Chapter 15
Next Generation Sequencing and Germplasm Resources

Paul Visendi, Jacqueline Batley and David Edwards

Contents

15.1 Introduction 370
15.2 Next Generation Sequencing 372
15.3 Sequencing Reference Genomes 373
15.4 Next Generation Diversity Analysis 379
15.4.1 Reference-Based Diversity Analysis 381
15.4.2 Non Reference-Based Diversity Analysis 381
15.5 SNPs 381
15.5.1 Copy Number Variation 383
15.6 Perspectives 384
References 384

Abstract DNA sequencing technology is advancing at an astounding rate, with rapid increases in data volumes and quality combined with reducing costs. The availability of this technology opens novel avenues for the analysis of plant germplasm resources. Where previous studies analysed a limited number of phenotypic or molecular genetic markers, it is now possible to re-sequence whole genomes to characterise diversity at a resolution of each nucleotide. Current approaches combine high resolution genetic markers with genome sequencing both for reference assembly and genotyping by sequencing. As next generation sequencing technologies continue to advance, we approach the potential to catalogue and characterise all genome variations across

D. Edwards (✉) · P. Visendi · J. Batley
University of Queensland, School of Agriculture and Food Sciences,
4072 Brisbane, QLD, Australia
e-mail: Dave.Edwards@uq.edu.au

D. Edwards · P. Visendi
Australian Centre for Plant Functional Genomics, University of Queensland,
4072 Brisbane, QLD, Australia

P. Visendi
e-mail: paul.muhindira@uqconnect.edu.au

J. Batley
e-mail: J.Batley@uq.edu.au

R. Tuberosa et al. (eds.), *Genomics of Plant Genetic Resources,*
DOI 10.1007/978-94-007-7572-5_15,

diverse germplasm to gain a greater understanding of how the genome contributes to the diversity seen in today's plants.

Keywords Genome sequencing · Pangenome · Illumina · Ion Torrent · AB SOLiD · Pacific biosciences · Oxford nanopore · Roche 454 · Single nucleotide polymorphisms · Bioinformatics

List of species

- *Arabidopsis thaliana*
- *Brachypodium distachyon*
- *Brassica juncea*
- *Brassica oleracea*
- *Brassica rapa*
- *Carica papaya*
- *Eucalyptus grandis*
- *Fragaria vesca*
- *Glycine max*
- *Lotus japonicus*
- *Malus domestica*
- *Medicago truncatula*
- *Mimulus guttatus*
- *Oryza sativa*
- *Populus trichocarpa*
- *Prunus persica*
- *Solanum tuberosum*
- *Sorghum bicolor*
- *Theobroma cacao*
- *Triticum aestivum*
- *Vitis vinifera*
- *Zea mays*

15.1 Introduction

Plant genomes are often both highly repetitive with ancient and more recent polyploidy. Because of these features, taxonomic analysis can be a challenge. To comprehensively analyse germplasm resources, allele and haplotype frequencies need to be studied. Next Generation Sequencing (NGS) techniques have greatly advanced genome analysis, enabling the elucidation of the genome or transcriptome sequence of an organism relatively quickly and cheaper in comparison to the traditional Sanger-based sequencing technologies (Edwards et al. 2013). This has opened avenues in plant research, by which genomic variations can be quickly and efficiently determined through computational and statistical analysis.

While NGS data is relatively inexpensive compared to traditional sequence data, sequence quality and read lengths are relatively poor. Accurate estimation of Single Nucleotide Polymorphism (SNP) calling and allele frequencies may result in false positives due to sequence or read mapping error. In some cases, a greater number of biological replicates rather than depth of coverage has been shown to produce better results when calling SNP and allele frequencies (Li 2011). These errors require downstream validation and confirmatory analysis, which usually involves re-sequencing (Kim et al. 2011).

NGS technologies have impacted on plant taxonomic systems, currently leveraging the Linnaean taxonomic methods with DNA barcoding (Pang et al. 2012) and evolutionary studies in plants (Darracq et al. 2010). Methods which make this possible for complex polyploid genomes include targeted sequence capture (Grover et al. 2012) and chromosome flow cytometry (Doležel et al. 2004). These approaches reduce the complexity of the assembly of NGS data as well as downstream analysis, enabling biologists to answer questions on diversity, polyploid origins, domestication and ancestry of many crops (Zhang et al. 2011).

With the development of advanced sequencing systems, sequence-based SNPs are becoming the marker of choice. The greatest genetic resolution is obtained through the analysis of SNPs and allele frequencies within a population. Two alleles associating with each other by random chance are said to be in Linkage Equilibrium (LE). If this association is found to be non-random either due to physical proximity or selection, they are said to be in Linkage Disequilibrium (LD). LD is the basis of Association Mapping (AM). Accurate AM is highly dependent on the extent of physical LD, population size and population structure (Duran et al. 2010). Small and highly structured populations often lead to elevated false positives in AM studies (Cuesta-Marcos et al. 2010).

The dynamic nature of living organisms, their interaction with the environment, with other species and with each other, affects adaptation and evolution. This led to the "Pangenome" concept (Tetz 2005). This concept proposes that genetic information of all living organisms belongs to a common system, the Pangenome, and that this system is both stable and fluid, and genetic elements such as DNA, RNA and plasmids among others traverse through the system, implying horizontal gene transfer between similar organisms and even across species. Gene transfer has been reported between bacteria (Yue et al. 2012), between viruses and eukaryotes (Wu and Zhang 2011), between plants and prokaryotes, and plants and other eukaryotes (Bock 2010). The concept further postulates that mechanisms of fluidity of the Pangenome include viruses and bacteria through infection, the food chain, death and decay. There has been little advancement of this concept in plants but several studies have been carried out in humans looking at the Asian, Caucasian and African human genomes (Li et al. 2010a) and in bacteria (Hall et al. 2010; Laing et al. 2010). It would be interesting to see applications of this concept to higher plants.

15.2 Next Generation Sequencing

DNA sequencing technology is undergoing a revolution and at the same time fuelling a revolution in genetics and genomics. Applications for Sanger-based sequencing remain, though the majority of DNA sequencing is now produced by one of a range of NGS technologies. NGS suffers from shorter reads and greater error rates than traditional Sanger sequencing, and its predominance is due to the ability to produce much larger volumes of data at a relatively low cost per sequenced base.

The first commercially available pyrosequencing system was commercialised by Roche (Basel, Switzerland) as the GS20, capable of sequencing over 20 million base pairs (Mbp) in just over 4 h (Margulies et al. 2005). This was replaced in 2007 by the GS FLX model, capable of producing over 100 Mbp of sequence in a similar amount of time. The current system, the GS FLX + can produce around 700 Mbp of data with read lengths of up to 1,000 bp with multiplexing of samples (www.my454.com). The Roche 454 FLX system performs amplification and sequencing in a high-throughput picoliter format. Emulsion PCR enables the amplification of a DNA fragment immobilized on a bead, generating sufficient DNA for the subsequent sequencing reaction. Beads are distributed onto the plate. DNA sequencing involves the sequential flow of both nucleotides and enzymes over the plate, which converts chemicals generated during nucleotide incorporation into a chemiluminescent signal that can be detected by a CCD camera. The light signal is quantified to determine the number of nucleotides incorporated during the extension of the DNA sequence. The output is in the form of 'flow space', which is converted to the traditional ACGT nucleotide sequence format. The sequence reads are much longer than most other NGS systems. The main error types are additional or reduced numbers of nucleotides around mononucleotide strings. These errors make the accurate calling of insertion/deletion (indel) differences a challenge.

The Illumina sequencing platforms use reversible terminator chemistry to generate up to 600 Gbp of sequence data per run, the greatest volume of data from any current NGS platform. Sequencing templates are immobilized on a flow-cell surface, and amplification generates clusters of up to 1,000 identical copies of each DNA molecule. Sequencing uses fluorescently labelled nucleotides to produce reads of up to 150 bp in length, though 100 bp is more common. Reads can be produced as pairs. The use of paired reads improves the accuracy of reference mapping, overcoming many of the limitations of short read lengths such as inaccurate resolution of repeats, indels and structural rearrangements. By using the distance between a read pair to infer an insertion or deletion in the reference or sample and to resolve repeats in *de novo* assembly, higher accuracy is achieved. Illumina sequencing is now becoming the platform of choice for resequencing, SNP discovery, whole-genome shotgun sequencing and *de novo* assembly (Imelfort et al. 2009b; Imelfort and Edwards 2009; Williams-Carrier et al. 2010; Dong et al. 2011; Shulaev et al. 2011).

The SOLiD System from Life Technologies (Applied Biosystems) enables parallel sequencing of amplified DNA fragments linked to beads. The method uses

sequential ligation of dye-labelled oligonucleotides, and the latest 5500xl system produces 20–30 Gbp of data per day, with read lengths of up to 75 bp (www.appliedbiosystems.com/). SOLiD data features a two-base encoding mechanism that interrogates each base twice providing a form of built-in error detection for SNP discovery when comparing reads to a reference.

Ion Torrent is a relatively new technology and uses a high-density array of semiconductor micro reaction chambers (www.iontorrent.com). Changes in pH are recorded as a result of the release of a hydrogen proton during the incorporation of a nucleotide during DNA synthesis. This produces reads of 100-200 bp, with up to 1 Gbp of data per run. The error profile of this system is still unknown, but the technology has potential for cost-effective re-sequencing and variant discovery with fast runs of 2 h.

Pacific Biosciences is one of the first 'third generation' sequencing systems to go on the market, and applies a novel single-molecule sequencing technique called SMRT^{TM} (Single Molecule Real Time) technology. Read lengths of around 1,000 bp have been reported (www.pacificbiosciences.com). As with the Ion Torrent system, little is known about the error profile of the system, but missing bases, and hence indel calling would be a likely challenge with this technology.

NGS technologies continue to evolve at an astounding rate and new technologies such as the Oxford Nanopore system are likely to continue to push the market forward over the coming years.

The vast quantities of data generated using NGS require the development of dedicated bioinformatics systems (Edwards et al. 2009; Marshall et al. 2010; Lai et al. 2012c; Lee et al. 2012). It was initially thought that bioinformatics systems would not be able to keep pace with sequencing developments, and while still a bottleneck exists in translating the data into biological information, the growth of bioinformatics has kept track with data production (Batley and Edwards 2009a; Lee et al. 2011b).

15.3 Sequencing Reference Genomes

There are now several optional approaches to sequence genomes, and the approach undertaken would depend on the use of the genome sequence (Imelfort et al. 2009a; Edwards and Batley 2010). Bacterial Artificial Chromosome (BAC) sequencing is still considered to be the most robust method for genome sequencing and involves the production of an overlapping tiling path of large genomic fragments maintained within BACs. Each BAC is shotgun sequenced, where many short reads are assembled to produce the sequence of the BAC. The whole genome sequence may then be reassembled based on sequence overlaps. This approach, while being the gold standard, remains prohibitively expensive and is unlikely to be undertaken for the majority of germplasm resources.

An alternative approach is the whole-genome shotgun (WGS) method, where the genome is fragmented into millions of smaller reads that are individually sequenced.

Computational algorithms then assemble the genome sequence, frequently requiring additional scaffolding or assembly to generate a representative genome. Scaffolding involves the assembly of several overlapping contigs into scaffolds and then assembling these further into a genome assembly. While WGS requires less time and resources than a BAC-by-BAC approach, assembly is often problematic due to repeats within the genome. This is particularly true for many plant species, with polyploid genomes and abundant repetitive elements (e.g. maize, wheat, etc.). With the decreasing cost of sequencing and rapid improvements in both data quality and read length from next and third generation technologies, WGS sequencing projects are likely to be established for many additional germplasm resources.

Arabidopsis thaliana was the first plant species to have a sequenced genome (Arabidopsis_ Genome_ Initiative 2000). Since this milestone, the number of plant genomes being sequenced continues to increase. While several species with small genome sizes as well as model species are now been sequenced, genome researchers are now starting to tackle some of the larger and more complex genomes (Berkman et al. 2011b; Feuillet et al. 2011; Edwards and Wang 2012; Berkman et al. 2013). These can act both as models to understand broad families of plants, as well as reference sequences for the mapping and comparison of unassembled genome sequence data from diverse germplasm resources.

Brassica species share extensive synteny with *Arabidopsis thaliana*, enabling comparative mapping and exploitation of the Arabidopsis genome sequence for Brassica crop improvement. Among the six cultivated species of Brassica, *B. rapa* (syn. *campestris*, AA, n = 10), *B. juncea* (AABB, n = 18) and *B. napus* (AACC, n = 19) are agronomically important oilseed crops, whereas *B. oleracea* (CC, n = 9) provides valuable leafy vegetables (e.g. broccoli, cauliflower, cabbage, kholrabi, etc.). The other two species, *B. nigra* (BB, n = 8) and *B. carinata* (BBCC, n = 17) are largely valued as condiments. Proprietary AA, CC and AACC genomes were sequenced in 2009 (http://tinyurl.com/brassicagenome), and recently the multinational Brassica genome project (MBGP) published the first public *B. rapa* genome (Mun et al. 2010; Wang et al. 2011; Edwards and Wang 2012).

Legumes represent the third largest plant family and are the second most important crop family for the human diet (Cannon et al. 2006). The model species *Medicago truncatula* is an annual diploid with eight chromosomes. It is closely related to tetraploid alfalfa (*M. sativa*). A combination of cytogenetic and BAC sequence data show that the *M. truncatula* genome is organized into distinct gene-rich euchromatin and repeat-rich pericentromeric regions, allowing the *M. truncatula* genespace to be efficiently sequenced using a BAC-by-BAC strategy (Young et al. 2011). Six chromosomes have been sequenced in a US project and two additional ones were sequenced by partners in Europe. Of special note is the *Medicago* HapMap Project, which aims to deep-sequence whole genomes of 30 inbred *Medicago* lines using the Illumina platform, and use the reference genome to determine SNPs and Indels. A current release (Mt3.5) of the genome is available at (www.medicagohapmap.org).

Lotus japonicus is a diploid self-fertile perennial pasture legume, with six chromosomes and a genome sequence of around 450 Mbp. Large-scale genome sequencing of variety Miyakojima MG-20 began in 2000. ESTs, cDNAs and gene segments

from *Lotus* and other legumes were used to select TAC clones, which were then sequenced using a shotgun approach (Sato et al. 2008). Data released is available at www.kazusa.or.jp/lotus. The genome sequences of *Medicago truncatula* and *Lotus japonicus* have provided invaluable resources for legume research given that they have both been sequenced using BAC clones making syntenic studies relatively simple and in turn providing evolutionary insights into other species such as *Arabidopsis* (George et al. 2008; Schlueter et al. 2008; Bertioli et al. 2009).

Glycine max (Soybean) is a major crop that accounts for 70 % of the world's edible plant-derived protein. Its 1.1 Gbp genome was sequenced using a WGS approach (Schmutz et al. 2010a), and the sequence is available through the phytozome database (http://www.phytozome.net/soybean). Re-sequencing of cultivated and wild varieties has enabled a detailed characterisation of genome variation in this species (Lam et al. 2010). Early and current studies on the evolution and domestication of soybean have shown a loss of genetic diversity as a result of domestication (Hyten et al. 2006; Li et al. 2010b). It has also been hypothesized that there was a single soybean domestication event, although this has not been confirmed since archaeological evidence suggests multiple domestication sites in East Asia (Lee et al. 2011a).

Trees, due to their long life span, have characteristics that distinguish them from annual, herbaceous plants. It is likely that many of these properties are based on a tree-specific genetic foundation. Poplar has been selected as a model system for trees because it has a relatively small genome. *Populus trichocarpa* (black cottonwood), with a paleopolyploid ($2n = 38$) genome of approximately 480 Mbp, was selected as the first tree genome to be sequenced. The *Populus trichocarpa* Nisqually 1 genotype was sequenced using WGS approaches (Tuskan et al. 2006). Approximately 7.6 million reads were assembled into 2,447 scaffolds containing 410 Mb of genomic DNA. The assembly was found to include more than 95 % of known cDNAs. Sequencing this genome allowed for the comparison between perennial and annual plant species on a whole genome basis for the first time and provides resources to help answer tree-specific questions about dormancy, development of a secondary cambium, juvenile-mature phase change and long-term host-pest interactions (Brunner et al. 2004; Tuskan et al. 2004).

Among the hardwoods, the most widely grown is *Eucalyptus*. *Eucalyptus* has high economic value due to its fiber, largely exploited for pulp, cellulose and paper. In addition, *Eucalyptus* has potential as a sustainable biofuel source, due to a fast growth rate. In June 2007 the DOE JGI initiated the *Eucalyptus grandis* genome-sequencing project. The project started in August 2007, and sequencing was completed by 2009. The project, coordinated by the *Eucalyptus* Genome Network (EUCAGEN), involved more than 130 scientists from 18 countries. A release of the first annotated assembly is available on phytozome (http://www.phytozome.net/), with an approximate genome size of 691 Mb, 4,952 scaffolds from 32,762 contigs and 300 scaffolds greater than 50 kb in size, representing approximately 94.2 % of the genome.

In 2004, the tomato genome was selected as the reference for sequencing within the Solanaceae genomics project. The sequencing of the domesticated tomato *(Solanum lycopersicum)* (Consortium 2012) marks the first step in bringing together genetic maps and genomes of all Solanaceae and related plants, including potato, eggplant,

pepper, petunia and coffee. The variety 'Heinz 1706' was initially selected as BAC resources were already available.,The 900 Mbp genome was sequenced using a combination of sanger and WGS approach. While the genome was known to consist of approximately three-quarters pericentromeric heterochromatin, known to be rich in repetitive sequences and poor in genes, the remaining one-quarter consisted of distal, euchromatic segments of chromosomes that contain mostly single copy sequences and more than 90 % of the genes. As such, only this portion was sequenced (Shibata 2005). To gain further insight into its evolution, its closest wild relative, *Solanum pimpinellifolium* LA1589 was also sequenced using illumina short reads (Tomato Genome Consortium 2012). The resulting 739 Mb draft genome showed a 0.6 % divergence from the domesticated variety. Comparative studies with potato (Xu et al. 2011c) revealed an 8.7 % nucleotide divergence. Further comparisons with the grape genome supported previous hypothesis that the rosid lineage diverged from a common eudicot ancestor following whole genome triplication. The tomato genome was however shown to have high synteny with potato, pepper, eggplant and nicotiana (Tomato Genome Consortium 2012).

Potato, another member of the family Solanaceae has a variety of ploidy levels, ranging from diploid (2n = 24) to hexaploid (6n = 72), with the cultivated potato varieties being tetraploid. Potato is an economically important food crop with 330 million tons produced globally in 2009 (http://www.fao.org). The size of the genome is 840 Mbp. The potato genome sequencing consortium (PGSC) sequenced the potato genome using a BAC-by-BAC approach (Xu et al. 2011b). Due to its high heterozygosity, the generation of a draft sequence required the use of a homozygous form, called a doubled monoploid (DM), and integration of the sequence with that of a heterozygous diploid form RH89-039-16. The two genomes were used to study the genome structure, with the heterozygous diploid resembling the cultivated tetraploid potato (http://solgenomics.net/organism/Solanum_tuberosum/genome).

The sequencing of grapevine *Vitis vinifera* (2n = 38) (Jaillon et al. 2007), was performed on the quasi-homozygous genotype PN40024 by a French-Italian collaborative project using a WGS approach. A total of 6.2 million reads representing 8.4x coverage of the genome were assembled with the Arachne12 assembler to produce 316 supercontigs, representing putative allelic haplotypes totaling 11.6 Mbp. The assembly of one of the haplotypes in each heterozygous region resulted in 19,577 contigs and 3,514 supercontigs totaling 487 Mb, consistent with an earlier predicted size of 475 Mb (Lodhi et al. 1995). A different approach to sequencing *V. vinifera* used a BAC-by-BAC approach (Zharkikh et al. 2008). In a comparison of the two approaches (Zharkikh et al. 2008) showed it was possible to sequence the highly heterozygous genome by using sufficient coverage and a variety of clones with different sizes. This approach yielded more informative data on repetitive elements and useful SNPs.

The monkey flower *Mimulus guttatus,* has become a model system for studying ecological and evolutionary genetics due to its diverse phenotypes, which include adaptations to desert and aquatic environments, selfing and outcrossing, annual and perennial forms and varied floral morphology. The DOE Joint Genomes Institute (JGI) commenced sequencing *Mimulus guttatus* in 2006 using a WGS approach.

In addition to the WGS sequence, JGI is sequencing 200,000 ESTs each from *M. guttatus* and *M. lewisii*. Additionally, WGS sequencing of IM62 inbred lines is ongoing. A draft release of the genome with 321.7 Mb of 2,216 scaffolds, 300.7 Mb of 17,831 contigs with gaps ~ 6.5 % and 512 scaffolds larger than 50 Kb, with 95.7 % of the genome represented in scaffolds greater than 50 Kbp is available on phytozome (http://www.phytozome.net/mimulus).

The papaya-sequencing project was founded by the centre for genomics, proteomics and bioinformatics research Initiative (CGPBRI) at the University of Hawaii in 2004. Papaya (*Carica papaya*) is the first fruit species and commercially important transgenic plant to be sequenced (Ming et al. 2008). Papaya has nine chromosomes and the size of the genome is 372 Mbp. A WGS approach produced a total of 2.8 million reads generated from a female transgenic cultivar "SunUp". After filtering, 1.6 million reads were assembled into contigs containing 271 Mb and scaffolds spanning 370 Mb. Validation of the assembly using 16,362 unigenes derived from expressed sequence tags (ESTs), showed 15,064 ESTs (92.1 %) matched the assembly.

Cocoa (*Theobroma cacao*; 2n = 2x = 20) was sequenced by the international cocoa genome sequencing consortium (ICGS), using a WGS approach with Sanger and Roche 454 technologies (Argout et al. 2011). Assembly was carried out using Newbler software, resulting in 25,912 contigs and 4,792 scaffolds, with an N50 of 473.8 Kb. Illumina reads (x44 coverage) were used to improve the assembly. The total assembly length was 326 Mb, representing approximately 76 % of the estimated genome size of cocoa. The sequencing of cocoa enabled the comparison of the grape, soybean, poplar and *A. thaliana* genomes with cocoa revealing 682 gene families (2,053 genes) unique to the cocoa genome. This indicated an expansion of some gene families during the evolution of cocoa. Some of the genes were annotated as flavonoid related, a contributing factor to the flavour and scent of chocolate. A re-assembly of the genome with an updated version of the Newbler assembler resulted in an assembly (ICGS Assembly 1.2) with an N50 of 5.624 Mb and the largest scaffold of 18.2 Mb. The new assembly covered 84.3 % of the genome.

Genetic diversity within the Rosaceae required the use of several model species as references for comparative analysis in this family. Model species identified for this purpose include strawberry (*Fragaria vesca*), peach (*Prunus persica*) and apple (Shulaev et al. 2008). Due to the complexity of the octoploid cultivated strawberry *F.* × *ananassa*, (2n = 8x = 56), the sequencing of its diploid progenitor, the woodland strawberry *F. vesca* ($2n = 2x = 14$), was undertaken (Shulaev et al. 2011). A *de novo* assembly of Roche/454, Life Technologies/SOLiD and Illumina/Solexa platform reads at 39x-combined coverage resulted in over 3,200 scaffolds with an N50 of 1.3 Mb. A total of 95 % (209.8 Mb) of the genome was represented in 272 scaffolds. Resequencing at 26x coverage with Illumina validated the assembly, with 99.8 % of the scaffolds and 99.98 % of bases perfectly matching with Illumina reads.

Sequencing of the apple genome (Velasco et al. 2010) followed a similar approach to that used to sequence the highly heterozygous grape genome *Vitis vinifera* cv Pinot Noir (Zharkikh et al. 2008) in which a combination of paired end reads produced by Sanger sequencing and unpaired reads produced by sequencing by synthesis was shown to be an efficient way of sequencing and assembling complex heterozygous

genomes. In total, 122,146 contigs provided a 16x coverage of the 603 Mb genome, Of these 103,076 were assembled into 1,629 meta-contigs. This assembly consisted of 26 % Sanger paired end reads and 74 % 454 sequencing by synthesis paired and unpaired reads.

Cereal crops diverged from a common ancestor some 60 million years ago and whole-genome organisation exhibits a high degree of synteny (Moore et al. 1995). Rice has the smallest genome size among major cereal crops, estimated at 430 Mbp (Goff et al. 2002a) and the genome sequences of rice provide a basis for integrating and comparing biological information from rice and related cereal crops (Goff et al. 2002b; Yu et al. 2002).

The International Rice Genome Sequencing Project (IRGSP) sequenced an inbred rice cultivar, *Oryza sativa* ssp. japonica cv. Nipponbare using a clone by clone approach with BACs and PACs. 3,401 BAC and PAC clones were sequenced and assembled, resulting in a high quality reference genome anchored to a genetic map (International Rice Genome Sequencing Project 2005). Gene content analysis estimated 37,544 genes to be present, of which 17,016 were supported by 25,636 full-length cDNAs. Using this reference, analysis of rice agronomic traits was carried out on 50 wild and domesticated rice accessions (Xu et al. 2012). 6.5 million SNPs were identified and population structure analysis determined the domestication origins of rice. *Brachypodium* is a close relative of the cool season grasses and in 2006 was sequenced by the US Department of Energy Joint Genome Institute (DOE JGI) to provide a genomic bridge between rice and other agronomically important cereals (International Brachypodium Initiative 2010).

The *Sorghum bicolor* genome consists of approximately 770 Mbp in 10 chromosomes ($2n = 20$). A WGS approach was applied within the DOE-JGI community sequencing program to sequence this genome. A validation of the assembly by comparison with 27 individually sequenced BACs indicated that the assembly was 98.46 % complete with an error rate of < 1 nucleotide per 10 kb (Paterson et al. 2009). Comparison of the genome sequence with Sorghum ESTs suggests that more than 95 % of known sorghum protein-coding genes are represented in this assembly. Sequencing the sorghum genome opened opportunities for comparative studies to be carried out in the grass family between rice, sorghum, maize and *Brachypodium* (Gu et al. 2009), including the identification of conserved noncoding sequences (CNSs) between maize, rice and sorghum (Salvi et al. 2007).

Maize was domesticated about 10,000 years ago from the grass teosinte (Doebley et al. 2006). The maize genome consists of about 2.5 Gbp of DNA maintained in 10 chromosomes, which are diverse due to changes in chromatin composition as a result of an increase in long terminal repeat retrotransposons (LTR retrotransposons) (SanMiguel et al. 1998). In 2005 the NSF, the United States Department of Agriculture (USDA), and the United States Department of Energy (DOE) provided 32 million dollars to the Washington University Genome Sequencing Centre, Cold Spring Harbor, the Arizona Genome Institute and Iowa State University, to undertake a maize genome sequencing project. B73 was selected as the maize variety to be sequenced and a BAC-by-BAC approach was chosen to complement the previous maize genome sequencing assessments. The draft release of the maize genome (Schnable et al. 2009) was sequenced using a minimum tiling path of BACs (n

= 16,848) and fosmid (n = 63) clones derived from integrated physical, genetic and optical maps. Shotgun sequencing of clones, to 4-6x coverage was completed and sequences manually improved. From this draft sequence, more than 32,000 genes were predicted in the genome, and 99.8 % of these were found to be on reference chromosomes. The majority of the genome space (85 %) was found to contain several hundred transposable element families, spread across the genome.

The size of the wheat (*Triticum aestivum*) genome is approximately 17,000 Mbp, much larger than related cereal genomes such as barley (*Hordeum vulgare*, 5,000 Mbp), rye (*Secale cereale*, 9,100 Mbp) and oat (*Avena sativa*, 11,000 Mbp). The size and hexaploid nature of the wheat genome create significant problems in elucidating its genome sequence. The International Wheat Genome Sequencing Consortium (IWGSC) (www.wheatgenome.org) was established in 2005 to facilitate and coordinate international efforts toward obtaining the complete sequence of the bread wheat genome. The IWGSC selected the cultivar Chinese Spring as the germplasm source for the project (Gill et al. 2004). A pilot project led by the French National Institute for Agricultural Research (INRA) was initiated in 2004 to assess the BAC fingerprinting of the largest hexaploid wheat chromosome 3B, which has been shown to carry QTLs for disease resistance and wheat quality (Börner et al. 2002; Carter et al. 2012). Using flow cytometry isolated chromosomes (Kubaláková et al. 2002; Doležel et al. 2004), a total of 68,000 BAC clones of a 3B chromosome-specific BAC library (Safar et al. 2004) was fingerprinted at the French National Sequencing Centre, Genoscope, and a minimal tiling path sequenced. This successful isolation and sequencing of chromosome 3B led extensive analysis of homoeologous gene composition and evolution, diversity, recombination and the generation of a physical map for chromosome 3B (Paux et al. 2006; Paux et al. 2008; Horvath et al. 2009; Saintenac et al. 2009; Breen et al. 2010; Hao et al. 2010; Carter et al. 2012). Similarly, chromosome specific BAC libraries have been constructed for chromosomes 1D, 4D and 6D (Janda et al. 2004). To complement these activities, individual flow sorted chromosome arms are being sequenced using Illumina shotgun sequencing (Berkman et al. 2011b; Hernandez et al. 2011; Berkman et al. 2012b; Berkman et al. 2013). Ultimately, under the International Wheat Genome Sequencing Consortium, all 34 wheat chromosome arms will be sequenced (Šafář et al. 2010). While currently these efforts have not produced a finished genome, the assemblies and syntenic builds of individual chromosome arms generated by comparison with related cereals, provides access to genomic sequence for all genes, while placing the majority of genes within an approximate order and orientation. Currently, only data for chromosome 7 is publically available at www.wheatgenome.info (Lai et al. 2012a), but this resource has already provided the basis for chromosome arm specific marker discovery (Nie et al. 2012) (Table 15.1).

15.4 Next Generation Diversity Analysis

While whole genome assemblies provide the most comprehensive resource for understanding an organism, it is currently inconceivable to attempt the *de novo* assembly of each and every plant species and variant. Following the model of human diversity

Table 15.1 Below summarises the crops sequenced to date. Though not exhaustive, it gives a glimpse into the breath of application of NGS to crop research

Species Name	Reference
BAC by BAC *Sequencing*	
Arabidopsis thalianaArabidopsis thaliana	The Arabidopsis Genome Initiative (2000)
Cajanus cajan (Pigeon pea)	Varshney et al. 2012
Lotus japonicusLotus japonicus	Sato et al. 2008
Medicago truncatulaMedicago truncatula	http://www.medicago.org/
Oryza sativa ssp. japonica (Nipponbare)	International Rice Genome Sequencing Project (2005)
Oryza sativa ssp. japonica (Nipponbare)	Barry 2001
Solanum lycopersicum (Tomato)	Tomato Genome Consortium 2012
Whole genome shotgun sequencing	
Arabidopsis lyrata (Rock cress)	Hu et al. 2011
Brachypodium distachyonBrachypodium distachyon	The International Brachypodium Initiative (2010)
Brassica rapa (Chiifu) (Chinese cabbage)	Wang et al. 2011
Carrica papaya (Papaya)	Ming et al. 2008
Cicer arietinum (Chickpea)	Varshney et al. 2013
Citrus sinensis (Sweet Orange)	http://www.phytozome.net/ orange)
Cucumis sativus (Cucumber)	Huang et al. 2009
Eucalyptus grandis (Eucalyptus)	Genome Network (EUCAGEN) (http://www.phytozome.net/)
Fragaria vesca (Woodland strawberry)	Shulaev et al. 2011
Glycine max (Soybean)	Schmutz et al. 2010b
Linum usitatissimum (Flax)	BGI (http://www.phytozome.net/)
Malus x domestica Borkh (Domesticated Apple)	Velasco et al. 2010
Manihot esculenta (Cassava)	http://www.phytozome.net/cassava
Mimulus guttatus (Monkey flower)	http://www.phytozome.net/mimulus
Oryza sativa ssp. indica (cv. 93-11 Rice)	Yu et al. 2002
Oryza sativa (Nipponbare)	Goff et al. 2002a
Phaseolus vulgaris (Common bean)	DOE-JGI (http:://www.phytozome.net/commonbean)
Populus trichocarpa (Black cottonwood)	Tuskan et al. 2006
Prunus persica (Peach)	http://www.rosaceae.org/peach/genome)
Ricinus communis (Castor bean)	Chan et al. 2010
Setaria italica (Foxtail Millet)	JCI(http://www.phytozome.net/foxtailmillet)
Solanum tuberosum (Potato)	Xu et al. 2011b
Sorghum bicolor (L.) Moench	Paterson et al. 2009
Theobroma cacao (Cocoa)	Argout et al. 2011
Vitis vinifera (ENTAV 115) (Grapevine)	Velasco et al. 2007
Vitis vinifera (PN40024) (Grapevine)	Jaillon et al. 2007
Zea mays (Palomero Toluqueno) (Corn)	Vielle-Calzada et al. 2009

analysis, after an initial set of diverse individuals was sequenced to provide a reference collection, the focus moved to study genome diversity using whole genome genotyping. With the rapid growth and plummeting cost of sequence data generation, the discovery, association and application of genome diversity information from NGS data is becoming increasingly attractive (Imelfort et al. 2009b; Berkman et al. 2012a). Such diversity studies using NGS data are not without challenges (Duran

et al. 2009b). These include the very large data volumes and the high error rates associated with this type of data. However, these challenges are being addressed, and NGS data mining is becoming a common approach for diversity analysis in a range of species (Seeb et al. 2011; Hayward et al. 2012b; Jiang et al. 2012; Kazakoff et al. 2012).

15.4.1 Reference-Based Diversity Analysis

Diversity analysis using a reference sequence is useful when a well-characterised and annotated genome sequence of a closely related species is available (Duran et al. 2009d). A good reference sequence would ideally be of very high quality, preferably a model organism for a particular genus. Limitations of reference-based diversity analysis include: low level of sequence coverage of the reference genome lowering the resolution and sensitivity with which variations can be identified, assembly and sequencing errors resulting from low-complexity regions and the alignment threshold used for mapping reads to reference. An alternative approach would negate the use of a reference sequence.

15.4.2 Non Reference-Based Diversity Analysis

Several approaches have been used to access diversity between genomes without the use of a reference. These generally involve sequence comparisons of sequence reads where diversity assessment only takes into account differences between assembled, cultivar-specific reads. This approach has been implemented using transcriptome data in AutoSNP and used to accurately call SNPs (Barker et al. 2003).

15.5 SNPs

SNPs are the ultimate form of molecular genetic markers, as a nucleotide base is the smallest unit of inheritance. A SNP represents a single nucleotide difference between two individuals at a defined location. There are three different forms of SNPs: transitions (C/T or G/A), transversions (C/G, A/T, C/A, or T/G) or small insertions/deletions (indels) (Edwards et al. 2007a; Hao et al. 2011). SNPs are direct markers as the sequence information provides the exact nature of the allelic variants. Furthermore, this sequence variation can have a major impact on how the organism develops and responds to the environment. SNPs represent the most frequent type of genetic polymorphism and may therefore provide a high density of markers near a locus of interest (Batley and Edwards 2007).

Studies of sequence diversity have recently been performed for a range of plant species. These have indicate that SNPs appear to be abundant in plant systems (Edwards et al. 2007b; Henry and Edwards 2009). SNPs are generally biallelic and only rarely triallelic. This disadvantage, when compared with multiallelic markers is compensated by their relative abundance. The low mutation rate of SNPs makes them excellent markers for the characterisation of germplasm resources (Syvanen 2001). The challenge of SNP discovery is not the identification of polymorphic nucleotide positions, but the differentiation of polymorphisms from abundant sequence errors. This is especially true for NGS data which has a higher error rate than Sanger DNA sequencing. These errors prevent the electronic mining of this data to identify potentially biologically relevant polymorphisms. A major source of sequence error comes from the balance between the need to produce the longest sequence length and the confidence that sequences are called correctly. Because of this, sequence trimming and filtering of sequence data is often performed to reduce the abundance of erroneous sequences (Kircher et al. 2011).

The identification of true polymorphisms in a background of sequence errors can be based on four methods: sequence quality values, redundancy of the polymorphism in an alignment, specificity of an allele call with a variety and co-segregation of SNPs to define a haplotype. By using the various measures of SNP confidence assessment, true SNPs may be identified with reasonable confidence from next generation DNA sequence data.

The frequency of occurrence of a polymorphism at a particular locus provides one of the best measures of confidence in the SNP representing a true polymorphism, and is referred to as the SNP redundancy score (Barker et al. 2003). By examining SNPs that have a redundancy score equal to or greater than two (two or more of the aligned sequences represent the polymorphism), the vast majority of sequencing errors are removed. True SNPs also co-segregate to define a conserved haplotype, however determining haplotypes from short-read data is challenging as sequence reads rarely include multiple SNPs. This is less of an issue for longer sequence reads from the Roche 454 system or in the application of paired reads from the Illumina or ABI SOLiD platforms.

There are many tools available for the discovery of SNPs from NGS data, but few have been designed specifically for plant populations (Appleby et al. 2009; Batley and Edwards 2009b; Duran et al. 2009b; Duran et al. 2013). One tool, based on autoSNP software (Barker et al. 2003; Batley et al. 2003) uses redundancy and haplotype co-segregation for SNP discovery. AutoSNPdb (Duran et al. 2009a) combines the SNP discovery pipeline of autoSNP with a relational database, hosting information on the polymorphisms, cultivars and gene annotations, to enable efficient mining and interrogation of the data. AutoSNPdb was originally developed for Sanger sequence data of rice, barley and Brassica (Duran et al. 2009c), but has also been applied to discover SNPs from wheat 454 data (Lai et al. 2012b) (http://autosnpdb.appliedbioinformatics.com.au/).

In one of the first examples of cereal SNP discovery from next generation genome sequence data, Barbazuk and co-workers identified more than 7,000 candidate SNPs between maize lines B73 and Mo17, with over 85 % validation rate (Barbazuk

et al. 2007). This success is particularly impressive considering the complexity of the maize genome and the early version of Roche 454 sequencing applied, which produced an average read length of only 101 bp.

The larger data volumes from the Illumina sequencing platform enable the confident discovery of very large numbers of genome wide SNPs (Imelfort et al. 2009b; Hayward et al. 2012a; Lorenc et al. 2012). More than 1 million SNPs have been identified between six inbred maize lines (Lai et al. 2010). SNPs are more prevalent in diverse germplasm. Around 3.6 million SNPs were identified by sequencing 517 rice landraces (Huang et al. 2010). This study allowed for the association of genome variation with complex traits in rice and is a model for future studies in other species. Allen and coworkers identified 14,078 putative SNPs across representative samples of UK wheat germplasm using Illumina GAIIx sequencing of cDNA libraries (Allen et al. 2011), with a portion of these SNPs validated using KASPar assays (Orrù et al. 2009). Data production for SNP discovery from large genomes remains costly and often requires the development of consortium approaches (Edwards et al. 2012).

15.5.1 Copy Number Variation

Phenotypic diversity in plants has been attributed to differences resulting from copy number variation (CNV) and SNPs. CNV refers to differences in the number of gene loci between species or cultivars. CNVs can occur at several scales, from single genes to whole genomes. CNVs may be a consequence of previous polyploidy events and are believed to be behind the phenotypic diversity of polyploids, as shown in maize inbred lines (Springer et al. 2009). In addition, CNVs have been shown to originate from gene amplification events (Xu et al. 2011a).

CNVs may account for a greater variation in nucleotide content than SNPs and are believed to be the greatest contributing factor to genetic diversity (Redon et al. 2006; Schnable et al. 2009). Approaches to study CNV and presence-absence variation (PAV) are varied and may involve the use of oligonucleotide microarrays, mRNA or DNA sequences (Springer et al. 2009). Probes from a reference sequence are hybridised with labeled DNA from cultivars of interest and an assessment of their differential hybridisation can be performed to predict copy number. A disadvantage of this approach is that the probes from the reference sequence inherently contain variations, thus confounding analysis. Due to limitations of microarrays, CNV by sequencing (CNV-seq) has been developed (Xie and Tammi 2009). In CNV-seq, DNA fragments of a reference and sample are sequenced and mapped to a template sequence. A sliding window algorithm is then used to count copy numbers per window position. PAV analysis is a relatively new approach to examine diversity between genomes, chromosomes or regions of interest. Unlike CNV, which focuses on quantitative differences of reads between individuals, PAV seeks to identify elements uniquely present or absent in one cultivar irrespective of the gene frequency.

15.6 Perspectives

As more plant genome sequences become available and the cost of sequencing drops further, attention is shifting to the analysis, interpretation and integration of sequence data through comparative studies. The biggest challenge for highly repetitive genomes remains the resolution of low-complexity regions. Although in the future this may be addressed by emerging sequencing technologies, methods that can effectively address this limitation will greatly advance the study and analysis of large complex polyploid plant genomes. Such methods focus on complexity reduction of repetitive regions. Examples include the sequencing and analysis of low-copy regions of individual wheat chromosome arms (Berkman et al. 2011a) and consensus calling of SNPs based on coverage (Azam et al. 2012), among others. With second generation sequencing technologies becoming cheaper and producing more reads per sequencing run and longer read lengths (Berkman et al. 2012a) coupled with emerging third generation single molecule sequencing technologies which promise even longer read lengths (Lieberman et al. 2010; Rasko et al. 2011) and the increased availability of diverse reference genomes, a broader comparison of germplasm and a greater understanding of plant genome evolution over the coming years will be possible. In addition, the presentation of organism-specific databases with detailed, integrated and intuitive summaries of varied comparative analysis will become more critical and offer plant breeders with highly curated and organised reference resources.

References

Allen AM, Barker GLA, Berry ST et al (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). Plant Biotech J 9:1086–1099

Appleby N, Edwards D, Batley J (2009) New technologies for ultra-high throughput genotyping in plants. In: Somers D, Langridge P, Gustafson J (eds) Plant Genomics. Humana Press (USA), pp 19–40

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Argout X, Salse J, Aury JM et al (2011) The genome of Theobroma cacao. Nat Genet 43:101–108

Azam S, Thakur V, Ruperao P et al (2012) Coverage-based consensus calling (CbCC) of short sequence reads and comparison of CbCC results to identify SNPs in chickpea (Cicer arietinum; Fabaceae), a crop species without a reference genome. Am J Bot 99:186–192

Barbazuk WB, Emrich SJ, Chen HD et al (2007) SNP discovery via 454 transcriptome sequencing. Plnat J 51:910–918

Barker G, Batley J, O'Sullivan H et al (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 19:421–422

Barry GF (2001) The use of the Monsanto draft rice genome sequence in research. Plant Physiol 125:1164–1165

Batley J, Barker G, O'Sullivan H et al (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132:84–91

Batley J, Edwards D (2007) SNP applications in plants. In: Oraguzie N, Rikkerink E, Gardiner S, De Silva H (eds) Association Mapping in Plants. Springer, New York, pp 95–102

Batley J, Edwards D (2009a) Genome sequence data: management, storage, and visualization. Biotechniques 46:333–336

Batley J, Edwards D (2009b) Mining for Single Nucleotide Polymorphism (SNP) and Simple Sequence Repeat (SSR) molecular genetic markers. In: Posada D (ed) Bioinformatics for DNA Sequence Analysis. Humana Press (USA), pp 303–322

Berkman PJ, Skarshewski A, Lorenc MT et al (2011a) Sequencing and assembly of low copy and genic regions of isolated Triticum aestivum chromosome arm 7DS. Plant Biotechnol J 9:768–775

Berkman PJ, Skarshewski A, Lorenc MT et al (2011b) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. Plant Biotechnol J 9:768–775

Berkman PJ, Lai K, Lorenc MT, Edwards D (2012a) Next-generation sequencing applications for wheat crop improvement. Am J Bot 99:365–371

Berkman PJ, Skarshewski A, Manoli S et al (2012b) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. Theor Appl Genet 124:423–432

Berkman PJ, Visendi P, Lee HC et al (2013) Dispersion and domestication shaped the genome of bread wheat. Plant Biotechnol J

Bertioli DJ, Moretzsohn MC, Madsen LH et al (2009) An analysis of synteny of Arachis with Lotus and Medicago sheds new light on the structure, stability and evolution of legume genomes. BMC Genomics 10:45

Bock R (2010) The give-and-take of DNA: horizontal gene transfer in plants. Trends Plant Sci 15:11–22

Börner AB, Schumann ES, Fürste AF et al (2002) Mapping of quantitative trait loci determining agronomic important characters in hexaploid wheat (*Triticum aestivum* L.). Theor Appl Genet 105:921–936

Breen J, Wicker T, Kong X et al (2010) A highly conserved gene island of three genes on chromosome 3B of hexaploid wheat: diverse gene function and genomic structure maintained in a tightly linked block. Bmc Plant Biol 10:98

Brunner AM, Busov VB, Strauss SH (2004) Poplar genome sequence: functional genomics in an ecologically dominant plant species. Trends Plant Sci 9:49–56

Cannon SB, Sterck L, Rombauts S et al (2006) Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. Proc Natl Acad Sci U S A 103:14959–14964

Carter A, Garland-Campbell K, Morris C, Kidwell K (2012) Chromosomes 3B and 4D are associated with several milling and baking quality traits in a soft white spring wheat (*Triticum aestivum* L.) population. Theor Appl Genet 124:1079–1096

Chan AP, Crabtree J, Zhao Q et al (2010) Draft genome sequence of the oilseed species Ricinus communis. Nat Biotechnol 28:951–956

Consortium TTG (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641

Cuesta–Marcos A, Szucs P, Close TJ et al (2010) Genome-wide SNPs and re-sequencing of growth habit and inflorescence genes in barley: implications for association mapping in germplasm arrays varying in size and structure. BMC Genomics 11:707

Darracq A, Varre JS, Touzet P (2010) A scenario of mitochondrial genome evolution in maize based on rearrangement events. BMC Genomics 11:233

Doležel J, Kubalakova M, Bartos J, Macas J (2004) Flow cytogenetics and plant genome mapping. Chromosome Res 12:77–91

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. Cell 127:1309–1321

Dong CH, Li C, Yan XH et al (2011) Gene expression profiling of Sinapis alba leaves under drought stress and rewatering growth conditions with Illumina deep sequencing. Mol Biol Rep 39:5851–7

Duran C, Appleby N, Clark T et al (2009a) AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. Nucleic Acids Res 37:D951–953

Duran C, Appleby N, Edwards D, Batley J (2009b) Molecular genetic markers: discovery, applications, data storage and visualisation. Curr Bioinform 4:16–27

Duran C, Appleby N, Vardy M et al (2009c) Single nucleotide polymorphism discovery in barley using autoSNPdb. Plant Biotechnol J 7:326–333
Duran C, Edwards D, Batley J (2009d) Genetic maps and the use of synteny. In: Somers D, Langridge P, Gustafson J (eds) Plant Genomics. Humana Press (USA), pp 41–56
Duran C, Eales D, Marshall D et al (2010) Future tools for association mapping in crop plants. Genome 53:1017–1023
Duran C, Singhania R, Raman H et al (2013) Predicting polymorphic EST-SSRs in silico. Mol Ecol Resour 13:538–45
Edwards D, Forster JW, Chagné D, Batley J (2007a) What are SNPs? In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN (eds) Association Mapping in Plants Springer NY, pp 41–52
Edwards D, Forster JW, Cogan NOI et al (2007b) Single Nucleotide Polymorphism Discovery. In: Oraguzie N, Rikkerink E, Gardiner S, De Silva H (eds) Association Mapping in Plants. Springer New York, pp 53–76
Edwards D, Hansen D, Stajich J (2009) DNA Sequence Databases. In: Edwards D, Hanson D, Stajich J (eds) Applied Bioinformatics. Springer (USA), pp 1–11
Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. Plant Biotechnol J 7:1–8
Edwards D, Wang X (2012) Genome Sequencing Initiatives. In: Edwards D, Parkin IAP, Batley J (eds) Genetics, Genomics and Breeding of Oilseed Brassicas. Science Publishers Inc., New Hampshire, (USA), pp 152–157
Edwards D, Wilcox S, Barrero RA et al (2012) Bread matters: a national initiative to profile the genetic diversity of Australian wheat. Plant Biotechnol J 10:703–708
Edwards D, Batley J, Snowdon R (2013) Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet 126:1–11
Feuillet C, Leach JE, Rogers J et al (2011) Crop genome sequencing: lessons and rationales. Trends Plant Sci 16:77–88
George J, Sawbridge TI, Cogan NO et al (2008) Comparison of genome structure between white clover and Medicago truncatula supports homoeologous group nomenclature based on conserved synteny. Genome 51:905–911
Gill BS, Appels R, Botha-Oberholster AM et al (2004) A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. Genetics 168:1087–1096
Goff SA, Ricke D, Lan TH et al (2002a) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Sci 296:92–100
Goff SA, Ricke D, Lan TH et al (2002b) A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). Sci 296:92–100
Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful tool for evolutionary analysis1. Am J Bot 99:312–319
Gu YQ, Ma Y, Huo N et al (2009) A BAC-based physical map of Brachypodium distachyon and its comparative analysis with rice and wheat. BMC Genomics 10:496
Hall BG, Ehrlich GD, Hu FZ (2010) Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiol 156:1060–1068
Hao C, Perretant M, Choulet F et al (2010) Genetic diversity and linkage disequilibrium studies on a 3.1-Mb genomic region of chromosome 3B in European and Asian bread wheat (*Triticum aestivum* L.) populations. Theor Appl Genet 121:1209–1225
Hao Z, Li X, Xie C et al (2011) Identification of functional genetic variations underlying drought tolerance in maize using SNP markers. J integrat plant biol 53:641–652
Hayward A, Dalton-Morgan J, Mason A et al (2012a) SNP discovery and applications in *Brassica napus*. J Plant Biotechnol (in press)
Hayward A, Vighnesh G, Delay C et al (2012b) Second-generation sequencing for gene discovery in the Brassicaceae. Plant Biotechnol J 10:750–759
Henry R, Edwards K (2009) New tools for single nucleotide polymorphism (SNP) discovery and analysis accelerating plant biotechnology. Plant Biotechnol J 7:311

Hernandez P, Martis M, Dorado G et al (2011) NGS and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. Plant J 69:377–386
Horvath A, Didier A, Koenig J et al (2009) Analysis of diversity and linkage disequilibrium along chromosome 3B of bread wheat (*Triticum aestivum* L.). Theor Appl Genet 119:1523–1537
Hu TT, Pattyn P, Bakker EG et al (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet 43:476–481
Huang S, Li R, Zhang Z et al (2009) The genome of the cucumber, Cucumis sativus L. Nat Genet 41:1275–1281
Huang XH, Wei XH, Sang T et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–976
Hyten DL, Song Q, Zhu Y et al (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci U S A 103:16666–16671
Imelfort M, Edwards D (2009) De novo sequencing of plant genomes using second-generation technologies. Brief Bioinform 10:609–618
Imelfort M, Batley J, Grimmond S, Edwards D (2009a) Genome sequencing approaches and successes. In: Somers D, Langridge P, Gustafson J (eds) Plant Genomics. Humana Press (USA), pp 345–358
Imelfort M, Duran C, Batley J, Edwards D (2009b) Discovering genetic polymorphisms in next-generation sequencing data. Plant Biotechnol J 7:312–317
International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768
International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800
Jaillon O, Aury JM, Noel B et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467
Janda J, Bartos J, Safar J et al (2004) Construction of a subgenomic BAC library specific for chromosomes 1D, 4D and 6D of hexaploid wheat. Theor Appl Genet 109:1337–1345
Jiang Q, Yen SH, Stiller J et al (2012) Diversity Analysis of the Tree Legume *Pongamia pinnata* using PISSRs (Pongamia Inter-Simple Sequence Repeats). J Plant Genome Sci (in press)
Kazakoff SH, Imelfort M, Edwards D et al (2012) Capturing the Biofuel Wellhead and Powerhouse: The Chloroplast and Mitochondrial Genomes of the Leguminous Feedstock Tree *Pongamia pinnata*. Plos One 7:51687
Kim SY, Lohmueller KE, Albrechtsen A et al (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinformatics 12:231
Kircher M, Heyn P, Kelso J (2011) Addressing challenges in the production and analysis of illumina sequencing data. BMC Genomics 12:382
Kubaláková M, Vrána J, Ĉ;íhalíková J et al (2002) Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). Theor Appl Genet 104:1362–1372
Lai JS, Li RQ, Xu X et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42:1027–1158
Lai K, Berkman PJ, Lorenc MT et al (2012a) WheatGenome.info: An integrated database and portal for wheat genome information. Plant Cell Physiol 53:1–7
Lai K, Duran C, Berkman PJ et al (2012b) Single nucleotide polymorphism discovery from wheat next-generation sequence data. Plant Biotechnol J 10:743–749
Lai K, Lorenc MT, Edwards D (2012c) Genomic databases for crop improvement. Agronomy 2:62–73
Laing C, Buchanan C, Taboada EN et al (2010) Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics 11:461
Lam HM, Xu X, Liu X et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053–1041
Lee GA, Crawford GW, Liu L et al (2011a) Archaeological soybean (Glycine max) in East Asia: does size matter? PLoS One 6:26720

Lee H, Lai K, Lorenc MT et al (2011b) Bioinformatics tools and databases for analysis of next generation sequence data. Briefings in Functional Genomics (in press)
Lee H, Lai K, Lorenc MT et al (2012) Bioinformatics tools and databases for analysis of next generation sequence data. Brief Funct Genomics 2:12–24
Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993
Li R, Li Y, Zheng H et al (2010a) Building the sequence map of the human pan-genome. Nat Biotechnol 28:57–63
Li YH, Li W, Zhang C et al (2010b) Genetic diversity in domesticated soybean (Glycine max) and its wild progenitor (Glycine soja) for simple sequence repeat and single-nucleotide polymorphism loci. The New phytologist 188:242–253
Lieberman KR, Cherf GM, Doody MJ et al (2010) Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. J Am Chem Soc 132:17961–17972
Lodhi MA, Daly MJ, Ye GN et al (1995) A molecular marker based linkage map of Vitis. Genome 38:786–794
Lorenc MT, Hayashi S, Stiller J et al (2012) Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. Biology 1:370–382
Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380
Marshall D, Hayward A, Eales D et al (2010) Targeted identification of genomic regions using db. Plant Methods 6:19
Ming R, Hou S, Feng Y et al (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452:991–996
Moore G, Devos KM, Wang Z, Gale MD (1995) Cereal Genome Evolution – Grasses, Line up and Form a Circle. Curr Biol 5:737–739
Mun J-H, Kwon S-J, Seol Y-J et al (2010) Sequence and structure of *Brassica rapa* chromosome A3. Genome Biol 11:94
Nie X, Li B, Wang L et al (2012) Development of chromosome-arm-specific microsatellite markers in *Triticum aestivum* (Poaceae) using NGS technology. Am J Bot 99:369–371
Orrù L, Catillo G, Napolitano F et al (2009) Characterization of a SNPs panel for meat traceability in six cattle breeds. Food Control 20:856–860
Pang X, Luo H, Sun C (2012) Assessing the potential of candidate DNA barcodes for identifying non-flowering seed plants. Plant Biol 14:839–844
Paterson AH, Bowers JE, Bruggmann R et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457:551–556
Paux E, Roger D, Badaeva E et al (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J 48:463–474
Paux E, Sourdille P, Salse J et al (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322:101–104
Rasko DA, Webster DR, Sahl JW et al (2011) Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med 365:709–717
Redon R, Ishikawa S, Fitch KR et al (2006) Global variation in copy number in the human genome. Nature 444:444–454
Safar J, Bartos J, Janda J et al (2004) Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. Plant J 39:960–968
Šafář J, Šimková H, Kubaláková M et al (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. Cytogenet Genome Res 129:211–223
Saintenac C, Falque M, Martin OC et al (2009) Detailed Recombination Studies Along Chromosome 3B Provide New Insights on Crossover Distribution in Wheat (*Triticum aestivum* L.). Genetics 181:393–403
Salvi S, Sponza G, Morgante M et al (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci 104:11376–11381

SanMiguel P, Gaut BS, Tikhonov A et al (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45
Sato S, Nakamura Y, Kaneko T et al (2008) Genome structure of the legume, Lotus japonicus. DNA Res 15:227–239
Schlueter JA, Scheffler BE, Jackson S, Shoemaker RC (2008) Fractionation of synteny in a genomic region containing tandemly duplicated genes across glycine max, *Medicago truncatula*, and *Arabidopsis thaliana*. J Hered 99:390–395
Schmutz J, Cannon SB, Schlueter J et al (2010a) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183
Schmutz J, Cannon SB, Schlueter J et al (2010b) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183
Schnable PS, Ware D, Fulton RS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115
Seeb JE, Carvalho G, Hauser L et al (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Mol Ecol Resour 11(1):1–8
Shibata D (2005) Genome sequencing and functional genomics approaches in tomato. J Gen Plant Pathol 71:1–7
Shulaev V, Korban SS, Sosinski B et al (2008) Multiple models for Rosaceae genomics. Plant Physiol 147:985–1003
Shulaev V, Sargent DJ, Crowhurst RN et al (2011) The genome of woodland strawberry (Fragaria vesca). Nat Genet 43:109–116
Springer NM, Ying K, Fu Y et al (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5:e1000734
Syvanen AC (2001) Accessing genetic variation: Genotyping single nucleotide polymorphisms. Nat Rev Genet 2:930–942
Tetz VV (2005) The pangenome concept: a unifying view of genetic information. Med Sci Monitor 11:HY24–29
Tuskan GA, DiFazio SP, Teichmann T (2004) Poplar genomics is getting popular: The impact of the poplar genome project on tree research. Plant Biol 6:2–4
Tuskan GA, Difazio S, Jansson S et al (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313:1596–1604
Varshney RK, Chen W, Li Y et al (2012) Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83–89
Varshney RK, Song C, Saxena RK et al (2013) Draft genome sequence of kabuli chickpea (*Cicer arietinum*): genetic structure and breeding constraints for crop improvement. Nat Biotechnol
Velasco R, Zharkikh A, Troggio M et al (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One 2:e1326
Velasco R, Zharkikh A, Affourtit J et al (2010) The genome of the domesticated apple (Malus x domestica Borkh.). Nat Genet 42:833–839
Vielle-Calzada JP, Martinez delaVO, Hernandez-Guzman G et al (2009) The Palomero genome suggests metal effects on domestication. Science 326:1078
Wang X, Wang H, Wang J et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1157
Williams-Carrier R, Stiffler N, Belcher S et al (2010) Use of Illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy Mutator lines of maize. Plant J 63:167–177
Wu DD, Zhang YP (2011) Eukaryotic origin of a metabolic pathway in virus by horizontal gene transfer. Genomics 98:367–369
Xie C, Tammi MT (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics 10:80
Xu JH, Bennetzen JL, Messing J (2011a) Dynamic Gene Copy Number Variation in Collinear Regions of Grass Genomes. Mol Biol Evol

Xu X, Pan S, Cheng S et al (2011b) Genome sequence and analysis of the tuber crop potato. Nature 475:189–195

Xu X, Pan S, Cheng S et al (2011c) Genome sequence and analysis of the tuber crop potato. Nature 475:189–194

Xu X, Liu X, Ge S et al (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotech 30:105–111

Young ND, Debelle F, Oldroyd GED et al (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature advance online publication

Yu J, Hu SN, Wang J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). Science 296:79–92

Yue WF, Du M, Zhu MJ (2012) High Temperature in Combination with UV Irradiation Enhances Horizontal Transfer of stx2 Gene from E. coli O157:H7 to Non-Pathogenic E. coli. PLoS One 7:e31308

Zhang Z, Belcram H, Gornicki P et al (2011) Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. Proc Natl Acad Sci U S A 108:18737–18742

Zharkikh A, Troggio M, Pruss D et al (2008) Sequencing and assembly of highly heterozygous genome of Vitis vinifera L. cv Pinot Noir: problems and solutions. J Biotechnol 136:38–43