of STSs and SSRs as rapid and reliable preselection tools is a marker assisted selection – backcross scheme. Plant Mol Biol Rep 15:154–162

Ribaut J-M, William HM, Khairallah M et al (2001) Genetic basis of physiological traits. In: Reynolds MP, Ortiz-Monasterio JI, McNab A (eds) Application of physiology in wheat breeding. CIMMYT, Mexico, DF, pp 29–47

Ribaut J-M, de Vicente MC, Delannay X (2010) Molecular breeding in developing countries: challenges and perspectives. Curr Opin Plant Biol 13:1–6

Rodi CP, Darnhofer-Patel B, Stanssens P et al (2002) A strategy for the rapid discovery of disease markers using the MassARRAY system. Biotechniques 32: S62–S69

Rogers JS (1972) Measure of genetic similarity and genetic distance; studies in genetics, VIII. Univ Texas Publ 2713:145–153

Ronaghi M, Karamohamed S, Patterson B et al (1996) Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242:84–89

Rutkoski JE, Heffner EL, Sorrels ME (2010) Genomic selection for durable stem rust resistance in wheat. Euphytica 179:161–173

Rymer PD, Rossetto M (2013) Applications of molecular markers in plant conservations. In: Henry RJ (ed) Molecular markers in plants. Wiley, Hoboken, pp 81–98

Saeed AI, Sharov V, White J et al (2003) TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34:374–378

Saiki R, Scharf S, Faloona FA et al (1985) Enzymatic amplification of b-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230:1350–1354

Salathia N, Lee HN, Sangster TA et al (2007) Indel arrays: an affordable alternative for genotyping. Plant J 51:727–737

Salvi S, Tuberosa R (2005) To clone or not to clone plant QTLs: present and future challenges. Trends Plant Sci 10:297–304

Salvi S, Tuberosa R (2007) Cloning QTLs in plants. In: Varshney RK, Tuberosa R (eds) Genomic assisted crop improvement, vol I. Springer, New York, pp 207–226

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd edn. Cold Springer Harbor Laboratory Press, Cold Springer Harbor

Sanchez AC, Brar DS, Huang N et al (2000) Sequence tagged site marker-assisted selection for three bacterial blight resistance genes in rice. Crop Sci 40:792–797

Sang X, Yang Z, Zhong B et al (2006) Assessment of purity of rice CMS lines using cpDNA marker. Euphytica 152:177–183

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467

Satgopan JM, Yandell BS, Newton MA et al (1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144:805–816

Sato T, Ueda T, Fukuta Y et al (2003) Mapping of quantitative trait loci associated with ultraviolet resistance in rice (*Oryza sativa* L). Theor Appl Genet 107:1003–1008

Savage D, Batley J, Erwin T et al (2005) SNPServer: a real-time SNP discovery tool. Nucleic Acids Res 33 (Web Server issue):W493–W495

Sax K (1923) Association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics 8:552–560

Scarcelli N, Cheverud JM, Schaal BA et al (2007) Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus. Proc Natl Acad Sci U S A 104:16986–16991

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19: R227–R240

Schendure J, Ji HL (2008) Next generation DNA sequencing. Nat Biotechnol 26:1135–1145

Schendure J, Porreca GJ, Reppas NB et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732

Schmierer DA, Kandemir N, Kudrna DA et al (2004) Molecular marker-assisted selection for enhanced yield in malting barley. Mol Breed 14:463–473

Schnable PS, Springer NM (2013) Progress toward understanding heterosis in crop plants. Annu Rev Plant Biol 64:71–88

Schneider K (2005) Mapping populations and principles of genetic mapping. In: Meksem K, Kahl G (eds) The handbook of plant genome mapping, genetic and physical mapping. WILEY-VCH Verlag GmbH & Co KGaA, Weinheim, pp 3–21

Schuster I (2011) Marker-assisted selection for quantitative traits. Crop Breed Appl Biotechnol S1:50–55

Seaton G, Haley CS, Knott SA et al (2002) QTL express: mapping quantitative trait loci simple and complex pedigrees. Bioinformatics 18:339–340

Segura V, Vilhjalmsson BJ, Platt A et al (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44:825–830

Servin B, Martin OC, Mezard M et al (2004) Toward a theory of marker-assisted gene pyramiding. Genetics 168:513–523

Shen L, Courtois B, McNally KL et al (2001) Evaluation of near-isogenic lines of rice introgressed with QTLs for root depth through marker-aided selection. Theor Appl Genet 103:75–83

Shi JQ, Li RY, Zou J et al (2011) A dynamic and complex network regulates the heterosis of yield correlated traits in rapeseed (*Brassica napus* L.). PLoS One 6: e21645

Silver J (1985) Confidence limits for estimates of gene linkage based on analysis of recombinant inbred strains. J Hered 76:436–440

Simpson SP (1989) Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. Theor Appl Genet 77:815–819

Singh BD (2009) Genetics, 2nd edn. Kalyani Publishers, New Delhi

Singh BD (2012a) Plant breeding, principles and methods, 9th edn. Kalyani Publishers, New Delhi

Singh BD (2012b) Biotechnology, expanding horizons, 4th edn. Kalyani Publishers, New Delhi

Singh AK, Rana MK, Singh S et al (2014a) CAAT box-derived polymorphism (CBDP): a novel promoter-targeted molecular marker for plants. J Plant Biochem Biotechnol 23:175–183

Singh VK, Singh AK, Kayastha AM et al (2014b) Bioinformatics for legume genomics research. In: Gupta S, Nadarajan N, Gupta DS (eds) Legumes in the omics era. Springer Science+Business Media, New York, pp 249–275

Smith S, Helentjaris T (1996) DNA fingerprinting and plant variety protection. In: Paterson AH (ed) Genome mapping in plants. RG Landes Co, Austin, pp 95–110

Sobrino B, Briona M, Carracedoa A (2005) SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Sci Int 154:181–194

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull 38:1409–1438

Sokolov BP (1990) Primer extension technique for the detection of single nucleotide in genomic DNA. Nucleic Acids Res 18:3671

Solberg TR, Sonesson AK, Woolliams JA et al (2008) Genomic selection using different marker types and densities. J Anim Sci 86:2447–2454

Soller M, Brody T (1976) On the power of experimental designs for the detection on linkage between marker loci and quantitative loci in crosses between inbred lines. Theor Appl Genet 47:35–39

Sonah H, Bastien M, Iquira E et al (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS One 8:e54603. doi:10.1371/journal.pone.0054603

Sorrells ME (1992) Development and application of RFLPs in polyploids. Crop Sci 32:1086–1091

Sosnowski O, Charcosset A, Joets J (2012) BioMercator V3: an upgrade of genetic map compilation and QTL meta-analysis algorithms. Http://bioinformatics.oxfordjournals.org/

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Spindel J, Begum H, Akdemir D et al (2015) Genomic selection and association mapping in rice (oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet 11: e1004982. doi:10.1371/journal.pgen.1004982

Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: JOINMAP. Plant J 3:739–744

Stam P, Zeven AC (1981) The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by backcrossing. Euphytica 30:227–238

Staub JE, Serquen FC, Manju G (1996) Genetic markers, map construction and their application in plant breeding. HortSci 31:729–741

Steele KA, Edwards G, Zhu J et al (2004) Marker-evaluated selection in rice: shifts in allele frequency among bulks selected in agricultural environments identify genomic regions of importance to rice adaptation and breeding. Theor Appl Genet 109:1247–1260

Steele KA, Price AH, Shahsidhar HE et al (2006) Marker-assisted selection to introgress rice QTLs controlling root traits into an Indian upland rice variety. Theor Appl Genet 112:208–221

Stich B, Melchinger AE, Frisch M et al (2005) Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor Appl Genet 111:723–730

Stich B, Mohring J, Piepho H et al (2008) Comparison of mixed-model approaches for association mapping. Genetics 178:1745–1754

Sticklen MB (2007) Feedstock crop genetic engineering for alcohol fuels. Crop Sci 47:2238–2248

Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc Ser B 64:479–498
Stuber CW, Polacco M, Senior ML (1999) Synergy of empirical breeding, marker-assisted selection and genomics to increase crop yield potential. Crop Sci 39:1571–1583
Stussey T (1990) Plant taxonomy. Columbia University Press, New York
Sundaram RM, Vishnupriya MR, Biradar SK et al (2008) Marker assisted introgression of bacterial blight resistance in Samba Mahsuri, an elite indica rice variety. Euphytica 160:411–422
Takagi H, Abe A, Yoshida K et al (2013) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J 74:174–183
Tan Y-D, Fu Y-X (2006) A novel method for estimating linkage maps. Genetics 173:2383–2390
Tanksley SD (1983) Molecular markers in plant breeding. Plant Mol Biol Rep 1:3–8
Tanksley SD (1993) Mapping polygenes. Annu Rev Genet 27:205–233
Tanksley SD, Nelson JC (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theor Appl Genet 92:191–203
Tanksley SD, Orton TJ (eds) (1983) Isozymes in plant genetics and breeding. Elsevier, Amsterdam
Tanksley SD, Rick CM (1980) Isozyme gene linkage map of the tomato: applications in genetics and breeding. Theor Appl Genet 57:161–170
Tanksley SD, Young ND, Patterson AH et al (1989) RFLP mapping in plant breeding: new tools for an old science. Biotechnology 7:257–263
Tanksley SD, Ganal MW, Martin GB (1995) Chromosome landing: a paradigm for map-based gene cloning in plants with large genomes. Trends Genet 11:63–68
Tautz D (1989) Hypervariability of simple sequences as a general source of polymorphic DNA markers. Nucleic Acids Res 17:6462–6471
Tessmer OL, Jiao Y, Cruz JA et al (2013) Functional approach to high-throughput plant growth analysis. BMC Syst Biol 7:S17–S29
Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science 327:818–822
Thiel T, Kota R, Grosse I et al (2004) SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development. Nucleic Acids Res 32:e5. doi:10.1093/nar/gnh006
Thimm O, Blasing O, Gibon Y et al (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37:914–939
Thoday JM (1961) Location of polygenes. Nature 191:368–370
Thompson JD, Gibson TJ, Plewniak F et al (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882. doi:10.1093/nar/25.24.4876
Thompson HJ, Zhao K, Wright M et al (2012) High-throughput single nucleotide polymorphism genotyping for breeding applications in rice using the BeadXpress platform. Mol Breed 29:875–886
Thongjuea S, Ruanjaichon V, Bruskiewich R et al (2009) RiceGeneThresher: a web-based application for mining genes underlying QTL in rice genome. Nucleic Acids Res 37(Database issue):D996–D1000
Toonen RJ, Puritz JB, Forsman ZH et al (2013) ezRAD: a simplified method for genomic genotyping in non-model organisms. PeerJ 1:e203. doi:10.7717/peerj.203
Tsuchihashi Z, Dracopoli NC (2002) Progress in high throughput SNP genotyping methods. Pharmacogenomics J 2:103–110
Tuinstra MR, Ejeta G, Goldsbrough PB (1997) Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci. Theor Appl Genet 95:1005–1011
UniProt Consortium (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res 41(Database issue):D43–D47
Utz HF, Melchinger AE (1996) PLABQTL: a program for composite interval mapping of QTL. JQTL 2(1), (http://wheat.pw.usda.gov/jag/papers96/paper196/utz.html)
Van Berloo R, Stam P (1998) Marker-assisted selection in autogamous RIL populations: a simulation study. Theor Appl Genet 96:147–154
van den Broeck D, Maes T, Sauer M et al (1998) Transposon display identifies individual transposable elements in high copy number lines. Plant J 13:121–129
Van Inghelandt D, Reif JC, Dhillon BS et al (2011) Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm. Theor Appl Genet 123:11–20
Van Ooijen JW (2006) JoinMap® 4, software for the calculation of genetic linkage maps in experimental populations. Kyazma B.V., Wageningen
van Tassell CP, Smith TPL, Matukumalli LK et al (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods 5:247–252
vanOrsouw NJ, Hogers RCJ, Janssen A et al (2007) Complexity reduction of polymorphic sequences (CRoPS™): a novel approach for large-scale polymorphism discovery in complex genomes. PLoS One 2: e1172. doi:10.1371/journal.pone.0001172
Verhoeven KJF, Jannink J-L, McIntyre LM (2006) Using mating designs to uncover QTL and genetic architecture of complex traits. Heredity 96:139–149
Veyrieras J-B, Goffinet B, Charcosset A (2007) MetaQTL: a package of new computational methods for the meta-analysis of QTL mapping experiments. BMC Bioinformatics 8:49–64
Vision TJ, Brown DG, Shmoys DB et al (2000) Selective mapping: a strategy for optimizing the construction of high-density linkage maps. Genetics 155:407–420

Vos P, Hogers R, Bleeker R et al (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414

Vosman B (1998) Variations on a theme. In: Karp A, Isaac PG, Ingram DS (eds) Molecular tools for screening biodiversity. Chapman and Hall, London, pp 262–264

Walter A, Studer B, Kolliker R (2012) Advanced phenotyping offers opportunities for improved breeding of forage and turf species. Ann Bot 110:1271–1279. doi:10.1093/aob/mcs026

Wang DG, Fang JB, Sio CJ et al (1998) Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Wang S, Basten CJ, Zeng Z-B (2005) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh

Wang J, Chapman SC, Bonnet DG et al (2007) Application of population genetic theory and simulation models to efficiently pyramid multiple genes via marker–assisted selection. Crop Sci 47:582–588

Wang J, Lin M, Crenshaw A et al (2009a) High-throughput single nucleotide polymorphism genotyping using nanofluidic Dynamic Arrays. BMC Genomics 10:561–573

Wang Q, Zhang B, Lu Q (2009b) Conserved region amplification polymorphism (CoRAP) a novel marker technique for plant genotyping in *Salvia miltiorrhiza*. Plant Mol Biol Rep 27:139–143

Wang Z, Gerstein M, Snyder M (2009c) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wang S, Basten CJ, Zeng Z-B (2012) Windows QTL Cartographer 2.5_011. Department of Statistics, North Carolina State University, Raleigh, http://statgen.ncsu.edu/qtlcart/WQTLCart.htm

Ware DH, Jaiswal P, Ni J et al (2002) Gramene, a tool for grass genomics. Plant Physiol 130:1606–1613

Weber JL, May PE (1989) Abundant class of human DNA polymorphism, which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18:7213–7218

Wenzl P, Carling J, Kudrna D et al (2004) Diversity arrays technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci U S A 101:9915–9920

White JW, Andrade-Sanchez P, Gore MA et al (2012) Field-based phenomics for plant genetics research. Field Crops Res 133:101–112

Whittaker JC, Thompson R, Denham MC (2000) Marker assisted selection using ridge regression. Genet Res 75:249–252

Williams JGK, Kubelik AR, Livak KJ et al (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:1631–1635

Williams MNV, Pande N, Nair M et al (1991) Restriction fragment length polymorphism analysis of polymerase chain reaction products amplified from mapped loci of rice (*Oryza sativa* L.) genomic DNA. Theor Appl Genet 82:489–498

Winzeler EA, Richards DR, Conway AR et al (1998) Direct allelic variation scanning of the yeast genome. Science 281:1194–1197

Witsenboer H, Vogel J, Michelmore RW (1997) Identification, genetic localization and allelic diversity of amplified microsatellite polymorphic loci in lettuce and wild relatives (*Lactuca* spp.). Genome 40:923–926

Wong CK, Bernardo R (2008) Genome wide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor Appl Genet 116:815–824

Woo NS, Badger MR, Pogson BJ (2008) A rapid non-invasive procedure for quantitative assessment of drought survival using chlorophyll fluorescence. Plant Methods 4:27–30

Wright S (1952) Quantitative genetics. Her Majesty's Stationery Office, London

Wright S (1978) Evolution and genetics of populations, vol IV. University of Chicago Press, Chicago

Wu R, Zeng ZB (2001) Joint linkage and linkage disequilibrium mapping in natural populations. Genetics 157:899–909

Wu KK, Burnquist W, Sorrells ME et al (1992) The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theor Appl Genet 83:294–300

Wu KS, Jones R, Dannaberger L et al (1994) Detection of microsatellite polymorphisms without cloning. Nucleic Acids Res 22:3257–3258

Wu J, Jenkins JN, McCarty JC et al (2011a) Comparisons of four approximation algorithms for large-scale linkage map construction. Theor Appl Genet 123:649–655

Wu Y, Close TJ, Lonardi S (2011b) Accurate construction of consensus genetic maps via integer linear programming. IEEE/ACM Trans Comput Biol Bioinform 8:381–394

Wu J-M, Li Y-R, Yang L-T et al (2013) cDNA-SCoT: a novel rapid method for analysis of gene differential expression in sugarcane and other plants. Aust J Crop Sci 7:659–664

Wurschum T (2012) Mapping QTL for agronomic traits in breeding populations. Theor Appl Genet 125:201–210

Xi ZY, He FH, Zeng RZ et al (2006) Development of a wide population of chromosome single-segment substitution lines in the genetic background of an elite cultivar of rice (*Oryza sativa* L.). Genome 49:476–484

Xu Y (2010) Molecular plant breeding. CAB International, Wallingford

Xu Y, Crouch JH (2008) Marker-assisted selection in plant breeding: from publications to practice. Crop Sci 48:391–407

Xu Y, Zhu L, Xiao J et al (1997) Chromosomal regions associated with segregation distortion of molecular markers in F_2, backcross, doubled haploid and

recombinant inbred populations in rice (*Oryza sativa* L). Mol Gen Genet 253:535–545
Xu J, Zhao Q, Du P et al (2010) Developing high throughput genotyped chromosome segment substitution lines based on population whole genome resequencing in rice (*Oryza sativa* L.). BMC Genomics 11:656–669
Yan L, Loukoianov A, Tranquilli G et al (2003) Positional cloning of the wheat vernalization gene *VRN1*. Proc Natl Acad Sci U S A 100:6263–6268
Yandell BS, Mehta T, Banerjee S et al (2007) R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. Bioinformatics 23:641–643
Yano M (2001) Genetic and molecular dissection of naturally occurring variation. Curr Opin Plant Biol 4:130–135
Yao H, Gray AD, Auger DL et al (2012) Genomic dosage effects on heterosis in triploid maize. Proc Natl Acad Sci U S A 110:265–269
Yashitola J, Sundaram RM, Biradar SK et al (2004) A sequence specific PCR marker for distinguishing rice lines on the basis of wild abortive cytoplasm from their cognate maintainer lines. Crop Sci 44:920–924
Yazdanbakhsh N, Fisahn J (2009) High throughput phenotyping of root growth dynamics, lateral root formation, root architecture and root hair development enabled by PlaRoM. Funct Plant Biol 36:938–946
Youens-Clark K, Buckler E, Casstevens T et al (2011) Gramene database in 2010: updates and extensions. Nucleic Acids Res 39(Database issue):D1085–D1094
Young ND (1999) A cautiously optimistic vision for marker-assisted breeding. Mol Breed 5:505–510
Young ND, Tanksley SD (1989) RFLP analysis of the size of chromosomal segments retained around the *Tm-2* locus of tomato during backcross breeding. Theor Appl Genet 77:353–359
Young ND, Zamir D, Ganal MW et al (1988) Use of isogenic lines and simultaneous probing to identify DNA markers tightly linked to the *Tm-2a* gene in tomato. Genetics 120:579–585
Yu SB, Li JX, Xu GC et al (1997) Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. Proc Natl Acad Sci U S A 94:9226–9231
Yu J, Pressoir G, Briggs WH et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208
Yu J, Holland JB, McMullen MD et al (2008) Genetic design and statistical power of nested association. Genetics 178:539–551
Zabeau M, Vos P (1993) European patent application. Publication no: EP0534858
Zamir D (2001) Improving plant breeding with exotic genetic libraries. Nat Rev Genet 2:983–989
Zamir D, Tadmar Y (1986) Unequal segregation of nuclear genes in plants. Bot Gaz 147:355–358
Zeng Z-B (1993) Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. Proc Natl Acad Sci U S A 90:10972–10976
Zeng Z-B (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468
Zhang H-Y, He H, Chen L-B et al (2008) A genome-wide transcription analysis reveals a close correlation of promoter INDEL polymorphism and heterotic gene expression in rice hybrids. Mol Plant 1:720–731
Zhao K, Aranzana MJ, Kim S et al (2007) An Arabidopsis example of association mapping in structured samples. PLoS Genet 3:71–82
Zhao Y, Gowda M, Liu W et al (2012a) Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet 124:769–776
Zhao Y, Gowda M, Longin FH (2012b) Impact of selective genotyping in the training population on accuracy and bias of genomic selection. Theor Appl Genet 125:707–713
Zhao Y, Gowda M, Liu W et al (2013) Choice of shrinkage parameter and prediction of genomic breeding values in maize elite breeding populations. Plant Breed 132:99–106
Zhong S, Dekkers JCM, Fernando RL et al (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182:355–364
Zhou X, Stephens M (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods 11:407–409. doi:10.1038/nmeth.2848
Zhu C, Gore M, Buckler ES et al (2008) Status and prospects of association mapping in plants. Plant Genome 1:5–20
Zhu Y, Wang Y, Shao M et al (2011) Estimating soil water content from surface digital image gray level measurements under visible spectrum. Can J Soil Sci 91:69–76
Zietkiewicz E, Rafalski A, Labuda D (1994) Genomic fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. Genomics 20:176–183
Ziska LH, Bunce JA (2007) Predicting the impact of changing CO_2 on crop yields: some thoughts on food. New Phytol 175:607–618
Zou F (2009) QTL mapping in intercross and backcross populations. In: DiPitrello K (ed) Cardiovascular genomics, methods and protocols. Humana Press (Springer Science+Business Media), LLC, New York, pp 157–173

Author Index

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0, © Author(s) 2015

Subject Index

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0, © Author(s) 2015