Compendium of Plant Genomes
Series Editor: Chittaranjan Kole

Xiaowu Wang · Chittaranjan Kole *Editors*

The *Brassica rapa* Genome

Springer

Compendium of Plant Genomes

Series editor

Chittaranjan Kole
Mohanpur, West Bengal
India

More information about this series at http://www.springer.com/series/11805

Xiaowu Wang · Chittaranjan Kole
Editors

The *Brassica rapa* Genome

Springer

Editors
Xiaowu Wang
Institute of Vegetables and Flowers
Chinese Academy of Agricultural Sciences
Beijing
China

Chittaranjan Kole
Department of Genetics and Plant Breeding
Bidhan Chandra Krishi Viswavidyalaya
Mohanpur, West Bengal
India

ISSN 2199-4781 ISSN 2199-479X (electronic)
Compendium of Plant Genomes
ISBN 978-3-662-47900-1 ISBN 978-3-662-47901-8 (eBook)
DOI 10.1007/978-3-662-47901-8

Library of Congress Control Number: 2015946594

Springer Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer-Verlag GmbH Berlin Heidelberg is part of Springer Science+Business Media (www.springer.com)

Preface

Brassica rapa is a crop species of the genus *Brassia*, which belongs to the plant family Brassicaceae (Cruciferae). It has ten chromosomes ($2n = 20$) and is widely cultivated as oil and vegetables crop across the world. It is also one of the basic species of the other two widely cultivated crops *Brassica napus* and *Brassica juncea.* Owning to its close relationship to the model species, *Arabidopsis* and relatively smaller genome size among the *Brassica* species, it was identified as the first *Brassica* species to be sequenced by the BrGSP. The consortium developed a plan based on the experience of *Arabidopsis* genome sequencing by using BAC to BAC strategy and a complete chromosome A03 was assembled by a Korea team using this traditional sequencing technology. However, the project was finally completed by using the second generation sequencing technology, which adopted completely different assembling strategies from that of traditional Sanger sequencing. The reporting of the whole genome sequence of *B. rapa* in 2011 was a milestone in the field of *Brassica* research. It was the first example of a complex genome with recent whole genome duplication (WGT) that was assembled using short sequencing reads generated by second generation sequencing technology. The assembly covered ~98 % of the gene region or 60 % of the ~500 Mb genome of *B. rapa* and the 40 % unassembled sequences were mainly due to the highly repetitive nature of this portion of the genome. With this genome assembly, dedicated tools were developed not only to compare—at the whole genome level—with the extensively studied *Arabidopsis* genome, but also to transfer the rich research information from the model plant to a cultivated crop. We thus could investigate the whole genome detail of the WGT event in the *Brassica* genome now. The subgenomes of *B. rapa* are not equally important. One of the three subgenomes generated from WGT has not only more genes, but also a higher gene expression level over the other two. Such a subgenome dominance phenomenon leads to a hypothesis that the WGT event was occurred through a two-step duplication process. Although the genome of *Arabidopsis* was frequently used to compare with the *Brassica* genome, it was inferred from the *B. rapa* genome that the direct diploid ancestor of *Brassica* has a genome structure of the translocation Proto-Calepineae Karyotype (tPCK), which is the same as for the extant species *Schrenkiella parvula.* The WGT created large number of duplicated genes which may evolve interactively and even concertedly through homoeologous recombination. It was detected that 8 %

of these duplicated genes were converted by one another after the divergence of *B. rapa* and *Brassica oleracea*. The release of the *B. rapa* genome reference promoted genome evolution research and facilitates gene discovery and functional studies, as well as the breeding of *Brassica* crops. However, the quality of the genome assembly still has room for improvement and the accuracy of the annotation will be increased with the accumulation of more mRNA-Seq data, and more tools and resources need to be established to help the breeders using the genomic data.

Xiaowu Wang

Contents

1 Economic/Academic Importance of *Brassica rapa* 1
Rifei Sun

2 Background History of the National and International *Brassica rapa* Genome Sequencing Initiatives 17
Ian Bancroft and Xiaowu Wang

3 Genomic Resources and Physical Mapping of the *B. rapa* Genome 25
Jeong-Hwan Mun, Hee-Ju Yu and Beom-Seok Park

4 De Novo Genome Assembly of Next-Generation Sequencing Data 41
Min Liu, Dongyuan Liu and Hongkun Zheng

5 Crop Genome Annotation: A Case Study for the *Brassica rapa* Genome 53
Erli Pang, Huifeng Cao, Bowen Zhang and Kui Lin

6 Miniature Transposable Elements (mTEs): Impacts and Uses in the *Brassica* Genome 65
Perumal Sampath, Jonghoon Lee, Feng Cheng, Xiaowu Wang and Tae-Jin Yang

7 Genomic Survey of the Hidden Components of the *B. rapa* Genome 83
Nomar Espinosa Waminal, Sampath Perumal, Ki-Byung Lim, Beom-Seok Park, Hyun Hee Kim and Tae-Jin Yang

8 The Common Ancestral Genome of the *Brassica* Species . . . 97
Feng Cheng, Martin A. Lysak, Terezie Mandáková and Xiaowu Wang

9 Genome Evolution after Whole Genome Triplication: the Subgenome Dominance in *Brassica rapa* 107
Feng Cheng, Jian Wu, Bo Liu and Xiaowu Wang

10 Genome Triplication Drove the Diversification of *Brassica* Plants 115
Feng Cheng, Jian Wu, Jianli Liang and Xiaowu Wang

11 Comparative Analysis of Gene Conversion Between Duplicated Regions in *Brassica rapa* and *B. oleracea* Genomes 121
Jinpeng Wang, Hui Guo, Dianchuan Jin, Xiyin Wang and Andrew H. Paterson

12 Molecular Mapping and Cloning of Genes and QTLs in *Brassica rapa* 131
Guusje Bonnema

13 Impact Molecular Marker and Genomics-Led Technologies on *Brassica* Breeding 145
Jianjun Zhao

14 The Database for *Brassica* Genome Studies—BRAD. 155
Feng Cheng, Xiaobo Wang, Jian Wu and Xiaowu Wang

15 Future Prospects. 165
Xiaowu Wang and Feng Cheng

Economic/Academic Importance of *Brassica rapa*

1

Rifei Sun

Abstract

Brassica rapa is a crop species of economic importance. It is cultivated worldwide as oil and vegetable crops. It belongs to the genus *Brassica*, tribe *Brassiceae* of the family *Brassicaceae*. The genus *Brassica* includes many important crops. Among them, relationship of six species formed the model of U's triangle, with three basic diploid species *B. rapa* (A genome, $n = 10$), *Brassica oleracea* (C genome, $n = 9$) and *B. nigra* (B genome, $n = 8$) gave rise to three amphidiploid species *Brassica napus* (AC genome, $n = 19$), *B. juncea* (AB genome, $n = 18$) and *B. carinata* (BC genome, $n = 17$). *Brassica* species are rich in diversities, which are sufficient resources for plenty of crop morphotypes. In this chapter, as the start of the book, we will introduce the origin, evolution history, diversified morphotypes, and features for breeding application of *B. rapa*, as well as its investigation progresses in disease resistances, quality improvements focusing on glucosinolate contents and creatings of new materials.

1.1 Introduction

Brassica rapa L. is a species of the genus *Brassia*, cruciferae family with chromosome $2n = 20$ and widely cultivated as oil and vegetable crops. The most important crops include: rapeseed or canola, turnip, Chinese cabbage. The species is also a widespread naturalized weed throughout temperate North America and elsewhere. The vast variations in this species are suitable materials for fundamental genetic, cytogenetic and genome research, and for using in applied plant breeding.

The species name *Brassica rapa* L. was published by Linnaeus (1753). *B. rapa* related to the turnip, an important vegetable and fodder crop in that era, and *Brassica campestris* L., another species published simultaneously by Linnaeus (1753), related to the weedy annual

R. Sun (✉)
Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China
e-mail: sunrifei@caas.cn

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_1

Brassica plants growing abundantly along roads and arable fields all over Europe. Metzger (1833) was the first author to combine both species and chose *B. rapa* as the name for the combined species. Therefore all those classification and nomenclature of *B. campestris* should be changed to *B. rapa.*

1.2 Origins of *Brassica rapa*

B. rapa is considered to have originated in the Mediterranean areas and Central Asia (Prakash and Hinata 1980), where this species is mainly recognized as the turnip. Based on comparative morphology, Sun (1946) proposed the existence of two races, the Western race comprising oilseed forms and turnip, and the Eastern race comprising vegetable forms. Molecular marker research (Song et al. 1998; Zhao et al. 2005; Takuno et al. 2007) substantiate the existence of two races. The most likely explanation is that these groups represent two independent centers of origin with Europe being the primary center for oleiferous forms. Turnip later traveled further eastward through the Middle East. Once the primitive or semi-differentiated types entered India and China, they developed toward oilseed forms in India and toward leafy forms in China, primarily south China. China is also the center of origin of a unique form of ssp. oleifera (oilseed form) (Li 1981a).

1.3 Biosystematic Relationship Among the *Brassica* Species

The *Brassica* genus comprises six species, each with considerable morphological variation. Through interspecific hybridizations in all possible combinations, three basic diploid species *Brassica rapa* (A genome, $n = 10$), *Brassica oleracea* (C genome, $n = 9$) and *B. nigra* (B genome, $n = 8$) gave rise to three amphidiploid species *Brassica napus* (AC genome, $n = 19$), *B. juncea* (AB genome, $n = 18$) and *B. carinata* (BC genome, $n = 17$) (Fig. 1.1) (U1935). In these species, *B. oleracea*, *B. rapa,* and *B. juncea* are highly polymorphic. Several varieties of *B. oleracea* provide vegetables [e.g., cabbage (var. *capitata*), cauliflower (var. *botrytis*), and broccoli (var. *italica*)]; fodder viz. (var. *acephala*); etc. Forms of *B. rapa* serve as sources of oilseed (var. *oleifera*), vegetables [e.g., Chinese cabbage (ssp. *pekinensis*), ssp. *chinensis*, *narinosa,* etc.]; and fodder, viz. turnip (ssp. *rapifera*). *B. juncea* (Indian mustard) is the predominant oilseed species of the Indian subcontinent. It is also widely grown as a vegetable in China, in addition to oilseed forms. Rapeseed (*B. napus*) is extensively cultivated in Canada, Europe, China, and Australia for its oil. Seeds of black mustard (*B. nigra*) are used as condiments, while cultivation of Ethiopian mustard (*B. carinata*) is very limited and is a dual-purpose crop used both as oilseed and fodder.

1.4 Crops of *Brassica rapa*

The cultivated forms of *B. rapa* consist of morphologically distinct infraspecific types, which are distinguished by the morphology of their edible or useful parts, such as swollen roots or fleshy leaves for vegetable use, or plentiful seed production for seed oil use. There are vast variations in *B. rapa*. Generally, the species could be subdivided into several subspecies.

***Brassica rapa* L. subsp. *rapifera* L.** This subspecies is commonly called turnip and thought to be an original type of *B. rapa* and is cultivated worldwide, where it has a wide genetic diversity in terms of root shape, size, and color. This crop is sometimes used as fodder in Europe. They grow best in a cool climate.

***Brassica rapa* L. subsp. *chinensis* (L.) Hanelt**. This subspecies is commonly called Pak-choi and also referred to as nonheading Chinese cabbage. Numerous nonheading leafy

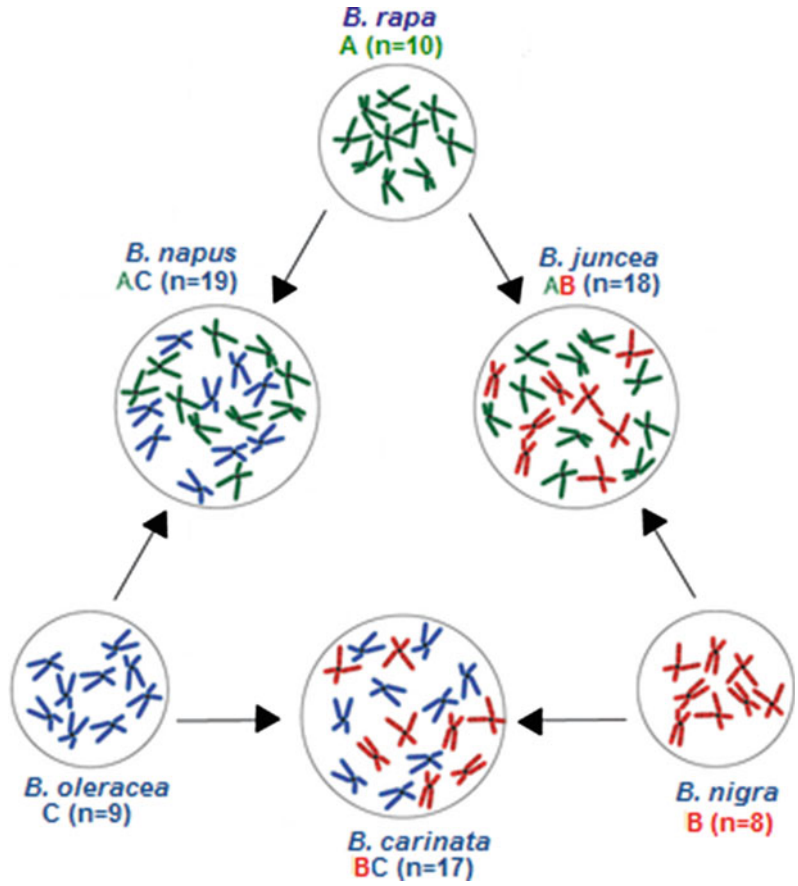

Fig. 1.1 Relationships between diploid and amphidiploid *Brassica* species (the 'triangle of U'; U, 1935)

vegetables with wide morphological variation are known in China, Korea, and Japan. Some of these varieties are only known locally. There are several recognized types: narrow petioles with round cross section versus wide petioles with flat cross section, white versus green petioles, variation in size, foliage color, etc. Varietal differentiations in this subspecies have been remarkably more pronounced than in any other leafy vegetables of the genus *Brassica*.

***Brassica rapa* L. subsp. *chinensis* (L.) Hanelt var. *parachinensis* (L.H. Bailey) Hanelt**. This var is commonly called Caixin and referred to flowering Chinese cabbage and considered as a derivative of subsp. *chinensis* because of similarities in petiole morphology. However, it readily bolts and branches profusely from the leaf axils. Flower buds with growing stems and leaves are used as cooked vegetables in southern and central China, in southeastern Asian countries such as Indonesia, Malaysia, Thailand and Vietnam, etc.

Brassica rapa* L. subsp. *chinensis* (L.) Hanelt var. *purpuraria* (L.H. Bailey) *Kitam. This var is commonly called Zicaitai and considered as a derivative of subsp. *chinensis* because of similarities in petiole morphology. Purple branches of stems flower buds and leaves are used as cooked vegetables in central China.

***Brassica rapa* L. subsp. *narinosa* (L.H. Bailey) Hanelt**. This subspecies is commonly called Taitsai and known as Chinese flat cabbage. The plant is generally low, compact and producing clusters of thick, often wrinkled leaves with broad, white petioles. This subspecies is well known for its cold tolerance. The crisp leaves and thick petioles are excellent for preparation as a boiled vegetable.

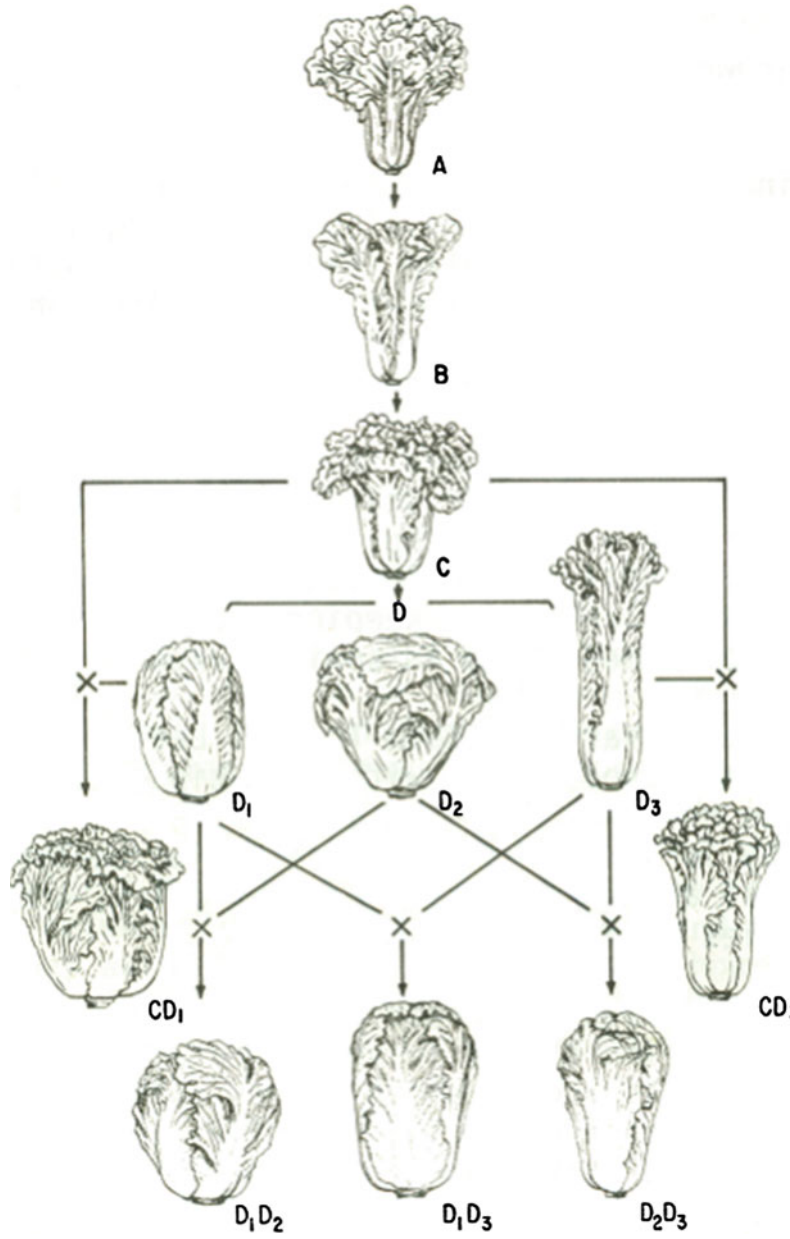

Fig. 1.2 The evolution of Chinese cabbage. *A* var dissoluta, *B* var infarcta, *C* var laxa, *D* var cephaiata, *D1* f ovata, *D2* f depressa, *D3* f cylindrica, *CD1* var laxa × f ovata, *CD3* var laxa f cylindrica, *D1D2* f ovate × f depressa, *D1D3* f ovate × f cylindrica, *D2D3* f depressa × f cylindrical (Li 1981b)

***Brassica rapa* L. subsp. *pekinensis* (Lour.) Hanelt**. This subspecies is commonly called Chinese cabbage. It comprises mainly heading types of Chinese cabbage. It is a native of China; it probably evolved from the natural crossing of Pak Choi (*Brassica rapa* L. subsp. *chinensis* (L.) Hanelt), which was cultivated in southern China for >1600 years, and turnip (*Brassica rapa* L. subsp. *rapa* L.), which was grown in northern China. Much of its variety differentiation took place in China during the past 600 years. An illustration of the headed shape of Chinese cabbage with wrapping leaves was first recorded in China in 1753. Chinese cabbage is now grown worldwide. Their head shapes vary in degree of compactness and may be divided into loose, semiheading and completely heading types. Further variations in head shape can be noted,

e.g. long, short, tapered, round or fliat top, wrapped-over or jointed-uip leaves, etc. (Fig. 1.2).

***Brassica rapa* L. subsp. *nipposinica* (L. H. Bailey) Hanelt**. This subspecies is one of the unique vegetables of Japan. This subspecies is typically characterized by an excess of basal branches and leaves. Two types are distinguished: 'Mizuna', with deeply dissected, bipinnate leaves–originally found as a local vegetable in Kyoto in central Japan, it is now widely grown throughout Japan, and has sometimes been found in China and Korea, and 'Mibuaa', with slender entire leaves. The formation of subsp. japonica is thought to have involved some introgression from the mustard group, *B. juncea*. This vegetable resembles *B. juncea* in petiole and sililque conformation and seed size; the flower stalks of both are not enclosed by leave.

***Brassica rapa* L. subsp. *oleifera* (DC.) Metzg**. This subspecies is an oil-yielding crop. The plants have many branches, with exceedingly well-developed siliques and seed. The oil seed types (ssp. *oleifera*) fall into different subgroups based on their growth habit (spring and winter types). One of them is possibly developed from Pak choi in southern China shows strong branching. In India oil seed types are divides into yellow Sarson (*Brassica rapa* L. subsp. *trilocularis* (Roxb.) Hanelt) and brown sarson (*Brassica rapa* L. subsp. *dichotoma* (Roxb.) Hanelt).

***Brassica rapa* L. subsp. *sylvestris* L. Janch. var. *esculenta* Hort**. This var is commonly called Brocoletto, locally known as friariello. It originates from southern Italy, and a similar crop is also grown in Portugal. The edible parts are the leaves, buds, and stems. The buds somewhat resemble broccoli. In both of these areas, it is highly likely that brocoletto would have been cultivated alongside *B. oleracea* crops such as kales, cabbages and broccolis, providing the necessary opportunities for inter-specific crosses to occur. Further work on the genetic diversity of the brocoletto crop type is required to verify such speculation.

1.5 Production of Doubled Haploids in *Brassica rapa*

Doubled-haploidy (DH) plants can be produced in a single generation through microspore culture, anther culture, and ovary/ovule culture. They could be further used in mutation breeding, genetic engineering, in vitro screening for complex traits like drought, cold, and salinity tolerance, and for developing mapping populations for linkage maps using molecular markers.

A high frequency spontaneous production of doubled haploid plants in microspore culture of *Brassica rapa* has been established. As Lichter (1982) first accounted the success of obtaining haploid plants from microspore culture in *B. napus*, microspore culture has been reported in *B. rapa*, including Chinese cabbage (*B. rapa* L. subsp. *pekinensis*, Sato et al. 1989a, b, 2002; Baillie et al. 1992), Pak Choi (*B. rapa* L. subsp. *chinensis*, Cao et al. 1994; Gu et al. 2003), Chinese flowering cabbage (*B. rapa* L. subsp. *chinensis* var. *parachinensis*, Wong et al. 1996) and Purple flowering stalk (*B. rapa* L. subsp. *chinensis* var. *purpurea*, Wang et al. 2009). Zhang et al. (2012) compared with the longer incubation period (72 h), heat shock treatment at 33 °C for 24 or 48 h was more beneficial for microspore embryogenesis. Microspore embryogenesis was improved with the addition of 6-benzylaminopurine and naphthalene acetic acid to the NLN medium. The highest frequency of microspore embryogenesis was achieved in 'SH-8' (21.1 embryos/bud), whereas 'SH-6' had the best regeneration ability (67.4 %). In total, 1112 regenerated plants were produced from 2025 embryos in a series of experiments using different modifications of microspore culture technique. The average spontaneous doubling frequency of the different genotypes was 57.5 %.

1.6 Reproduction Biology

1.6.1 Self-incompatibility

Self-incompatibility (SI) is an elaborate breeding system for securing outcrossing and maximum recombination in angiosperms. SI was controlled by S locus, which has multiple alleles in self-recognition reaction. Until now the sporophytic SI system has been successfully used for the seed production of F_1 hybrid cultivars in the cruciferous vegetables, such as cabbage, Chinese cabbage, broccoli, cauliflower and radish.

The specificity of pollen-stigma interaction in self-incompatibility is controlled by multiple alleles of the S locus. More than 30 S alleles and more than 40 S alleles have been identified in *Brassica rapa* (Nou et al. 1993) and *B. oleracea* (Ockendon 1974), respectively. During the last decade, the S-locus region has been dissected to identify highly polymorphic SI genes–SLG (a gene for S-locus glycoprotein), SRK (a gene for S-receptorkinase), and SP11/SCR (a gene for S-locus protein 11 or S-locus cysteine-rich protein) (Watanabe et al. 2001). SLG encodes a secreted glycoprotein (Takayama et al. 1987); SRK encodes a transmembrane receptor like kinase, which consists of an extracellular SLG-likedomain (S domain), a transmembrane domain, and acytoplasmic kinase domain (Stein et al. 1991); and SP11/SCR encodes a small cysteine-rich protein (Schopfer et al. 1999; Suzuki et al. 1999). Based on the sequence diversity of the SI genes, S haplotypes are classified into two classes–I and II. The amino acid sequence similarity of SLG and SRK is almost 65 % between classes and 80–90 % within classes (Watanabe et al. 2001). Class-I Shaplotypes are dominant over those of class II on the pollen side (Hatakeyama et al. 1998).

1.6.2 Male Sterility

Although the sporophytic SI system is widely adopted, there are many concomitant problems, such as parent reproduction, inbred depression, involved in its use. These problems translate into high seed costs to growers. Plant breeders have, therefore, been greatly interested in finding alternative schemes for hybrid production—ideally, one that is more stable, simple and straightforward, and at the same time more economical than the sporophytic SI system. The emphasis on exploitation of male sterility as a possible alternative to self-incompatibility has gained momentum in recent years.

1.6.2.1 Genic Male Sterility

Genic male sterility (GMS) is manifested through the nuclear genes inhibiting the normal development of anthers and pollens. Several genic male sterile plants were found in *Brassica rapa* such as brown sarson and yellow sarson (Chowdhury and Das 1966; Das and Pandey 1961) and Chinese cabbage (Niu et al. 1980). F_1 hybrid Chinese cabbage has been produced in China using single recessive genic male steriles (Niu et al. 1980). However, hybrid seed production, through genic male sterility, is an inefficient system because a pure population of genic male-sterile plant cannot be produced; hand-roguing of heterozygous male-fertile plants from female lines is required prior to the onset of bee pollination. Even assuming that an effective system for selecting out fertile plants from female lines can be found, e.g. seedling genetic marker closely linked to male-sterile gene, the system does not provide for a cost-effective seed production since only half of the female plants are used for producing hybrid seed. The relative interest, therefore, in the use of genic male sterility for hybrid production is limited.

Zhang et al. (1990) bred two types of genic male sterile systems: in Type I the male sterility was seems to be controlled by monogenic recessive gene, and in Type II the male sterility was seems to be controlled by monogenic dominant male gene. A 100 % male sterile population was obtained when male sterile plants from Type I were pollinated by fertile plant from Type II. And this was explained by a interactive epistatic model. Feng et al. (1995, 1996) had obtained similar male sterile systems and the inheritance of that male sterility explained by a

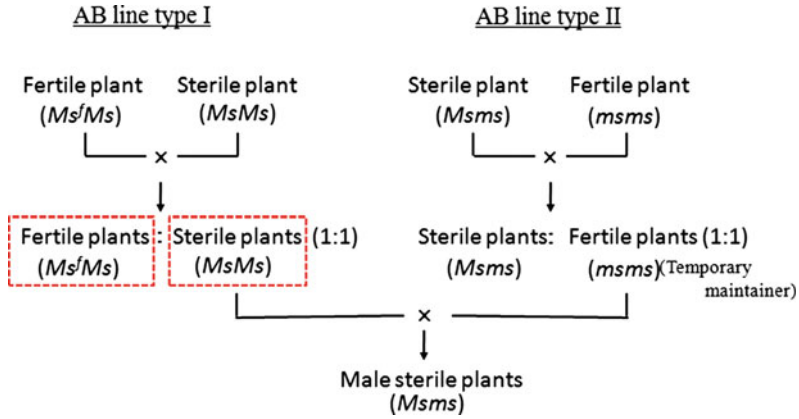

Fig. 1.3 Genetic model of the genic multiple-allele inherited male sterile line in Chinese cabbage. Male sterility could be controlled by three different genes at one locus. Msf, Ms, and ms represent dominant restorer, dominant sterile, and recessive fertile genes, respectively. Correlation of dominance and recessiveness among these genes is Msf > Ms > ms. *Dotted boxes* indicate plants used in this study (Feng et al. 1995, 1996)

multiple allele model (Fig. 1.3). Those 100 % male sterile GMS lines have been successfully utilized in commercial Chinese cabbage hybrid seed production in China.

1.6.2.2 Cytoplasmic Male Sterility

Cytoplasmic male sterility (CMS) is controlled by the cytoplasm, but may be influenced by the nuclear genes. The sterile cytoplasm often results from the introduction of nuclear chromosomes into a foreign cytoplasm. Since the cytoplasm is transferred only through the egg, cytoplasmic male sterility is transmitted only through the mother plant.

Ogura (1968) first described the male sterility system in radish. This male sterility inducing cytoplasm was successfully transferred to *B. oleracea* and *B. napus* (Bannerot et al. 1974). Improved 'Ogura' cytoplasms were then transferred to vegetable *B. rapa* (Heath et al. 1994) and cabbage (Sigareva and Earle 1997). Protoplast fusion was also an efficient way of combining atrazine-resistant chloroplasts in *B. rapa* with the CMS trait of *B. nigra* cytoplasm in *B. oleracea* (Stein et al. 1991).

Using stocks of Rlaacc obtained from Bannerot in 1975, Heyn repeatedly backcrossed sarson, *B. rapa* ssp *trilocularis* (Aaa.t) selecting for female fertility in the R1 cytoplasm. After four back-crosses, moderately fertile substituted sarson (Rlaa.t) was further crossed for five generations to the rapidly cycling stocks of *B. rapa* (Aaa). Rlaa stocks have been used in a backcrossing program to various subspecies of *B. rapa* including Chinese cabbage. But those lines showed low nectary function and chlorosis (Williams and Heyn 1981).

1.6.2.3 Polima CMS

Polima CMS was discovered by Fu in 1972 (Fu et al. 1997) and this male sterility system has been successfully used in three/two-line hybrid production in rapeseed (*B. napus* L.). Barsby et al. (1987) transferred polima CMS in one step from spring to winter lines of oilseed rape (*B. napus* L.) by protoplast fusion. Amphidiploid individuals with normal female fertility were recovered. The regenerated plants retained the winter habit. All traits studied were stable through subsequent sexual generations. Verma et al. (2000) also transferred polima CMS *B. napus* 'ISN 706' to five different cultivars of *B. rapa* by repeated backcrossing. The cultivars 'Candle' and 'ATC 94211' possessed the restorer gene and restoration is controlled by a single dominant gene.

Ke et al. (1992) transferred polima CMS to Chinese cabbage by repeated backcrossing. Polima CMS Chinese cabbage line CMS3411-7 was successfully used to produce F_1 hybrids. However this Polima CMS Chinese cabbage line may produce minor pollens under extreme low or high temperature.

1.7 Resistance to Disease

1.7.1 Resistance to *Turnip Mosaic Virus*

Turnip mosaic virus (TuMV), a member of the *Potyvirus* genus, is one of the most damaging viruses of vegetables in the world. It causes the most important disease affecting field *Brassica* vegetables in North China and Taiwan, and the second most important in the UK (Tomlinson 1987).

TuMV is difficult to control because of its wide host range and non-persistent mode of transmission by aphids. Natural plant resistance is likely to be the most effective and environmentally friendly method of controlling TuMV. Recently identified resistances in *B. rapa* appear to be effective against a broad range of TuMV isolates (Provvidente 1980, 1981; Walsh et al. 2002). Several different modes of inheritance of TuMV resistance in *B. rapa* have been described (Yoon et al. 1993; Suh et al. 1995).

Hughes et al. (2002) evaluated 26 *B. rapa* lines for resistance to TuMV isolates representing the three major pathotypes in Europe. Of these lines, 13 showed resistance. Genetic models for inheritance of resistance were proposed for four *B. rapa* lines: Jong Bai No. 2 had dominant resistance to pathotype 1 conferred by a single allele; PI418957C and Jin G 55 had recessive resistance to pathotype 4 where a single allele was required; PI418957C also had recessive resistance to pathotype 3 where a model with one of two epistatic, unlinked loci was proposed. Jong Bai No. 1 also had recessive resistance to pathotype 3, apparently conferred by alleles at three loci, where any two of the three loci were epistatic and required for resistance.

The extreme resistance to Turnip mosaic virus observed in the Chinese cabbage line, BP8407, is monogenic and recessive. A candidate gene, Bra035393 was predicted within the mapped resistance locus. This gene was analysed in four resistant and three susceptible lines. A correlation was observed between the amino acid substitution (Gly/Asp) in the eIF(iso)4E protein and resistance/susceptibility (Fig. 1.4; Qian et al. 2013).

Recessive strain-specific resistance to a number of plant viruses in the *Potyvirus* genus has been found to be based on mutations in the eukaryotic translation initiation factor 4E (*eIF4E*) and its isoform, *eIF*(*iso*)*4E*. Nellist et al. (2014) identified three copies of *eIF*(*iso*)*4E* in a number of *B. rapa* lines. Broad-spectrum resistance to the potyvirus TuMV due to a natural mechanism based on the mis-splicing of the *eIF*(*iso*)*4E* allele in some TuMV-resistant *B. rapa* var. *pekinensis* lines. Of the splice variants, the most common results in a stop codon in intron 1 and a much truncated, non-functional protein. The existence of multiple copies has enabled redundancy in the host plant's translational machinery, resulting in diversification and emergence of the resistance (Fig. 1.4). Deployment of the resistance is complicated by the presence of multiple copies of the gene. Our data suggest that in the *B. rapa* subspecies *trilocularis*, TuMV appears to be able to use copies of *eIF*(*iso*)*4E* at two loci. Transformation of different copies of *eIF*(*iso*)*4E* from a resistant *B. rapa* line into an *eIF*(*iso*)*4E* knockout line of *Arabidopsis thaliana* proved misleading because it showed that, when expressed ectopically, TuMV could use multiple copies which was not the case in the resistant *B. rapa* line. The inability of TuMV to access multiple copies of *eIF*(*iso*)*4E* in *B. rapa* and the broad spectrum of the resistance suggest it may be durable.

1.7.2 Resistance to Clubroot

Clubroot disease, caused by the soilborne, obligate plant pathogen *Plasmodiophora brassicae* Wor.

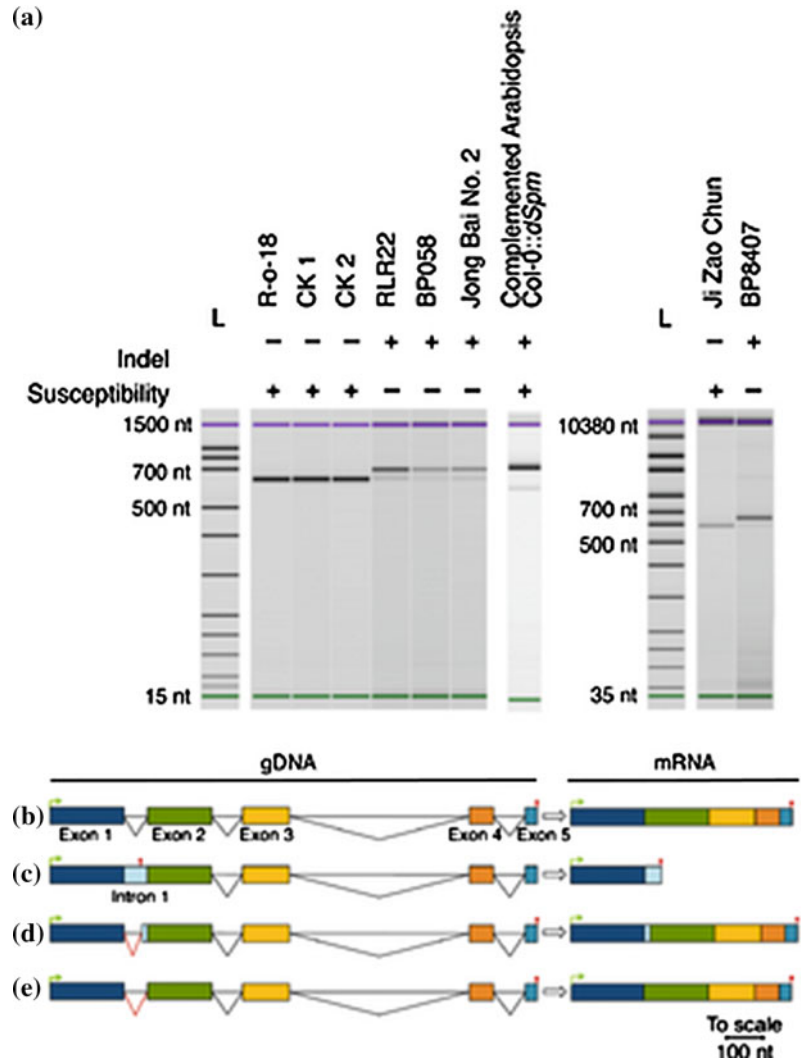

Fig. 1.4 *BraA.eIF*(*iso*)*4E.a* is mis-spliced in resistant *Brassica rapa* and in transgenic *Arabidopsis thaliana*. **a** Detection of a single transcript in susceptible plants, corresponding to correctly spliced *BraA.eIF*(*iso*)*4E.a*, splice variants in resistant *B. rapa* plants and splice variants from an Arabidopsis Col-0::*dSpm* plant complemented with the RLR22 allele of *BraA.eIF*(*iso*)*4E.a*, separated using an Agilent 2100 Bioanalyser. **b** Correctly spliced *BraA.eIF*(*iso*)*4E.a* detected in susceptible plants (**a**). **c** The most common mis-spliced variant retained the extra G and the whole of intron 1 resulting in a premature stop codon at the 234 nucleotide (nt) position, detected in resistant lines and Arabidopsis Col-0::*dSpm* complemented by the RLR22 allele of *BraA.eIF*(*iso*)*4E.a*. **d** Variant retaining the last 15 nt of intron 1 (in-frame), detected in RLR22 and Col-0::*dSpm* complemented by *BraA.eIF*(*iso*)*4E.a* (**a**). **e** Variant lacking the last 3 nt of exon 1 (in-frame) and with a substitution, detected in RLR22 and some F_2 plants from the RLR22, CK 1 cross (Nellist et al. 2014)

infects all cruciferous vegetable and oil crops, including *B. rapa*, *B. oleracea*, *B. napus*, and other *Brassica* species. This disease is one of the most economically important diseases of *Brassica* crops worldwide.

In *B. rapa*, genetic resources with clubroot resistance (CR) were found only in European turnips (Yoshikawa 1993) as was also suggested by Crute et al. (1980). European turnips such as cvs. Siloga, Gelria, Milan White, and Debra have

been used for breeding different CR cultivars. Some of these turnip cultivars have also been used to select the differential hosts ECD-02 (from cv. Gelria) and ECD-01 (from cv. Debra) of the ECD set (Crute 1986).

A number of CR loci in *B. rapa* have been identified and distinguished with the aid of molecular markers. Kuginuki et al. (1999) first reported linkage markers for a clubroot resistance locus known as Crr1. Five resistance loci, CRa, Crr1, Crr2, Crr3, and CRb, were found as major genes (Matsumoto et al. 1998; Suwabe et al. 2003; Hirai et al. 2004; Piao et al. 2004; Saito et al. 2006). Among them, Crr1 and Crr2 were mapped in linkage groups R8 and R1, respectively. In contrast, Crr3 and CRb were mapped in R3. Because the genome regions for Crr1, Crr2, and CRb correspond to chromosome 4 of *A. thaliana*, they may differentiate from the same part of the ancestral genome (Suwabe et al. 2006). The triplicated positions are all active as clubroot resistance genes in the *B. rapa* genome. Crr3 was found to correspond to chromosome 3 of *A. thaliana*.

1.8 Glucosinolates

Glucosinolates (GS) are a group of amino acid-derived secondary metabolites found throughout the Cruciferae family. Glucosinolates and their degradation products play important roles in pathogen and insect interactions, the special flavors and tastes and human health.

When crucifer tissue is damaged, GS are hydrolyzed by myrosinase into degradation products such as isothiocyanate, nitrile, thiocyanates, cyano-epithioalkanes and oxazolidine-2-thiones, which have different bioactivities (Bones and Rossiter 2006; Rask et al. 2000; Wittstock et al. 2003). Degradation products of GS are thought to inhibit carcinogenesis by effecting cell cycle arrest and stimulating apoptosis (Hayes et al. 2008). Sulforaphane, the isothiocyanate derived from glucoraphanin (GRA), exhibits strong anticarcinogenic properties (Keck and Finley 2004). Indole-3-carbinol, a derivative of glucobrassicin, also has anticarcinogenic properties (Choi et al. 2010). In addition, phenethyl isothiocyanate can block the conversion of several carcinogens to their carcinogenic forms (Hecht 2000).

Several experiments were done to determine the kinds and contents of GS in *B. rapa*. Chong et al. (1982) analysed glucosinolates (determined by quantifying their hydrolysis products, goitrin (5-vinyl-oxazolidine-2-thione), volatile isothiocyanates, and the thiocyanate ion) in marketable roots of 10 summer turnip cultivars. The levels of total GS in the turnip seeds are 30–110 times than those in the vegetative parts; the contents of glucosinolate products generally increased with increasing maturity dates.

In 12 cultivars of vegetable turnip rape, two GSL compounds, gluconapin, and glucobrassicanapin, were mainly found in all the edible parts. The total GSL content of edible parts, recalculated by using the percentage of dry weight to the sum weight, ranged from 60 to 80 mmol kg^{-1} DW in all the cultivars (Kim et al. 2003).

In leaves of 113 varieties of turnip greens from northwestern Spain grown at two sites, sixteen glucosinolates were identified, being the aliphatic glucosinolates, gluconapin and glucobrassicanapin the most abundant. Other aliphatic glucosinolates, such as progoitrin, glucoalyssin, and gluconapoleiferin were relatively abundant in varieties with a different glucosinolate profile. Indolic and aromatic glucosinolate concentrations were low and showed few differences among varieties. Differences in total glucosinolate content, glucosinolate profile and bitterness were found among varieties, with a total glucosinolate content ranging from 11.8 to 74.0 μmol g^{-1} dw at one site and from 7.5 to 56.9 μmol g^{-1} dw at the other site (Padilla et al. 2007). Eight GSL were identified, being two aliphatic GSL, gluconapin (84.4 % of the total GSL) and glucobrassicanapin (7.2 % of the total GSL) the most abundant. Indolic and aromatic GSL content were low but also showed significant differences among varieties. Total GSL content ranged from 19 to 37.3 μmol g^{-1} dw in early and extra-late groups, respectively, and

from 19.5 to 36.3 μmol g^{-1} dw for turnips and turnip greens groups (María 2012).

In 129 DH lines of Chinese cabbage, eight main glucosinolates were detected in all the accessions including 3 aliphatic glucosinolates, 4 indolic glucosinolates and 1 aromatic glucosinolate. Principal component analysis showed aliphatic glucosinolates, gluconapin (NAP), glucobrassicanapin (GBN) and progoitrin (PRO) were the main glucosinolate profiles with the highest ratio of about 60.0 %. Combined variance analysis showed that there were significant differences among varieties in aliphatic glucosinolates (NAP, GBN and PRO) (Liao et al. 2012).

In recent years the genes for the glucosinolate biosynthetic pathway in *B. rapa* were explored. Lou et al. (2008) identified quantitative trait loci (QTL) for glucosinolate accumulation in leaves of *B. rapa* from two novel segregating double haploid (DH) populations: DH38, derived from a cross between yellow sarson R500 and pak choi variety HK Naibaicai; and DH30, from a cross between yellow sarson R500 and Kairyou Hakata, a Japanese vegetable turnip variety. An integrated map of 1068 cM with 10 linkage groups, assigned to the international agreed nomenclature, is developed based on the two individual DH maps with the common parent using amplified fragment length polymorphism (AFLP) and single sequence repeat (SSR) markers. Eight different glucosinolate compounds were detected in parents and F_1s of the DH populations and found to segregate quantitatively in the DH populations. QTL analysis identified 16 loci controlling aliphatic glucosinolate accumulation, three loci controlling total indolic glucosinolate concentration and three loci regulating aromatic glucosinolate concentrations. Both comparative genomic analyses based on *Arabidopsis–B. rapa* synteny and mapping of candidate orthologous genes in *B. rapa* allowed the selection of genes involved in the glucosinolate biosynthesis pathway that may account for the identified QTL.

Wang et al. (2011) conducted comparative genomic analyses of *A. thaliana* and *B. rapa* on a genome-wide level. 102 putative genes in *B. rapa* were identified as the orthologs of 52 GS genes in *A. thaliana*. All but one gene was successfully mapped on 10 chromosomes. Most GS genes exist in more than one copy in *B. rapa*. A high co-linearity in the glucosinolate biosynthetic pathway between *A. thaliana* and *B. rapa* was also established. The homologous GS genes in *B. rapa* and *A. thaliana* share 59–91 % nucleotide sequence identity and 93 % of the GS genes exhibit synteny between *B. rapa* and *A. thaliana*. Moreover, the structure and arrangement of the *B. rapa* GS (BrGS) genes correspond with the known evolutionary divergence of *B. rapa*.

1.9 Transfer of Genes Between *B. rapa* and *B. napus*

B. rapa (AA) shares a common set of chromosomes with *B. napus* (AACC), which is one of the most valuable oilseeds. Resynthesis of *B. napus* is an important tool for the broadening of genetic diversity as well as crop cultivar improvement in oilseed rape by crossing the original ancestors, *B. oleracea* and *B. rapa*. This has the potential not only to increase genetic variability with a view to hybrid breeding but also to broaden the genetic base with respect to pest and disease resistances.

Mithen and Magrath (1992) generated synthetic lines of *B. napus* carrying resistance to blackleg disease derived from *B. rapa* via embryo culture. The resistance was then integrated successfully into spring canola, resulting in the release of the cv. Surpass in the late 1990s and subsequent efforts to introgress this resistance into winter oilseed rape material. Two blackleg resistance genes, LepR1 and LepR2, from *B. rapa* subsp. *Sylvestris* were transferred into *B. napus* (Yu et al. 2012).

The yellow-seed trait has been introduced to *B. napus* from *B. rapa* and others (Chen et al. 1988; Meng et al. 1998; Rahman 2001). The interest in developing yellow-seeded oilseed rape relates to the fact that this character is associated with higher oil and protein contents and less fibre

content, which are desirable food and livestock feed agronomic goals (Shirzadegan and Röbbelen 1985).

Resynthesised rapeseed also represents an interesting source of genetic variation for quality improvement in oilseed rape (Chen and Heneen 1989). For example, Crosses between *B. rapa* ssp. *trilocularis* (Yellow Sarson) and several selected cauliflowers (*B. oleracea* convar. *botrytis* var. *botrytis*) were made to create new oilseed rape germplasm with a high erucic acid content (Seyis et al. 2003).

Hilgert-Delgado et al. (2014) made twenty-four different one-sided crosses between six accessions of winter *B. rapa* L. ssp. *oleifera* (DC) Metzg. f. biennis (winter turnip rape) and four accessions of winter *B. oleracea* L. var. acephala (DC.) (winter curly kale). Successful germination of the embryos was achieved in 23 combinations by ovule cultures. On average, 0.34 embryos were obtained per single bud in twenty-four different combinations. Hybrid morphological characteristics of true leaves were more perceptible after transfer to in vivo in the larger and more developed plants.

Muangprom (2006) transferred a dwarf gene (Brrga1-d), which was a single, semi-dominant, gibberellin insensitive dwarf mutant, from *B. rapa* to Oilseed *B. napus* by using interspecific hybridization of *B. rapa* and *B. oleracea* and embryo rescue to resynthesize *B. napus* containing the Brrga1-d dwarf gene. The dwarf gene was backcrossed into two parents of a commercial hybrid combination and evaluated as inbred and hybrid lines in field experiments. The Brrga1-d gene reduced plant height and lodging in inbred and hybrid lines of *B. napus*, even when present as a single dose in heterozygous genotypes.

References

Baillie AMR, Epp DJ, Hutcheson D, Keller WA (1992) In vitro culture of isolated microspores and regeneration of plants in *Brassica campestris*. Plant Cell Rep 11:234–237

Bannerot H, Boulidard L, Cauderon Y, Tempe J (1974) Transfer of cytoplasmic male sterility from *Raphanus sativus* to Brassicaoleracea. In: Proceedings of EUCARPIA meeting Cruciferae, Dundee, pp 52–54

Barsby TL, Yarrow SA, Kemble RJ, Grant I (1987) The transfer of cytoplasmic male sterility to winter-type oilseed rape (*Brassica napus* L.) by protoplast fusion. Plant Sci 53(3):243–248

Bones A, Rossiter J (2006) The enzymic and chemically induced decomposition of glucosinolates. Phytochemistry 67(11):1053–1067

Cao MQ, Li Y, Liu F, Doré C (1994) Embryogenesis and plant regeneration of pak choi (*Brassica rapa* L. ssp. chinensis) via in vitro isolated microspore culture. Plant Cell Rep 13:447–450

Chen BY, Heneen WK (1989) Fatty acid composition of resynthesized *Brassica napus* L., *B. rapa* L. and *B. Alboglabra* Bailey with special reference to the inheritance of erucic acid content. Heredity 63:309–314

Chen BY, Heneen WK, Jonsson R (1988) Resynthesis of *Brassica napus* L. through interspecific hybridization between *B. alboglabra* Bailey and *B. campestris* L. with special emphasis on seed colour. Plant Breed 101:52–59

Choi H, Cho M, Lee H, Yoon D (2010) Indole-3-carbinol induces apoptosis through p53 and activation of caspase-8 pathway in lung cancer A549 cells. Food Chem Toxicol 48(3):883–890

Chong C, Ju H-Y, Bible BB (1982) Glucosinolate composition of turnip and rutabaga cultivars. Can J Plant Sci 62(2):533–536

Chowdhury J, Das K (1966) Male sterility in yellow sarson. Indian J Genet Plant Breed 26(3):374–380

Crute IR (1986) The relationship between *Plasmodiophora brassicae* and its hosts: the application of concepts relating to variation in inter-organismal associations. Adv Plant Pathol 5:1–52

Crute IR, Gray AR, Crisp P, Buczacki ST (1980) Variation in *Plasmodiophora brassicae* and resistance to clubroot disease in Brassicas and allied crops—a critical review. Plant Breed Abstr 50:91–104

Das K, Pandey B (1961) Male sterility in brown sarson. Indian J Genet Plant Breed 21:185–190

Feng H, Wei YT, Zhang SN (1995) Inheritance of and utilization model for genic male sterility in Chinese cabbage (*Brassica pekinensis* Rupr.). Acta Horticiltu-rae 402:133–140

Feng H, Wei YT, Ji SJ, Jin G, Jin JS (1996) Multiple allele model for genic male sterility in Chinese cabbage. Acta Horticulturae 467:133–142

Fu TD, Yang GS, Yang XN, Ma CZ (1997) Discovery, study and utilization of polima cytoplasmic male sterility in *Brassica napus* L. Prog Nat Sci 5:169–177

Gu HH, Zhou WJ, Hagberg P (2003) High frequency spontaneous production of doubled haploid plats in microspore culture of *Brassica rapa* ssp. chinensis. Euphytica 134:239–245

Hatakeyama K, Watanabe M, Takasaki T, Ojima K, Hinata K (1998) Dominance relationships between S-alleles in self incompatible *Brassica campestris* L. Heredity 80:241–247

Hayes J, Kelleher M, Eggleston I (2008) The cancer chemopreventive actions of phytochemicals derived from glucosinolates. Eur J Nutr 47:73–88

Heath DW, Earle ED, Dickson MH (1994) Introgressing cold-tolerant Ogura cytoplasm from rapeseed into pak choi and Chinese cabbage. HortScience 29(3): 202–203

Hecht S (2000) Inhibition of carcinogenesis by isothiocyanates 1. Drug Metab Rev 32(3–4):395–411

Hilgert-Delgado A, Klíma M, Viehmannová I, Urban MO, Fernández-Cusimamani E et al (2014) Efficient resynthesis of oilseed rape (*Brassica napus* L.) from crosses of winter types *B. rapa* × *B. oleracea* via simple ovule culture and early hybrid verification. Plant Cell, Tissue Organ Culture (Published online: 9 Aug 2014)

Hirai M, Harada T, Kubo N, Tsukada M, Suwabe K et al (2004) A novel locus for clubroot resistance in *Brassica rapa* and its linkage markers. Theor Appl Genet 108:639–643

Hughes SL, Green SK, Lydiate DJ, Walsh JA (2002) Resistance to *Turnip mosaic virus* in *Brassica rapa* and *B. napus* and the analysis of genetic inheritance in selected lines. Plant Pathol 51:567–573

Ke GL, Zhao ZY, Song YZ, Zhang LG, Zhao LM (1992) Breeding of alloplasmic male sterile line CMS3411-7 in heading Chinese cabbage. Acta Horticulturae Sinica 19(4):333–340

Keck A, Finley J (2004) Cruciferous vegetables: cancer protective mechanisms of glucosinolate hydrolysis products and selenium. Integr Cancer Ther 3(1):5–12

Kim SJ, Kawaguchi S, Watanabe Y (2003) Glucosinolates in vegetative tissues and seeds of 12 cultivars of vegetable turnip rape (*Brassica rapa* L.). Soil Sci Plant Nutr 49:337–346

Kuginuki Y, Yoshikawa H, Hirai M (1999) Variation in virulence of *Plasmodiophora brassicae* in Japan tested with clubroot-resistant cultivars of Chinese cabbage (*Brassica rapa* L. ssp. pekinensis). Eur J Plant Pathol 105:327–332

Li CW (1981a) The origins and evolution of vegetable crops in China. Sci Agr Sin 14:90–95

Li CW (1981b) The origin, evolution, taxonomy, and hybridization of Chinese cabbage. In: Talekar NS, Griggs TD (eds) Chinese cabbage. Proceedings of first international symposium. Asian Vegetable Research and Development Center, Shanhua, Tainan, pp 3–10

Liao Y, Song M, Wang H, Xu D, Wang X (2012) Glucosinolate profile and accumulation in *Brassica campestris* L. ssp. pekinensis. Acta Horticulturae Sinica 38(5):963–969

Lichter R (1982) Induction of haploid plants from isolated pollen of *Brassica napus*. Z Pflanzenphysiol 105:427–434

Linnaeus C (1753) Species plantarum, 1st edn. Stockholm

Lou P, Zhao J, He H, Hanhart C, Carpio DPD et al (2008) Quantitative trait loci for glucosinolate accumulation in *Brassica rapa* leaves. New Phytol 179:1017–1032

Matsumoto E, Yasui C, Ohi M, Tsukada M (1998) Linkage analysis of RFLP markers for clubroot resistance and pigmentation in Chinese cabbage (*Brassica rapa* ssp. pekinensis). Euphytica 104:79–86

Meng JL, Shi SW, Gan L, Li ZY, Qu XS (1998) The production of yellow-seeded *Brassica napus* (AACC) through crossing interspecific hybrids of *B. campestris* (AA) and *B. carinata* (BBCC) with *B. napus*. Euphytica 103:329–333

Metzger J (1833) Systematische Beschreibung der kultivirten Kohlarten. Heidelberg

Mithen RF, Magrath R (1992) Glucosinolates and resistance to *Leptosphaeria maculans* in wild and cultivated *Brassica* species. Plant Breed 108:60–68

Muangprom A, Mauriera I, Osborn TC (2006) Transfer of a dwarf gene from *Brassica rapa* to oilseed *B. napus*, effects on agronomic traits, and development of a 'perfect' marker for selection. Mol Breeding 17:101–110

Nellist CF, Qian W, Jenner CE, Moore JD, Zhang S et al (2014) Multiple copies of eukaryotic translation initiation factors in *Brassica rapa* facilitate redundancy, enabling diversification through variation in splicing and broad-spectrum virus resistance. Plant J 77:2

Niu XK, Wu FY, Zhong HH, Li XS (1980) The selection and utilization of Chinese cabbage (B. pekinensis Rupr.) of male sterile AB line. Acta Horticulturae Sinica 7:25–32

Nou IS, Watanabe M, Isuzugawa K, Isogai A, Hinata K (1993) Isolation of S-allele from a wild population of *Brassica campestris* L. at Balcesme, Turkey and their characterization by S-glycoprotein. Sex Plant Reprod 6:71–78

Ockendon DJ (1974) Distribution of self-incompatability alleles and breeding structure of open-pollinated cultivars of Brussels sprouts. Heredity 33:159–171

Ogura H (1968) Studies of a new male-sterility in Japanese radish, with special reference to the utilization of this sterility towards the practical raising of hybrid seeds Mem. Fac Agric Kagoshima Univ 6:39–78

Padilla G, Cartea ME, Velasco P (2007) Variation of glucosinolates in vegetable crops of *Brassica rapa*. Phytochemistry 68:536–545

Piao ZY, Deng YQ, Choi SR, Park YJ, Lim YP (2004) SCAR and CAPS mapping of CRb, a gene conferring resistance to *Plasmodiophora brassicaein* Chinese cabbage (*Brassica rapa* ssp. pekinensis). Theor Appl Genet 108:1458–1465

Prakash S, Hinata K (1980) Taxonomy, cytogenetics and origin of crop *Brassica*, a review. Opera Botanica 55:1–57

Provvidente R (1980) Evaluation of Chinese cabbage cultivars from Japan and the People's Republic of China for resistance to turnip mosaic and cauliflower mosaicvuirus. HortScience 105:571–573

Provvidente R (1981) Sources of resistance to turnip mosaic virus in Chinese cabbage. In: Talekar NS, Griggs TD (eds) Chinese cabbage. The Asain Vegetable Research and Development Centre, Taiwan, pp 423–430

Qian W, Zhang S, Zhang S, Li F, Zhang H et al (2013) Mapping and candidate-gene screening of the novel Turnip mosaic virus resistance gene retr02 in Chinese cabbage (*Brassica rapa* L.). Theor Appl Genet 126:1–179

Rahman MH (2001) Production of yellow-seeded *Brassica napus* through interspecific crosses. Plant Breed 120:463–472

Rask L, Andreasson E, Ekbom B, Eriksson S, Pontoppidan B et al (2000) Myrosinase: gene family evolution and herbivore defense in Brassicaceae. Plant Mol Biol 42(1):93–114

Saito M, Kubo N, Matsumoto S, Suwabe K, Tsukada M et al (2006) Fine mapping of the clubroot resistance gene Crr3 in *Brassica rapa*. Theor Appl Genet 114:81–91

Sato T, Nishio T, Hirai M (1989a) Plant regeneration form isolated microspore culture of Chinese cabbage (*Brassica rapa* ssp. pekinensis). Plant Cell Rep 8:486–488

Sato T, Nishio T, Hirai M (1989b) Culture conditions for the initiation of embryogenesis from isolated microspores in Chinese cabbage (*Brassica rapa* L.). Bull. Nat. Res. Inst. Veg. ornam. Plants Tea Japan Ser A 3:55–65

Sato S, Katoh N, Iwai S, Hagimori M (2002) Effect of low temperature pretreatment of buds or inflorescence on isolated microspore culture in *Brassica rapa* (syn. *B. rapa*). Breed Sci 52:23–26

Schopfer CR, Nasrallah ME, Nasrallah JB (1999) The male determinant of self-incompatibility in *Brassica*. Science 286:1697–1700

Seyis F, Snowdon R, Luhs W, Friedt W (2003) Molecular characterization of novel resynthesized rapeseed (*Brassica napus*) lines and analysis of their genetic diversity in comparison with spring rapeseed cultivars. Plant Breed 122:473–478

Shirzadegan M, Röbbelen G (1985) Influence of seed color and hull proportion on quality properties of seeds in *Brassica napus* L. Fette Seifen Anstrichmitte l87:235–237

Sigareva M, Earle E (1997) Direct transfer of a cold-tolerant Ogura male-sterile cytoplasm into cabbage (*Brassica oleracea* ssp. *capitata*) via protoplast fusion. Theor Appl Genet 94(2):213–220

Song KM, Osborn TC, Williams PH (1998) *Brassica* taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs). 2. Preliminary analysis of subspecies within *B. rapa* (syn. campestris) and *B. oleracea*. Theor Appl Genet 76:593–600

Stein JC, Howlett B, Boyes DC, Nasrallah ME, Nasrallah JB (1991) Molecular cloning of a putative receptor protein kinase gene encoded at the self-incompatibility locus of *Brassica oleracea*. Proc Natl Acad Sci USA 88:8816–8820

Suh SK, Green SK, Park HG (1995) Genetics of resistance to five strains of turnip mosaic virus in Chinese cabbage. Euphytica 81:71–77

Sun VG (1946) The evaluation of taxonomic characters of cultivated *Brassica* with a key to species and varieties. I. The characters. Torrey Bot Club Bull 73:244–281

Suwabe K, Tsukazaki H, Iketani H, Hatakeyama K, Fujimura M et al (2003) Identification of two loci for resistance to clubroot (*Plasmodiophora brassicae* Woronin) in *Brassica rapa* L. Theor Appl Genet 107:997–1002

Suwabe K, Tsukazaki H, Iketani H, Hatakeyama K, Kondo M et al (2006) Simple sequence repeat-based comparative genomics between *Brassica rapa* and *Arabidopsis thaliana*: the genetic origin of clubroot resistance. Genetics 173:309–319

Suzuki G, Kai N, Hirose T, Fukui K, Nishio T et al (1999) Genomic organization of the S locus: identification and characterization of genes in SLG/SRK region of S9 haplotype of *Brassica campestris* (syn. rapa). Genetics 153:391–400

Takayama S, Isogai A, Tsukamoto C, Ueda Y, Hinata K et al (1987) Sequences of S-glycoproteins, products of the *Brassica campestris* self-incompatibility locus. Nature 326:102–105

Takuno S, Kawahara T, Ohnishi O (2007) Phylogenetic relationships among cultivated types of *Brassica rapa* L. em. Metzg. as revealed by AFLP analysis. Genet Resour and Crop Ev 54(2):279–285

Tomlinson JA (1987) Epidemiology and control of virus diseases of vegetables. Ann Appl Biol 110:661–681

Verma JK, Sodhi YS, Mukhopadhyay A, Arumugam N, Gupta V, Pental D, Pradhan AK (2000) Identification of stable maintainer and fertility restorer lines for 'Polima' CMS in *Brassica campestris*. Plant Breed 119(1):90–92

Walsh JA, Rusholme RL, Hughes SL et al (2002) Different classes of resistance to Turnip mosaic virus in *Brassica rapa*. Eur J Plant Pathol 108:15–20

Wang TT, Li HX, Zhang JH, Ouyang B et al (2009) Initiation and development of microspore embryogenesis in recalcitrant purple flowering stalk (*Brassica campestris* ssp. Chinesis var. purpurea Hort.). Sci Hortic 121:419–424

Wang H, Wu J, S Sun, Bo Liu, Feng C et al (2011) Glucosinolate biosynthetic genes in *Brassica rapa*. Gene 487(2):135–142

Watanabe M, Hatakeyama K, Takada Y, Hinata K (2001) Molecular aspects of self-incompatibility in *Brassica* species. Plant Cell Physiol 42:560–565

Williams PH, Heyn FW (1981) The origin and development of cytoplasmic male sterile Chinese cabbage. In: Proceeding of the first international symposium on Chinese cabbage. Asign Vegetable Research and Development Center, Taiwan, pp 293–300

Wittstock U, Kliebenstein D, Lambrix V, Reichelt M, Gershenzon J (2003) Glucosinolate hydrolysis and its impact on generalist and specialist insect herbivores. Recent Adv Phytochem 37:101–125

Wong RSC, Zee SY, Swanson EB (1996) Isolated microspore culture of Chinese flowering cabbage

(*Brassica campestris* ssp. parachinensis). Plant Cell Rep 15:396–400

Yoon JY, Green SK, Opena RT (1993) Inheritance of resistance to turnip mosaic virus in Chinese cabbage. Euphytica 69:103–108

Yoshikawa H (1993) Studies on breeding of clubroot resistance in cole crops. Bull Natl Res Inst Veg Ornam Plants Tea Jpn Ser A7:1–165

Yu F, Lydiate DJ, Gugel RK, Sharpe AG, Rimmer SR (2012) Introgression of *Brassica rapa* subsp. *sylvestris* blackleg resistance into *B. napus*. Mol Breed 30 (3):1495–1506

Zhang SF, Song Z, Zhao X (1990) Breeding of interactive genic male sterile line in Chinese cabbage (*Brassica pekinensis* Rupr.) and utilization model. Acta Horticulturae Sinica **17**(2): 117–125

Zhang Y, Wang A, Liu Y, Wang Y, Feng H (2012) Improved production of doubled haploids in *Brassica rapa* through microspore culture. Plant Breed 131:164–169

Zhao J-J, Wang X-W, Deng B, Lou P, Wu J et al (2005) Genetic relationships within *Brassica rapa* as inferred from AFLP fingerprints. Theor Appl Genet 110:1301–1314

Background History of the National and International *Brassica rapa* Genome Sequencing Initiatives

2

Ian Bancroft and Xiaowu Wang

Abstract

Whole genome sequencing of *Brassica rapa* was first launched by the multinational group of *Brassica rapa* Genome Sequencing Project (BrGSP). The group planned to perform the assembly using the method called "bacterial artificial chromosome (BAC) by BAC" in the initial stage. However, the progress was limited and only chromosome A03 was finished under this method. Along with the development of the second generation sequencing technology, the Chinese suggest to adopt this new sequencing method and initiative assembled the *B. rapa* genome in short time by SOAP-denovo, which integrated the data of pair-ends short reads generated from the Illumina sequencing platform and the data of BAC sequences from BrGSP. This well assembled whole genome sequences of *B. rapa*—verified by the comparison to the A03 assembled by BAC sequenced—was then serves as the genome reference for the evolution, gene mapping and function studies of *B. rapa*.

The Steering Group for the Multinational *Brassica* Genome Project published a concept note in 2003 for the first *Brassica* Genome Sequencing Project (http://brassica.nbi.ac.uk/brassica_genome_sequencing_concept.htm). *B. rapa* was selected first as it has the smallest genome among the cultivated *Brassica* species and fewer transposon-related sequences are interspersed between the genes than are found in *B. oleracea*, for example (Town et al. 2006; Yang et al. 2006; Cheung et al. 2009). The project aimed initially to produce, from bacterial artificial chromosome (BAC) clones, "Phase 2" sequence (i.e. fully oriented and ordered sequence but some small sequence gaps and low quality sequences) for the gene space of the ca. 500 Mb genome of *B. rapa* subspecies *pekinensis*, cultivar Chiifu. The activity was named the

I. Bancroft (✉)
Department of Biology, University of York, York YO10 5DD, UK
e-mail: ian.bancroft@york.ac.uk

X. Wang (✉)
Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China
e-mail: wangxiaowu@caas.cn

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_2

"*Brassica rapa* Genome Sequencing Project" (BrGSP).

Early activities of the BrGSP Consortium were centered on establishing mechanisms for information exchange, agreeing upon on the mapping populations to be used for anchoring sequences to genetic linkage maps and agreeing upon on the BAC libraries to be used for sequencing. Much of the early activities were led by the groups in South Korea, where major genomics research programs in *B. rapa* were already underway. Two mapping populations were agreed upon; both derived from the crosses between the cultivars Chiifu and Kenshin (CKDH and CKRI). Two BAC libraries were selected initially: KBrH and KBrB, both constructed in South Korea. Each library consists of 144 × 384-well plates; made using *Hind*III (KBrH) or *Bam*HI (KBrB) digested genomic DNA. In all, approximately 20-fold redundant representation of the genome was made available.

The initial sequencing strategy was defined as a BAC-by-BAC approach, starting from seed BACs anchored genetically with extension based on overlaps between clones identified using BAC-end sequences. Several countries with major research programs involving *Brassica* species participated in the sequencing. Initially, the BAC libraries were end-sequenced (by groups in South Korea, Canada, UK, Australia and Germany), seed BACs were sequenced and mapped (largely by groups in South Korea), and the task of sequencing the genome was allocated on a chromosome-by-chromosome basis (involving groups in South Korea, UK, China, Canada and Australia). Later on, a complete BAC-based physical map was constructed (Mun et al. 2008) to improve the rate of progress.

The BAC-by-BAC approach to genome sequencing was based on capillary sequencing technology. Over 1000 BAC clones were sequenced, annotated and placed rapidly in the public domain, underpinning early insights in the sequence-level structure of *Brassica* genomes (Mun et al. 2009). However, several countries failed to fund the sequencing of chromosomes allocated to them as part of the BrGSP, so only chromosome A03 was completed by the strategy (Mun et al. 2010).

By 2009, advances in sequencing technology made strategies for sequencing complete genomes based on capillary sequencing obsolete. *Brassica* species had seemed unpromising subjects for the deployment of "Next Generation Sequencing" (NGS) technologies, which produced massively parallel but relatively short sequence reads, as extensive triplication had long been evident in *Brassica* genomes (O'neill and Bancroft 2000), potentially confounding assembly. However, increases in sequence read length and improvements in computational strategies overcame this potential barrier. In China, a whole-genome NGS approach was taken up in the Chinese Initiative of *B. rapa* sequencing, which involved sequencing of the genome of *B. rapa* cv. Chiifu, and produced excellent results. The BrGSP Consortium agreed in 2009 to abandon the BAC-by-BAC approach, focusing efforts on using insights from the higher-quality data to optimize the NGS-based approach and analysis (http://brassica.nbi.ac.uk/pdf/BrGSP_aug_2009.pdf).

The Chinese initiative assembled the *B. rapa* genome by SOAP-denovo (Li et al. 2010). They generated seven libraries with insertion size ranging from 184 bp to 10 Kb (Table 2.1). Three libraries ranging from 184 to 500 bp were used to assemble contigs while fourlibraries with large inserts ranging from 2 to 10 Kb were used to link the contigs to scaffords. To make full use of the existing resources and complement the disadvantage of the limited insertion size of Illumina sequencing DNA libraries, the Chinese initiative adopted a strategy combining the Illumina GAII data with BAC sequence data generated by the BrGSP. The assembly achieved by the Chinese initiative has an N50 contig size over 27 Kb and scaffold size over 339 Kb. Combining the assembled contigs from the Illumina GAII data with BAC sequence data, it has been produced 39 super-scaffolds with an N50 of over 1.97 Mb (Table 2.2). Excluding the highly abundant satellite sequences, the assembled sequence accounted for 284 Mb, of which 255 Mb (~90 % of the 284 Mb) has been anchored onto

Table 2.1 Summary of Illumina sequencing data for *B. rapa* genome

Sequence data	Library insert size	Total length (Gb)	Sequence depth (X)	Read length (bp)
Illumina reads	184 bp	2.482	5.045	101
	200 bp	14.940	30.366	44, 75
	500 bp	7.810	15.874	44, 75
	2 Kb	3.580	7.276	44
	5 Kb	3.210	6.524	45
	8 Kb	2.460	5.000	44
	10 Kb	1.522	3.093	44
Total		36.004	72.36	

Table 2.2 Summary of the final assembly statistics

	Contig size	Contig number	Scaffold size	Scaffold number
N90	5593	10,564	357,979	159
N80	10,984	7292	773,703	104
N70	15,947	5308	1,257,653	77
N60	21,229	3874	1,452,355	56
N50	27,294	2778	1,971,137	39
Total Size	264,110,991		283,823,632	
Total Number (>100 bp)		60,521		40,549
Total Number (>2 Kb)		14,207		794

the ten chromosomes, covering 58.5 % of the estimated 485 Mb genome and about 98 % of the gene space. The number of predicted gene models for *B. rapa* is 41,174, about half as much again as *Arabidopsis* (Table 2.3).

Because the *B. rapa* var. Chiifu chromosome A03 assembly (BAC A03) reportedby Mun et al. (2010) was completely based on the sequence data generated from traditional Sanger sequencer, it provided a perfect reference for the evaluation of the quality of *B. rapa* genome assembly by whole-genome shotgun (WGS) based on NGS data. After the Chinese team released the WGS assembly, several teams performed the evaluation by comparing the two A03 assemblies. The comparison showed very high level of agreement between both the Sanger sequenced BAC-by-BAC approach and the WGS approach. There are only minor discrepancies between Sanger and the NGS data. The total sizes of WGS A03 and BAC A03 are approximately 31.72 and 32.70 Mb, respectively, with slightly more repeat sequences assembled using the BAC approach (9.82 Mb in BAC A03 and 5.68 Mb in WGS A03). There were more gaps observed in BAC A03 (1035/1,358,889 bp, number of gaps/total size of gaps) than in WGS A03 (858/844,319 bp). Forty-four obvious inversions (>1 kb) between the two assemblies were verified by mapping the paired-end reads. The depth of the mapped reads and gaps at the boundaries for 38 inversions supported the WGS assembly, and six inversions remained ambiguous (Fig. 2.1).

Based on this assembly, two groups did extensive gene synteny analysis of it with *Arabidopsis*. Xiaowu Wang's group developed a gene synteny analysis pipeline specifically adapted to the closely-related species for identifying the accurate syntenic genes between *Brassica* and *Arabidopsis* (Cheng et al. 2012). Mike Freeling's group analyzed synteny using CoGe (Tang and Lyons 2012). Both the analyses confirmed that the

Table 2.3 General statistics for predicted protein-coding genes

Gene set	Number	Average length of transcribed region (bp)	Average length of CDS (bp)[1]	# Exons per gene	Average length of exon (bp)	Average length of intron (bp)
A. thaliana	45,483	1642	950	3.92	242	237
C. papaya	45,992	1369	835	3.51	238	213
P. trichocarpa	46,063	1436	845	3.65	231	223
V. vinifera	38,783	1664	949	4.08	233	232
O. sativa	34,875	1685	1009	3.96	255	229
Genscan	40,614	4150	1293	5.87	220	562
Augustus	47,460	1886	1123	4.99	225	191
Brassica_FLcDNA	9028	1265	622	3.14	198	183
Brassica_95k_EST	84,953	3166	645	3.46	186	170
*B_rapa*_unigene	25,219	1250	770	3.55	217	171
*B_rapa*_EST	8614	1596	562	2.84	198	191
GLEAN	41,174	2015	1172	5.03	233	209

genome of *B. rapa* had undergone genome triplication subsequent to the last genome duplication observed in the genome of *A. thaliana* (the α duplication). Moreover, they found that the extent of gene loss (fractionation) among triplicated genome segments varies, with one copy containing a greater proportion of genes expected to have been present in its ancestor (70 %) than the remaining two (46 and 36 %). With this, they proposed a "two-step" hypothesis for *B. rapa* genome evolution, whereby one hybridization between diploid species occurred, following which genome fractionation occurred for a period of time before hybridization with a further diploid species, after which fractionation proceeded on all three subgenomes (Wang et al. 2011; Cheng et al. 2012; Tang et al. 2012).

One of the important goals of sequencing the *B. rapa* genome was to explain the extreme plasticity of the morphological variations, which can be found in *B. rapa* and other Brassica crops (Teutonico and Osborn 1994; Gustafson et al. 2006; Wittkop et al. 2009; Liu et al. 2012). Three possible factors contributing to the rich morphological polymorphism in the species were identified. The first factor may be a general increase in nucleotide substitution rates. The relatively recent polyploidizations in *B. rapa* may also have contributed to accelerated evolution due to genomic instability and gene redundancy. The third factor is the expansion of auxin-related gene families, as auxin controls many plant growth and morphological developmental processes. *B. rapa* has also experienced striking amplification of the plant-specific TCP transcription factor gene family, important in the evolution and specification of plant morphology.

Brassica Genome Gateway (http://brassica.nbi.ac.uk/), Brassica.Info (www.brassica.info) and http://www.brassica-rapa.org were the three most important web-based genome database for *Brassica* community when the BAC-by-BAC sequencing project was being conducted. Brassica.Info and *Brassica* Genome Gateway kept on updating regularly the data of the BACs being sequenced by the BrGSP consortium. *Brassica* Genome Gateway provided further annotation data of the sequenced BACs. The web site, http://www.brassica-rapa.org, hosted by the National Institute of Agricultural Biotechnology (NIAB) provided also annotated BAC information, mapping data and the physical map, which were generated in South Korea. After the Chinese Initiative finished the NGS sequencing project, Institute of Vegetables and Flowers (IVF) set up the *Brassica* database (BRAD, http://brassicadb.

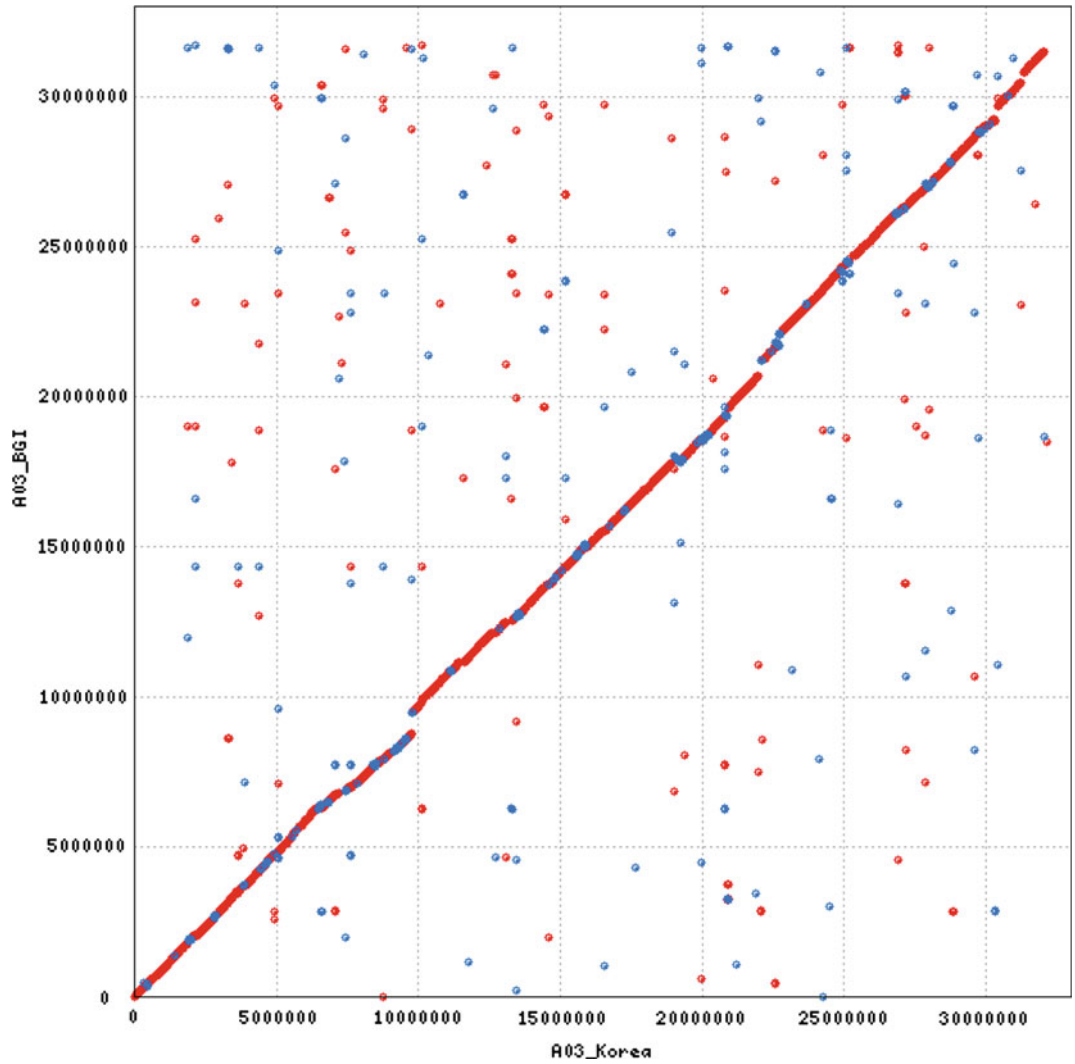

Fig. 2.1 Mummer plot of pseudochromosome A03 from Mun et al. versus that from the WGS assembly

org), which provided services of the complete annotated *Brassica* A genome sequence (Cheng et al. 2011). It marked the completion of the *B. rapa* Genome Sequencing Project.

Important events of the BrGSP Consortium

Jan. 2000: A Brassica Session was separated from the Arabidopsis Workshop for the Plant and Animal Genome Meeting held at San Diego, CA, USA during … (provide web site).

Apr. 2002: During the 13th Crucifer Genetics Workshop, there was acceptance of the requirement for bringing together various national projects under the banner of "Multinational Brassica Genome Project" (MBGP).

Jun. 2003: Steering Group for Multinational Brassica Genome Project was established and announced "Concept note for the Brassica Genome Sequencing Project". The project aimed initially to produce, from BAC clones, "Phase 2" sequence for the ca. 500 Mb genome of *B. rapa* subspecies *pekinensis* and planned to finish the sequencing of the genome by the end of 2007.

Important events of the Chinese Initiative of *B. rapa* Sequencing

Oct. 2008: IVF and BGI initiated *B. rapa* genome sequencing by NGS. IVF signed an agreement with BGI and started *B. rapa* genome sequencing activities.

Jan. 2009: The Chinese initiative produced the first draft assembly of the *B. rapa* genome with purely Solexa reads and sent the results to BrGSP Consortium members. BrGSP Consortium decided to have a meeting with the Chinese initiative members.

Mar. 2009: BrGSP Consortium members had a meeting with the Chinese initiative members. It was reported that the Chinese Initiative will finish the assembling of the *B. rapa* genome before July 2009. BrGSP Consortium decided to evaluate the quality of the Chinese assembly when it is finished.

May. 2009: **Oil Crop Research Institute** (OCRI) decided to join the Chinese initiative.

Jul. 2009: Chinese initiative sent the assembly of Chromosome A02, A03, A08 and A09 to BrGSP Consortium for evaluation.

Aug. 2009: The Chinese Initiative reported the *B. rapa* genome assembly based NGS in the *Brassica* Genome Sequencing meeting in Saskatoon, Canada. During the meeting a decision was made to accept the *B. rapa* genome assembly of the Chinese Initiative as the reference for the *Brassica* research community. BrGSP Consortium abandoned the BAC-by-BAC sequencing activities.

Jul. 2010: The *Brassica* Database (BRAD, http://brassicadb.org/), hosted by IVF/CAAS and providing searching and downloading services of all *B. rapa* genome sequences, was online, indicating the release of *B. rapa* genome sequence to the public (Cheng et al. 2011).

Aug. 2011: The *B. rapa* genome was published as "The genome of the mesopolyploid crop species *B. rapa*" in Nature Genetics (Wang et al. 2011).

References

Cheng F, Liu S, Wu J, Fang L, Sun S et al (2011) BRAD, the genetics and genomics database for *Brassica* plants. BMC Plant Biol 11:136

Cheng F, Wu J, Fang L, Wang X (2012) Syntenic gene analysis between *Brassica rapa* and other *Brassicaceae* species. Front Plant Sci 3:198

Cheung F, Trick M, Drou N, Lim YP, Park J-Y et al (2009) Comparative analysis between homoeologous genome segments of Brassica napus and its progenitor species reveals extensive sequence-level divergence. Plant Cell Online 21:1912–1928

Gustafson J, Badani AG, Snowdon RJ, Wittkop B, Lipsa FD et al (2006) Colocalization of a partially dominant gene for yellow seed colour with a major QTL influencing acid detergent fibre (ADF) content in different crosses of oilseed rape (*Brassica napus*). Genome 49:1499–1509

Li R, Zhu H, Ruan J, Qian W, Fang X et al (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272

Liu L, Stein A, Wittkop B, Sarvari P, Li J et al (2012) A knockout mutation in the lignin biosynthesis gene *CCR1* explains a major QTL for acid detergent lignin content in *Brassica napus* seeds. Theor Appl Genet 124:1573–1586

Mun J-H, Kwon S-J, Yang T-J, Kim H-S, Choi B-S et al (2008) The first generation of a BAC-based physical map of *Brassica rapa*. BMC Genom 9:280

Mun J-H, Kwon S-J, Yang T-J et al (2009) Genome-wide comparative analysis of the *Brassica rapa* gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. Genome Biol 10:R111

Mun J-H, Kwon S-J, Seol Y-J, Kim JA, Jin M et al (2010) Sequence and structure of *Brassica rapa* chromosome A3. Genome Biol 11:R94

O'neill CM, I Bancroft (2000) Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. Plant J 23:233–243

Tang H, Lyons E (2012) Unleashing the genome of *Brassica rapa*. Front Plant Sci 3:172

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS et al (2012) Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. Genetics 190, 1563–1574.

Teutonico R, Osborn T (1994) Mapping of RFLP and qualitative trait loci in *Brassica rapa* and comparison to the linkage maps of *B. napus*, *B. oleracea*, and *Arabidopsis thaliana*. Theor Appl Genet 89:885–894

Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ et al (2006) Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss,

fragmentation, and dispersal after polyploidy. Plant Cell Online 18:1348–1359

Wang X, Wang H, Wang J, Sun R, Wu J et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Wittkop B, Snowdon R, Friedt W (2009) Status and perspectives of breeding for enhanced yield and quality of oilseed crops for Europe. Euphytica 170:131–140

Yang T-J, Kim JS, Kwon S-J, Lim K-B, Choi B-S et al (2006) Sequence-level analysis of the diploidization process in the triplicated FLOWERING LOCUS C region of *Brassica rapa*. Plant Cell Online 18:1339–1347

Genomic Resources and Physical Mapping of the *B. rapa* Genome

3

Jeong-Hwan Mun, Hee-Ju Yu and Beom-Seok Park

Abstract

The genus *Brassica* includes the most extensively cultivated dicotyledonous vegetable crops worldwide. Investigation of the *Brassica* genome presents excellent challenges to study plant genome evolution and divergence of gene function associated with polyploidy and genome hybridization. Among the *Brassica* crops, *Brassica rapa* has been an ideal model for genomic studies on the *Brassica* species. *B. rapa* (AA genome) has a relatively compact diploid genome (529 Mb), compared to *Brassica nigra* (BB genome, 632 Mb) and *Brassica oleracea* (CC genome, 696 Mb). There is also a large collection of cultivars and a broad array of available genomic resources including five large-insert bacterial artificial chromosome (BAC) libraries providing 53-fold genome coverage, end sequences of approximately 146,000 BAC clones, >150,000 ESTs from 33 cDNA libraries, successful shotgun sequencing of 886 euchromatic region-tiling BACs, and a BAC-based physical map. These genomic resources provided fundamental basis of the genome sequencing project and contributed to successful assembly of the whole genome sequences.

3.1 Introduction

The Brassicaceae family includes approximately 3700 species in 350 genera. The species have diverse characteristics, many of which are of agronomic importance as vegetables, condiments, fodder, and oil crops (Beilstein et al. 2006). Economically, *Brassica* species contribute to approximately 10 % of the world's vegetable crop produce and approximately 12 % of the worldwide edible oil supplies (Economic Research Service 2008). The tribe Brassiceae, which is one of 13–19 tribes in the Brassicaceae,

J.-H. Mun (✉)
Department of Bioscience and Bioinformatics, Myongji University, Yongin 449-728, Korea
e-mail: munjh@mju.ac.kr

H.-J. Yu
Department of Life Science, The Catholic University of Korea, Bucheon 420-743, Korea
e-mail: yuheeju@catholic.ac.kr

B.-S. Park
The Agricultural Genome Center, National Academy of Agricultural Science, Rural Development Administration, Wanju 565-851, Korea
e-mail: pbeom@korea.kr

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_3

consists of ~240 species and contains most crop species of *Brassica*. Species of particular importance are *Brassica napus* and *Brassica juncea* as sources of canola oil, *B. rapa* and *Brassica oleracea* as vegetable colecrops, and *Brassica nigra* as a source of the mustard condiment. In addition to the crop species, many of the wild species in the tribe Brassiceae have potential as new crops, sources of condiments, industrial oil, and other diverse products and/or host systems for molecular farming. Wild relative species possess a number of useful agronomic traits, including nuclear and cytoplasmic male sterility, resistance to disease, insect, and nematode pests, tolerance of cold, salt, and drought stresses. For this reason, an understanding of the genetic potential of Brassiceae wild relatives is critical for the establishment of long-term breeding programs of these crops.

Brassica crops are characterized by diverse morphologies including, in some cases, enlarged vegetative and floral meristems (Lukens et al. 2004). These characteristics have long been the targets of breeding programs worldwide, and it is in this regard that the study of genomics could be most beneficial. Comparative genetic mapping has revealed colinear chromosome segments in the Brassicaceae family (Schmidt et al. 2001), and conserved linkage arrangements between *Arabidopsis* and *Brassica* diverged from a common ancestor approximately 14.5–20.4 million years ago (Mya) (Bowers et al. 2003). The genomes of *Brassica* species contain triplicated homoelogous counterparts of the corresponding segments of *Arabidopsis* genome, due to triplication of the entire genome that occurred approximately 13–17 Mya (O'Neill and Bancroft 2000; Town et al. 2006; Yang et al. 2006). Furthermore, an additional natural allopolyploidization event, which happened during the last 10,000 years and resulted in a change of the chromosome numbers and genome size, played a role in the diversification of *Brassica* crops (Nagaharu 1935; Johnston et al. 2005). Of the six widely cultivated *Brassica* species, *B. rapa* (AA, $2n$ = 20, 529 Mb), *B. nigra* (BB, $2n$ = 16, 632 Mb), and *B. oleracea* (CC, $2n$ = 18, 696 Mb) are the monogenomic diploids. The interspecific breeding between these three diploid species resulted in the creation of three new species of allotetraploid hybrids, namely *B. juncea* (AABB, $2n$ = 36, 1068 Mb), *B. napus* (AACC, $2n$ = 38, 1132 Mb), and *B. carinata* (BBCC, $2n$ = 34, 1284 Mb). Thus, investigation of the *Brassica* genome provides substantial opportunities to study the divergence of gene function and genome evolution associated with polyploidy, extensive duplication, and hybridization.

The close phylogenetic relationship between the *Brassica* species and the model plant, *Arabidopsis thaliana*, implies that knowledge transfer from the results of studies on *Arabidopsis* for crop improvement in the Brassicaceae could be straightforward. The complex genome organization of the *Brassica* species as a result of multiple rounds of polyploidy and genome hybridization, however, renders the identification of orthologous relationships of genes between the genomes highly difficult. Genome triplication, subsequent extensive interspersed gene-loss or gene-gain events, and large-scale chromosomal rearrangements including segmental duplications or deletions in the *Brassica* lineage all complicate the orthologous relationships of the loci between the two genomes (Town et al. 2006; Yang et al. 2006). For this reason, the genomes of several *Brassica* crop species including *B. rapa*, *B. oleracea*, *B. nigra*, and *B. napus* have been characterized in detail over the past decade.

B. rapa is native to East Asia and Europe, where the existence of a large native *B. rapa* population has provided an important resource for the breeding program. There are wide morphological variations in *B. rapa*, including the leafy type (Chinese cabbage and pakchoi), turnip type (vegetable turnip), and oil type (yellow sarson). This great morphological plasticity of *B. rapa* has led to its domestication and selective breeding into a range of different crop types. Several *B. rapa* species are of regional agricultural importance, either as vegetable or oil crops. As a consequence of its native distribution and agronomic usage, *B. rapa* has great potential for use as a model for study of both basic and applied aspects of plant biology. In particular, *B. rapa* ssp. *pekinensis* (Chinese cabbage), one of

the most widely cultivated annual vegetable crops in Northeast Asia, exhibits characteristics that are useful for the study of genome characteristics such as diploidy and small genome size (529 Mb). In response to the need for a simple genetic system with favorable genetic characteristics for research among *Brassica* species, *B. rapa* has played a central role as a model species representing the *Brassica* "A" genome, and is the focus of genome sequencing projects. Genomic studies on *B. rapa* ssp. *pekinensis* cv. *Chiifu* began in 2003 when a bacterial artificial chromosome (BAC) library was constructed, and a collection of BAC-end sequences was initiated. Shortly afterward, a cytogenetic study, the generation of molecular markers, the construction of high-density genetic and physical maps, and eventually large-scale genome sequencing based on the clone-by-clone strategy along with whole-genome shotgun sequencing using the next-generation sequencing techniques were initiated. The information and genomic resources produced during this course of investigations were highly useful for understanding the genetic system of *B. rapa* and sequencing the whole genome. Moreover, these resources have been beneficial for *Brassica* crop breeding, because they enable comparative genomic studies and subsequent transfer of knowledge from *B. rapa* to other *Brassica* crop species. In this chapter, we summarize the genomic resources and physical map pertaining to the *B. rapa* genome as conducted mainly by the Korea *Brassica rapa* Genome Sequencing Project (KBGP).

3.2 Genome Structure

3.2.1 Karyotype

Cytogenetic analyses have showed that the *B. rapa* genome is organized into relatively small, compact chromosomes, with genes concentrated in the euchromatic space, unlike centromeric repeat sequences and rDNAs in the heterochromatin, which are arranged as tandem arrays (Fig. 3.1) (Lim et al. 2005, 2007). The individual chromosome size ranges from 2.1 to 4.5 μm, with a total chromosome length of 32.6 μm. An alternative cytogenetic map based on a pachytene DAPI (4′,6-diamidino-2-phenylindole dihydrochloride) and fluorescent in situ hybridization (FISH) karyogram was also constructed. The mean lengths of 10 pachytene chromosomes were determined by DAPI analysis to range from 23.7 to 51.3 μm, with a total chromosome length of 385.3 μm. Thus, chromosomes in the meiotic prophase stage are 12 times longer than those in the mitotic metaphase, and display a well-differentiated pattern of brightly fluorescing heterochromatin segments (Koo et al. 2004).

The genetic and cytogenetic correspondences between the linkage groups and chromosomes were determined based on the relative positions of 33 sequence tagged site (STS) markers and five repetitive sequences on the metaphase chromosomes using FISH analysis (Lim et al. 2005; Kim et al. 2006). In addition, more than 100 gene-containing BAC clones were analyzed by FISH, and the distribution of five major centromeric and pericentromeric repeats were determined. By the cytogenetic method, all chromosomes can be identified based on their lengths, centromere positions, heterochromatin patterns, and positions of various repeat sequences. Centromeric satellites were estimated to encompass approximately 30 % of the total chromosomes in the mitotic metaphase, particularly in the core centromeric blocks of all the chromosomes. Moreover, in the pachytene FISH, the total length of the pericentromeric heterochromatic regions was estimated to be 38.2 μm, which is approximately 10 % of the total chromosome length (Koo et al. 2004). Thus, overall, heterochromatin comprises approximately 40 % of the total *B. rapa* genome. All gene-containing BACs are localized to the euchromatin. The relationship between the cytogenetic pachytene FISH and the sequence contig distance at several locations in the euchromatin has been estimated to range from 400 to 500 kb per μm. The FISH karyotype has created a rational basis for integrating the molecular, genetic, and cytogenetic maps of *B. rapa*. More importantly, it provides a basis for intelligently targeting the gene space of

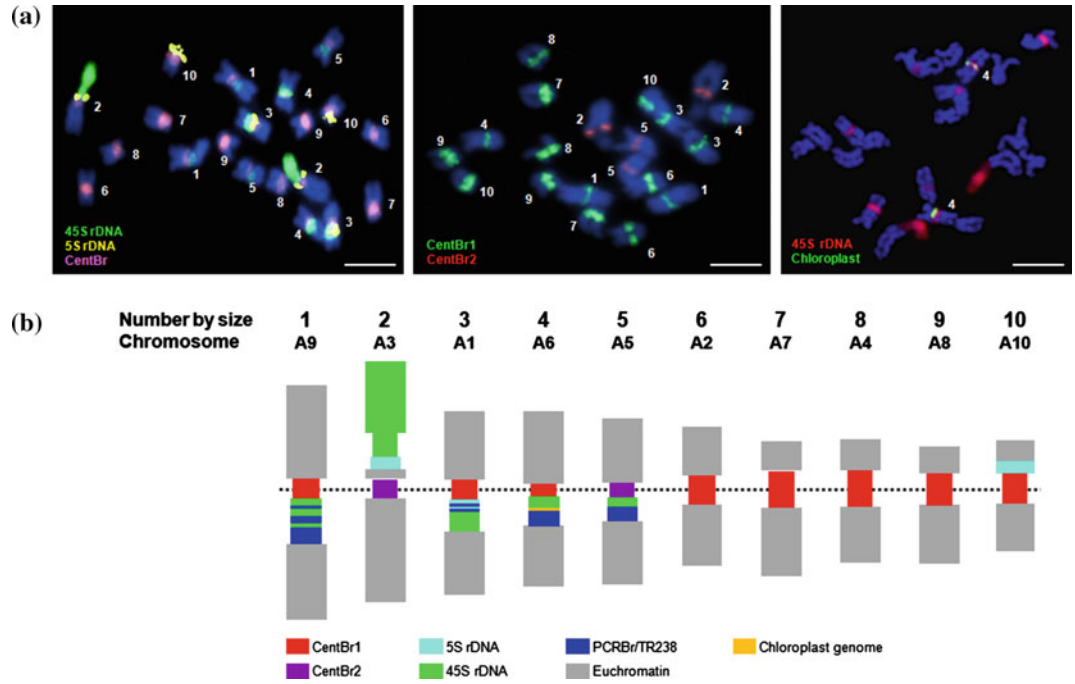

Fig. 3.1 Karyotype of *B. rapa* ssp. *pekinensis* cv. *Chiifu*. **a** Metaphase FISH using fluorescent probes of various repetitive sequences or chloroplast sequence. Chromosome numbers and colored probes are presented. Bars are 5 μm. **b** Idiogram of *B. rapa* chromosomes represented by euchromatic arms and heterochromatic repetitive sequence blocks

B. rapa and for strategic assembly of genome sequences that will facilitate the discovery of most genes.

3.2.2 Heterochromatin

The rapidly evolving centromere structure consists of highly repetitive sequences, such as centromere-specific retrotransposons and tandem satellite repeats. The centromeric repeats characterized in plant genomes often extend over several millions of nucleotides, with 150–180 motifs such as the pAL1 satellite in *Arabidopsis* (Copenhaver et al. 1999), CentC in maize (Ananiev et al. 1998), CentO in rice (Zhang et al. 2004), and MtRs in *Medicago truncatula* (Kulikova et al. 2004). The composition of the repetitive sequences in the *B. rapa* genome has been surveyed by a similarity search of 10,204 BAC-end sequences, with a previously reported tandem repeat that contains a *Hin*dIII site at both ends (Harrison and Heslop-Harrison 1995). Approximately 30 % of the repetitive sequences showed a high similarity, and two kinds of 176 bp tandem repeats were classified with 82 % sequence similarity to each other (Lim et al. 2005). From the FISH data, these 176 bp tandem repeats were localized on the centromeric regions of specific chromosomes, and named as centromeric tandem repeats of *Brassica* 1 & 2 (CentBr1 & CentBr2). The CentBr1 occupies centromeres on eight chromosomes (A1, A2, A4, and A6–A10) and CentBr2 resides on only two chromosomes (A3 and A5). This observation may reflect the predominance of CentBr1 in the *B. rapa* genome. The CentBr repeats were also present in other *Brassica* species studied, particularly in the *B. oleracea* whole-genome shotgun sequences, indicating that these repeats are the major components of the centromeric sequences of the *Brassica* genomes. Furthermore, sequence analysis of the heterochromatin-specific BAC clones identified additional repeat classes, including centromere-specific Ty1-copia-like retrotransposon (CRB), 238 bp-long degenerate

tandem repeat (TR238) arrays, rDNAs, and pericentromere-specific Ty3-gypsy-like retrotransposons (PCRBr) (Lim et al. 2007). CRB was one of the major components of all centromeres in the three diploid *Brassica* species and their three reciprocal allotetraploid hybrids; however, TR238 and PCRBr were A-genome specific. Characterization of these specific centromeric or pericentromeric repeat elements may be important in identifying the heterochromatin/euchromatin borders. In addition to the centromeric or pericentromeric repeats searches, the BAC-end sequence (BES) search also identified many *B. rapa*-specific sequences (~50 % of BES) that had no similarity with any of the sequences of *Arabidopsis*. This finding suggested that the amplification of *B. rapa*-specific sequences, many of which probably form heterochromatic blocks of transposons or tandem repeats, might contribute the genome expansion of *B. rapa* (Lim et al. 2005; Yang et al. 2006). Moreover, the nuclear genome of *B. rapa* includes segments of the chloroplast genome located within the centromeric region of chromosome A6 showing transfer of chloroplast DNA to nuclear genome.

Additional unique group of non-autonomous LTR retrotransposons, terminal-repeat retrotransposons in miniature (TRIM) elements, were identified. TRIM elements are characterized by terminal repeats (TR), ranging from 100 to 250 bp in length, encompassing an internal domain of ~300 bp and creating 5 bp target site duplications. The internal sequence begins with a complement of the primer-binding site of tRNA-methionine and ends with typical polypurine tract motifs. From 96 Mb BAC-end sequences of *B. rapa*, four distinct lineages of TRIMs (Br1–Br4), with lengths ranging from 364 to 1311 bp, were identified (Yang et al. 2007). The estimated copy number of Br TRIMs was more than six-times greater in the *Brassica* species than in *Arabidopsis*, suggesting that various TRIM elements were inserted into the *Brassica* genome after divergence from the *Arabidopsis* lineage. Similar to the case of *Arabidopsis* TRIM elements, many of the Br TRIMs are located in the euchromatic region. The abundant TRIMs in the euchromatin of the *B. rapa* genome are expected to play an important role not only in the reconstruction of the host genome but also in the modification of the gene features by insertion of promoter or terminator sequences residing inside the elements. This modification may act as the driving force for gaining new function, even among the duplicated genes in the *Brassica* genome.

The euchromatic distribution and higher insertion polymorphisms of the Br TRIMs have potential as molecular markers to distinguish the various *Brassica* crop species. For instance, a transposon-display system using the unique sequences of Br1 and Br2 TRIMs was developed. The TRIM display system successfully accessed the genetic diversity in the Brassicaceae family, and effectively identified 16 commercial F_1 hybrids of *B. rapa* and other *Brassica* crops (Kwon et al. 2007).

3.3 Genomic Resources

3.3.1 BAC Libraries and BAC-end Sequences

Various genomic resources are indispensable for genomic study of any crop species. The genomic resources available for *B. rapa* are summarized in Table 3.1. A successful structural genomic study of the *B. rapa* genome relies on the quality and availability of detailed large-insert genomic libraries. Five large-insert BAC libraries of *B. rapa* ssp. *pekinensis* cv. *Chiifu* are publicly available, providing approximately 53-fold genome coverage overall. These libraries were constructed using the restriction enzymes *Eco*RI, *Bam*HI, *Hin*dIII, and *Sau*3AI.

Using these BAC libraries, a total of 260,637 BES have been generated from 146,688 BAC clones (~203 Mb), as a collaborative outcome of the multinational *B. rapa* Genome Sequencing Project consortium. Analysis of BES, combined with BAC sequence surveys, enabled the outlining of the features of the whole-genome structure of *B. rapa*. BLAST search of BES identified that up to 25 % of BES were estimated

Table 3.1 Summary of the genomic resources available for the genomic study of *B. rapa* ssp. *pekinensis* cv. *Chiifu*

Resources	Material	Characteristics	Number[a]	Reference[b]
BAC library	5 libraries		234,544 clones (53.4X)	Mun et al. (2008)
KBrB	*Bam*HI	avr. insert size 120 kb	55,296 clones (12.5X)	
KBrE	*Eco*RI	avr. insert size 139 kb	23,040 clones (6X)	
KBrH	*Hin*dIII	avr. insert size 120 kb	56,448 clones (12.8X)	
KBrS1	*Sau*3AI	avr. insert size 100 kb	55,296 clones (10.5X)	
KBrS2	*Sau*3AI	avr. insert size 132 kb	46,464 clones (11.6X)	
BAC-end sequence	146,688 clones	both end	260,637 reads (203 Mb)	Mun et al. (2009)
KBrB	55,296 clones	single-pass sequence	97,912 reads (75 Mb)	BrGSP (2011)
KBrE	23,040 clones		43,168 reads (33 Mb)	
KBrH	50,688 clones		88,951 reads (73 Mb)	
KBrS1	6144 clones		8117 reads (5 Mb)	
KBrS2	11,520 clones		22,489 reads (17 Mb)	
BAC sequence	KBrB, KBrE, KBrH, KBrS clones	BAC shotgun sequence	886	Mun et al. (2010)
Physical map	67,468 BAC fingerprints	HICF map	1428 contigs (717 Mb)	Mun et al. (2008)
Expressed sequence tag	33 cDNA libraries	cDNA single-pass sequence	152,253 ESTs (91 Mb)	KBGP, NCBI
Microarray	3 NimbleGen chips			
KBGP-24K	24,000 unigenes	60mer, 6 probes/gene		Lee et al. (2008)
KBGP-50K	32,000 unigenes, 17,000 CDS	60mer, 7 probes/gene		KBGP
Br300K	47,584 unigenes	60mer, 7 probes/gene		Dong et al. (2013a, b)

[a]Total sequence length or genome coverage is represented in parenthesis. Genome coverage was estimated based on the haploid genome equivalent of *B. rapa* as 529 Mb

[b]BrGSP, The *B. rapa* genome sequencing project consortium. KBGP, the Korea *B. rapa* Genome Project. NCBI, the National Center for Biotechnology Information

to contain centromeric or pericentromeric repetitive sequences. An additional 15 % of BES were matched with transposons and other repeat sequences. Based on this data and FISH analyses (see Sect. 3.2.2), the heterochromatic region was postulated to occupy >40 % of the *B. rapa* genome, while the euchromatic gene space was postulated to constitute <60 % of the *B. rapa* genome (Yang et al. 2005; Lim et al. 2007). Comparison of the BES with the *B. rapa* ESTs and *Arabidopsis* CDS led to the recognition of approximately 11 % of the BES as protein-coding genes. Assuming that the average CDS size of the BAC survey sequences ranges from 1.1 to 1.3 kb, the estimated number of genes in the whole genome would be approximately 45,000, which is roughly similar to the number of protein coding genes predicted from the genome assembly v1.0 (The *Brassica rapa* Genome Sequencing Project Consortium 2011).

Comparison of the BES with the *Arabidopsis* genome was an efficient application of the BES dataset to select the seed BAC clones for *Brassica* genome sequencing. In silico comparative sequence matching of 91,000 *B. rapa* BES to the *A. thaliana* chromosome sequences identified approximately 50 % of the BES showing significant sequence similarity along with overall

colinearity to counterpart *Arabidopsis* chromosomal regions. Based on the comparative genetic map and microcolinearity between the two genomes, the BAC clones mapped onto the *Arabidopsis* genome can be chosen as seed BAC clones, even before a complete physical map is established. Practically, approximately 890 seed BAC clones were selected and sequenced by this method. The details of BES matching and seed BAC clone selection are explained in the section on BAC sequencing (see Sect. 3.3.3). Almost all the BAC-end sequenced clones were also fingerprinted, and the BES data and physical contig information was combined in the BAC extension program of genome sequencing to select the BAC clones from a seed point to be sequenced (Mun et al. 2009). This tool uses the BLAST algorithm to match one or more input sequences with any sequence associated with the fingerprinted clones, similar to the Blast Some Sequence (BSS) tool in the Finger Printed Contig (FPC) program.

3.3.2 cDNA Libraries and Expressed Sequence Tags

Expressed sequence tags (ESTs), which are short sequences obtained by analysis of cDNA clones, are highly important to support genome annotation and functional study. A total of 152,253 *B. rapa* ESTs have been sequenced enabling rapid access to the sequences of important genes. These ESTs are obtained from 33 cDNA libraries representing a variety of organs and development stages and most of them have been deposited in public databases (Table 3.2). These ESTs were clustered into 39,095 unique sequences (unigenes), including 16,898 tentative consensus sequences. The collection of *B. rapa* ESTs consistently represents the whole *B. rapa* genome, with the *Arabidopsis* genome as a reference. In silico mapping of the 39,095 unigenes on the *Arabidopsis* genome identified 85 % of the unigenes, covering 75 % of the overall *Arabidopsis* counterpart coding sequences. It was found that the remaining 15 % of the *B. rapa* unigenes were not homologous with any gene in *A. thaliana*, rather, this 15 % represented novel *B. rapa*-specific genes. Gene ontology analysis of the ESTs did not show any significant overestimation of specific categories in the *B. rapa* genome, as compared to the *Arabidopsis* genome. Analysis of the EST collections identified 21,409 full-length cDNA sequences. It is anticipated that these sequences contribute to the formation of a *B. rapa*-specific set of sequences, not only for gene prediction programs for *B. rapa* and other *Brassica* species, but also to evaluate structure and alternative splicing of the predicted gene models.

The corresponding cDNA clones have been used to construct microarrays for expression profiling during development, under various biotic- and abiotic-stress conditions. Two early version of microarrays, KBGP-24K and KBGP-50K, were developed using the NimbleGen platform. Both microarrays included the six (KBGP-24K) or seven (KBGP-50K) 60-nucleotide-long probes per gene. The 24K chip covered approximately 24,000 unigenes clustered by 127,144 ESTs from 20 cDNA libraries, whereas the 50K chip doubled the gene contents by including an additional 8500 unigenes, plus 17,500 genes predicted from the seed BAC sequences in the genome sequencing pipeline. These microarrays examined the changes in the genome-wide gene expression of *B. rapa*, in response to transcriptional changes. Using the KBGP-24K chip, genome-wide transcriptome analysis was conducted in response to three abiotic stresses that significantly affect the productivity of *Brassica* crops: salt, cold, and drought (Lee et al. 2008). This analysis successfully identified stress-related genes along with novel transcription factor genes, suggesting the existence of a *B. rapa*-specific signaling pathway that works together with the common stress-response pathway under abiotic stress conditions. After the finish of the genome sequencing project, an advanced version of microarray, Br300K Nimble Gen micro array chip, was manufactured. The Br300K chip covered 47,584 unigenes, of which 32,395 genes were from representative sequences with

Table 3.2 Summary of *B. rapa* ssp. *pekinensis* cv. *Chiifu* ESTs generated from the various libraries

Library	Tissue source	Number of ESTs	NCBI accession
KBAY	Anther, young anther	1859	Ex015357–Ex017215
KBCD	Whole plant, cold treated	6732	Ex017216–Ex023947
KBCG	Callus, developing callus	11,847	Ex023948–Ex035794
KBFL	Floral bud, >2 mm in size	10,332	Ex035795–Ex046126
KBFS	Floral bud, <2 mm in size	9102	Ex046127–Ex055228
KBLS	Whole plant, salt treated	6894	Ex055229–Ex062122
KBLW	Non-photosynthetic mature leaf	2379	Ex062123–Ex064501
KBL	Mature pollen	3587	Ex064502–Ex068088
KBS	Seedling, 1 week old	4174	Ex068089–Ex072262
KBRT	Root, 1 month old	4682	Ex072263–Ex076944
KBSP	Silique, 1–10 days after pollination	6226	Ex076945–Ex083170
KBSQ	Silique, 10–25 days after pollination	933	Ex083171–Ex084103
KBST	Floral stem, bolting	1577	Ex084104–Ex085680
KCOV	Ovule, before pollination	421	Ex115626–Ex116046
KCOW	Ovule, 5–10 days after pollination	1225	Ex116047–Ex127271
KFFB	Floral bud, open flower	8771	Ex085681–Ex094451
KFPC	Leaf, *Pectobacterium carotoborum* infected	3708	Ex094452–Ex098159
KFRT	Root, mixture of 1, 3, and 7 weeks old	3555	Ex098160–Ex101714
KFSD	Mature seed and 2 days old germinating seed	4689	Ex101715–Ex106403
KFYP	Young plant, 3 weeks old	8408	Ex106404–Ex114811
KHCT	Cotyledon, in greening stage	2733	Ex117272–Ex120004
KHLD	Defected leaf	2037	Ex120005–Ex122041
KHLM	Mature green leaf	3423	Ex122042–Ex125464
KHLW	Non-photosynthetic mature leaf	2541	Ex125465–Ex128005
KHOS	Ovule and silique	2390	Ex128006–Ex130395
KHRT	Root, mixed stage and treatment	8265	Ex130396–Ex138660
KLPS	Seedling, 1 week old, etiolated	3,840	Ex138661–Ex142500
NRFB	Floral bud and open flower, normalized	814	Ex114812–Ex115625
KFRC	Root, *Plasmodiophora brassicae* infected	5169	
KFXT	Whole plant, heat treated	6481	
KFFO	Floral organ, mixed stages	8654	
KFFC	Carpel, regular library	2409	
KFFA	Anther, regular library	2396	
Total	33 cDNA libraries	152,253	127,144

cDNA/EST support, and 17,458 genes were predicted from BAC sequences without cDNA/EST support. Seven probes of 60-nucleotide long were designed to cover 150 bp (starting at 60 bp upstream and finishing at 90 bp downstream of the stop codon) of the 3′-region of the gene. This microarray has been used to investigate Ogura-CMS (Dong et al. 2013b) and genic male sterility (Dong et al. 2013a) in *B. rapa*.

3.3.3 BAC Sequences

Cytogenetic studies based on extensive FISH analysis of both metaphase and pachytene chromosomes as described above have provided a detailed insight into the organization of heterochromatic and euchromatic regions. This work demonstrated that the genome of *B. rapa* is organized into distinct regions of pericentromeric heterochromatin, which are rich in repetitive sequences, and gene-rich euchromatin, which are not highly interspersed with heterochromatin. Moreover, sequencing of several BAC clones, chosen because of preliminary evidence that they were gene-rich, has confirmed that the gene density in *B. rapa* is relatively high on the order of 1 gene per 3–4 kb (Yang et al. 2006). Each of the gene-rich BAC clones examined by FISH (>100 BACs) was found to be localized to the visible euchromatic region of the genome (Mun et al. 2009). These data indicate a genome organization where the overwhelming majority of the *B. rapa* euchromatic space can be sequenced in an efficient manner by the clone-by-clone strategy.

The clone-by-clone strategy typically involves a "minimum tiling path" of large-insert (~100 kb) clones, such as BACs of a known order, which is determined using a combination of genetic, physical, and/or cytogenetic mapping. The clone-by-clone sequencing starts from the defined seed points and builds outward. The fingerprint-based physical map, combined with BES and genetic anchoring data, provides a basis for selecting seed BAC clones and for creating a draft tiling path. Alternatively, comparative approaches using an already sequenced closely related model genome, such as the comparative tiling method, can be used as a backbone for in silico clone validation of seed BACs, even before the availability of a physical map. This method includes in silico mapping of BES to the *Arabidopsis* sequences based on unique, significant ($<1E^{-6}$), and directional matches of the paired ends sequences of each BAC with a complement match within the 30–500 kb interval (Yang et al. 2005). Using this method, a total of 4317 BAC clones were mapped onto the *Arabidopsis* chromosomes (Mun et al. 2009). These *B. rapa* BAC clones spanned 93 Mb of the *A. thaliana* sequences, representing ~78 % of the total *Arabidopsis* genome. BAC-FISH and genetic mapping using BES of selected BAC clones positioned on the counterpart *Arabidopsis* chromosomes showed the real euchromatic locations of the BAC clones scattered on the *B. rapa* chromosomes. A single *B. rapa* BAC clone was calculated to span an average of 147 kb of the *A. thaliana* counterpart sequence. Theoretically, 500 contiguous BAC clones can cover approximately 80 Mb of the euchromatic regions of the *Arabidopsis* genome, when assuming the average insert size of a BAC clone is 120 kb. Therefore, if minimally overlapping BAC clones mapped onto the *Arabidopsis* genomes were selected and scattered onto the *B. rapa* chromosomes, they could provide a seed point for bidirectional outward genome sequencing. As a result, 589 minimally tiled *B. rapa* BAC clones spanning 75 Mb of the *A. thaliana* genome were sequenced in phase 3 (finished sequences) or phase 2 (sequences that are fully oriented and ordered, but contain some small sequence gaps and low-quality regions). Most of them were distributed onto the 10 *B. rapa* chromosomes by genetic mapping, FISH analysis, and physical contig information. The 589 sequenced BAC clones were provided to the genome sequencing project as seed BACs and used as starting points for chromosome sequencing and references for validation of whole genome sequence assembly. Integration of seed BACs with the physical map provides "gene-rich" contigs spanning ~160 Mb (see Sect. 3.4).

The structures of the potential protein coding genes in the BAC clones were predicted ab initio using FGENESH (www.softberry.com) based on a *B. rapa* matrix (Mun et al. 2009). The structural features of the protein coding genes in *B. rapa* showed a smaller average gene length (1.6 kb) as compared to *Arabidopsis* (2.2 kb). This difference appears to amount to almost one less exon per gene (4.7 and 5.8 exons per gene in *B. rapa* and *A. thaliana*, respectively) along with a shorter exon (225 bp in *B. rapa* and 268 bp in *A. thaliana*, respectively) and intron length (141 bp in *B. rapa* and 165 bp in *A. thaliana*, respectively) in *B. rapa*. The gene density in the sequenced BAC clones of the *B. rapa* genome (one gene per 4.2 kb) is higher than that in the *Arabidopsis* genome (one gene per 4.5 kb), indicating the compact organization of the euchromatic region of *B. rapa*. A similarity search of the protein coding genes of *B. rapa* against public databases indicated that approximately 18 % of the predicted genes showed no significant similarity with any of the genes reported. This result is roughly consistent with the results of previous studies conducted using EST analysis (see Sect. 3.3.2) and synteny analysis of *FLC* regions (Yang et al. 2006). Repetitive sequence analysis revealed that 6 % of the seed BAC sequences are composed of transposons, 2-fold higher proportions than those identified in the counterpart *Arabidopsis* euchromatic genome, presumably due to the increased number of LTRs and long interspersed elements (LINEs). In addition, low complexity repetitive sequences were quite abundant in the *B. rapa* euchromatic region, representing *B. rapa*-specific expansion of repetitive sequences. The distribution of repetitive sequences and transposons along the chromosomes was not uneven. This result suggests that transposons were partly responsible for genome expansion in *B. rapa.*

As of December 2011, a total of 107 Mb from 886 BAC clones (401 phase 3, 445 phase 2, and 40 phase 1 clones) were sequenced and deposited in the HTGS database of NCBI. Besides the reported BAC clones, KBGP sequenced additional 884 BAC clones as a result of sequencing of the chromosomes A3 and A9. Using 670 BAC clones, the sequence scaffolds for the two chromosomes were constructed. In the case of A3, 348 minimum tiled BAC sequences generated nine sequence scaffolds spanning 31.9 Mb (Mun et al. 2010); in the case of A9, 289 minimally tiled BAC clones generated 15 scaffolds comprising 28.5 Mb. These sequences provided near-complete chromosome sequences and served as references of the whole genome shotgun sequence assembly.

3.4 Physical Mapping

3.4.1 High-Information Content Fingerprinting

The availability of a genome-wide, sequence-ready physical map is one of the crucial components for a successful clone-by-clone strategy because a map generated by genetic techniques is rarely sufficient for directing the sequencing phase of a genome project. A physical map not only makes it possible to determine clones for genome sequencing with comprehensive coverage and reduced sequencing redundancy, but also enables one to simplify the sequence assembly by arranging the sequence contigs in order. Physical mapping uses molecular biology techniques to examine DNA molecules directly in order to construct maps showing the positions of sequence features. Thus far, the utility of physical maps has been reported by major genome sequencing projects on the human (The International Human Genome Mapping Consortium 2001), *A. thaliana* (Marra et al. 1999), rice (Chen et al. 2002), and *M. truncatula* (Mun et al. 2006). These physical maps were constructed using a combination of techniques, including restriction enzyme-digested BAC fragment fingerprinting on agarose gels and assembly of the fingerprints using the FPC software package (Soderlund et al. 2000). The agarose method has been successful, but has limited throughput because of the need for manual band calling. Alternatively, fluorescence-labeled

fingerprinting methods have been used to make larger and more accurate contigs, with increased throughput using an automatic capillary sequencer (Gregory et al. 1997; Ding et al. 2001; Luo et al. 2003; Xu et al. 2004). Fluorescence-labeled capillary electrophoresis methods include the 3-enzyme method and the high-information content fingerprinting (HICF) methods, which use type IIS restriction enzymes or SNaPshot labeling techniques, respectively. As compared to the agarose method, the automatic workflow and higher resolution of these methods facilitates better physical map construction, both in terms of throughput and the quality of fingerprinting. The first genome-wide plant HICF physical map was constructed for maize (Nelson et al. 2005).

3.4.2 Physical Map Based on the HICF Method

A genome-wide BAC-based physical map of *B. rapa* was constructed by the SNaPshot HICF method (Mun et al. 2008). Figure 3.2 shows a workflow of physical map construction based on HICF method. To create a robust sequence-ready physical map, a total of 99,456 BAC clones from the three independent BAC libraries (~22.5X coverage) were fingerprinted by digestion with combinations of five restriction enzymes (*Eco*RI, *Bam*HI, *Xba*I, *Xho*I, and *Hae*III), followed by SNaPshot reagent labeling of four colors at the 3′-ends of the restriction fragments, then sizing on the ABI 3730xl capillary sequencer. Of the

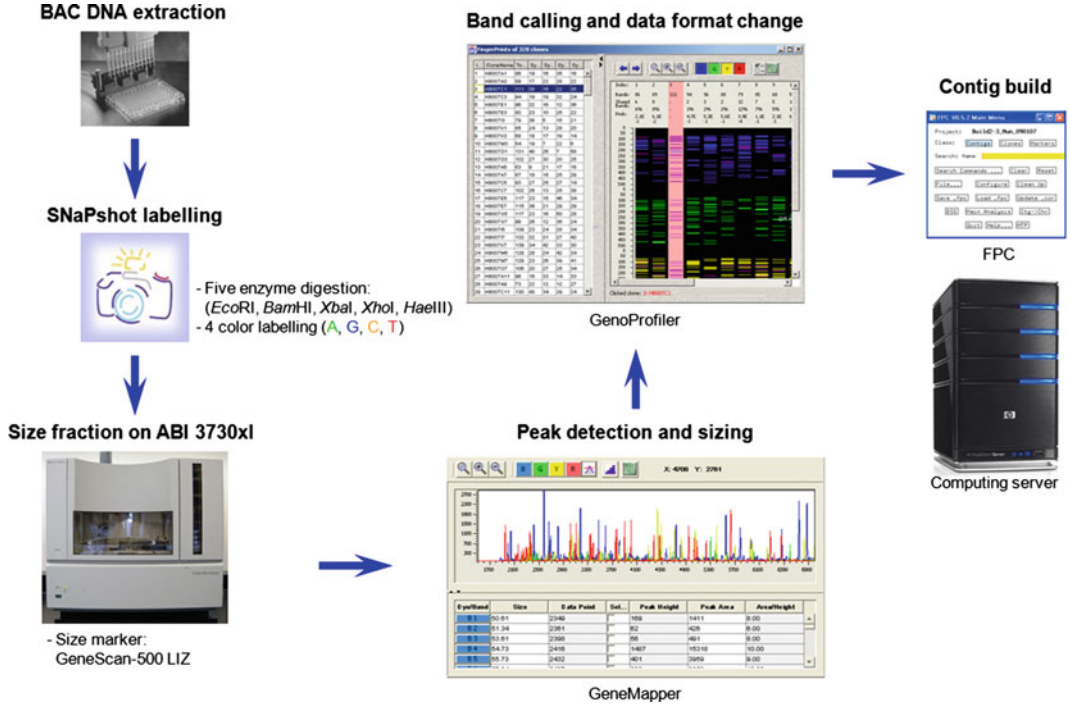

Fig. 3.2 A workflow of physical map construction based on the HICF method. BAC DNAs extracted from BAC clones are fingerprinted by digestion with five restriction enzyme combinations (*Eco*RI, *Bam*HI, *Xba*I, *Xho*I, and *Hae*III) followed by SNaPshot reagent labeling of four colors at the 3′-ends of the restriction fragments and sizing on the ABI 3730xl capillary sequencer. The size of DNA fragments from the capillary fingerprinting chromatograms is collected by GeneMapper. The fingerprint data is then imported to GenoProfiler to change data format suitable for FPC analysis. The fingerprints are assembled into contigs, which are accurately ordered contiguous overlapping clone sets

fingerprints, a total of 93,689 clones (94.2 %) were successfully fingerprinted to be used for contig assembly. From the initial dataset, 26,221 BAC clones containing heterochromatic repetitive sequences were removed from the contig assembly, which significantly enriches the euchromatic contigs in the resulting build. The physical contig map was assembled using 67,468 high-quality, heterochromatic repeat-free BAC fingerprints from the initial dataset. These BAC clones represent 15.2X coverage of the *B. rapa* genome; they were condensed into 1,417 contigs, and the resulting contigs were manually edited to validate reliability. With the results of the contig evaluation, manual editing of the initial contig build yielded 1428 contigs, with an average length of 512 kb spanning 717 Mb, 1.3X coverage of the genome (Table 3.3). An unsatisfactory aspect of this assembly was its large number of Q clones, as the Q clones in this assembly corresponded to 15 % of the clones. However, three specific deep contigs contributed to 48.3 % of all the Q clones in the build. Thus, when these deep contigs of the initial build were excluded due to false-positive overlaps, the Q clones in the remaining contigs correspond to 7.7 % of all clones. This ratio is similar to the ratios reported in catfish (7.3 %) (Quiniou et al. 2007) and maize (11 %) HICF maps (Nelson et al. 2005).

The contigs produced in the course of the fingerprinting work were tagged with 315 anchored STS markers. Practically, the important aspect of this was the integration of a physical map into a genetic map, enabling the positioning of 242 gene-rich contigs to specific locations on 10 chromosomes and providing seeds for the genome sequencing effort. An example of such a contig is shown in Fig. 3.3. The number of contigs associated with genetic loci was ~160.7 Mb, or 30 % of the total genome. The total coverage of the physical contigs suggests that most contigs do not have sufficient overlaps, and the gaps between the contigsneed to be filled by additional fingerprinting. To improve the map, additional fingerprinting of approximately 30,000 clones of two new BAC libraries (KBrE and KBrS2) was performed. This data is being merged into the current build to continue the refinement of the physical map. Linkage analysis of SSRs and single nucleotide polymorphisms (SNPs) in the seed BACs and BESs has been carried out to provide more anchoring points for the physical map. It is noteworthy that the current assembly of the *B. rapa* genome did not cover the heterochromatic region (The *Brassica rapa* Genome Sequencing Project Consortium 2011). In this regard, physical contigs along with BAC clones can be valuable resources to investigate the heterochromatin that has not yet been characterized. Refinement of the physical map and information of physical contigs will play an important role in further study of the *B. rapa* genome.

Table 3.3 Summary of the *B. rapa* physical map constructed by the HICF method

Number of clones used for the map construction	67,468	
Number of singletons	14,816	
Number of contigs	1428	
>200 clones	32	
101–200 clones	73	
51–100 clones	176	
26–50 clones	244	
10–25 clones	284	
1–9 clones	619	
Physical length of the contigs (Mb)	717	
Number and length of contigs anchored to chromosome[a]	242	(160.7)
R1	18	(13.6)
R2	19	(13.2)
R3	57	(36.5)
R4	6	(2.0)
R5	18	(9.0)
R6	17	(13.8)
R7	14	(12.7)
R8	17	(12.2)
R9	66	(39.7)
R10	10	(8.1)

[a]Length of physical contigs anchored to each chromosome is represented as Mb in parenthesis

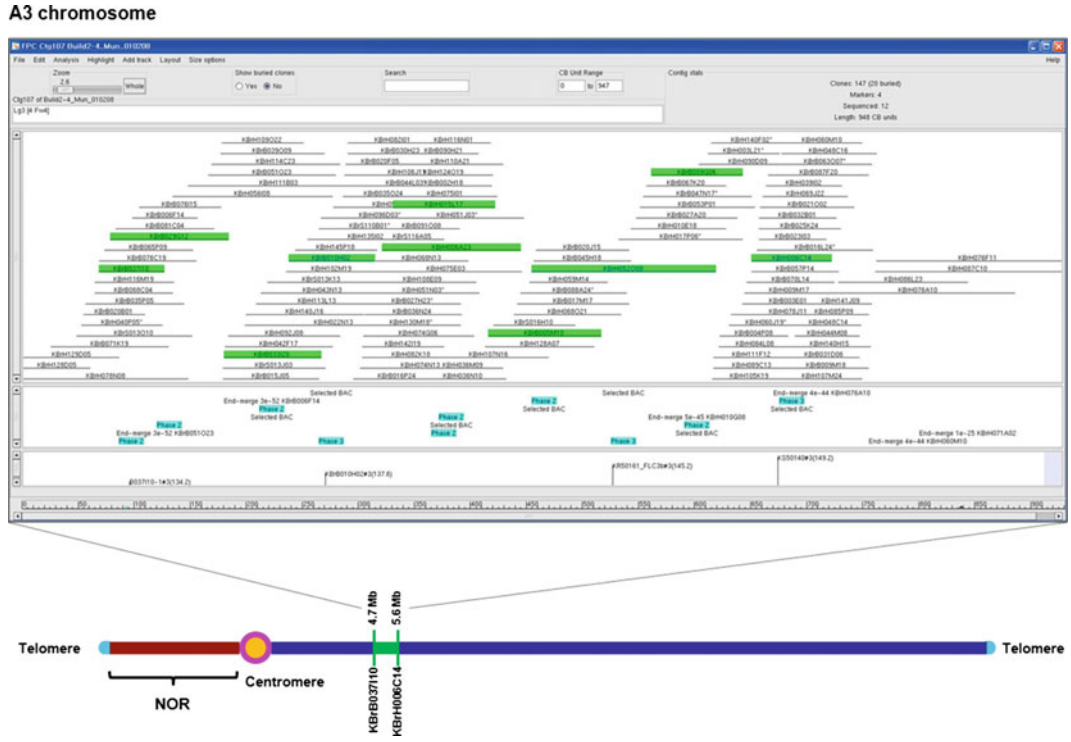

Fig. 3.3 An example of a physical contig used as a platform in BAC selection for sequencing. This contig consists of 147 BAC clones from 3 BAC source libraries and is estimated to cover approximately 1.1 Mb. All the green highlighted BAC clones were fully sequenced. This contig was anchored to the region around 4.7–5.6 Mb of the chromosome A3 based on sequence comparison. The clones prefixed with KBrH are from the *Hin*dIII library, those prefixed with KBrB are from the *Bam*HI library, and those prefixed with KBrS are from the *Sau*3AI library (Table 3.1). NOR, nucleolar organizer region

3.5 Conclusion

B. rapa provides an excellent reference genome sequence, as well as valuable information for understanding the genetic systems of the *Brassica* crop species. The most significant beneficiaries of *B. rapa* sequences will be *Brassica* crop researchers, from breeders to plant biologists. A physical map, BAC clones, BAC-end sequences, and associated sequence information have accelerated genome mapping and whole genome sequencing. Furthermore, a large number of *B. rapa* EST collections, cDNA clones, and the genome sequences will enable the identification and isolation of the genes of interest for plant biology and agriculture. The comparative genomics approach for *B. rapa* will also benefit genomic investigation of closely related *Brassica* crops, including *B. oleracea*, *B. nigra*, *B. napus*, and even *Raphanus* and *Sinapis* species. In fact, quantitative trait loci (QTL) or association mapping of valuable phytochemical-related genes in *B. napus* has utilized the available *B. rapa* genome sequence data and resources. In addition, genome sequencing of other *Brassica* crops, particularly the construction of sequence assemblies and scaffolds of *B. napus*, has been improved owing to the presence of information on the *B. rapa* genome. Therefore, there is no doubt that increasing genomic resources will facilitate the molecular genetic studies on *B. rapa* and will eventually contribute to improving *Brassica* crops.

Acknowledgments This work was supported by grants from the Next-Generation Biogreen21 program (PJ011086), Rural Development Administration, Korea.

References

Ananiev EV, Phillips RL, Rines HW (1998) Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions. Proc Natl Acad Sci USA 95:13073–13078

Beilstein MA, Al-Shehbaz IA, Kellogg EA (2006) Brassicaceae phylogeny and trichome evolution. Am J Bot 93:607–619

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438

Chen M, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B et al (2002) An integrated physical and genetic map of the rice genome. Plant Cell 14:537–545

Copenhaver GP, Nickel K, Kuromori T, Benito MI, Kaul S et al (1999) Genetic definition and sequence analysis of *Arabidopsis* centromeres. Science 286:2468–2474

Ding Y, Johnson MD, Chen WQ, Wong D, Chen YJ et al (2001) Five-color-based high-information content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. Genomics 74:142–154

Dong X, Feng H, Xu M, Lee J, Kim Y et al (2013a) Comprehensive analysis of genic male sterility related genes in *Brassica rapa* using a newly developed Br300K oligomeric chip. PLoS One 8:e72178

Dong X, Kim W, Lim Y-P, Kim Y-K, Hur Y (2013b) Ogura-CMS in Chinese cabbage (*Brassica rapa* ssp. *pekinensis*) causes delayed expression of many nuclear genes. Plant Sci 199–200:7–17

Economic Research Service, USDA (2008) Vegetables and melons outlook. http://www.ers.usda.gov/Publications/VGS/Tables/World.pdf

Gregory SG, Howell GR, Bentley DR (1997) Genome mapping by fluorescent fingerprinting. Genome Res 7:1162–1168

Harrison GE, Heslop-Harrison JS (1995) Centromeric repetitive DNA sequences in the genus *Brassica*. Theor Appl Genet 90:157–165

Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J et al (2005) Evolution of genome size in Brassicaceae. Ann Bot 95:229–235

Kim JS, Chung TY, King GJ, Jin M, Yang TJ et al (2006) A sequence-tagged linkage map of *Brassica rapa*. Genetics 174:29–39

Koo DH, Plaha P, Lim YP, Hur Y, Bang JW (2004) A high-resolution karyotype of *Brassica rapa* ssp. *pekinensis* revealed by pachytene analysis and multicolor fluorescence in situ hybridization. Theor Appl Genet 109:1346–1352

Kulikova O, Geurts R, Lamine M, Kim DJ, Cook DR et al (2004) Satellite repeats in the functional centromere and pericentromeric heterochromatin of *Medicago truncatula*. Chromosoma 113:276–283

Kwon SJ, Kim DH, Lim MH, Long Y, Meng JL et al (2007) Terminal repeat retrotransposon in miniature (TRIM) as DNA markers in *Brassica* relatives. Mol Genet Genom 278:361–370

Lee SC, Lim MH, Kim JA, Lee SI, Kim JS et al (2008) Transcriptome analysis in *Brassica rapa* under the abiotic stresses using *Brassica* 24K oligo microarray. Mol Cells 26:595–605

Lim KB, de Jong H, Yang TJ, Park JY, Kwon SJ et al (2005) Characterization of rDNAs and tandem repeats in the heterochromatin of *Brassica rapa*. Mol Cells 19:436–444

Lim KB, Yang TJ, Hwang YJ, Kim JS, Park JY et al (2007) Characterization of the centromere and peri-centromere retrotransposons in *Brassica rapa* and their distribution in related *Brassica* species. Plant J 49:173–183

Lukens LN, Quijada PA, Udall J, Pires JC, Schranz ME et al (2004) Genome redundancy and plasticity within ancient and recent *Brassica* crop species. Biol J Linn Soc Lond 82:665–674

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S et al (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82: 378–389

Marra M, Kucaba T, Sekhon M, Hillier L, Martienssen R et al (1999) A map for sequence analysis of the *Arabidopsis thaliana* genome. Nat Genet 22:265–270

Mun J-H, Kim DJ, Choi HK, Gish J, Debelle F et al (2006) Distribution of microsatellites in the genome of *Medicago truncatula*: a resource of genetic markers that integrate genetic and physical maps. Genetics 172:2541–2555

Mun J-H, Kwon SJ, Yang TJ, Kim HS, Choi BS et al (2008) The first generation of a BAC-based physical map of *Brassica rapa*. BMC Genom 9:280

Mun JH, Kwon SJ, Yang TJ, Seol YJ, Jin M et al (2009) Genome-wide comparative analysis of the *Brassica rapa* gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. Genome Biol 10:R111

Mun JH, Kwon SJ, Seol YJ, Kim JA, Jin M et al (2010) Sequence and structure of *Brassica rapa* chromosome A3. Genome Biol 11:R94

Nagaharu U (1935) Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. Jpn J Bot 7:389–452

Nelson WM, Bharti AK, Butler E, Wei F, Fuks G et al (2005) Whole-genome validation of high-information-content fingerprinting. Plant Physiol 139:27–38

O'Neill CM, Bancroft I (2000) Comparative physical mapping of segments of the genome of *Brassica*

oleracea var. *alboglabra* that are homoelogous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. Plant J 23: 233–243

Quiniou SMA, Waldbieser GC, Duke MV (2007) A first generation BAC-based physical map of the channel catfish. BMC Genom 8:40

Schmidt R, Acarkan A, Boivin K (2001) Comparative structural genomics in the Brassicaceae family. Plant Physiol Biochem 39:253–262

Soderlund C, Humphray S, Dunham I, French L (2000) Contigs built with fingerprints, markers, and FPC V4.7. Genome Res 11:934–941

The Brassicarapa Genome Sequencing Project Consortium (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1040

The International Human Genome Mapping Consortium (2001) A physical map of the human genome. Nature 409:934–941

Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ et al (2006) Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. Plant Cell 18:1348–1359

Xu Z, Sun S, Covaleda L, Ding K, Zhang A et al (2004) Genome physical mapping with large-insert bacterial clones by fingerprint analysis: methodologies, source clone genome coverage, and contig map quality. Genomics 84:941–951

Yang TJ, Kim JS, Kwon SJ, Lim KB, Choi BS et al (2006) Sequence-level analysis of the diploidization process in the triplicated *FLOWERING LOCUS C* region of *Brassica rapa*. Plant Cell 18:1339–1347

Yang TJ, Kim JS, Lim KB, Kwon SJ, Kim JA et al (2005) The Korea *Brassica* Genome Projects: a glimpse of the *Brassica* genome based on comparative genome analysis with *Arabidopsis*. Compar Funct Genom 6:138–146

Yang TJ, Kwon SJ, Choi BS, Kim JS, Jin M et al (2007) Characterization of terminal-repeat retrotransposon in miniature (TRIM) in *Brassica* relatives. Theor Appl Genet 114:627–636

Zhang Y, Huang Y, Zhang L, Li Y, Lu T et al (2004) Structural features of the rice chromosome 4 centromere. Nucl Acid Res 32:2023–2030

De Novo Genome Assembly of Next-Generation Sequencing Data

4

Min Liu, Dongyuan Liu and Hongkun Zheng

Abstract

With rapid development of next-generation sequencing (NGS) technologies, de novo genome assembly appears increasingly common. However, inherent features of NGS data pose great challenges for de novo genome assembly. Many genomes, such as *Brassica rapa*, having undergone three paleo-polyploidy events, contain high content repeats, makes genome assembly of NGS data tougher. In past several years, numerous algorithms have been developed to address the challenges in de novo genome assembly from NGS reads. Here we summarize the main approaches for genome assembly. We also describe several algorithms for each approach. In addition, we compare the performance of existing assemblers in the accuracy and contiguity of assemblies. The comparative analysis shows that there is not any assembler that performs best in all the observed measures, which are also dependent on the dataset used.

4.1 Introduction

Rapid development of next-generation sequencing (NGS) technologies has greatly reduced DNA sequencing costs and made genome assembly increasingly common. However, inherent features of NGS data also pose new challenges for de novo genome assembly. In the past several years, numerous algorithms have been developed to cope with the challenges in de novo genome assembly from NGS reads. Most adopt the de Bruijn graph approach (Li et al. 2012), where a vertex represents a unique length-k substring called k-mer, and an edge connects two vertices if they appear consecutively in a read (Compeau et al. 2011). A few use the overlap–layout–consensus (OLC) approach, such as Edena (Hernandez et al. 2008) and string graph assembler (SGA) (Simpson and Durbin 2012). There are also some extension-based algorithms available for NGS reads, which do extension from 5′ or 3′ terminal of read by k-mer or read, such as SSAKE (Warren et al. 2007) and JR-Assembler (Chu et al. 2013).

M. Liu · D. Liu · H. Zheng (✉)
Biomarker Technologies Corporation, Beijing 101300, China
e-mail: zhenghk@biomarker.com.cn

M. Liu
e-mail: lium@biomarker.com.cn

D. Liu
e-mail: liudy@biomarker.com.cn

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_4

NGS reads are very short and error-prone, compared with traditional Sanger sequencing. To assemble this new kind of sequencing data, several assemblers, represented by MaSuRCA (Zimin et al. 2013), firstly construct longer reads (super-read), then assembly super-reads into contigs (Butler et al. 2008; Gnerre et al. 2011; Zimin et al. 2013). The basic construction process of super-reads is to extend each original read forwards and backwards, base by base, as long as the extension is unique. All reads that extend to the same super-read are replaced by that super-read. This allows subsequent computation quick and thus reduces memory requirements. Different from MaSuRCA, allpaths-lg uses an overlapping paired-end library with a suitable insert size to generate super-reads for contig assembly (Butler et al. 2008; Gnerre et al. 2011). These types of super-reads allow assembler to use OLC strategies with a few representative reads or de Bruijn graph approach with big *k*-mer.

Despite considerable progress made in the past years, genome assembly remains challenging. For example, recent completion of *Brassica rapa* genome sequencing has revealed that there are high content of repeats in *B. rapa* accession chiifu-401-42, which make many mate-pair library can't span the two side of unique region. Hence, the assembly had to use 199,452 BAC-end Sanger sequences, which have very long insert size to construct the super scaffold. Despite these efforts, the N50 of the assembled genome only reaches 1.9 Mb, with 283.8 Mb of the total size (Wang et al. 2011), much smaller than the real genome size ($2n = 2x = 529$ Mb). Therefore, there is a great need to provide novel algorithms and assemblers for de novo genome assembly of NGS data.

4.2 The Challenge of Genome Assembly

The primary difficulty in genome assembly is to merge overlapping reads along continuous sequences. First, contigs that the assembly algorithms produce are not complete and do not cover the entire chromosomes, due to sequencing errors and the existence of unsequenced parts. Even with high coverage, there is still a nonzero probability for the existence of unsequenced parts and sequencing errors. Second, repeats and heterozygous sequences will further complicate the assembly.

In order to assemble a genome, we first need to sequence random DNA fragments from the whole genome. The rapidly decreasing costs of NGS allow us to rapidly obtain vast amounts of DNA sequence data at a low cost. Unfortunately, the sequence length of NGS data is much shorter than that of the genomes or genomic features being studied, which commonly spans tens of thousands to billions of base pairs. Hence, many analyses starting with the computational process of sequence assembly that joins together the many sequence fragments the NGS generates. The workflow of a typical assembly algorithm is shown in Fig. 4.1.

Second, assembly algorithm will merge sequence fragments into contigs. A sequence contig is a contiguous, overlapping sequence read, which is assembled with the small DNA fragments generated by bottom-up sequencing strategies. Contig assembly is difficult in the

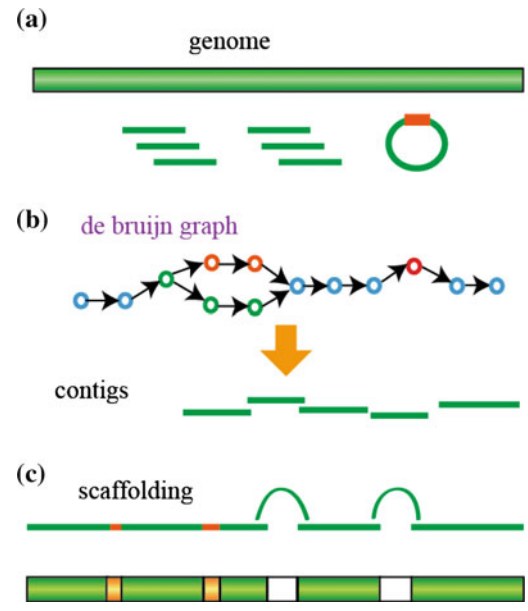

Fig. 4.1 The workflow of a typical assembly algorithm

process of genome assembly. Assemblers use the overlapping information of fragment to search contiguous paths. The sequence will be broken when faced with the branch, which may come from the sequencing error or repeats in the genome. Heterozygous sequences also can produce branches. For diploid species, there may be two paths for one single nucleotide polymprphism (SNP). For polyploidy, there may be many paths for different regions. Because of these enormous difficulties, contig assembly is important.

Third, contigs will be ordered to construct scaffold. Paired-end read libraries are useful in genome assembly. These data can help to extend contigs and resolve repeat areas. If one end of a paired-read is assembled in a contig and the other end in a second contig, it can be inferred that these contigs are adjacent in the final assembly. Because there may are erroneous links, assemblers need to filter out low weight links. For example, many assemblers only keep the links, which contain at least three or five paired-reads. There also exists the strategy, which firstly uses high weight links, and then makes use of low-weight links to form scaffolds. Recently, optical mapping is increasingly being used to order contigs or scaffolds.

Finally, gap closing will be used to fill the gaps in the scaffolds. After scaffolding, many assemblers will remap pair-end reads onto contigs and get linking information between them. The local un-assembly reads will be retrieved. Unaligned reads of the single aligned pair-end always can align multiple regions of genome. In small local regions, read overlapping information will be used to form sequences with much lenient standard.

4.3 De Novo Assembly Algorithm

4.3.1 Classification of De Novo Assembly Algorithm

Most of existing de novo assembly tools for NGS platforms utilize the de Bruijn graph approach (Li et al. 2012). In the de Bruijn graph, a vertex represents a unique length-*k* substring called *k*-mer, and an edge connects two vertices if they appear consecutively in a read (Compeau et al. 2011). There also exist some assemblers, which apply the overlap–layout–consensus (OLC) approach for handling NGS reads, such as Edena (Hernandez et al. 2008) and SGA (Simpson and Durbin 2012). Additionally, some extension-based assemblers also appeared to assemble NGS reads, which do extension from 5′ or 3′ terminal of read by *k*-mer or read, such as SSAKE (Warren et al. 2007) and JR-Assembler (Chu et al. 2013).

4.3.2 The Overlap Layout Consensus Approach

The sequence assembly problem can be taken as a graph problem by making an overlap graph of reads. In the overlap graph, reads are presented as nodes, and the existing overlap between two reads is presented as an edge between corresponding nodes. A modified version of the Smith-Waterman dynamic programming algorithm is usually used to find overlapping reads in almost all assemblers (Gnerre et al. 2011; Zimin et al. 2013).

In an overlap graph, assembling the reads into the genome is equivalent to finding a Hamiltonian path, a path that contains each node exactly once. Unfortunately, finding a Hamiltonian path is an NP-complete problem, which cannot be done in polynomial time.

Overlap layout consensus methods are based on graph theory. In these methods, an overlap graph is built from reads and the assembly problem is simplified to find a Hamiltonian path in the graph. ARACHNE (Metzker 2010), Celera (Li et al. 2012) and its revised version for short-reads (Peng et al. 2012), CAP3 (Jaillon et al. 2007), and Newbler (Zhang et al. 2012) use this method as their core idea.

An OLC algorithm starts by searching overlaps between reads (or graph nodes) (Fig. 4.2). In fact, it must check possible overlaps between any two reads in the input read set. The layout step will simplify the overlap graph by removing redundant information and will put these reads together using identified overlaps. The final step

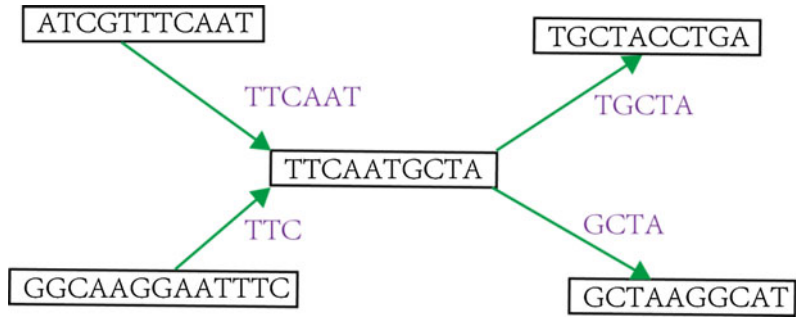

Fig. 4.2 An OLC assembly graph. Nodes are complete reads, and edges connect reads that overlap. Note that in an actual OLC assembly graph, reads and overlaps would be much larger. Here, theoretical reads and overlaps are shortened for clarity

is to find a consensus for the existing layout. The overlap step is computationally intensive. Therefore, this approach is more suitable for whole genome shotgun sequencing reads that Sanger sequencing technology produces. Similarly, the Hamiltonian path problem is an NP-complete problem in itself. It needs heuristic solutions.

4.3.3 De *Bruijn Graph Approach*

In 2001, Pevzner and Tang introduced a method based on the Eulerian path approach for assembling NGS reads (Pevzner and Tang 2001). In the new approach, reads are cut into smaller but regular pieces, called *k*-mers, which are then used to create a de Bruijn graph. By reducing the fragment assembly to a de Bruijn graph, the NP-complete Hamiltonian path is transformed to seek a Eulerian path in a de Bruijn graph. This approach avoids the complicated step of searching all overlaps between reads, which are required to form an overlap graph in the case of overlap layout consensus approach. There are polynomial time algorithms for finding Eulerian path problems. However, in practice, there may be several Eulerian paths in de Bruijn graphs. Finding the shortest Eulerian super path is still NP-hard (Zerbino and Birney 2008). Existing algorithms use heuristic methods to compute this super path by modifying the Eulerian graph. In addition, De Bruijn graph approach also simplifies the sequence repeat issue.

A *k*-dimensional de Bruijn graph is a directed graph whose nodes are all possible length-*k* sequences of *m* symbols. Obviously, each *k*-dimensional de Bruijn graph of *m* symbols has *mk* vertices. A de Bruijn graph is a representation based on all *k*-mers (length *k* words), which makes it suitable for high-coverage, very short-read data.

An edge in de Bruijn graphs connects two vertices (*k*-mers), if one vertices postfix of length $k - 1$ is equal to the prefix of the other one with the same length. The edge is directed, and the direction is from the *k*-mer, including the postfix to the *k*-mer including the prefix.

Given a sequence (GGATCGTTTCGTAAT), one can make a de Bruijn graph of it. To create a de Bruijn graph, it is enough to put the directed edges in the graph according to the sequence. The de Bruijn graph for this set is shown in Fig. 4.3.

In de Bruijn graph approach assembly algorithms, the graphs of input reads are created and then paths in graphs are used to detect contigs. Finding Eularian paths is the key to finding contigs in this step. Optionally, the algorithm may use other data, such as paired-end data, in order to make longer contigs and complete the assembly process. The need for predefined *k*-value, and also errors in reads that lead to a complex graph structure, are of issues in de Bruijn graph-based assembly algorithms.

The Euler assembler (Kent 2002) is the first algorithm that uses the de Bruijn approach for handling sequence assembly problems.

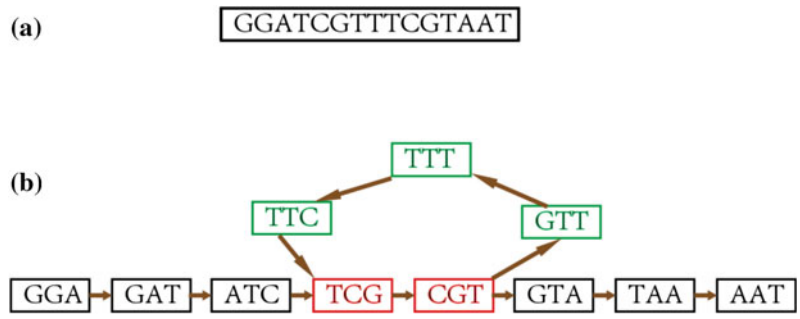

Fig. 4.3 The de Bruijn graph of an input set (GGATCGTTTCGTAAT)

Velvet (Zerbino and Birney 2008), Euler-USR (Chaisson et al. 2009), AllPaths (Butler et al. 2008; Maccallum et al. 2009), Abyss (Simpson et al. 2009), and IDBA (Peng et al. 2012) are some other assembly algorithms that use this approach.

4.3.4 *Extension*-Based Approach

The shotgun sequence assembly problem was first formalized by finding the shortest common superstring of the set of all reads (Delcher et al. 2002). Since this algorithm is computationally NP-complete, greedy approaches were introduced to solve the problem. The greedy approach uses a greedy idea, that is, to merge two reads with maximum overlap score at the time (Fig. 4.4). Reads and overlaps are considered to be nodes of graph and edges between nodes in a graph, respectively. Now the problem is simplified to find a Hamiltonian path in the graph.

Greedy algorithms for read assembly can be written in the following steps:

1. Calculate pairwise alignments of all fragments.
2. Choose two fragments with the largest overlap.
3. Merge chosen fragments.
4. Repeat steps 1, 2 and 3 until only one fragment is left.

The main problem of this approach is getting stuck in local maxima, as in the cases of all greedy algorithms. A local maxima can occur if the current contig takes on reads that would help further contigs grow even larger. Examples

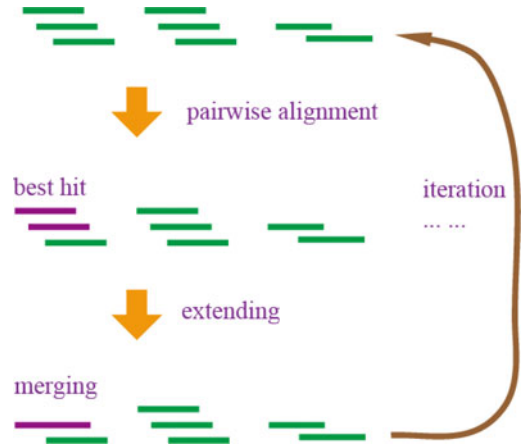

Fig. 4.4 The main steps in greedy algorithms for genome assembly

of algorithms using a greedy approach are PE-Assembler (Ariyaratne and Sung 2011), SSAKE (Warren et al. 2007), SHARCGS (Dohm et al. 2007), and VCAKE (Miller et al. 2010).

4.4 Comparison of Algorithms

4.4.1 Datasets and Assemblers

SPAdes 3.0 (Bankevich et al. 2012), MaSuRCA 2.2.1 (Zimin et al. 2013), SOAPdenovo2 (Li et al. 2010; Luo et al. 2012), and ALLPATHS-LG 44683 (Butler et al. 2008; Gnerre et al. 2011) were compared with nine bacterial data sets. ABySS 1.2.6 (Simpson et al. 2009), Edena 2.1.1 (Hernandez et al. 2008), SOAPdenovo 1.0.5,

Table 4.1 The basic information of the nine bacterial data sets of next generation sequence used for the assembly comparison

Species	Genome size (bp)	GC (%)	Library			Coverage (Gb)	SRA
			No.	Size	Type		
Acinetobacter baumannii NIPH24	3,893,975	39.16	1	180	PE	1.1	SRX236318
			2	5k	MP	1.1	SRX221053
Acinetobacter indicus CIP110367	3,211,639	45.34	1	180	PE	0.52	SRX342013
			2	180	PE	0.51	SRX342012
			3	5k	MP	0.71	SRX342014
			4	5k	MP	0.66	SRX342011
Enterobacter cloacae UCICRE12	5,210,535	55.59	1	180	PE	1.2	SRX342585
			2	5k	MP	1.5	SRX286723
Enterococcus faecium BM4538	3,133,897	38.07	1	180	PE	1.2	SRX341265
			2	5k	MP	1.8	SRX341264
Escherichia coli BIDMC 39	4,882,922	51.01	1	180	PE	0.33	SRX277757
			2	180	PE	0.57	SRX277758
			3	5k	MP	1.03	SRX277759
Klebsiella pneumoniae BIDMC41	5,702,446	26.95	1	180	PE	0.69	SRX277856
			2	180	PE	0.39	SRX277855
			3	5k	MP	1.18	SRX277857
Mucispirillum schaedleri ASF457	2,332,248	57.08	1	180	PE	1.2	SRX332194
			2	5k	MP	1.5	SRX332193
Pseudomonas aeruginosa CF614	6,797,445	31.01	1	180	PE	0.99	SRX366180
			2	5k	MP	0.81	SRX366181
			3	5k	MP	0.82	SRX366179
Streptococcus intermedius ATCC 27335	1,951,449	66.08	1	180	PE	1.1	SRX297066
			2	5k	MP	1.47	SRX297065

SOAPdenovo2 (Li et al. 2010; Luo et al. 2012), JR-Assembler 1.0 (Chu et al. 2013), and Velvet 1.0.19 (Zerbino and Birney 2008) were compared with median data sets.

The NGS datasets of *Streptomyces roseosporus*, *Neurospora crassa*, *Plasmodium falciparum*, and *Saprolegnia parasitica* genomes were downloaded from the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA) (www.ncbi.nlm.nih.gov/sra) under accession numbers: SRX016044, SRX026747, SRX030834, SRX022535, SRX016057, and SRX016059. Another dataset covering nine bacterial genomes (*Staphylococcus aureus* and *Rhodobacter sphaeroides*) was also downloaded from NCBI SRA; the accession numbers are listed in Table 4.1. The *Escherichia coli* reference genome was retrieved from GenBank under accession no. NC_000913.

4.4.2 Performance Comparison Using Medium-Sized Genomes

The performance comparison of these assemblers was evaluated using four medium-sized dataset, including a bacterial genome (*S. roseosporus*, genome size 7.7 Mb) and three fungal genomes (*N. crassa*, *P. falciparum*, and *S. parasitica*, genome sizes 37.1, 22.9, and 53.1 Mb, respectively). The rank method was used to evaluate the assembler (Chu et al. 2013). No assembler outperformed other assemblers in total contig

Table 4.2 Assembly statistics of four median genomes

Species	Assembler	No. of contigs[a]	Total size (Mb)[b]	Max (bp)[c]	Mean (bp)	N50 (bp)[d]
Streptomyces roseosporus	JR-Assembler	1189	7.68	40,501	6461	11,374
	ABySS	1127	7.73	55,078	**6859**	**12,499**
	Velvet	1192	7.49	**61,423**	6286	11,075
	SOAPdenovo	2453	7.65	24,303	3120	4691
Neurospora crassa	JR-Assembler	12,244	38.61	**58,672**	3153	6074
	ABySS	13,420	38.05	45,381	2835	6350
	Velvet	10,187	36.11	45,599	**3544**	**6781**
	SOAPdenovo	16,261	40.25	31,423	2475	5029
	Edena	17,083	39.95	42,952	2338	4534
Saprolegnia parasitica	JR-Assembler	40,587	46.09	**119,543**	**1135**	**1510**
	ABySS	52,087	38.26	94,931	734	740
	Velvet	53,736	47.38	91,073	881	1021
	SOAPdenovo	66,456	45.59	30,400	686	712
	Edena	62,357	44.13	41,473	707	746
Plasmodium falciparum	JR-Assembler	13,352	11.02	7939	**825**	**975**
	ABySS	16,658	11.80	7934	708	826
	Velvet	16,423	11.91	**7940**	725	848
	SOAPdenovo	17,424	11.93	7939	684	786
	Edena	16,531	11.76	7936	711	831

The top two best values of each assembly metrics are marked in bold
[a]Contigs of length <300 bp were not counted
[b]"Total" refers to the total number of bases in the contigs
[c]"Max" and "Mean" refer to the length of the longest contig and the mean length of contigs, respectively
[d]N50 is the size of the smallest contig such that 50 % of the assembled bases are in the contigs of size equal to or larger than the N50 value
A contig is misassembled if it cannot be aligned in full-length to the reference genome

number, total contig size, the maximal contig length, the mean contig length, or N50 length (Table 4.2). With *S. parasitica and P. falciparum*, the N50 lengths and mean contig length were longer than other assemblers. With *S. roseosporus* and *N. crassa,* ABySS and velvet exhibit a relatively good performance, respectively.

4.4.3 Performance Comparison Using Nine Bacterial Genomes

The performance of four commonly-used assemblers was evaluated using nine genome datasets with high coverage. The raw sequence data were derived from a strain for which the assembly level is either scaffolds or contigs. Every raw datum contains at least two libraries: one paired ends library and one mate-pair library (Table 4.1).

The N50 contig sizes are summarized for all nine of these datasets in Table 4.3. For all assemblers, good or nearly-good assembly can be obtained. All data sets except the *Mucispirillum schaedleri* dataset were able to produce a high-contig N50 from 200 to 500 kb. These results are better than those produced by datasets, which produced only one-paired ends library (Salzberg et al. 2012; Magoc et al. 2013). Because of the mate-pair library, better scaffold N50s were also produced by most datasets. The best scaffold N50s ranged from 1.4 to 5.7 Mb in size, which span more than the half of the

Table 4.3 Assembly statistics of nine bacterial genomes

Species	Assembler	Total length	No. scaffolds	Scaffold N50	No. contigs	Contig N50
Acinetobacter baumannii NIPH 24	allpaths-lg	3,899,709	18	2,378,052	35	343,910
	Soapdenovo2	3,881,660	20	2,379,771	30	**586,913**
	SPAdes	4,420,053	68	538,328	68	538,328
	MaSuRCA	4,051,564	46	**2,414,221**	60	438,417
Acinetobacter indicus CIP 110367	allpaths-lg	3,188,830	8	**2,659,306**	46	130,481
	Soapdenovo2	3,192,750	68	**1,741,014**	129	133,591
	SPAdes	3,178,658	38	266,989	38	**266,989**
	MaSuRCA	3,061,054	28	914,711	65	133,907
Enterobacter cloacae UCICRE 12	allpaths-lg	5,167,463	24	**2,910,535**	83	152,889
	Soapdenovo2	5,167,151	55	2,892,397	132	154,254
	SPAdes	5,663,090	70	247,654	70	**247,654**
	MaSuRCA	5,141,681	48	**4,489,688**	102	227,675
Enterococcus faecium BM4538	allpaths-lg	3,131,274	7	954,529	49	126,411
	Soapdenovo2	3,068,531	66	767,650	156	87,971
	SPAdes	3,431,198	58	266,407	58	**266,407**
	MaSuRCA	3,165,896	102	**2,119,662**	168	85,115
Escherichia coli BIDMC 39	allpaths-lg	4,905,456	25	2,678,791	110	121,904
	Soapdenovo2	4,903,160	94	2,497,784	203	216,794
	SPAdes	4,902,401	65	284,858	65	**284,858**
	MaSuRCA	4,842,876	36	**3,718,131**	78	262,190
Klebsiella pneumoniae BIDMC 41	allpaths-lg	5,661,146	9	4,151,878	59	187,007
	Soapdenovo2	5,702,239	46	1,971,292	105	240,104
	SPAdes	5,751,074	34	813,379	35	**813,379**
	MaSuRCA	5,714,443	37	**4,562,366**	84	299,706
Mucispirillum schaedleri ASF457	allpaths-lg	2,311,286	20	741,189	82	60,693
	Soapdenovo2	2,337,314	62	594,782	136	68,178
	SPAdes	2,348,799	71	149,716	72	**149,716**
	MaSuRCA	2,335,222	62	**1,891,207**	106	84,260
Pseudomonas aeruginosa CF614	allpaths-lg	6,807,352	9	1,431,211	25	429,413
	Soapdenovo2	6,818,856	51	**5,774,020**	112	378,571
	SPAdes	6,751,614	31	965,679	31	**965,679**
	MaSuRCA	6,797,296	21	1,933,268	39	**689,346**
Streptococcus intermedius ATCC 27335	allpaths-lg	1,892,452	10	634,497	16	**284,930**
	Soapdenovo2	1,929,122	12	**910,762**	19	260,557
	SPAdes	1,918,222	11	277,339	11	**277,339**
	MaSuRCA	2,017,461	60	548,293	65	239,440

*The top two best values of each assembly metrics are marked in bold

Table 4.4 Assembly accuracy of assemblers evaluated by using REAPR

Species	Rank (FCD gap)		
	Allpaths-lg	MaSuRCA	SOAPdenovo2
Acinetobacter baumannii NIPH 24	3(4)	1(12)	3(4)
Acinetobacter indicus CIP 110367	1(9)	3(7)	2(8)
Enterobacter cloacae UCICRE 12	3(4)	1(27)	2(5)
Enterococcus faecium BM4538	3(4)	1(22)	2(20)
Escherichia coli BIDMC 39	2(16)	3(3)	1(25)
Klebsiella pneumoniae BIDMC 41	3(4)	1(17)	2(16)
Mucispirillum schaedleri ASF457	3(1)	1(22)	2(16)
Pseudomonas aeruginosa CF614	3(0)	2(1)	1(7)
Streptococcus intermedius ATCC 27335	3(1)	3(1)	1(2)
	24	16	16

genome. Results produced by all of the assemblers for *M. schaedleri* are far more fragmented than those of other datasets, with contig N50 sizes ranging from 60 to 149 kb. For this genome, the choice of assembler seems to have a large impact on the quality of the resulting assembly.

No assembler ranked highest among all metrics (Table 4.3). For this reason, a ranking approach was used to evaluate the overall performance of each assembler. For each assembly metric and dataset, the top two values are marked in bold (Table 4.3). The number of marked values was determined and used as the voting score for each assembler. For N50 length of scaffold, the scores for allpaths-lg (Gnerre et al. 2011), SOAPdenovo2 (Luo et al. 2012), SPAdes (Bankevich et al. 2012), and MaSuRCA (Zimin et al. 2013) were 26, 24, 9, and 31, respectively. In this way, MaSuRCA were found to have the best overall performance. For N50 length of contig, the scores were 15, 19, 34, and 22, respectively. In this way, SPAdes were found to have the best overall performance.

The accuracy of assembly was evaluated by REAPR (Table 4.4). REAPR uses the per-base error of the fragment coverage distribution (FCD) to detect assembly errors without the need for a reference sequence and provides corrected assembly statistics allowing the quantitative comparison of multiple assemblies (Hunt et al. 2013). For each data set, no assembler produced the lowest number of FCD gaps, and MaSuRCA was also found to produce highly accurate results in some datasets (*Acinetobacter indicus CIP 110367*, *E. coli BIDMC 39*, and *Streptococcus intermedius ATCC 27335*). Hence, a ranking approach was used to evaluate the overall performance in assembly accuracy: the higher the score, the more accurate the assembly (Chu et al. 2013). Because SPAdes only produced the contig assembly, it does not participate in the comparative analysis. Allpaths-lg produced the highest ranking score (24), which is much higher than the scores that MaSuRCA (16) and SOAPdenovo2 (16) produced.

4.5 Discussion

In this section, the main approaches for genome assembly are presented. For each approach, several algorithms are explained. There are three main categories for assembly algorithms: extension-based algorithms, overlap-layout consensus algorithms and de Bruijn graph algorithms. Overlap-layout consensus algorithms are based on overlap graphs and Hamiltonian path-finding. de Bruijn graph algorithms are based on de Bruijn graphs and Eulerian path-finding in assembly graphs. de Bruijn graph

methods show more strength for short-reads and in resolving repeats. Overlap graph methods are more suitable for Sanger shotgun data. While extension-based methods seemed applicable just on long sequences, some tricks used in new algorithms, which use paired-end reads, such as the PE-Assembler, could apply the greedy idea efficiently for short reads. The evaluation of performances of an assembly algorithm is bases on both the accuracy and contiguity of assemblies, and there is always a trade-off between different measures of assembly performances. It is not a trivial task to compare assembly algorithms.

In fact, the assembly results also depend on the dataset used, besides assembly algorithm. An algorithm may do well with a dataset but not with other datasets. For a new dataset, we cannot exactly predict which algorithm would produce a better assembly just based on previous assembly results, due to the difference in dataset used for assembly. In addition, some algorithms, such as parameter k in de Bruijn graph-based methods, require users to predefine assembly parameter. The use of parameters makes it more difficult to compare assembly algorithms, for the final assembly result is definitely dependent on the parameter chosen for the assembly task. There are some metrics available for comparing assembly algorithms, but the availability of a good metric that is not dependent on the reference genome is still missing from the literature.

References

Ariyaratne PN, Sung WK (2011) PE-assembler: de novo assembler using short paired-end reads. Bioinformatics 27:167–174

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M et al (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK et al (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18:810–820

Chaisson MJ, Brinza D, Pevzner PA (2009) De novo fragment assembly with short mate-paired reads: does the read length matter? Genome Res 19:336–346

Chu TC, Lu CH, Liu T, Lee GC, Li WH, Shih AC (2013) Assembler for de novo assembly of large genomes. Proc Natl Acad Sci USA 110:E3417–E3424

Compeau PE, Pevzner PA, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. Nat Biotechnol 29:987–991

Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30:2478–2483

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res 17:1697–1706

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108:1513–1518

Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res 18:802–809

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD (2013) REAPR: a universal tool for genome assembly evaluation. Genome Biol 14:R47

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12:656–664

Li R, Zhu H, Ruan J, Qian W, Fang X et al (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272

Li Z, Chen Y, Mu D, Yuan J, Shi Y et al (2012) Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief Funct Genomics 11:25–37

Luo R, Liu B, Xie Y, Li Z, Huang W et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18

Maccallum I, Przybylski D, Gnerre S, Burton J, Shlyakhter I et al (2009) ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. Genome Biol 10:R103

Magoc T, Pabinger S, Canzar S, Liu X, Su Q et al (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics 29:1718–1725

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46

Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315–327

Peng Y, Leung HC, Yiu SM, Chin FY (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–1428

Pevzner PA, Tang H (2001) Fragment assembly with double-barreled data. Bioinformatics 17:S225–S233

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T et al (2012) GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Res 22:557–567

Simpson JT, Durbin R (2012) Efficient de novo assembly of large genomes using compressed data structures. Genome Res 22:549–556

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123

Wang X, Wang H, Wang J, Sun R, Wu J et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Warren RL, Sutton GG, Jones SJ, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23:500–501

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

Zhang T, Luo Y, Chen Y, Li X, Yu J (2012) BIGrat: a repeat resolver for pyrosequencing-based re-sequencing with Newbler. BMC Res Notes 5:567

Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. Bioinformatics 29:2669–2677

Crop Genome Annotation: A Case Study for the *Brassica rapa* Genome

5

Erli Pang, Huifeng Cao, Bowen Zhang and Kui Lin

Abstract

Genome annotation is crucial for the bridging the gap between sequence and biology. Nonetheless, it is also a dynamic and continuous improvement process for better understanding of the molecular biology of the genome. With the deep RNA-sequencing of eight *Brassica rapa* tissues, it should be able to predict protein-coding genes with more accuracy when incorporating this type of RNA information into analysis. In doing so, we used our built annotation pipeline to re-annotate the *B. rapa* genome on the levels of repetitive elements, protein-coding genes and non-coding RNA genes, respectively. In total, we identified 139.9 MB repetitive elements, 6,088 non-coding RNA genes and 45,149 protein-coding genes, respectively. These results, together with those published previously, would provide a valuable resource for further understanding of *B. rapa*.

5.1 Introduction

The first genome (*Haemophilus influenzae*) was sequenced at 1995 (Fleischmann et al. 1995). Since then, the genome era was started. As the development of new sequencing technologies, thousands of genomes have been sequenced. Nevertheless the genome sequence, as a string of nucleotides, has no meaning for biologists before it is transformed into biological features. Genome annotation, a process of identifying potential functional features from the genomic sequences, bridges the gap from the sequence to the biology of the organism (Stein 2001). Therefore, the

E. Pang · H. Cao · B. Zhang · K. Lin (✉)
College of Life Sciences, Beijing Normal University, Beijing 100875, China
e-mail: linkui@bnu.edu.cn

E. Pang
e-mail: pangerli@bnu.edu.cn

H. Cao
e-mail: hfcao@mail.bnu.edu.cn

B. Zhang
e-mail: billz3187@gmail.com

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_5

value of the genome depends largely on the quality of genome annotations.

There are generally four steps in the process of annotating a genome, namely repeat identification, protein-coding genes and non-coding RNAs prediction, functional annotation and visualization of annotation results.

Repeats, here, are used to describe two different types of DNA sequences: tandem repeats and transposable elements. They usually complicate genome annotation and need to be identified and masked before annotating genes. RepeatMasker (http://www.repeatmasker.org) is a tool to mask repeats. Repeat identification is considered as a complicated process due to poor conservation of repeats sequences. Thus, in order to identify and mask repeats accurately, one or more repeat libraries specific to the genome being annotated are required. In general, two types of repeat libraries can be used: existing libraries, such as Repbase (Jurka et al. 2005) and TIGR (Ouyang and Buell 2004), and libraries constructed by de novo methods and/or tools. There are many de novo tools developed to create repeat libraries, such as LTR_FINDER (Xu and Wang 2007), PILER (Edgar and Myers 2005), RepeatScout (Price et al. 2005), RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) and so on.

After having cleaned the annoying repeats in the genome sequence, protein-coding genes are discovered using different methodologies. The prediction of protein-coding genes is the core of genome annotation (Liang et al. 2009). At present, there are various programs developed to predict protein-coding genes. Generally, they can be roughly divided into three groups of prediction approaches.

The approaches from the first group are based on evidence alignment that uses external evidence to identify genes and their intron-exon structures. The basic principle of these approaches is to align the protein sequences, expressed sequence tags (ESTs) or/and RNA-seqs against the genome sequence (Wang et al. 2009). These proteins, ESTs and RNA-seqs are usually from the organism whose genome is being annotated, otherwise from other closely related organisms (Curwen et al. 2004; Haas et al. 2008). Obviously, considering that there is more conserved for protein sequences than for their coding DNAs even over longer evolutionary time, the protein sequences from other organisms sometimes are also used. Tools used to align the evidence at the DNA level to the genomic sequences including BLAST (Altschul et al. 1990), BLAT (Kent 2002), GMAP (Wu and Watanabe 2005), EST_GENOME (Mott 1997), AAT (Huang et al. 1997), BWA (Li and Durbin 2009, 2010), Bowtie (Langmead et al. 2009) and so on. On the other hand, for the evidence at the protein level, there are also many tools can be used to find protein-coding genes in the genome sequence, such as GeneWise (Birney et al. 2004), Scipio (Keller et al. 2008), TWINSCAN (Korf et al. 2001). To our best knowledge and practical experience, we believe that GeneWise is the most accurate and important protein-to-genome alignment program, as it is the central part of Ensembl (Flicek et al. 2014) gene annotation pipeline.

The second group consists of de novo protein-coding gene prediction methods, which are based on various mathematical models to identify putative protein-coding genes and their intron-exon structures. The Hidden Markov models (HMMs) are the most frequently and importantly used models in bioinformatics. There are numerous de novo protein-coding gene predictors based on HMMs including GENSCAN (Burge and Karlin 1997), GENE-ID (Parra et al. 2000), GlimmerHMM (Majoros et al. 2004), GeneMark.hmm (Lukashin and Borodovsky 1998), FGENESH (Salamov and Solovyev 2000), Augustus (Stanke and Waack 2003) and so forth. In 2007, another model was also introduced into genome annotation called conditional random fields (CRFs), and then two gene predictors were developed to discover putative protein-coding genes (DeCaprio et al. 2007; Gross et al. 2007). Usually, de novo gene predictors need first to be trained on the genome to capture respective genomic traits. Therefore, the quality of the training data sets is important for any of supervised de novo gene predictors. In addition, it should be noted that most, if not all, of the de novo gene predictors can't be reported

alternatively spliced transcripts and 5′- and 3′-untranslated regions.

For protein-coding gene prediction, in addition to the two groups of methods abovementioned, combined or integrated methodology is preferable in genome annotation. This group of methods integrates all evidences from various predictors either from alignment-based or de novo approaches and reports a most probably correct gene structure. Due to their integration of many different types of evidences, they output the results more reasonably (Brent 2008). Currently, there are many combined predictors being developed. Among them, the more widely used are JIGSAW (Allen and Salzberg 2005), GLEAN (Elsik et al. 2007), Ensembl (Curwen et al. 2004), MAKER (Campbell et al. 2014; Cantarel et al. 2008) and EvidenceModeler (EVM) (Haas et al. 2008). The Ensembl is an automatic annotation system, which integrated evidences derived from known proteins, full-length cDNA (FLcDNA) and EST sequences (Curwen et al. 2004). The accuracy of the annotations depends largely on the number and quality of available FLcDNAs. Unfortunately, they are expensive for us to obtain enough FLcDNAs for annotation. MAKER was developed to allow scientists to annotate eukaryotic genomes, in particular for plant genomes. Nonetheless, it only integrates SNAP predictor and alignments of Exonerate (Slater and Birney 2005) and BLAST. Recently, in order to better support plant genome annotation efforts, a more efficient version for parallel computing called MAKER-P was developed and published (Campbell et al. 2014; Cantarel et al. 2008). It can deal with large repeat-rich plant genomes, which is one of most challenges in sequencing many non-model plants. The EVM is a genome annotation system integrating evidence from both de novo gene predictors and protein and transcript alignments into consensus gene structures by using a non-stochastic weighted evidence combining technique (Haas et al. 2008). The protein-coding genes annotation of our local genome annotation pipeline was built with the EVM framework.

Although non-coding RNA annotation is developing rapidly, it is still incomplete compared with protein-coding gene annotation. Non-coding RNAs are different in sequence conservation. For rRNAs, nucleotide homologies are used to detect them for sequence conservation such as BLAST. For tRNAs, poor conserved at the primary sequence level, conserved secondary structures are commonly used such as tRNAscan-SE (Lowe and Eddy 1997). Infernal (Nawrocki et al. 2009) is another approach to annotate possible non-coding RNAs accompanying with the Rfam (Gardner et al. 2011) database. The current version of MAKER, MAKER-P, has integrated ncRNA tools for identifying non-coding RNAs (Campbell et al. 2014). Finally, all of the outputs of a genome annotation, including repeats, noRNAs and protein-coding genes, are commonly organized and formatted into the GenBank, GFF3, or EMBL formats.

When we have identified the putative genes for the genome sequence, the next step is to annotate them or their products (for protein-coding genes) at the level of molecular function. This also is a challenging task contemporarily. For protein-coding genes, their transcripts could be easily translated into the respective protein sequences. On one hand, these protein sequences can be used to assess the quality of our annotation of protein-coding genes by search against well-annotated protein databases, for example UniProt (Apweiler et al. 2013). On the other hand, we need annotate their function(s) with as many as possible. Currently, functional annotation is usually at two different levels: whole proteins or parts of each protein, or both. At the whole protein level, besides to search against to the most well-annotated Swiss-Prot and then transform the corresponding function or knowledge of the best subjects hit to the query proteins, we may also consider to annotate their function(s) based on the Gene Ontology (GO) (Ashburner et al. 2000) using the Blast2Go tool (Conesa et al. 2005) which is based on similarity searches with statistical analysis. If there is no any homolog matched at the whole protein level, we have to consider what parts/domains may be embedded in the protein sequences being annotated. To the best of our

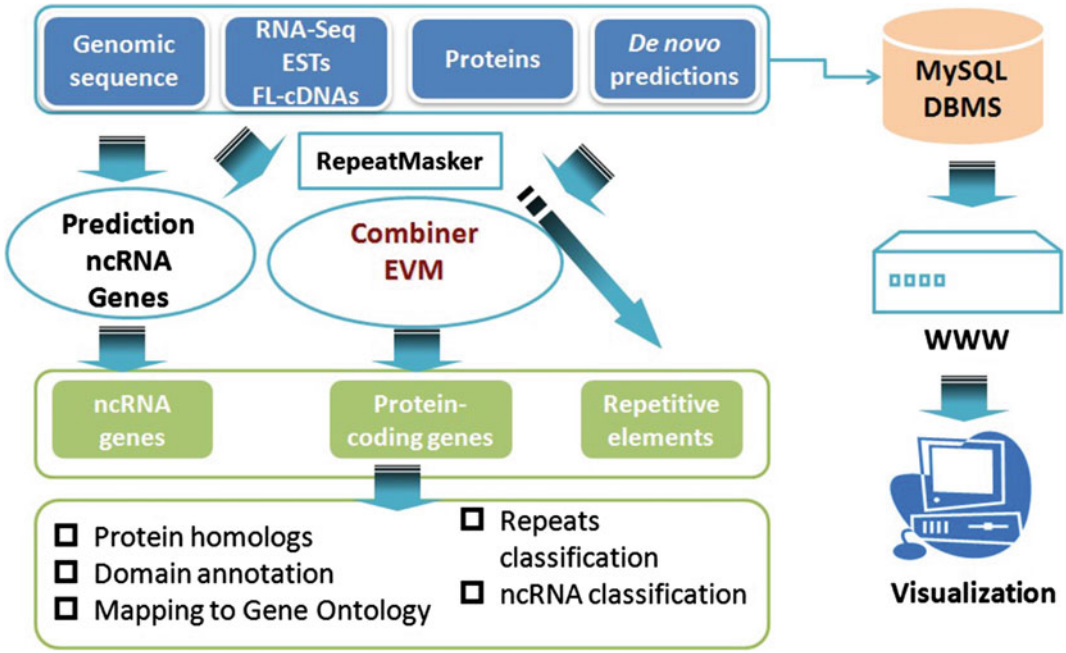

Fig. 5.1 Schematic diagram of our genome annotation pipeline

knowledge, commonly used known domains and/or other signatures are identified by profiling the sequences with the InterProScan tool (Jones et al. 2014), or simply by searching the most current Pfam database (Punta et al. 2012). Finally, the outputs of genome annotation are usually visualized using GBrowse (Stein et al. 2002) and JBROWSE (Skinner et al. 2009) packages.

In this study, we demonstrated to annotate the *B. rapa* genome using our own annotation pipeline that was based mainly on EVM strategy. Figure 5.1 shows the overall framework of our pipeline, and the detailed annotation process.

5.2 The Processes of a Genome Annotation

5.2.1 Filtering the Sequences and Masking Putative Repeats

To make sense of biological property, for each scaffolds being annotated, we only considered those scaffolds with length more than 2 kb and its percentage of gaps less than 50 %. Then, these genomic scaffolds were masked using RepeatMasker (http://www.repeatmasker.org, version 3.2.6). The underlying library was combined from Repbase TE library (release September 2011), TIGR (release November 2008) and de novo libraries. In addition, we also used some tools to create de novo repeat libraries, such as LTR_FINDER, PILER and RepeatScout. The numbers of the sample sequences within Repbase, TIGR and de novo library are 10,758, 214,020 and 7,207, respectively. The process of masking repeats is shown in Fig. 5.2.

5.2.2 Prediction of Protein-Coding Genes

EVidenceModeler (EVM, versionr03062010), which is a nonstochastic weighted evidence evaluation system to produce consensus gene structure, was used to combine the alignments of proteins and transcripts to the genomic sequences, and various de novo predictions into a predicted gene set.

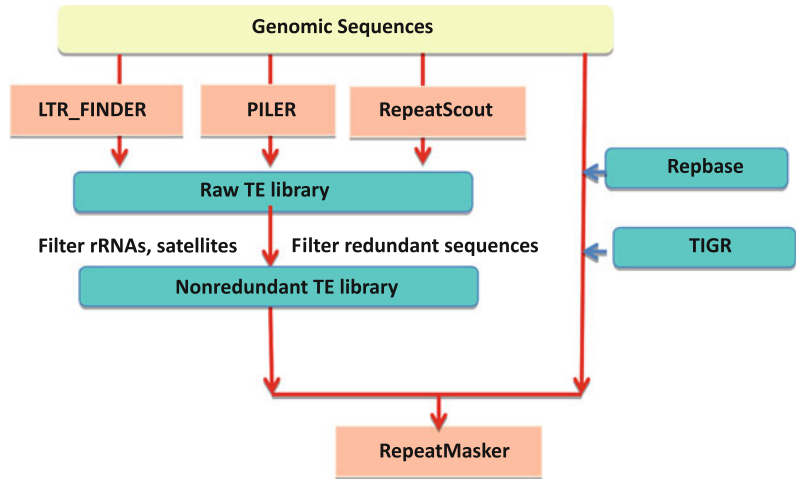

Fig. 5.2 Process of masking genomic repeats

A more detailed explanation is as follow. Firstly, we processed evidence at the transcript level. Spaln (version 1.4.3) (Gotoh 2008) was used to mapp the plant ESTs downloaded from TIGR plant transcript (Childs et al. 2007) onto our assembled genome. And PASA (version rJAN_09_2011) (Haas et al. 2003) was used to map the *B. rapa* ESTs, FLcDNAs and transcripts assembled by RNA-Seq reads of *B. rapa* using inchworm (Grabherr et al. 2011). Furthermore, Cufflinks (Trapnell et al. 2010) was used to assemble the RNA-Seq reads alignments obtained by TopHat (Trapnell et al. 2009) into transcripts. This process by the three programs produced the dataset of putative intron-exon boundaries. Meanwhile, the protein-coding gene models were produced by PASA aligning the ESTs, FLcDNAs and assembled transcripts using inchworm to the genome. Basing on the set of gene models, we constructed a training set, which will be used by de novo predictors, by selecting the genes with complete structures and at least 100 % mapping rate for UniProt plant (Bairoch et al. 2005) proteins, and filtering out the redundant genes with more than 70 % sequence identity by CD-HIT (version 4.1.1) (Li and Godzik 2006).

Secondly, we focused on the evidence at protein sequence level. The protein sequences from UniProt-SwissProt plant proteins (release2010_07), *Arabidopsis thaliana* proteins (TAIR9, Augustus 2009 release) and *Oryza sativa* proteins (TIGR Release 5.0, January 2007) were mapped onto the genomic sequence using Spaln (version 1.4.4) (Gotoh 2008), Scipio (Keller et al. 2008) and TBLASTN (Altschul et al. 1990). The putative intron-exon boundaries were generated by Spaln and Scipio. For TBLASTN mapping, we performed the four procedures: (1) for each protein, joining all of the HSPs (1e-5) with the gap of 5000 bp into a consecutive region; (2) selecting the region when the overlapping coverage of its HSPs with the protein is greater or equal to 80 %; (3) extending 2000 bp at both ends of the region; (4) applying GeneWise onto the region to identify the putative intron-exon boundaries of the predicted gene.

Thirdly, we collected protein-coding evidence by de novo predictors. At this process, Augustus (version 2.4) (Stanke et al. 2008), Geneid (version 1.4.4) (Parra et al. 2000), GeneMark-ES (version 2.3) (Ter-Hovhannisyan et al. 2008), GlimmerHMM (version 3.0.1) (Majoros et al. 2004) and SNAP (2006-07-28) (Korf 2004) were used. Besides GeneMark-ES, the others used the masked genomic sequences. August, GlimmerHMM and Snap are supervised predictors with the training set generated by PASA abovementioned, while Geneid utilized the parameters of *A. thaliana*.

Finally, all evidences for protein-coding genes collected by the methods abovementioned was

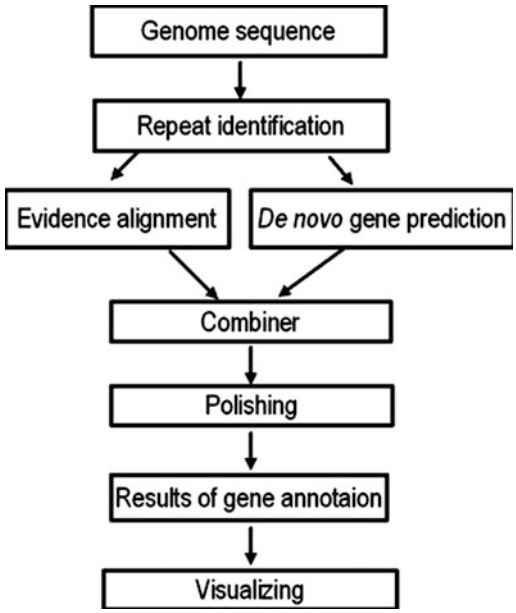

Fig. 5.3 Flow chart for protein-coding gene annotation

combined into a consensus protein-coding gene models by EVM. In addition, based this set of gene models and EST dataset, we also used PASA to polish the gene models by adding untranslated regions (UTRs), correcting gene models, and generating all possible alternatively spliced isoforms at the mRNA level. The flowchart of predicting protein-coding is described in Fig. 5.3.

5.2.3 Identification of Non-coding RNAs

The ncRNAs were identified by our non-coding RNAs analysis pipeline, which integrates four programs. The six types of non-coding RNAs detected were: rRNAs identified by BLAST (version 2.2.26); tRNAs predicted by tRNAscan-SE (version 1.3.1); Box C/D snoRNAs identified by snoscan (version 0.9b) (Lowe and Eddy 1999); antisense RNAs, snRNAs, miRNAs, rRNAs, and cis-regulatory Riboswitch RNAs detected by INFERNAL (version 1.1rc1) which were based on the Rfam database (release 10.1). We only used 333 plant covariance models curated in the Rfam database. The process of identifying non-coding RNAs is show in Fig. 5.4.

5.2.4 Functional Annotation of Proteins and Genome Visualization

The putative functions of proteins predicted are annotated based on two levels: protein domains and gene ontology (GO) terms. The protein domains were assigned by Pfam27.0 (Punta et al. 2012). The GO annotation was implemented by Blast2GO (version 2.5) (Conesa et al. 2005) based on the GO version of 2012-12-09. The visualization of genome and its annotation results were based on the GBrowse system (version 2.39) (Stein et al. 2002), which could be downloaded (http://sourceforge.net/projects/gmod/files/GenericGenomeBrowser/) and installed locally.

5.3 The Annotation Results of the Reference Sequence

5.3.1 Repetitive Elements in the Reference Genome

We only annotated those scaffolds, whose lengths were more than 2 KB and the proportions of gaps were less than 50 %. De novo repeat libraries were constructed by multiple de novo methods. Then a repeat library was obtained combing Repbase, TIGR and de novo libraries. We identified 139.9 MB repetitive elements using the combing repeat library, which occupy ~43 % of genome. Among them 85.7 % could be classified based on known repeats. The long terminal repeat (LTR) retrotransposons were the majority of the transposable elements and occupied 28.7 % of the genome. Other retrotransposons such LINE and SINE comprised 5.95 and 0.14 % of the genome, respectively. DNA transposable elements comprised 4.38 % of the genome (Table 5.1).

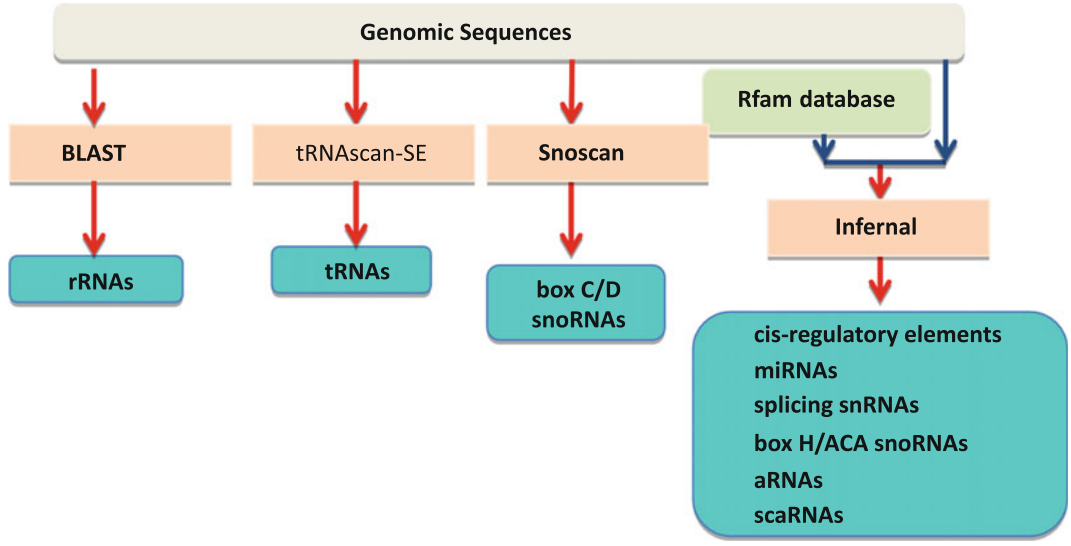

Fig. 5.4 Process of identifying non-coding RNAs

Table 5.1 Repetitive elements in the genome

Type	#Copies	Length (bp)	Percent of the genome (%)
DNA transposon	45,628	14,233,624	4.38
Retrotransposon			
LTR	202,092	93,197,939	28.66
LINE	53,399	19,343,903	5.95
SINE	2,838	443,936	0.14
Unclassified	83,463	20,045,924	6.16
Total	387,420	139,859,118	43.00

5.3.2 Protein-Coding Genes and NcRNAs

Based on these 4,283 scaffolds, we used evidence alignments and de novo gene predictors to identify protein-coding genes and then built a consensus gene structure by combing all of the results. We predicted 45,149 genes with a mean sequences size of 2,039 bp. These genes product 55,290 transcripts with a mean coding sequences size of 1,433 bp and an average 5.22 exons per gene (Table 5.2). About 84.7 % (46,832/55,290) the proteins have homologs (e-value 10^{-5}) in the plant UniProt database (released April, 2013), and 76.7 % (42,407/55,290) the proteins contain PfamA domains (Pfam27.0 version) including 3,897 domains. Using Blast2Go, 29,862 proteins were annotated by 8,883 GO terms (Table 5.3). Addition to protein-coding genes, we identified 1,617 rRNA fragments and 1,254 tRNA, 230 splicing snRNA, 1,216 small nucleolar RNA (snoRNA), two scaRNA and 1,714 miRNA genes in the genome (Table 5.4). Figure 5.5 shows an example of the results of annotation viewed by GBrowse.

5.3.3 Summary of the Previous Annotation of *Brassica Rapa*

In the first reference genome of *B. rapa* published in 2011(Wang et al. 2011), they annotated the genome at the level of repetitive elements and protein-coding genes. To annotate repetitive elements, they constructed de novo libraries using three different software packages: PILER,

Table 5.2 Statistics of predicted protein-coding genes

Feature	Value
Gene number	45,149
Gene on Watson	22,949
Gene on Crick	22,200
Gene density (/Mb)	139
Average gene length (bp)	2,039
Transcripts number	55,290
Average transcript length (bp)	1,433
Multiple exon	44,616
Single exon	10,674
Average exon number	5.22
Transcripts having UTRs	41,372
Average UTR length	224
5′UTR number	38,703
Average 5′UTR length	171
3′UTR number	38,492
Average 3′UTR length	276

Table 5.3 The numbers of proteins with functional annotation

Total proteins predicted	Database	Annotated number	Percent (%)
55,290	Plant UniProt	46,832	84.7
	Pfam	42,407	76.7
	Gene ontology	29,862	54.0

Table 5.4 The non-coding RNA genes (ncRNAs) in the genome

Type	Number of copies	Length (bp)	Percent of the genome (%)
Antisense RNA	41	2,725	0.00084
Cis-regulatory RNA element			
IRES	1	172	0.00005
Frameshift element	2	111	0.00003
Riboswitch	13	1,039	0.00032
snRNA			
Splicing snRNA	230	27,767	0.00854
Box C/D snoRNA	983	122,223	0.03685
Box H/ACA snoRNA	231	23,458	0.00715
scaRNA	2	258	0.00008
tRNA	1,254	92,824	0.02854
miRNA	1,714	162,427	0.04557
rRNA	1,617	1,043,890	0.32094
Total	6,088	1,476,890	0.42179

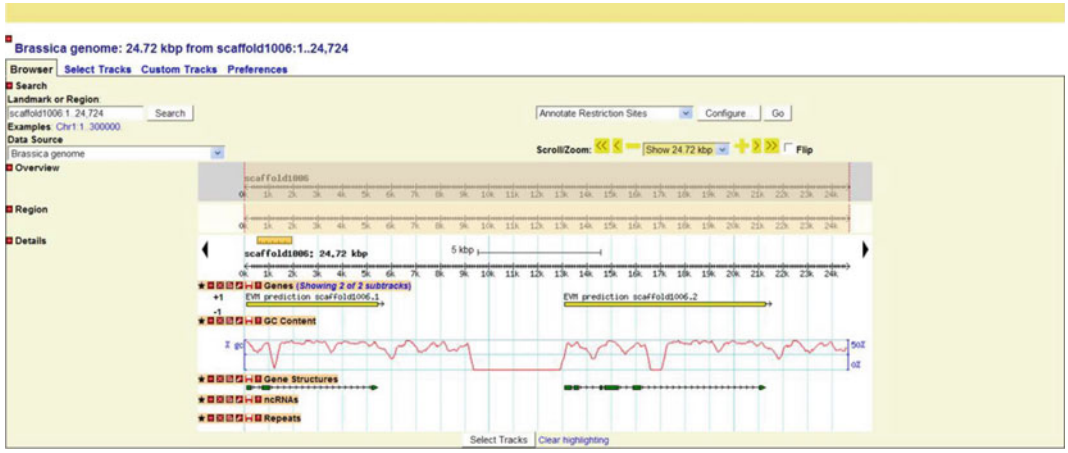

Fig. 5.5 Visualization the genome annotation results using GBROWSE

RepeatScout, and LTR_FINDER. Then repeat library was obtained combing Repbase, TIGR and de novo libraries. Finally, they identified transposon-related sequences occupying 39.5 % of the genome, and the proportions of retrotransposons, DNA transposons and long interspersed elements were 27.1, 3.2 and 2.8 %, respectively. For protein-coding genes, they used GLEAN to integrated de novo gene sets and other homology-based gene sets and transcript evidence. The de novo predictors were GENSCAN and Augustus. Their gene model parameters trained from *A. thaliana*. Applying the pipeline, they predicted 41,174 putative protein-coding genes with an average transcript length of 2,015 bp and coding length of 1,172 bp.

5.3.4 Comparison of the Results of This Study with Those of the Previous Annotation

We here compared the annotation results of our genome annotation pipeline with the results of the work when the first reference genome of *B. rapa* published in 2011 (Wang et al. 2011). The comparison of annotation results simply focused on three aspects, including repeats, protein-coding genes, and non-coding RNA genes.

At the repeats level, the repetitive element sequences masked by our pipeline occupied 43.00 % of the genome, with the proportions of retrotransposons (with long terminal repeats), DNA transposons and long interspersed elements being 28.66, 4.38 and 5.95 %, respectively. This shows that there are more repetitive elements identified by our analysis compared to that of the results published in 2011, in which the repetitive elements only occupied 39.5 % of the whole genome with the proportions of retrotransposons, DNA transposons and long interspersed elements being 27.1, 3.2 and 2.8 %, respectively.

From the viewpoint of the protein-coding genes identified, we predicted 45,149 putative protein-coding genes, which may produce 55,290 alternative spliced transcripts. The number of protein-coding genes we predicted is somewhat more than that of protein-coding genes, 41,174, predicted when the genome published (Wang et al. 2011). The average length of the transcripts we identified is about 1,433 bp, much longer than that (1,172 bp) of the prediction in the published results in 2011. However, the average exons of per protein-coding gene are similar, 5.22 versus 5.03. More protein-coding genes were predicted by our pipeline might have benefited from the new type of transcriptomic data we used. Eight RNA-Seq datasets from different tissues were used and integrated when we predicted the putative protein-coding genes. As we know, this type of transcriptomes produced by the next-generation

RNA-Seq technology may greatly facilitate prediction and annotation of protein-coding genes within a genome (Denoeud et al. 2008; Li et al. 2011). On the other hand, the full-length cDNAs (9,364), ESTs (902,518), and other transcript sequences (3,895,049) downloaded from the TIGR plant transcriptdatabase, and all of these data might also have some contribution to the difference of protein-coding genes. Additionally, our pipeline also predicted 5′/3′-UTR regions and putative alternative transcripts, while they were not included in the results published in 2011. On the other hand, the published work in 2011 did not predict putative non-coding RNA genes; nonetheless, our pipeline identified totally 6,088 ncRNA genes in this demonstration, which were classified into six categories (Table 5.4). In summary, by integrating identification of protein-coding genes, ncRNAs, and various types of repeats, our genome annotation pipeline, and others like MAKER-P (Campbell et al. 2014; Cantarel et al. 2008), will help us better characterize the architectures of sequencing genomes, this should facilitate corresponding downstream researches, such as genome evolution, crop breeding, and biological adaptation.

5.4 Conclusion

In summary, by applying our genome annotation pipeline, which is built mainly upon the EVM framework, to the *B. rapa* reference genome sequence, we have demonstrated the basic processes about how to annotate a given genome sequence. As we know, the results from an annotation task, either for protein-coding genes, or non-coding RNA genes, or repeats, only provide the first step and a basis for further research on the genome biology of the organism/species of interest. Meanwhile, we also need to point out that any genome annotation is a continually improvement process with new evidence coming from various aspects, including new DNA sequencing and/or resequencing data, RNA-Seq data, and even the new or updated information from any of closely related species.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant: 31171235). We thank to the people who have contributed to the building and maintaining of the genome annotation pipeline in the Laboratory of Computational Molecular Biology of the Beijing Normal University.

References

Allen JE, Salzberg SL (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics 21:3596–3603

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y et al (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res 41:D43–D47

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B et al (2005) The universal protein resource (UniProt). Nucleic Acids Res 33:D154–D159

Birney E, Clamp M, Durbin R (2004) GeneWise and genomewise. Genome Res 14:988–995

Brent MR (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet 9:62–73

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Campbell MS, Law M, Holt C, Stein JC, Moghe GD et al (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol 164:513–524

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E et al (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18:188–196

Childs KL, Hamilton JP, Zhu W, Ly E, Cheung F et al (2007) The TIGR plant transcript assemblies database. Nucleic Acids Res 35:D846–D851

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M et al (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E et al (2004) The Ensembl automatic gene annotation system. Genome Res 14:942–950

DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M et al (2007) Conrad: gene prediction using conditional random fields. Genome Res 17:1389–1398

Denoeud F, Aury J-M, Da Silva C, Noel B, Rogier O et al (2008) Annotating genomes with massive-scale RNA sequencing. Genome Biol 9:R175

Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21: I152–I158

Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS et al (2007) Creating a honey bee consensus gene set. Genome Biol 8:R13

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Flicek P, Amode MR, Barrell D, Beal K, Billis K et al (2014) Ensembl 2014. Nucleic Acids Res 42:D749–D755

Gardner PP, Daub J, Tate J, Moore BL, Osuch IH et al (2011) Rfam: wikipedia, clans and the "decimal" release. Nucleic Acids Res 39:D141–D145

Gotoh O (2008) Direct mapping and alignment of protein sequences onto genomic sequence. Bioinformatics 24:2438–2444

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

Gross SS, Do CB, Sirota M, Batzoglou S (2007) CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. Genome Biol 8:R269

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK et al (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31:5654–5666

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE et al (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol 9:R7

Huang X, Adams MD, Zhou H, Kerlavage AR (1997) A tool for analyzing and annotating genomic sequences. Genomics 46:37–45

Jones P, Binns D, Chang H-Y, Fraser M, Li W, et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O et al (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Keller O, Odronitz F, Stanke M, Kollmar M, Waack S (2008) Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinformatics 9:278

Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12:656–664

Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5:59

Korf I, Flicek P, Duan D, Brent MR (2001) Integrating genomic homology into gene structure prediction. Bioinformatics 17:S140–S148

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659

Li Z, Zhang Z, Yan P, Huang S, Fei Z et al (2011) RNA-Seq improves annotation of protein-coding genes in the cucumber genome. BMC Genom 12:540

Liang CZ, Mao L, Ware D, Stein L (2009) Evidence-based gene predictions in plant genomes. Genome Res 19:1912–1923

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:0955–0964

Lowe TM, Eddy SR (1999) A computational screen for methylation guide snoRNAs in yeast. Science 283:1168–1171

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26:1107–1115

Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20:2878–2879

Mott R (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 13:477–478

Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25:1335–1337

Ouyang S, Buell CR (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res 32:D360–D363

Parra G, Blanco E, Guigo R (2000) GeneID in Drosophila. Genome Res 10:511–515

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21:I351–I358

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J et al (2012) The Pfam protein families database. Nucleic Acids Res 40:D290–D301

Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res 10:516–522

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH (2009) JBrowse: a next-generation genome browser. Genome Res 19:1630–1638

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31

Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19:II215–II225

Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24:637–644

Stein L (2001) Genome annotation: from sequence to biology. Nat Rev Genet 2:493–503

Stein LD, Mungall C, Shu S, Caudy M, Mangone M et al (2002) The generic genome browser: a building block for a model organism system database. Genome Res 12:1599–1610

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res 18:1979–1990

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wang X, Wang H, Wang J, Sun R, Wu J et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265–W268

Miniature Transposable Elements (mTEs): Impacts and Uses in the *Brassica* Genome

6

Perumal Sampath, Jonghoon Lee, Feng Cheng, Xiaowu Wang and Tae-Jin Yang

Abstract

Transposable elements occupy large portions of eukaryotic genomes and play an important role in genome evolution. Terminal repeat retrotransposons in miniature (TRIMs), short interspersed elements (SINEs) and miniature inverted-repeat transposable elements (MITEs) are representative forms of so-called miniature transposable elements (mTEs), which are present in very high-copy numbers, stable, widely distributed and in close association with genic regions in plant genomes. These features make mTEs useful for applications such as developing marker systems, functional characterization of associated genes, and elucidating the contribution of TEs to gene evolution. Here, we summarize the characteristics, copy numbers and distribution patterns of five TRIM families, 14 short interspersed elements (SINE) families and 20 MITE families in the *Brassica rapa* genome. We also show the comparative distribution pattern of paralogous mTE family members in *Brassica oleracea* and 11 *B. rapa* accessions. In addition, we describe putative roles for mTEs in the evolution of the triplicated *Brassica* genome and discuss the utility of mTEs for analysis of genome evolution and for developing practical marker systems.

6.1 Introduction

Transposable elements, also known as "mobile genetic elements", are DNA sequence fragments that move or are copied from one location to another in the genome either directly, by a cut-and-paste mechanism (class II DNA transposons), or indirectly, by a copy-and-paste mechanism through an RNA intermediate (class

P. Sampath · J. Lee · T.-J. Yang (✉)
Department of Plant Science, Plant Genomics and Breeding Institute, and Research Institute of Agriculture and Life Sciences, College of Agriculture and Life Sciences, Seoul National University, Seoul 151-921, Republic of Korea
e-mail: tjyang@snu.ac.kr

F. Cheng · X. Wang
Chinese Academy of Agricultural Sciences, Institute of Vegetables and Flowers, Zhongguancun Southern Street 12, 100081 Beijing, China

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_6

I retrotransposons; Wicker et al. 2007; Feschotte et al. 2002) (Fig. 6.1). Transposition of both classes of elements may result in a heritable increase in copy number within the genome; hence, individual TE types are found in multiple copies (often referred to as a TE family) and constitute the majority of the repetitive fraction of eukaryotic genomes. The large-scale sequencing of eukaryotic genomes has revealed that TEs are the most abundant component of most eukaryotic genomes, are present ubiquitously, and occupy large fractions of genomes: TEs account for 40 % of *Oryza sativa* (rice) (Feschotte 2008), 50 % of *Glycine max* (soybean) (Schmutz et al. 2010), and >80 % of *Zea mays* (maize), *Triticum aestivum* (wheat) and *Hordeum vulgare* (barley) (Paterson et al. 2009; Wicker et al. 2009; Bennett and Smith 1976) genomes. Whole-genome analyses estimated that ~40 % of the *Brassica rapa* ($2n = 2x = 529$ Mb) and *B. oleracea* ($2n = 2x = 696$ Mb) genomes are occupied by transposon-related sequences (Wang et al. 2011; Liu et al. 2014). High proportions of TEs are intact in *B. rapa* and *B. oleracea* genomes (68 and 98 %, respectively), although TEs have been continuously amplified in both genomes since at least 4.6 million years ago (MYA) (Liu et al. 2014). Compared to *B. rapa*, *B. oleracea* has many younger TEs, which are responsible for its increased genome size (Liu et al. 2014). Amplification of TEs in the genome can not only cause an increase in genome size but also help to drive the evolution of genes and genomes (Feschotte et al. 2002; Bire and Rouleux-Bonnin 2012; Feschotte 2008; Alzohairy et al. 2013), although most TEs are inactive

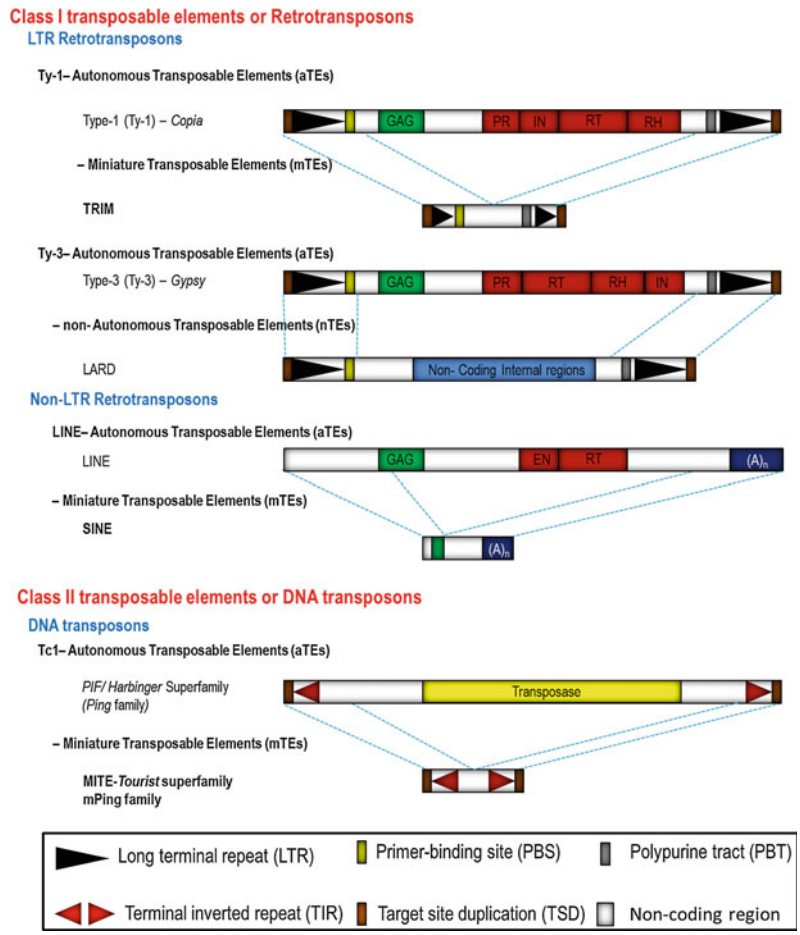

Fig. 6.1 Classification of TEs and mTEs. *LARD* large retrotransposon derivative; *TRIM* terminal-repeat retrotransposons in miniature; *LINE* long interspersed nuclear element; *SINE* short interspersed nuclear element; *GAG* a structural protein for virus-like particles; *PR* protease; *IN* integrase; *RT* reverse transcriptase; *RH* RNAse H; *EN* endonuclease

and mainly controlled by epigenetic mechanisms (e.g., DNA and histone methylation) (Hollister and Gaut 2009; Lisch 2009, 2012; Casacuberta and Santiago 2003).

TEs containing their own functional genes for transposition are referred to as autonomous transposable elements (aTEs), whereas TEs that lack coding genes and therefore cannot produce their own transposase or reverse transcriptase are termed nonautonomous or noncoding transposable elements (nTEs) (Casacuberta and Santiago 2003). The nTEs, such as large retrotransposon derivatives (LARDs), terminal repeat retrotransposons in miniature (TRIMs), short interspersed elements (SINEs) and miniature inverted-repeat transposable elements (MITEs), are generally deletion derivatives of aTEs and require a *trans*-acting transposase from their corresponding autonomous partner elements for transposition. TRIMs, SINEs and MITEs are examples of miniature transposable elements (mTEs; Fig. 6.1) (Casacuberta and Santiago 2003; Feschotte and Pritham 2007; Wessler et al. 1995; Okada et al. 1997), and families belonging to each type of mTE have had significant influences on gene and genome evolution (Wessler 2006; Witte et al. 2001). In this chapter, we summarize the characteristics, copy numbers and comparative distribution of 39 mTE families in *B. rapa* and *B. oleracea*. We go on to discuss the utility of mTEs for genomics-assisted breeding and evolutionary studies.

6.2 Characteristics and Distribution of mTEs

mTEs have unique structural characteristics and are ubiquitously present in eukaryotic genomes. The important characteristics of mTEs are summarized in Table 6.1.

Table 6.1 Characteristics of mTEs

Characteristics	TRIM	SINE	MITE
Element size	~900 bp (up to 2.5 Kb)	<700 bp	~800 bp (up to 2 Kb)
Structure			
Terminal repeats (bp)	TDR (100–350 bp)	No	TIR (10–1000 bp)
Target site duplication (TSD)	5 bp	No	2–11 bp (7 families identified)
Origin	Type-1 Copia LTR-RT	LINE	DNA-TE
Copy number	Moderately high	High	Very high
Copies in rice genome	~350	~2500	22,000–36,000
Copies in *Brassica* genome	~2000	~5000	~48,000
First identified element First identified organism (copies)	*Katydid* *Solanum tuberosum* (540)	*TS* *Nicotiana tabacum* (5400)	*Zm-1* *Zea mays* (1000–10,000)
Applications	Insertion polymorphism markers TRIM display	Insertion polymorphism markers SINE display	Insertion polymorphism markers MITE display
Discovery tools			
Structure based	LTR_Finder, LTR_STRUC, LTR_MINER, TRANSPO	SINEDR	MITEHunter, MITE Digger
Homology based	HMMER, repbase, repeat masker	SINEBase, repbase, repeat masker	P-MITE, repbase, repeat masker

6.2.1 TRIMs

TRIM elements are nonautonomous terminal direct repeat (TDR) retrotransposons with similar, but smaller, structural characteristics to long terminal repeat (LTR) retrotransposons. The TDR ranges from 100 to 350 bp, and has a 5-bp target site duplication (TSD). The internal region of 150–500 bp begins with a tRNA-methionine primer binding site (PBS) and ends with a poly-purine tract (PPT) motif (Witte et al. 2001; Yang et al. 2007; Fig. 6.1). It has been suggested that TRIM elements are mobilized through copy-and-paste mechanisms and that the proliferation of TRIMs has occurred with the help of *trans*-acting autonomous partner elements like Type-1/COPIAs, with which TRIMs share the characteristic signature structure, but there are no reports confirming this as yet (Witte et al. 2001; Kalendar et al. 2008).

TRIM elements are abundant and widely distributed and have been identified in both monocot and dicot plants and rarely in animals. *Katydid* was the first TRIM identified, originally in *Solanum tuberosum* (potato) and subsequently in the *Arabidopsis* genome (Witte et al. 2001). *Cassandra*, a unique TRIM family harboring 5S rDNA sequence, has been identified in more than 50 plant species including ferns and trees and is thought to have evolved 250 MYA (Kalendar et al. 2008; Sampath and Yang 2014a, b). Among the five TRIM families found in *Brassica* (TB-1-5), a total of 1393 and 1639 members were identified 283 Mb *B. rapa* sequences (including 256 Mb pseudo-chromosome and 27 Mb unanchored scaffold sequences) and 385 Mb *B. oleracea* pseudo-chromosome sequences. TRIM families are distributed throughout the *B. rapa* and *B. oleracea* genomes (Murukarthick et al. 2014). Though the insertions of five *Brassica* TRIM families were random, Genome-wide characterization showed that 619 (44 %) and 656 (40 %) of the members of the five *Brassica* TRIM families reside in or within 2 kb of a gene in the *B. rapa* and *B. oleracea* genome, respectively (Yang et al. 2007; Murukarthick et al. 2014; Table 6.2). Most-TRIM families are present in relatively similar copy numbers between *Brassica* species but the *Cassandra* family appears to show high divergence between *B. rapa* and *B. oleracea*, based on members found in counterpart paralogous sequences of pseudo-chromosomes (Sampath and Yang 2014a, b).

6.2.2 SINEs

SINEs are relatively short (75–662 bp), nonautonomous, non-LTR retrotransposons. SINEs have a unique structure composed of a head, body and tail. The head (5' end), which consists of an internal promoter, is derived from cellular RNA usually tRNA, 7SL RNA and/or 5S rRNA. SINEs lack a TSD but contain monopolymer tails. The internal promoter provides a transcription signal for transcription of the SINE by RNA polymerase III. The body of the SINEs originates either from autonomous partner elements or from distant SINE families. The 3' end (tail) of SINEs consists mostly of simple repeats. It has been suggested that SINEs also need a *trans*-acting partner (most probably LINEs) for amplification and mobilization (Okada et al. 1997; Kramerov and Vassetzky 2011a, b; Kramerov and Vassetzky 2005) (Fig. 6.1).

SINEs are abundant, present in high copy number and occupy a significant fraction of eukaryotic genomes. For instance, Alu, a well-characterized SINE from primates, exists in >1,500,000 copies in the human genome, covers >11 % of the total genome and played an important role in human population genetics and evolution (Venter et al. 2001; Batzer and Deininger 2002). In the *Brassicaceae*, 16 SINE families have been reported, including 1270 and 2364 members in the *B. rapa* and *B. oleracea* genomes, respectively (Vassetzky and Kramerov 2013). SINEs are distributed in various genomic locations of *B. rapa* and *B. oleracea*, with 599 (47.1 %) and 1154 (48.8 %) of the members, respectively, present in close association with genic regions with <2 kb of a gene (Murukarthick et al. 2014).

Table 6.2 Distribution of the members of 39 mTE families in the *B. rapa* and *B. oleracea* pseudo-chromosome sequences and 1× WGS data

mTE no.	mTE type	mTE ID[a]	Unit size (bp)	*B. rapa*			*B. oleracea*		
				Copies in genome assembly[b]	Estimated copies based on 1× WGS[c]	Range (copies)[c]	Copies in genome assembly[d]	Estimated copies based on 1× WGS[e]	Range (copies)[e]
1	TRIM	TB-1	364	72	120 ± 14.6	95–156	69	147 ± 15.5	131–162
2	TRIM	TB-2	387	21	61 ± 10.7	45–81	4	105 ± 2.5	102–107
3	TRIM	TB-3	1313	1	10 ± 1.4	7 ± 12	1	21 ± 1.5	19–22
4	TRIM	TB-4	598	43	132 ± 6.6	122–143	55	276 ± 15.5	260–291
5	TRIM	TB-5	781	19	128 ± 36.5	71–208	131	190 ± 39.5	150–229
6	SINE	SB-1	171	0	0 ± 0.3	0–1	14	50 ± 2	48–52
7	SINE	SB-2	149	0	23 ± 3.8	15–30	0	33 ± 3	30–36
8	SINE	SB-3	297	19	30 ± 7.6	17–47	86	184 ± 2	182–186
9	SINE	SB-5	162	0	51 ± 19.2	28–89	112	196 ± 7.5	188–203
10	SINE	SB-6	297	55	58 ± 15.6	25–81	93	279 ± 23.5	255–302
11	SINE	SB-7	352	16	63 ± 11.5	39–81	43	155 ± 11	144–166
12	SINE	SB-8	95	0	0 ± 0	0–0	64	57 ± 0.5	56–57
13	SINE	SB-9	212	80	83 ± 27.1	56–154	101	120 ± 11	109–131
14	SINE	SB-10	159	5	5 ± 1.9	2–9	76	70 ± 10	60–80
15	SINE	SB-11	170	0	1 ± 0.5	1–2	32	38 ± 1	37–39
16	SINE	SB-12	170	3	5 ± 1.1	3–7	41	46 ± 6	40–52
17	SINE	SB-13	225	0	0 ± 0.5	0–1	12	40 ± 14.5	25–54
18	SINE	SB-14	156	11	31 ± 10.4	18–54	50	74 ± 0.5	73–74
19	SINE	SB-15	206	0	1 ± 0.4	0–2	1	0 ± 0	0–0
20	MITE	BraSto-1	267	16	32 ± 22.8	5–92	50	249 ± 44	205–293
21	MITE	BraSto-2	260	401	155 ± 97.2	32–392	210	671 ± 137	534–808
22	MITE	BraSto-3	242	6	2 ± 1.3	0–5	2	2 ± 1	1–3
23	MITE	BraSto-4	558	97	38 ± 35.7	6–138	336	489 ±146.5	342–635
24	MITE	BraTo-2	212	8	44 ± 22.5	11–92	127	986 ± 135.5	850–1121
25	MITE	BraTo-1	366	61	25 ± 9.8	9–40	60	117 ± 25.5	91–142
26	MITE	BraTo-3	252	245	152 ± 70.4	43–291	116	212 ± 22	190–234
27	MITE	BraTo-4	160	287	217 ± 66.2	88–335	36	91 ± 10.5	80–101
28	MITE	BraTo-5	286	118	52 ± 28.6	14–120	37	93 ± 15.5	77–108
29	MITE	BraTo-6	257	60	14 ± 5.3	5–24	76	142 ± 35.5	106–177
30	MITE	BraTo-7	366	54	25 ± 14.3	4–53	199	471 ± 95.5	375–566
31	MITE	BraTo-8	348	29	14 ± 9	2–35	26	66 ± 26	40–92
32	MITE	BraTo-9	264	20	72 ± 24.4	31–127	32	30 ± 9.5	20–39
33	MITE	BraTo-10	255	35	28 ± 14.2	8–58	50	105 ± 19.5	85–124
34	MITE	BraTo-11	305	4	2 ± 1.5	0–6	5	5 ± 2	3–7
35	MITE	BraTo-12	273	66	31 ± 17	7–62	67	81 ± 7	74–88

(continued)

Table 6.2 (continued)

mTE no.	mTE type	mTE ID[a]	Unit size (bp)	*B. rapa*			*B. oleracea*		
				Copies in genome assembly[b]	Estimated copies based on 1× WGS[c]	Range (copies)[c]	Copies in genome assembly[d]	Estimated copies based on 1× WGS[e]	Range (copies)[e]
36	MITE	BraTo-13	268	74	35 ± 15.4	11–62	85	177 ± 26	151–203
37	MITE	BraHAT-1	439	24	15 ± 7.7	5–30	55	193 ± 37.5	155–230
38	MITE	BraHAT-2	248	16	23 ± 12.9	6–58	19	23 ± 1	22–24
39	MITE	BraMu-1	271	24	12 ± 7.2	2–28	16	87 ± 21	66–108
			Total	1990	1790		2589	6371	

[a]SB-4 and SB-16 are not shown here because they are not present in either *B. rapa* or *B. oleracea* but are present in *A. thaliana*

[b]mTE copies were identified from the available 283 Mb whole-genome pseudo-chromosome sequences of *B. rapa* with 80 % sequence similarity

[c]Mean, standard deviation (SD) and range were calculated based on members estimated from 1× (529 Mb) WGS of 11 *B. rapa* accessions

[d]mTE copies were identified from the available 385 Mb whole-genome pseudo-chromosome sequences of *B. oleracea* with 80 % sequence similarity

[e]Mean, standard deviation (SD) and range were calculated based on members estimated from 1× (696 Mb) WGS of 2 *B. oleracea* accessions

6.2.3 MITEs

MITEs are class II nonautonomous TEs characterized by relatively small size (<800 bp), AT-rich sequences, and flanking terminal inverted repeats (TIRs) ranging from 10 to 200 bp (Sampath and Yang 2014a, b). Insertion of a MITE can produce TSD ranging from 2 to 11 bp depending on the MITE superfamily involved (Fig. 6.1) (Bureau and Wessler 1994; Lu et al. 2012). The TIRs are more conserved than their respective internal sequences, and act as a recognition site for endonucleases for integration of TEs via transposition (Casacuberta and Santiago 2003). The TIRs are complementary to each other, leading to the formation of a secondary loop structure, which can be a source of small RNA and may act in gene regulation (Mo et al. 2012; Sampath et al. 2013; Sarilar et al. 2011). The internal sequences of MITEs have sequence diversity due to the influence of unrelated autonomous TEs during transposition (Sampath et al. 2013; Yaakov and Kashkush 2012). Unlike TRIMs and SINEs, transposition of MITEs occurs by cut-and-paste mechanisms, and MITEs can be amplified in the genome by abortive gap repair, bursts of amplification, or as yet unknown mechanisms under stress (Fattash et al. 2013; Casacuberta 2013).

Different MITE families are classified based upon TSD length, structure, and sequence similarity to the putative transposase of the corresponding DNA transposon. MITEs were first identified in the maize genome and later in various other plant and animal genomes (Bureau and Wessler 1994, 1992; Feschotte and Wessler 2002). So far, seven MITE superfamilies have been identified in plants, although 15 superfamilies of DNA transposons have been reported (Fattash et al. 2013). MITEs comprise two major families, namely *Stowaway*-like (with TA as the TSD) and *Tourist*-like (with TAA as the TSD), as well as several other minor families including

Table 6.3 Plant materials used for re-sequencing and mTE analysis

	ID	Morphotype	Species	Sub species	Accession (cultivar)	Genome	WGS reads for mTE analysis	
							Amounts (Mbp)	Coverage (x)
1	Br-1	Chinese cabbage	*B. rapa*	ssp. *pekinensis*	Chiifu	AA	2321.4	4.4
2	Br-2	Chinese cabbage	*B. rapa*	ssp. *pekinensis*	Kenshin	AA	1498.9	2.8
3	Br-3	Chinese cabbage	*B. rapa*	ssp. *pekinensis*	DF10C062	AA	1410.9	2.7
4	Br-4	Chinese cabbage	*B. rapa*	ssp. *pekinensis*	Z16	AA	1496.2	2.8
5	Br-5	Turnip Asian	*B. rapa*	ssp. *rapifera*	Yoya	AA	1495.9	2.8
6	Br-6	Rapini-Caixin	*B. rapa*	ssp. *parachinensis*	L58	AA	1492.5	2.8
7	Br-7	Pak Choi	*B. rapa*	ssp. *chinensis*	Suzhouqing	AA	1495.3	2.8
8	Br-8	Canola	*B. rapa*	ssp. *oleifera*	R-o-18	AA	1497.8	2.8
9	Br-9	Mizuna	*B. rapa*	ssp. *nipposinica*	Mizuna	AA	1497.5	2.8
10	Br-10	Turnip Europe	*B. rapa*	ssp. *rapifera*	Manchester	AA	1484.5	2.8
11	Br-11	Canola-rapid cycling	*B. rapa*	ssp. *oleifera*	L144	AA	1496.6	2.8
12	Bo-1	Cabbage	*B. oleracea*	ssp. *capitata*	C1176	CC	1541	2.2
13	Bo-2	Cabbage	*B. oleracea*	ssp. *capitata*	C1220	CC	1606.8	2.3

1–11 WGS of *B. rapa* accessions were kindly provided by Xiaowu Wang (Key Laboratory of Horticultural Crops Genetic Improvement of Ministry of Agriculture, Sino-Dutch Joint Lab of Horticultural Genomics Technology, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China). *12–13* WGS of *B. oleracea* was generated with support of a grant from the *Golden Seed Project (Center for Horticultural Seed Developmenet,* ***No. 213003-04-3-SB430****), Ministry of Agriculture, Food and Rural Affairs(MAFRA)*

hAT-like (with 5, 6, or 8 bp TSDs), *MULE* (with 9–10 bp TSDs), and *En/Spm* (3-bp TSDs) MITEs (Oki et al. 2008).

MITEs include the mTE members with the most copies, distributed throughout the genome. MITE family members occupy different proportions in plant and animal genomes, reaching up to 10 % in rice, 8 % in *Medicago*, 4 % in *B. rapa,* 0.71 % in *A. thaliana* and 16 % in *Aedes aegypti* (yellow fever mosquito) (Paterson et al. 2009; Schmutz et al. 2010; Lu et al. 2012; Nene et al. 2007; Chen et al. 2013). *In silico* analysis reveals 174 families with more than 45,821 members including *Tourist*, *Stowaway* (Feschotte and Pritham 2007), *Mutator* (Alzohairy et al. 2013) and *CACTA* (Wicker et al. 2007) in the *B. rapa* genome. Furthermore, 20 MITE families including two novel families were identified in *B. rapa* and *B. oleracea*, and comparative analysis providing useful information for genomics and breeding (Chen et al. 2013; Sampath et al. 2014).

6.3 Identification of mTEs

There are various bioinformatics tools available for mining of mTEs within genomes, each with its own advantages and drawbacks (Janicki et al. 2011). Sequence similarity-based analysis tools require a known repeat library for sequence searches, whereas structure-based mTE mining tools promote identification of novel families but can have false positive rates of up to 86 % (Han and Wessler 2010). Currently, 40 different mTE families, including five TRIM, 16 SINE and 20 MITE families, and their member distribution in

B. rapa and *B. oleracea* are listed in a recently developed database, BrassicaTED (http://im-crop.snu.ac.kr/BrassicaTED/index.php). BrassicaTED also includes tools for mining and characterization of mTEs and TEs (Murukarthick et al. 2014).

Genomic tools currently available for genome-wide surveys of TRIM elements include LTR_finder, LTR_STRUC, LTR_MINER, TRANSPO, and find_ltr, which use structure-based approaches, as well as Repbase and Repeatmasker based on sequence similarity.

For SINE families, comprehensive, up-to-date information about structural characteristics and copy numbers can be found in the SINEBase database (Vassetzky and Kramerov 2013). SINEBase, Repbase and Repeatmasker can be used for homology searches of SINE families. Most novel SINE families have been identified using SINE consensus sequences, such as the tRNA-related portions. The SINEDR tool has also been used to identify novel SINE families.

Mining of MITEs on the genome scale can be done using various genomics tools. For instance, FINDMITE (Tu 2001), MUST (Chen et al. 2009), MITE Hunter (Han and Wessler 2010), RSBP (Lu et al. 2012) and MITE digger (Yang 2013) are available online to identify MITEs based on signature structures such as the TIR and TSD. Repbase, Repeatmasker, Inverted Repeat Finder, REPuter, RECON, Micropeats and STAN can also be used to mine the MITEs based on sequence similarity (Smit and Hubley 1996; Warburton et al. 2004; Jurka 2000; Lerat 2009). A recently developed database for plant MITEs (P-MITE) contains MITEs from 40 different species including *Brassica*.

6.4 Influence of mTEs on Evolution of the Triplicated *Brassica* Genome

The mTEs can play roles in remodeling gene structures by exon shuffling and in gene expression modification by providing new transcription start sites, splicing sites and poly-A sites (Casacuberta and Santiago 2003; Witte et al. 2001; Yang et al. 2007; Lu et al. 2012; Antonius-Klemola et al. 2006; Havecker et al. 2004; Benjak et al. 2009). TRIM elements are actively involved in rearrangements of the highly duplicated *Brassica* genomes. It has been suggested that TRIM insertion into genes has mediated the gain of different functions, expression and evolution of triplicated genes (neofunctionalization) in *Brassica* genomes (Yang et al. 2007). Our previous comparative analysis of *Cassandra* (TB-5)family members in *A. thaliana, B. rapa* and *B. oleracea* suggested that some *Cassandra* elements have been commonly retained during the last 20 million years in the three species and that some elements have uniquely evolved in specific *Brassica* species (Sampath and Yang 2014a, b) (Fig. 6.2a).

Many studies have demonstrated that SINEs play a significant role in plant genetic variation and genomic evolution, including changing gene function and/or expression by providing regulatory elements such as alternative splicing sites and poly adenylation signals for functional RNA genes (Kramerov and Vassetzky 2011a, b; Ben-David et al. 2013). SINEs also cause insertional mutations and alter the methylation pattern in the genome (Batzer and Deininger 2002).

MITEs also play an important role in gene regulation and rearrangement and expression (Sampath et al. 2013, 2014). Transposition of MITEs into genes have been found to modify gene structure and function by deletion, point mutation, and affecting the transcriptional activity (Wessler et al. 1995; Mo et al. 2012; Sarilar et al. 2011; Shirasawa et al. 2012). Specifically, MITE transposition into introns in triplicated *B. rapa* genes appears to underlie their differential expression patterns (Sampath et al. 2013) (Fig. 6.2b). Although most MITEs are associated with genic regions, they are generally not found in exons. An exception is a *tourist* family of MITEs from *B. rapa*, *BraTo-9*, which is preferentially present in the exons of triplicated *B. rapa* genes. *BraTo-9* has provided new exons for functional genes of *B. rapa* (Sampath et al. 2014). When *BraTo-9* insertion occurred in triplicated or duplicated genes of *B. rapa*, the element was always found in

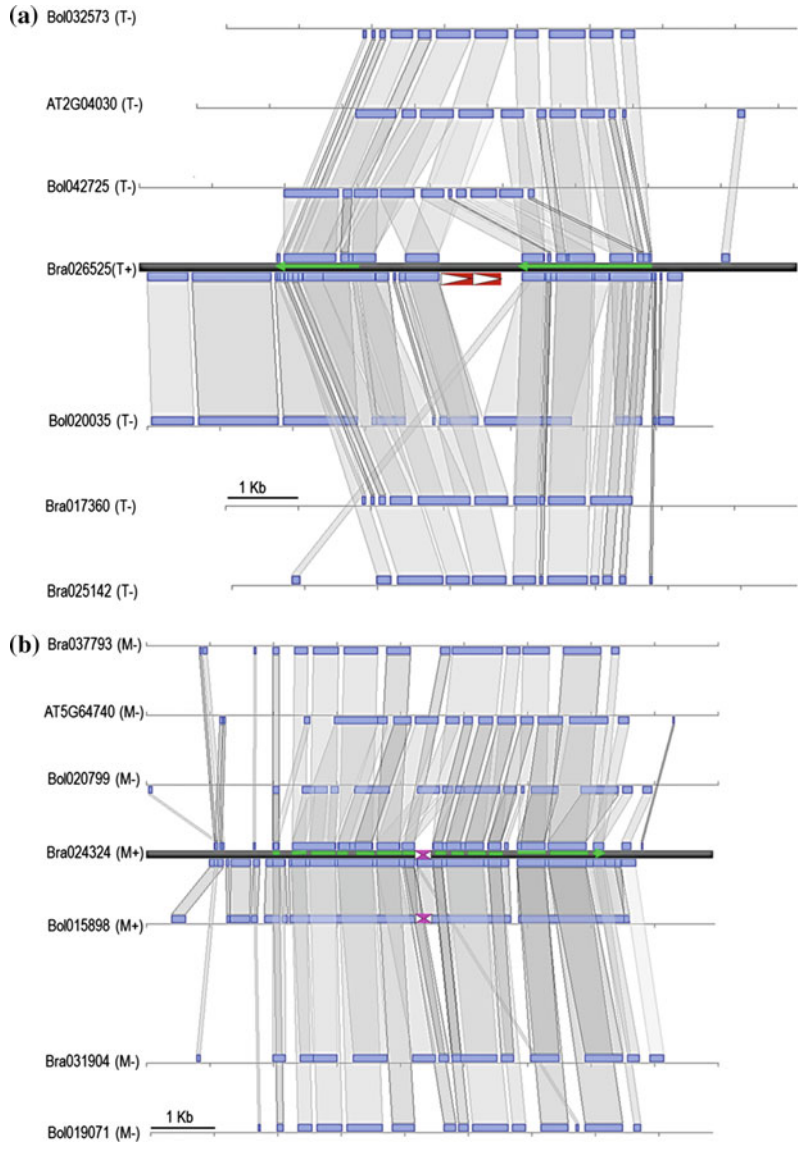

Fig. 6.2 Microsynteny comparison of *B. rapa* genomic regions containing TRIM (*TB-5*) and MITE (*BraSto-2*) elements with their noninserted paralogs (NIPs) and noninserted orthologs (NIOs) in *A. thaliana* and *B. oleracea*. **a** Unique *TB-5* (Br) element insertion in *B. rapa*. Microsynteny comparison of the genomic region of *TB-5* (Br) (Bra026525) with its NIPs (Bra025142, Bra017360) and NIOs from *B. oleracea* (Bol020035, Bol032573, Bol042725) and *A. thaliana* (AT2G04030). **b** Microsynteny between the genomic region showing shared insertion of *BraSto-2* in genes of *B. rapa* (Bra024324) and *B. oleracea* (Bol015898) compared with those of its NIPs of *B. rapa* (Bra037793, Bra031904) and *B. oleracea* (Bol020799, Bol019071) and its NIO from *A. thaliana* (AT5G64740). Exons and gene direction are indicated with green arrows. mTEelement insertions are shown as *red* and *pink bars*. + and − indicate genes with TRIM (T) or MITE (M) insertion and non-insertion, respectively. The *gray bars* connecting boxes on genome sequences indicate syntenic blocks present in both sequences. The map was generated based on nucleotide sequence similarity determined by BLASTn search

only one of the duplicated or triplicated genes, suggesting that the *BraTo-9* members were actively amplified in *B. rapa* after divergence with *B. oleracea* 4.6 MYA. MITE excision has caused gene knockout or silencing, and up- or down-regulation of gene expression by gene

rearrangement, *trans* duplication, and footprint mutation (Benjak et al. 2009; Shirasawa et al. 2012; Naito et al. 2009). In addition, MITEs are sources of small interfering RNA and can control genes in their vicinity (Piriyapongsa and Jordan 2007; Piriyapongsa et al. 2007; Kuang et al. 2009). *Trans* duplication in MITEs (i.e., MITEs with host gene sequence captured during excision) increases the likelihood of generating siRNA, which can influence gene regulation (Benjak et al. 2009; van Leeuwen et al. 2003). For instance, a MITE-based siRNA represses the expression of nearby genes by acting as a functional regulator triggering DNA methylation, and thereby affects agronomic traits such as leaf angle, plant height and inflorescence morphology (Wei et al. 2014). MITEs also have the ability to escape from silencing more efficiently than other TEs (Benjak et al. 2009; Parisod et al. 2010).

6.5 Copy Number Variation of mTE Members Between Accessions and Species

We estimated the copy numbers for each mTE family for each mTE in *B. rapa* compared to those in *B. oleracea* (Table 6.2) based on mapping of 1× coverage whole-genome sequence (WGS) reads using the criteria of 80 % sequence similarity with 80 % coverage against representative members of 39 mTE families. This analysis showed that *B. oleracea* has more mTE copies than does *B. rapa*, with 3-fold differences observed. In addition, copy numbers of some mTEs vary up to 2-fold among accessions of *B. rapa*. Together, these data suggest that mTE members were greatly amplified after the *B. rapa* and *B. oleracea* diversification about 4.6 MYA (Table 6.2; Fig. 6.3) (Mun et al. 2009).

The copy numbers of the five TRIM families (TB-1-5) ranged between 375 and 541 among *B. rapa* accessions, suggesting that TRIM elements have been active recently in different accessions (Fig. 6.4a; Table 6.2). Differential amplification of TRIM elements within or between species will be important target for identification of functional genes associated with mTE insertions and also for molecular breeding purposes (Fig. 6.5). More analysis related to gene function and association of TRIM elements could also promote the identification of agriculturally important genes.

Our analysis based on 14 SINE families shows that SINEs have been differentially amplified, with between 207 and 541 copies in *B. rapa* accessions but a larger range (207–1341) between *B. rapa* and *B. oleracea* (Fig. 6.4b; Table 6.2). Most of the SINE families exhibit similar amplification between both species, but some families (SB-Wicker et al. 2007; Liu et al. 2014; Bire and Rouleux-Bonnin 2012; Feschotte 2008; Alzohairy et al. 2013; Hollister and Gaut 2009; Lisch 2009) are more abundant in *B. oleracea* than in *B. rapa* (Table 6.2). This suggests that SINE elements also have proliferated with recent activation in both genomes, but more so in *B. oleracea* than in *B. rapa* (Fig. 6.5). These

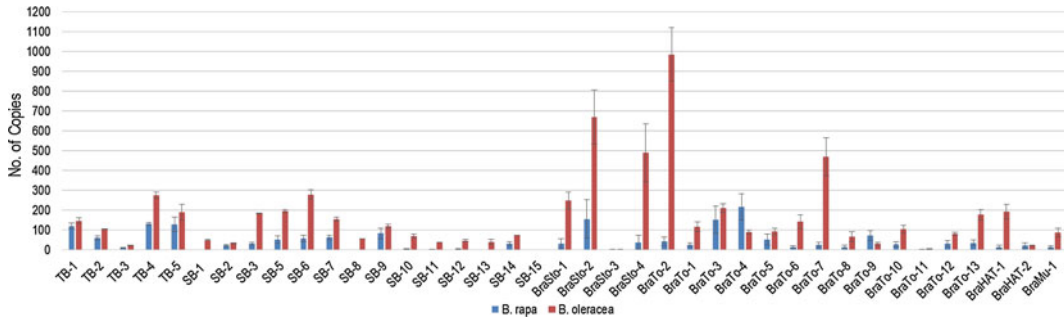

Fig. 6.3 Copy numbers of 39 mTE families in *B. rapa* and *B. oleracea* genomes based on 1× WGS data. The number of mTE members (average) and standard deviation were calculated from 11 *B. rapa* and 2 *B. oleracea* accessions

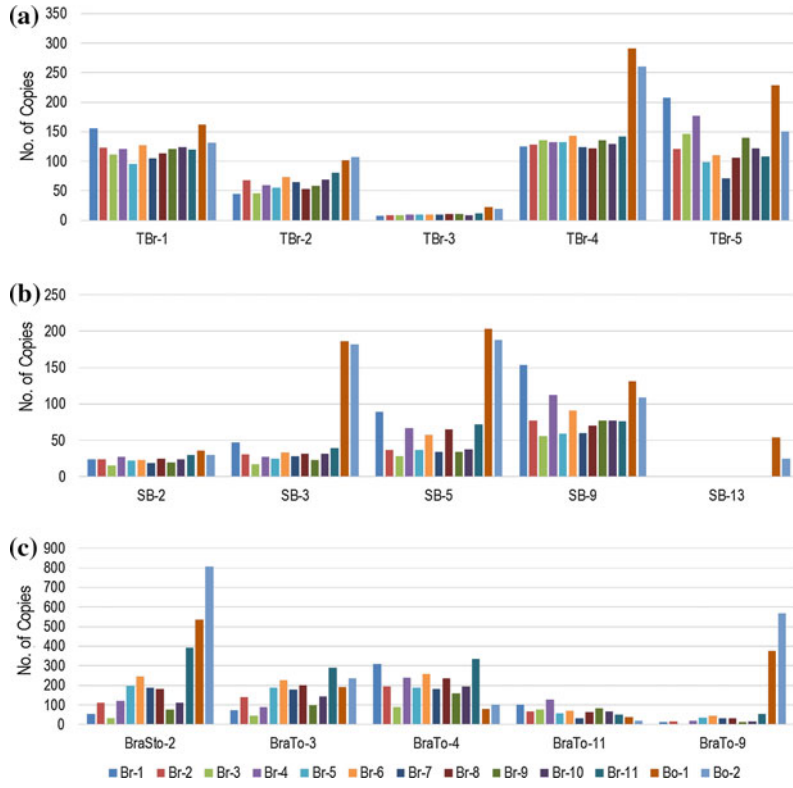

Fig. 6.4 Distribution of mTE family members in *B. rapa* and *B. oleracea* genomes based on 1× WGS data. The *graph* shows the total members from five families of TRIM (**a**), SINE (**b**) and MITE (**c**) elements based on 10 *B. rapa* accessions (Br-1 ~ 10) and two *B. oleracea* accessions (Bo 1, 2). The total copy numbers of mTE families were calculated based on 1× WGS read mapping. The accession names are listed in Table 6.3

diversely amplified members can be important resources for molecular and evolutionary studies.

MITEs also show significant divergence in copy number between *B. rapa* accessions (range 337–1968) and between *B. rapa* and *B. oleracea* (range 337–5076) (Fig. 6.4c; Table 6.2). This indicates that, like the other mTEs, MITEs remained highly active in both *Brassica* genomes. Thus, differential accumulation of MITEs within or between species should also be an important target for molecular breeding purposes and evolutionary studies (Fig. 6.5).

6.6 Utility of mTEs as Molecular Markers

DNA markers are used for a wide range of genomic applications such as construction of genetic linkage maps, genome-wide association studies and evolutionary studies (Casa et al. 2000; Kwon et al. 2007; Purugganan and Wessler 1995; Yaakov et al. 2012). TEs have been used to develop molecular markers such as those for inter-retrotransposon amplified polymorphism (IRAP), REtrotransposn-microsatellite amplified polymorphism (REMAP), sequence-specific amplification polymorphism (S-SAP), retrotransposon-based insertion polymorphism (RBIP), inter-MITE polymorphism (IMP) and transposon display (TD) (Agarwal et al. 2008). TE-based markers have been successfully utilized for various genomics purposes such as analysis of genetic diversity, inspection of clonal variation and breeding. TE markers are also useful to identify unambiguous gene flow between closely related species (Bire and Rouleux-Bonnin 2012; Carrier et al. 2012; Deragon and Zhang 2006). The principle characteristics of mTEs, namely their abundance, small size, stability, and distribution in genic regions, are advantageous for DNA marker development in both plants and animals. Thus, so-called mTE

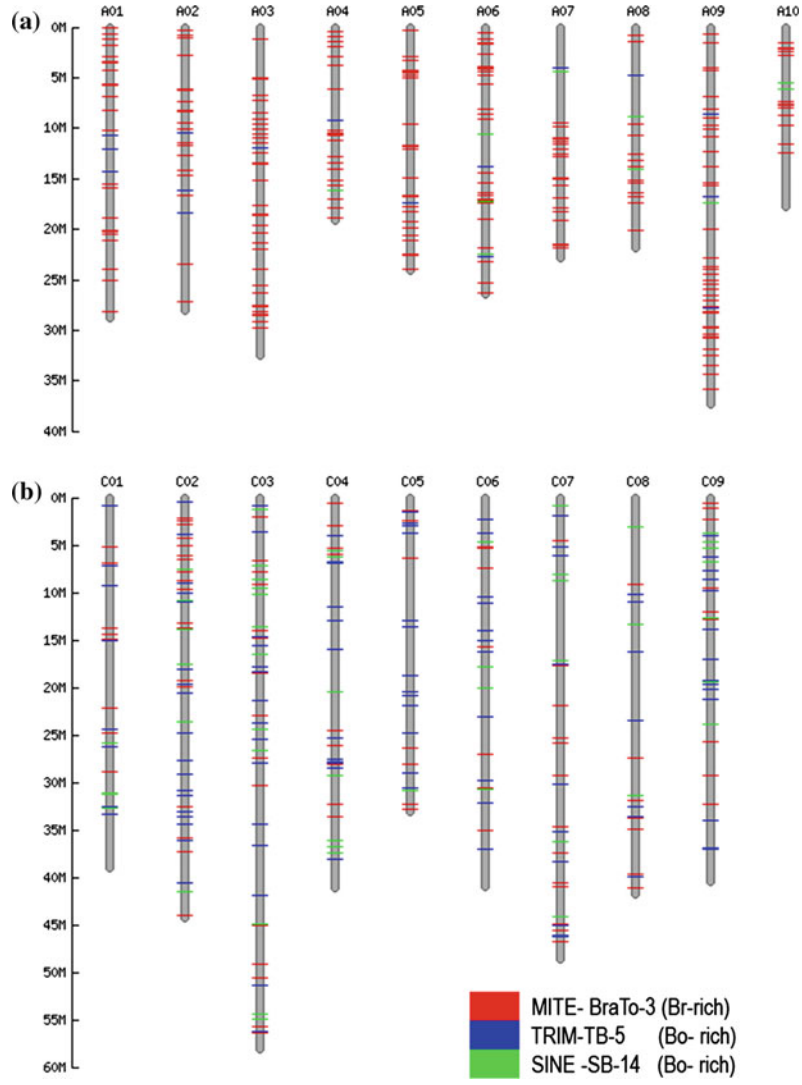

Fig. 6.5 Differential distribution of mTE family members in *B. rapa* and *B. oleracea*. mTE families with intact members were used for in silico map construction on the 256-Mb *B. rapa* (**a**) and the 385-Mb *B. oleracea* (**b**) pseudo-chromosome sequences based on physical positions. The physical position information for the mTE families of *B. rapa* and *B. oleracea* can be found in BrassicTED (24)

markers have been developed from mTEs such as TRIMs (Witte et al. 2001; Kwon et al. 2007), SINEs (Deragon and Zhang 2006; Shedlock and Okada 2000; Tatout et al. 1999) and MITEs (Shirasawa et al. 2012; Casa et al. 2000).

6.6.1 Insertion *Polymorphism of mTEs*

The presence (inserted site) or absence (empty site) of an mTE at a particular locus can be different among accessions, and this insertion polymorphism (IP) can be surveyed (Kwon et al. 2007; Yaakov et al. 2012) by PCR analysis using primers designed from the mTE flanking region (Fig. 6.6a). The mTE markers have an advantage over other types of markers because the stability and high copy numbers of mTEs allows development of abundant markers (Monden et al. 2009). IP markers represent co-dominant alleles at a single locus and can be used for applications such as identification of genome duplication or allopolyploidization events, genetic diversity analysis among related accessions or species, and development of markers and mapping using

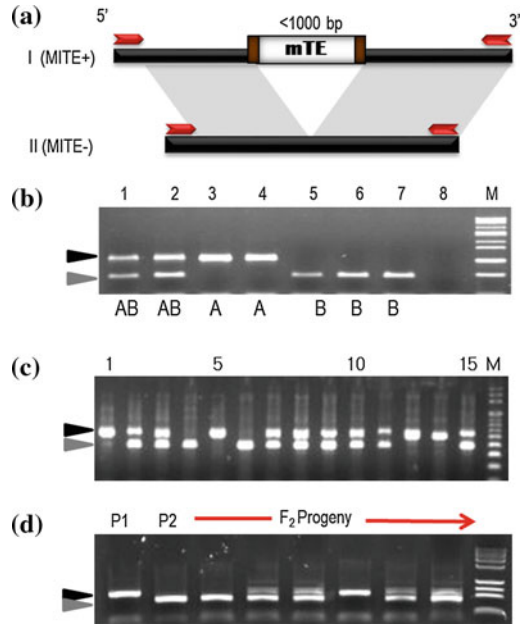

Fig. 6.6 Utility of mTEs as molecular markers**. a** MITE insertion polymorphism analysis using flanking primers. Comparison of DNA fragments showing the presence or absence of MITE insertion. MITE-flanking primer positions are indicated as *red arrowheads*. **b** Polymorphism profile by MIP analysis of 7 *Brassica* accessions based on *BraMi-1*, a *Brassica* MITE. *AB* insertion and non-insertion (Heterozygous insertion); *A* Insertion (Homozygous insertion); *B* non-insertion (Homozygous non-insertion). The list of accessions used and their ploidy are given in Sampath et al. (33). **c** Diversity analysis using different *B. oleracea* commercial cultivars. **d** Genotyping analysis of 94 *B. oleracea* F_2 plants from a cross between parental lines C1234 (P1) and C1184 (P2)

segregating populations between parental lines (Fig. 6.6b–d) (Sampath et al. 2013).

IP markers can be produced from sequences harboring TRIM, SINE, or MITE elements near the genes of interest. TRIM insertion polymorphism (TIP) markers were successfully developed to analyze genetic diversity and insertion time through divergent appearance of various TRIM elements in different *Brassica* accessions (Yang et al. 2007). SINE insertion polymorphism (SIP) markers have been used for construction of genetic maps, candidate-gene association studies and analysis of evolution within the *Glycine* genus. More than 52 % (77/146) of SIP markers developed from members of the GmAu1 mTE family displayed polymorphism (Shu et al. 2011). In addition, MITE insertion polymorphism (MIP) markers have been extensively studied in rice using a *Tourist* family MITE, *mPing*, the first active MITE identified in eukaryotes (Monden et al. 2009). MIP markers based on three MITEs (*Hbr, zmv1, Ins2*) were successfully used to study genetic diversity and identify a new candidate gene for flowering time variation in maize (Casa et al. 2000). MIP markers also have been used for high-resolution genetic diversity analysis and to elucidate the evolutionary history of *Triticum* (Yaakov et al. 2012). A MIP survey of three different *Brassica* accessions revealed high levels of inter- and intra-species polymorphism, at 52 % (150 markers) and 23 % (66 markers), respectively (Sampath et al. 2014). Transposition of MITEs and evolutionary dynamics were also evaluated in *Brassica* species using a MIP approach (Sampath et al. 2013). Thus, mTEs can be valuable targets from which to produce high numbers of successful markers quickly. It is important to note that high IP ratios are dependent on recent activation and high copy numbers for the target mTEs. Our analysis shows that compared with TRIMs and SINEs, MITEs are high copy and differentially amplified inter- and intra-species, suggesting that MITEs are particularly good candidates as targets for plant genome analysis.

6.6.2 Transposon *Display (TD) for mTEs*

TD is a modification of the AFLP method to target TEs and amplify most of the insertion sites of TEs. TD is be an efficient approach for rapid marker development because multiple insertion sites can be simultaneously amplified using conserved sequences of target mTEs that are distributed throughout the genome. TD was first developed and used for the maize *heartbreaker* MITE family (Casa et al. 2004). TD can be performed with primers targeting conserved regions of mTEs, such terminal inverted repeats (TIRs) for MITEs. TD-based markers have been effectively utilized for examining genetic diversity, phylogenetic analysis, genetic mapping,

identification of activation time of TEs based on divergence time and evolutionary studies (Kwon et al. 2007; Monden et al. 2009; Naito et al. 2006). mTE-based display, termed mTE-TD, has been applied for genome-wide detection of insertion sites that are polymorphic between or within species such as rice, maize, *Brassica*, *Vitis vinifera* (grapevine) and mosquito (Naito et al. 2006, 2009; Kwon et al. 2007; Zhang et al. 2000). The mTE-TD approach has advantages over AFLP because mTEs are more widely distributed in genome, especially in euchromatin regions. In addition, mTEs are closely associated with genic regions, which may help to develop markers related to agronomically important traits. Reports have also suggested that mTE-TD identifies a higher proportion of polymorphisms than does AFLP. TRIM-TD using conserved regions of *TRIM-Br1&2* revealed various insertion sites and were used for genetic mapping and high-resolution genetic diversity analysis of various *Brassica* relatives (Kwon et al. 2007).

Next-generation sequencing (NGS) technology produces numerous short DNA reads at relatively low cost and in a short period of time. NGS has a wide range of applications and has revolutionized the use of genomic data for crop improvement (Wei et al. 2013). The combination of TD with NGS technology allows the use of different high copy mTE families to detect insertion polymorphism among accessions. This approach will be a powerful tool for molecular breeding and evolutionary analysis.

6.7 Conclusion

Although mTEs cannot transpose by themselves due to their lack of protein-coding genes, mTEs have played important roles in plant genome evolution. Understanding the characteristics and member distribution of mTEs will promote their effective utilization to analyze genome evolution, dynamics, and plasticity as well as to identify the relevant genetic components of germplasm with agronomically important traits in the *Brassica* genome. The mTE-based markers are valuable resources for high-density genetic mapping, diversity analysis and evolution studies. Furthermore, insertion polymorphism surveys and NGS combined with TD are potential tools for marker systems aimed at high throughput marker development with minimum time and cost.

Acknowledgments This research was carried out with the support by Golden Seed Project (Center for Horticultural Seed Development, No. 213003-04-3-SB430), Ministry of Agriculture, Food and Rural Affairs (MAFRA), Ministry of Oceans and Fisheries (MOF), Rural Development Administration (RDA) and Korea Forest Service (KFS), Republic of Korea.

References

Agarwal M, Shrivastava N, Padh H (2008) Advances in molecular marker techniques and their applications in plant sciences. Plant Cell Rep 27(4):617–631

Alzohairy AM, Gyulai G, Jansen RK, Bahieldin A (2013) Transposable elements domesticated and neofunctionalized by eukaryotic genomes. Plasmid 69(1):1–15

Antonius-Klemola K, Kalendar R, Schulman AH (2006) TRIM retrotransposons occur in apple and are polymorphic between varieties but not sports. TAG Theor Appl Genet Theoretische und angewandte Genetik 112(6):999–1008

Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. Nat Rev Genet 3(5):370–379

Ben-David S, Yaakov B, Kashkush K (2013) Genome-wide analysis of short interspersed nuclear elements SINES revealed high sequence conservation, gene association and retrotranspositional activity in wheat. Plant J Cell Mol Biol 76(2):201–210

Benjak A, Boue S, Forneck A, Casacuberta JM (2009) Recent amplification and impact of MITEs on the genome of grapevine (*Vitis vinifera* L.). Genome Biol Evol 1:75–84

Bennett MD, Smith J (1976) Nuclear DNA amounts in angiosperms. Philos Trans R Soc Lond B Biol Sci 274 (933):227–274

Bire S, Rouleux-Bonnin F (2012) Transposable elements as tools for reshaping the genome: it is a huge world after all! Methods Mol Biol 859:1–28

Bureau TE, Wessler SR (1992) Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4(10):1283–1294

Bureau TE, Wessler SR (1994) Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6(6):907–916

Carrier G, Le Cunff L, Dereeper A, Legrand D, Sabot F, Bouchez O et al. (2012) Transposable elements are a major cause of somatic polymorphism in *Vitis vinifera* L. PloS One**7**(3):e32973

Casa AM, Brouwer C, Nagel A, Wang L, Zhang Q, Kresovich S et al (2000) The MITE family heartbreaker (Hbr): molecular markers in maize. Proc Natl Acad Sci USA 97(18):10083–10089

Casa AM, Nagel A, Wessler SR (2004) MITE display. Methods Mol Biol 260:175–188

Casacuberta JM (2013) MITEs, miniature elements with a major role in plant genome evolution. Plant Trans Elem Impact Genome Struct Funct 24:113

Casacuberta JM, Santiago N (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene 311:1–11

Chen Y, Zhou F, Li G, Xu Y (2009) MUST: a system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi. Gene 436(1–2):1–7

Chen J, Hu Q, Zhang Y, Lu C, Kuang H (2013) P-MITE: a database for plant miniature inverted-repeat transposable elements. Nucleic acids research 42(D1): D1176–D1181

Deragon J-M, Zhang X (2006) Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. Syst Biol 55(6):949–956

Fattash I, Rooke R, Wong A, Hui C, Luu T, Bhardwaj P et al (2013) Miniature inverted-repeat transposable elements: discovery, distribution, and activity1. Genome Natl Res Counc Can Genome Conseil Natl de Rech Can 56(9):475–486

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9 (5):397–405

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Ann Rev Genet 41:331

Feschotte C, Wessler SR (2002) Mariner-like transposases are widespread and diverse in flowering plants. Proc Natl Acad Sci USA 99(1):280–285

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3(5):329–341

Han Y, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38(22):e199

Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. Genome Biol 5(6):225

Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19(8):1419–1428

Janicki M, Rooke R, Yang G (2011) Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. Chromosome Res Int J Mol Supramol Evolutionary Aspects Chromosome Biol 19(6):787–808

Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends Genetics 16(9): 418–420

Kalendar R, Tanskanen J, Chang W, Antonius K, Sela H, Peleg O et al (2008) Cassandra retrotransposons carry independently transcribed 5S RNA. Proc Natl Acad Sci USA 105(15):5833–5838

Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. Int Rev Cytol 247:165–221

Kramerov DA, Vassetzky NS (2011) Origin and evolution of SINEs in eukaryotic genomes. Heredity 107 (6):487–495

Kramerov DA, Vassetzky NS (2011b) SINEs. Wiley interdisciplinary reviews. RNA 2(6):772–786

Kuang H, Padmanabhan C, Li F, Kamei A, Bhaskar PB, Ouyang S, et al.(2009) Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. Genome Res 19 (1):42–56

Kwon SJ, Kim DH, Lim MH, Long Y, Meng JL, Lim KB et al (2007) Terminal repeat retrotransposon in miniature (TRIM) as DNA markers in *Brassica* relatives. Mol Genet Genomics MGG 278(4):361–370

Lerat E (2009) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity 104(6):520–533

Lisch D (2009) Epigenetic regulation of transposable elements in plants. Ann Rev Plant Biol 60:43–66

Lisch D (2012) Regulation of transposable elements in maize. Curr Opin Plant Biol 15(5):511–516

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IA et al (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun 5:3930

Lu C, Chen J, Zhang Y, Hu Q, Su W, Kuang H (2012) Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in *Oryza sativa*. Mol Biol Evol 29 (3):1005–1017

Mo Y-J, Kim K-Y, Shin W-C, Lee G-M, Ko J-C, Nam J-K et al.(2012) Characterization of Imcrop, a Mutator-like MITE family in the rice genome. Genes Genomics 34(2):189–198

Monden Y, Naito K, Okumoto Y, Saito H, Oki N, Tsukiyama T et al.(2009) High potential of a transposon mPing as a marker system in japonica x japonica cross in rice. DNA Res Int J Rapid Publ Rep Genes Genomes 16(2):131–140

Mun JH, Kwon SJ, Yang TJ, Seol YJ, Jin M, Kim JA et al.(2009) Genome-wide comparative analysis of the *Brassica rapa* gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. Genome Biol 10 (10):R111

Murukarthick J, Sampath P, Lee SC, Choi BS, Senthil N, Liu S et al.(2014) BrassicaTED—a public database for utilization of miniature transposable elements in *Brassica* species. BMC Res Notes **7**(1):379

Naito K, Cho E, Yang G, Campbell MA, Yano K, Okumoto Y et al (2006) Dramatic amplification of a rice transposable element during recent domestication. Proc Natl Acad Sci USA 103(47):17620–17625

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO et al. (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461 (7267):1130–1134

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ et al (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. Sci Signal 316(5832):1718

Okada N, Hamada M, Ogiwara I, Ohshima K (1997) SINEs and LINEs share common 3′ sequences: a review. Gene 205(1):229–243

Oki N, Yano K, Okumoto Y, Tsukiyama T, Teraishi M, Tanisaka T (2008) A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, *Oryza sativa* ssp. japonica. Genes Genet Sys 83 (4):321–329

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C et al (2010) Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol 186(1):37–45

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457(7229):551–556

Piriyapongsa J, Jordan IK (2007) A family of human microRNA genes from miniature inverted-repeat transposable elements. PloS One 2(2):e203

Piriyapongsa J, Marino-Ramirez L, Jordan IK (2007) Origin and evolution of human microRNAs from transposable elements. Genetics 176(2):1323–1337

Purugganan MD, Wessler SR (1995) Transposon signatures: species-specific molecular markers that utilize a class of multiple-copy nuclear DNA. Mol Ecol 4 (2):265–269

Sampath P, Yang T-J (2014) Comparative analysis of cassandra TRIMs in three *Brassicaceae* genomes. Plant Genet Resour Charact Utilization 12(S1):S146–S150

Sampath P, Yang TJ (2014b) Miniature inverted-repeat transposable elements (MITEs) as valuable genomic resources for evolution and breeding of Brassica crops. Plant Breed Biotechnol 2:322–333

Sampath P, Lee S-C, Lee J, Izzah NK, Choi B-S, Jin M et al. (2013) Characterization of a new high copy stowaway family MITE, BRAMI-1 in *Brassica* genome. BMC Plant Biol 13(1):56

Sampath P, Murukarthick J, Izzah NK, Lee J, Choi HI, Shirasawa K et al. (2014) Genome-wide comparative analysis of 20 miniature inverted-repeat transposable element families in *Brassica rapa* and *B. oleracea*. PloS One 9(4):e94499

Sarilar V, Marmagne A, Brabant P, Joets J, Alix K. BraSto (2011) A stowaway MITE from *Brassica*: recently active copies preferentially accumulate in the gene space. Plant Mol Biol 77(1–2):59–75

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W et al. (2010) Genome sequence of the palaeopolyploid soybean. Nature 463(7278):178–183

Shedlock AM, Okada N (2000) SINE insertions: powerful tools for molecular systematics. BioEssays News Rev Mol Cell Dev Biol 22(2):148–160

Shirasawa K, Hirakawa H, Tabata S, Hasegawa M, Kiyoshima H, Suzuki S et al (2012) Characterization of active miniature inverted-repeat transposable elements in the peanut genome. Theor Appl Genet 124 (8):1429–1438

Shu Y, Li Y, Bai X, Cai H, Ji W, Ji Z et al. (2011) Identification and characterization of a new member of the SINE Au retroposon family (GmAu1) in the soybean, *Glycine max* (L.) Merr., genome and its potential application. Plant Cell Rep 30(12):2207–2213

Smit A, Hubley R (1996) Repeat modeler open- 3.0. Available at http://www.repeatmasker.org/

Tatout C, Warwick S, Lenoir A, Deragon J-M (1999) SINE insertions as clade markers for wild crucifer species. Mol Biol Evol 16(11):1614

Tu Z (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae. Proc Natl Acad Sci 98 (4):1699–1704

van Leeuwen H, Monfort A, Zhang H-B, Puigdomènech P (2003) Identification and characterisation of a melon genomic region containing a resistance gene cluster from a constructed BAC library. Microcolinearity between *Cucumis melo* and *Arabidopsis thaliana*. Plant Mol Biol 51(5):703–718

Vassetzky NS, Kramerov DA (2013) SINEBase: a database and tool for SINE analysis. Nucleic Acids Res 41(Database issue):D83–D89

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG et al. (2001) The sequence of the human genome. Science 291(5507):1304–1351

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S et al. (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43(10):1035–1039

Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G (2004) Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome Res 14 (10a):1861–1869

Wei L, Xiao M, Hayward A, Fu D (2013) Applications and challenges of next-generation sequencing in *Brassica* species. Planta 238:1005–1024

Wei L, Gu L, Song X, Cui X, Lu Z, Zhou M et al (2014) Dicer-like 3 produces transposable element-associated 24-nt siRNAs that control agricultural traits in rice. Proc Natl Acad Sci USA 111(10):3877–3882

Wessler SR (2006) Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci USA 103(47):17600–17601

Wessler SR, Bureau TE, White SE (1995) LTR-retrotransposons and MITEs: important players

in the evolution of plant genomes. Curr Opin Genet Dev 5(6):814–821

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8(12):973–982

Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M et al (2009) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. Plant J 59 (5):712–722

Witte C-P, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci 98(24):13778–13783

Yaakov B, Kashkush K (2012) Mobilization of Stowaway-like MITEs in newly formed allohexaploid wheat species. Plant Mol Biol 80(4–5):419–427

Yaakov B, Ceylan E, Domb K, Kashkush K (2012) Marker utility of miniature inverted-repeat transposable elements for wheat biodiversity and evolution. TAG Theor Appl Genet Theoretische und angewandte Genetik 124(7):1365–1373

Yang G (2013) MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements. BMC Bioinfor 14:186

Yang TJ, Kwon SJ, Choi BS, Kim JS, Jin M, Lim KB et al. (2007) Characterization of terminal-repeat retrotransposon in miniature (TRIM) in *Brassica* relatives. TAG Theor Appl Genet Theoretische und angewandte Genetik 114(4):627–636

Zhang Q, Arbuckle J, Wessler SR (2000) Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family Heartbreaker into genic regions of maize. Proc Natl Acad Sci USA 97(3):1160–1165

Genomic Survey of the Hidden Components of the *B. rapa* Genome

7

Nomar Espinosa Waminal, Sampath Perumal, Ki-Byung Lim, Beom-Seok Park, Hyun Hee Kim and Tae-Jin Yang

Abstract

The sequencing of the *Brassica rapa* genome has enabled better understanding of its structure and evolution, and created numerous opportunities for exploration of genome function and breeding applications. Nevertheless, the currently available completed genome sequences are estimated to cover only about 60 % of the genome, while the remaining 40 % is unassembled mainly due to the highly repetitive nature of this portion of the genome. Elucidation of the nature and distribution of repeat elements in the context of the entire genome would enhance our understanding of their role in genome structure, function, and evolution. In this chapter, we review the genomic distribution, characterization and evolutionary implications of currently identified repeat elements comprising the 'hidden' portion of the *B. rapa* genome. Low-coverage whole-genome sequence (WGS) was used to survey the major genomic repeats and their proportion in the *B. rapa* genome. Coupling this with molecular cytogenetics, we characterized the abundance and genomic distribution of seven major repeats, namely centromeric tandem repeats 1 and 2, centromeric retrotransposons, pericentromeric retrotransposons, 5S rDNA, 45S rDNA, and subtelomeric tandem repeats. These repeats accounted for approximately 20 % of the *B. rapa* genome, which is much more than the <1 % covered by repeats in the currently available genome

N.E. Waminal · S. Perumal · T.-J. Yang (✉)
Department of Plant Sciences, Plant Genomics and Breeding Institute and Research Institute of Agriculture and Life Sciences, College of Agriculture and Life Sciences, Seoul National University, Seoul 151-921, Republic of Korea
e-mail: tjyang@snu.ac.kr

K.-B. Lim
Department of Horticultural Science, Kyungpook National University, Daegu 702-701, Korea

B.-S. Park
National Academy of Agricultural Science, Rural Development Administration, 150 Suinro, Suwon 441-707, Republic of Korea

H.H. Kim
Department of Life Science, Plant Biotechnology Institute, Sahmyook University, Seoul 139-742, Republic of Korea

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_7

assembly. We also compared their distributions among different *B. rapa* accessions and in the close relative *Brassica oleracea,* for better understanding of the plasticity of the *Brassica* genomes.

7.1 Introduction

Knowledge of genome sequences has a huge impact in plant biology (Schadt et al. 2010). The number of plant genomes being sequenced is rising (Michael and Jackson 2013) due to the rapid advancement of genome sequencing technologies, including those that allow high-throughput sequencing of longer reads and high-resolution assembly algorithms (Edwards and Batley 2010; Metzker 2010; Schatz et al. 2012). However, a common hurdle is assembly accuracy, especially considering the highly repetitive nature of plant genomes (Macas et al. 2007; Schatz et al. 2012). For example, bread wheat, which has one of the largest genomes among those sequenced from plants (17,000 Mbp; Brenchley et al. 2012), has an estimated repeat content of 80 % and the sequences assembled into scaffolds covered only 22 % of the genome (Brenchley et al. 2012; Michael and Jackson 2013). Even for Chinese cabbage (*Brassica rapa*), which has a relatively small genome of 529 Mbp, only about 60 % of the genome was assembled into pseudo-chromosome sequences, with the remaining 40 % made up mainly of repeat elements (Johnston et al. 2005; Wang et al. 2011; Michael and Jackson 2013).

Repetitive components of genomes are responsible for the extensive genome size variation in higher plants (Hardman 1986; Pagel and Johnstone 1992; Macas et al. 2007) and used to be considered 'junk' (Doolittle and Sapienza 1980; Nowak 1994; Shapiro and von Sternberg 2005). However, many recent studies have shown that repetitive elements have diverse functions within cells (Biémont and Vieira 2006; Biémont 2010), from involvement in maintaining chromosome integrity (Nowak 1994), and gene expression (Biémont and Vieira 2006), to changing phenotypes (Biémont and Vieira 2006). Therefore, characterization of these components in relation to genome assemblies is fundamental to understanding the holistic landscape and deciphering the complexity of plant genomes (Biémont 2010).

Despite their importance, repetitive sequences have hindered genome assembly and increased costs in terms of both time and money (Schatz et al. 2012). They remain largely unexplored and unassembled in many sequenced plant genomes (Wang et al. 2011; Michael and Jackson 2013; Liu et al. 2014), because most assembly algorithms are designed for less complex sequences (Schatz et al. 2012). However, the large amount of information that could be gathered from these repeats would be useful for understanding genome structure and evolution (Biémont 2010).

In the assembled genome sequences, most of the repetitive elements that occupy ~40 % of the *B. rapa* genome are transposon related (Wang et al. 2011; Michael and Jackson 2013). However, more redundant repeats such as centromeric and pericentromeric LTR retrotransposons (CRBs and PCRBrs, respectively; Lim et al. 2007), centromeric tandem repeats (including CentBr1 and CentBr2; Lim et al. 2005), and subtelomeric tandem repeats (STRs; Koo et al. 2011), in addition to the rDNA arrays were not included in the assembled genome sequence. Less than 1 % of these repeats are included in the currently available 283 Mbp assembled sequences (Table 7.1) despite coverage of >98 % of the euchromatic regions (Wang et al. 2011). This discrepancy demonstrates the difficulty of anchoring repeats in the assembly. Characterizing, quantifying and cytogenetically mapping these elements should aid in the final refinement of the genome structure.

Table 7.1 Comparison of major repeat composition identified in the reference genome assembly of *B. rapa* 'Chiifu' (Wang et al. 2011) with that found in 1x WGS sequence of 11 *B. rapa* accessions

Repeat element	Unit length (bp)	Reference genome (283 Mbp) GR[a]						1x WGS (529 Mbp)[b]		GP by FISH (%)	Genome appearance (%)[d]
		256 Mbp pseudo-molecule		Unanchored scaffold		Total (a + b) (kb) (A)	GP[c] (%)	GR (Kbp) (B)	GP (%)		
		Copy	(kb) (a)	Copy	(kb) (b)						
CENTBr1	176	48	8.1	51	8.8	16.9	0.006	34,700 (±8568)	6.56	11.4	0.0
CENTBr2	176	147	25.3	93	16.1	41.4	0.015	7095 (±3177)	1.34	2.3	0.5
STR	351	831	272.6	135	43.3	315.9	0.112	5908 (±2942)	1.12	2.4	2.6
45S rDNA	7764	0	4.2	0	3.2	7.4	0.003	42,534 (±13,265)	8.04	5.9	0.0
5S rDNA	501	12	5.9	5	2.5	8.4	0.003	2631 (±1160)	0.50	1.7	0.2
CRB	5908	1	5.9	0	0.0	5.9	0.002	4098 (±613)	0.77	2.5	0.2
PCRBr	8395	0	0.0	0	0.0	0.0	0.000	11,221 (±4087)	2.12	3.3	0.0
Total		1039	322.4	284	73.9	395.9	0.140	108,186 (±15,415)	20.45	29.5	0.3

[a]Genome representation: average read depth × contig length
[b]Average value from 11 *Brassica rapa* accessions
[c]Genome proportion: (total GR/reference genome size in kbp) × 100
[d]Appearance in genome sequence based on the GR value of Chiifu WGS (%) = (A/B) × 100

In this chapter, we describe a genomic survey for major repeats of *B. rapa* using 1x whole-genome sequence (WGS) that captured a substantial portion of previously reported repeats and allowed us to characterize others. We also review the possible evolutionary roles of the identified repetitive elements in shaping the *B. rapa* genome. We further demonstrate the utility of combining in silico mapping of low-coverage WGS and fluorescence in situ hybridization (FISH) techniques to localize and estimate the genomic distribution and abundance of each repeat family. Finally, we discuss exciting applications and future prospects for this approach, especially for large and repeat-replete genomes and resource-deficient plant species.

7.2 The Hidden Genome: Characterization of Major Repeats

Knowing the distribution of repetitive elements within a genome is important in understanding genome organization, evolution, and function (Harrison and Heslop-Harrison 1995). In *B. rapa*, analysis using mitotic chromosome spreads demonstrated that heterochromatin is mostly concentrated in the centromeric and pericentromeric regions (Lim et al. 2005). These regions were later shown to contain major repetitive elements including the centromeric tandem repeats CentBr1 and CentBr2, centromeric retrotransposon of *Brassica* (CRB) and peri-centromeric retrotransposon of *B. rapa* (PCRBr; Harrison and Heslop-Harrison 1995; Lim et al. 2000, 2005; Koo et al. 2004). Repeats that are not concentrated in the centromeric regions have also been characterized (Wang et al. 2011; Liu et al. 2014). In addition to the tandemly repeated housekeeping 5S and 45S rRNA genes, a tandem repeat named STR based on its localization in the subtelomeric regions of several *Brassica* species was recently discovered (Koo et al. 2011). Collectively, these elements constitute the major repeat components of the hidden portion of the *B. rapa* genome (Table 7.1).

Most of these repeats have been identified by capture and characterization of single or a few elements via various efforts by independent groups; thus, global and comparative analyses of repetitive elements among related genomes has been limited (Macas et al. 2007). Oftentimes, considerable time and resources were spent to characterize these elements. For example, CentBr1, CentBr2, CRB, and PCRBr were isolated after identification of patterns in restriction enzyme digestion, screening several thousand BAC clones, downstream cloning of isolated sequences, sequencing, and cytogenetic mapping (Harrison and Heslop-Harrison 1995; Koo et al. 2004; Lim et al. 2005, 2007). With the current availability of NGS technology, a huge amount of information now awaits capture and utilization without the tedium and expense of more traditional approaches.

7.2.1 Reconstruction of Nuclear rDNA Units

Owing to the vital function they play in protein biosynthesis and cellular function, ribosomal RNA genes are highly conserved across plant species (Hershkovitz and Zimmer 1996; Martins and Wasko 2004; Waminal et al. 2014). However, the spacers between each rDNA repeat unit are more divergent among species, making them an excellent tool for phylogenetic studies (Martins and Wasko 2004). Additionally, they have been exploited as cytogenetic FISH markers for studies related to genome dynamics and evolution (Roa and Guerra 2012; Waminal et al. 2012). However, complete sequences of *B. rapa* nuclear rDNAs have not yet been reported. Using de novo assembly of low-coverage WGSs (dnaLCW; Kim et al. 2015), we obtained the complete 5S unit without gaps and 45S rDNA unit sequences with small gaps in the intergenic spacer (IGS) for *B. rapa*.

The complete 5S rDNA unit was 501 bp, comprising a 120-bp 5S rRNA gene and 381-bp IGS (Fig. 7.1a). Based on mapping of raw reads

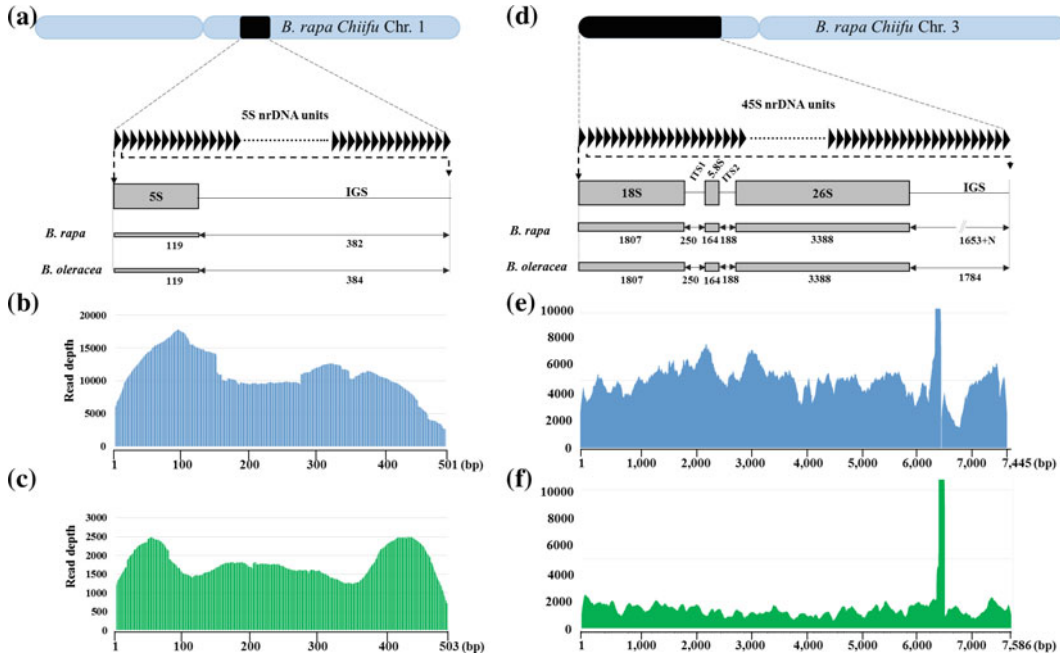

Fig. 7.1 Structure of 5S and 45S rDNAs of *B. rapa* 'Chiifu' and *B. oleracea* C1234 and raw read mapping. **a** Structure of the complete 5S rDNA unit of *B. rapa* and *B. oleracea* assembled based on the dnaLCW method (Kim et al. 2015). **b**, **c** Coverage of the 5S rDNA unit based on raw read mapping against the 1x genomes of *B. rapa* (genbank no: KM538957) and *B. oleracea* (genbank no: KM538957), respectively. **d** Structure of the 45S rDNA unit of *B. rapa* (partial) (genbank no: KM538957) and *B. oleracea* (complete) (genbank no: KM538957) assembled based on the dnaLCW method. **e**, **f** Coverage of 45S rDNA unit based on raw read mapping against the 1x genome of *B. rapa* and *B. oleracea*, respectively

to the complete 5S rDNA contig (Fig. 7.1b), it was estimated that there were 16,756 copies of the 5 rDNA unit in the haploid 'Chiifu' genome (Fig. 7.2c). Likewise, the complete 5S rDNA unit for *Brassica oleracea* 'C1234' totaled 503 bp with 119-bp genic and 384-bp IGS regions. However, only 1743 copies were estimated to be present in the *B. oleracea* genome based on raw read mapping (Fig. 7.1c); a value much lower than that in the *B. rapa* 'Chiifu' genome, and supported by FISH analysis (Fig. 7.3a, b, e, f). Obtaining the complete unit of the 45S rDNA sequence for *B. rapa* 'Chiifu' was hindered by GC-rich repeats in the IGS region. Due to the abundant subrepeat regions and possible heterogeneous sequences in the IGS, gap-filling methods were ineffective, leaving a small gap in the 45S rDNA unit of 7764 bp for *B. rapa* 'Chiifu' (Fig. 7.1d). Nevertheless, using the same methods we successfully obtained a complete 7586-bp 45S rDNA unit for *B. oleracea*. Mapping 1x NGS reads to 45S rDNA sequences of *B. rapa* 'Chiifu' and *B. oleracea* C1234 (Fig. 7.1e, f) revealed 8709 and 1339 copies, respectively (Fig. 7.2c).

7.2.2 Exploring the Hidden Portion of the Genome

A few studies have reconstructed and estimated the genomic content of major repeats using low-coverage NGS sequences (Hawkins et al. 2006; Macas et al. 2007; Swaminathan et al. 2007). This approach allowed the identification of up to 48 % of the 75–97 % repeats in the 4300 Mbp *Pisum sativum* genome (Macas et al. 2007). Even though not all of the repeats were captured in silico, enough information was available to carry out comparative studies among closely-related species. Coupled with FISH, this

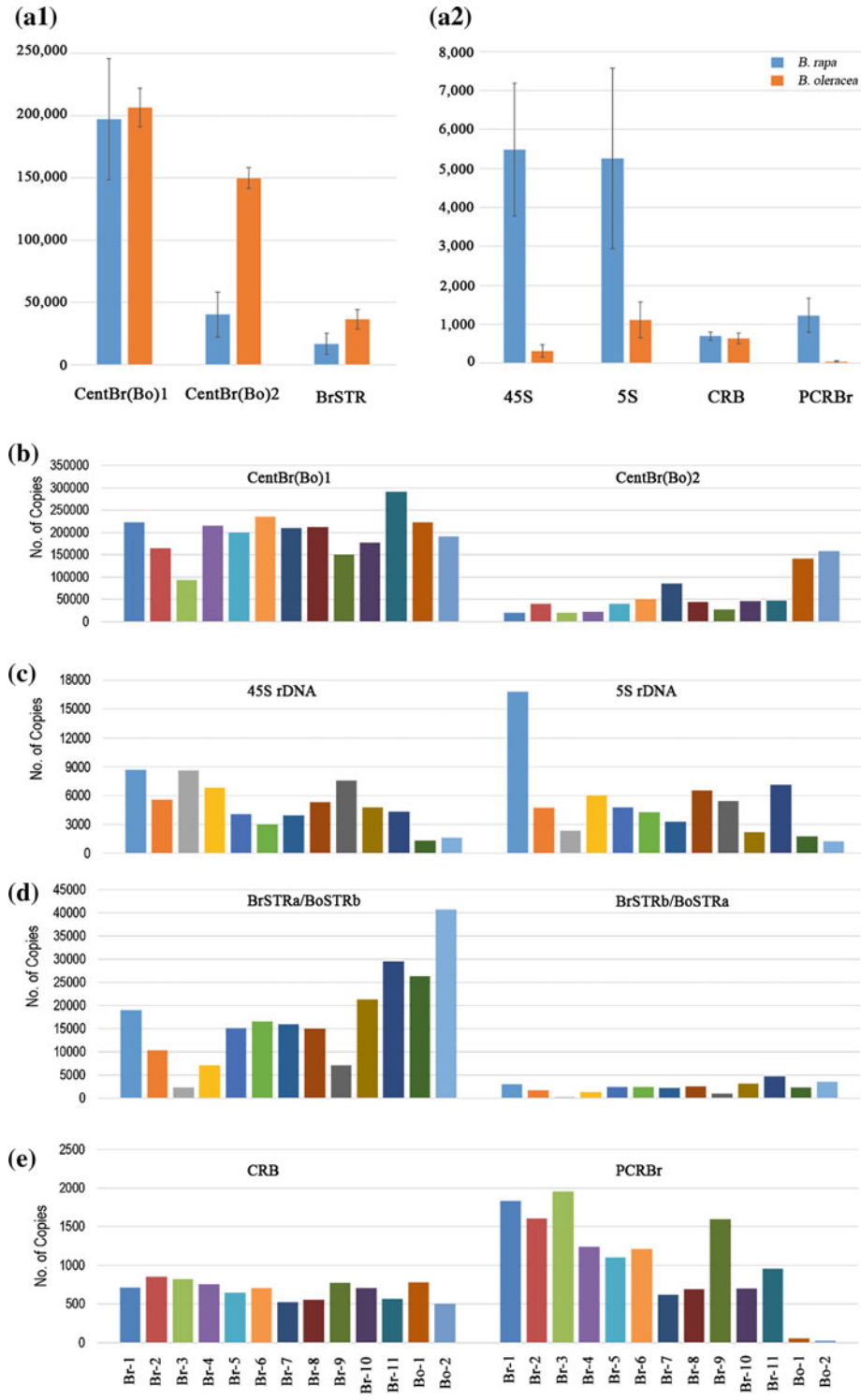

Fig. 7.2 Sequences identified via genomic survey of major repeats among 11 different *B. rapa* and two *B. oleracea* accessions for comparison. **a1** Comparison of centromeric and subtelomeric tandem repeat copy numbers and **a2** rDNA and centromeric retrotransposons between *B. rapa* and *B. oleracea*. Error bars represent standard deviation. Copy numbers of **b** centromeric tandem repeats of *B. rapa* (CentBr), **c** ribosomal DNA (rDNA), **d** *B. rapa* and *B. oleracea* subtelomeric satellite repeats (BrSTR and BoSTR, respectively), and **e** centromere-specific retrotransposon of *Brassica* (CRB) and peri-centromeric retrotransposon of *B. rapa* (PCRBr)

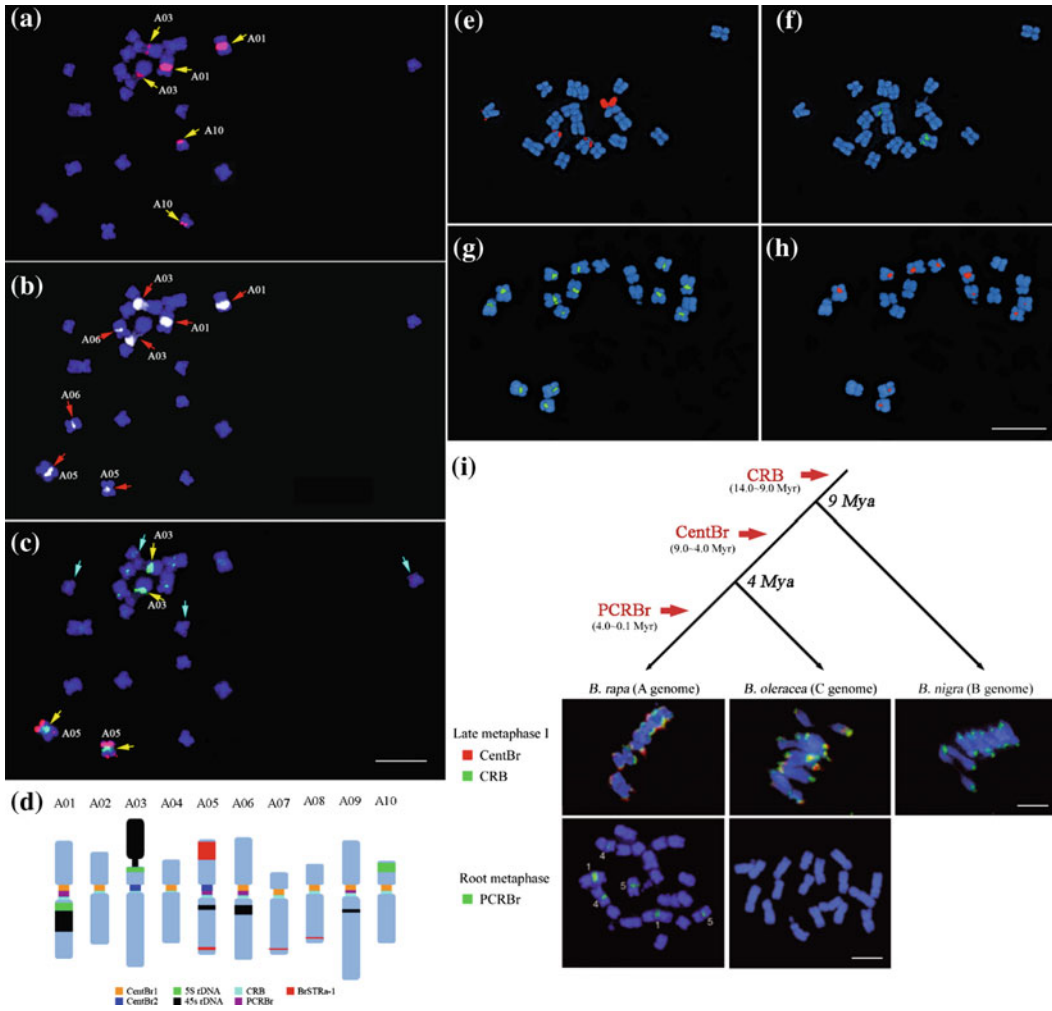

Fig. 7.3 Cytogenetic mapping and evolution of *B. rapa* and *B. oleracea* major repeats. *B. rapa* **a** FISH signals of 5S rDNA (*yellow arrows*), **b** 45S rDNA (*red arrows*), **c** CentBr2 (*green signals, yellow arrows* indicate 4 major signals) and BrSTR (*red*, on both arms of chromosome A05, *blue arrows* indicate weak BrSTR signals) in *B. rapa* root metaphase chromosomes. **d** Karyotype idiogram showing the cytogenetic distribution of major repeats. Chromosome numbering is according to Xiong and Pires (2011). **e**–**h** *B. oleracea*. **e** 45S rDNA **f** 5S rDNA **g** CentBo1, and **h** CentBo2. **i** Genome-specific evolution of *Brassica* centromeric repeats showing lineage divergence (Mya) at nodes and repeats with corresponding estimated insertion and amplification time (Myr). Bars in **a**–**h** = 10 μm, **i** = 5 μm

approach was able to reveal the distribution of the newly identified tandem repeats, providing a better picture of their actual location and abundance in the genome.

In *B. rapa*, several types of major repeats have been characterized, including the centromeric and pericentromeric LTR retrotransposons, CRB and PCRBr, respectively (Lim et al. 2007), centromeric tandem repeats CentBr1 and CentBr2 (Lim et al. 2005), and subtelomeric tandem repeat STR (Koo et al. 2011). We used these publicly available sequences along with the *B. rapa* rDNA sequences we assembled herein to survey the abundance of each element using 1x Illumina WGS data with at least 80 % sequence similarity as a criterion. As stated above,

repetitive elements currently identified in the *B. rapa* pseudo-chromosome sequences covered less than 1 % of the total assembled sequence (Wang et al. 2011). Here, we identified repetitive elements representing more than 20 % of the genome. Accordingly, only 0.3 % of these sequences are represented in the current genome assembly (Table 7.1).The most abundant repeats in the *B. rapa* genome were 45S rDNA (8 %), followed by CentBr1 (7 %) and PCRBr (2 %).

In *B. rapa* (A genome), CentBr1 is more abundant than CentBr2 (Fig. 7.2a, b), unlike their orthologous sequences in *B. oleracea* (C genome), CentBo1 and CentBo2, which are present in similar copy numbers (Fig. 7.2a, b; Lim et al. 2007; Koo et al. 2011). This was supported by our 1x WGS survey of 11 *B. rapa* and two *B. oleracea* accessions that revealed large copy number differences between CentBr1 and CentBr2, but not much difference between CentBo1 and CentBo2 (Fig. 7.2b).

The 1x WGS survey also identified >5000 and >300 times more 45S and 5S rDNA, respectively, than what was included in the assembled pseudo-chromosome sequences (Table 7.1). When compared to *B. oleracea*, *B. rapa* had 5 and 17 times more copies of 5S and 45S rDNA, respectively (Fig. 7.2a), which was consistent with FISH results (Fig. 7.3a, b, e, f; Xiong and Pires 2011).

Previous reports have identified two classes of subtelomeric tandem repeats in *Brassica*, STRa and STRb which share 89 % sequence identity (Koo et al. 2011). More sequences were identified from the 1x WGS reads when searching with BrSTRa compared to BrSTRb, suggesting that BrSTRa type TR sequences are more abundant than BrSTRb type sequences in both the *B. rapa* and *B. oleracea* genomes (Fig. 7.2d). In addition, different accessions of *B. rapa* and *B. oleracea* showed orders of magnitude difference in abundance for other repeat elements, indicative of genome plasticity which may reflects phenotypic polymorphism among accessions (Fig. 7.2b–e).

There was not much copy number variation for CRB elements among different *B. rapa* and *B. oleracea* accessions (Fig. 7.2e), supporting their common existence in the genus *Brassica* (Lim et al. 2007). By contrast, PCRBr was significantly more abundant in *B. rapa* compared with the negligible amount found in *B. oleracea* (Fig. 7.2e), supporting the observation of Lim et al. (2007) that PCRBr is specific to the A genome.

7.2.3 Cytogenetic Mapping of Repetitive Elements

FISH is an invaluable tool in genetic and genomic studies. It has allowed confirmation of chromosomal segment inversions (van der Knaap et al. 2004; Huang et al. 2009; Cabo et al. 2014), localization of centromeric repeats (Lee et al. 2005; Wolfgruber et al. 2009), visualization of transposons (Yu et al. 2007; Neumann et al. 2011) and repetitive elements (Lamb et al. 2007a; Macas et al. 2007; Suzuki et al. 2012), and even detection of single genes (Khrustaleva and Kik 2001; Lamb et al. 2007b) and transgenes (Santos et al. 2006; Park et al. 2010). Macas et al. (2007) demonstrated the utility of FISH to cytogenetically map the major repeats identified in the pea genome in a survey of 454 NGS sequence data. Additionally, there are some limitations in identifying these repetitive elements through computational analysis, which may not always accurately report the proportion of repeats that resides in that genome (Macas et al. 2007; Schatz et al. 2012).

With our analysis of the *B. rapa* genome, FISH data afforded us a better view of the genomic proportion of each repetitive element. Whereas about 20 % of the total repetitive elements were captured using in silico analysis, FISH generally revealed about 29 % of all the repetitive elements in the genome (Table 7.1). We consider the FISH signal likely to represent an overestimate because it only detects two-dimensional hybridization signals from the three-dimensional chromosome structure.

In *B. rapa*, CentBr1 and CentBr2 show about 85 % sequence similarity and are separately distributed to eight and two chromosome pairs, respectively (Lim et al. 2007). However, in *B. oleracea*, there is less distinct separation between

the chromosomal locations of CentBo1 and CentBo2, which show co-localization in several centromeres (Fig. 7.3g, h; Lim et al. 2007; Koo et al. 2011; Liu et al. 2014). This is consistent with there being little copy number difference between CentBo1 and CentBo2 compared to CentBr1 and CentBr2 in the 1x WGS survey (Fig. 7.2a). This also suggests that there was a different rate of homogenization of centromeric tandem repeats between *B. rapa* and *B. oleracea* genomes as well as among centromeres within each genome, as observed in some Brassicaceae species (Hall et al. 2005).

CentBr arrays are intermingled with a major centromeric LTR retrotransposon, CRB. Although CRB is common to the three basic *Brassica* lineage A, B, and C genomes, CentBr is present only in the A and C genomes (Lim et al. 2007). Additionally, the A genome-specific retrotransposon PCRBr hybridized to *B. rapa* chromosomes, but not to those of *B. oleracea* and *B. nigra* (Fig. 7.3i; Lim et al. 2007). It localized to four chromosomes with major heterochromatin blocks in *B. rapa*, which could explain the relatively high genomic proportion of PCRBr identified based on the 1x WGS survey (Fig. 7.2a, e). In addition, although Koo et al. (2011) reported three loci on three separate chromosomes for BrSTR, our data showed two loci on both arms of chromosome A05, with a major locus on the short arm, and two other very weak loci on two short chromosomes (Fig. 7.3c). This may be explained by the different sensitivity of FISH experiments, or different cytotypes used in the experiments, noting that *Brassica* genomes are highly dynamic and polymorphic (Koo et al. 2011). This was also demonstrated by Xiong and Pires (2011), who showed different numbers of 5S rDNA loci between different *B. rapa* accessions 'Chiifu' and the double haploid *B. rapa* IMB218. Taken together, the satellite repeat distribution in *B. rapa* further supports the general observation that centromeric and subtelomeric regions are havens for satellite repeats (Charlesworth et al. 1994).

Although in silico analysis identified more 45S rDNA than CentBr1, FISH showed that 45S rDNA was second to CentBr1 in terms of genomic abundance (Table 7.1). This suggests that some CentBr1 may not have been thoroughly captured despite their relative abundance; this is likely true for the other types of sequence as well considering that our analysis identified only half of the 40 % unassembled sequences.

There are more 5S and 45S rDNA loci in *B. rapa*, three and five, respectively, (Lim et al. 2005; Koo et al. 2011; Xiong and Pires 2011) compared with *B. oleracea*, which has only one and two (Liu et al. 2014). This underlies the higher genomic proportion of rDNA in *B. rapa* relative to that in *B. oleracea* (Fig. 7.2a, c).

A summary of the cytogenetic distribution of *B. rapa* repeats is presented in Fig. 7.3d. Genome composition of the eight major repeats studied in this study account for about half of the unassembled sequence based on mapping of 1x WGS reads, indicating that more repeats such as DNA transposons still remain hidden in the genome and could be further identified through a refined dnaLCW method (Table 7.2).

7.3 Functions and Evolutionary Implications of Repetitive Elements

The differential accumulation of repetitive elements, rather than gene sequences, is mainly responsible for the differences in C-value in plant genomes (Wei et al. 2013), a phenomenon commonly known as the C-value paradox (Hardman 1986; Pagel and Johnstone 1992; Macas et al. 2007). A growing amount of evidence supports the importance of these repeats in genome functions and evolution (Nowak 1994; Pardue and DeBaryshe 2003; Hall et al. 2005; Shapiro and von Sternberg 2005; Biémont and Vieira 2006; Wei et al. 2013).

Transposable elements (TE) are now known to possess characteristics that help shape the structure and evolution of genomes. They help regulate genes, defend genomes from retrotransposon proliferation and retrovirus invasion, cause

Table 7.2 Summary of different *B. rapa* and *B. oleracea* accessions used in this survey

	ID	Morphotype	Species	Sub species	Accession (cultivar)	Genome	WGS reads for repeat analysis	
							Amounts (Mbp)	Coverage (x)
1	Br-1	Chinese cabbage	*B. rapa*	ssp. *pekinensis*	Chiifu	AA	2321.4	4.4
2	Br-2	Chinese cabbage	*B. rapa*	ssp. *pekinensis*	Kenshin	AA	1498.9	2.8
3	Br-3	Chinese cabbage	*B. rapa*	ssp. *pekinensis*	DF10C062	AA	1410.9	2.7
4	Br-4	Chinese cabbage	*B. rapa*	ssp. *pekinensis*	Z16	AA	1496.2	2.8
5	Br-5	Turnip Asian	*B. rapa*	ssp. *rapifera*	Yoya	AA	1495.9	2.8
6	Br-6	Rapini-Caixin	*B. rapa*	ssp. *parachinensis*	L58	AA	1492.5	2.8
7	Br-7	Pak Choi	*B. rapa*	ssp. *chinensis*	Suzhouqing	AA	1495.3	2.8
8	Br-8	Canola	*B. rapa*	ssp. *oleifera*	R-o-18	AA	1497.8	2.8
9	Br-9	Mizuna	*B. rapa*	ssp. *nipposinica*	Mizuna	AA	1497.5	2.8
10	Br-10	Turnip Europe	*B. rapa*	ssp. *rapifera*	Manchester	AA	1484.5	2.8
11	Br-11	Canola-rapid cycling	*B. rapa*	ssp. *oleifera*	L144	AA	1496.6	2.8
12	Bo-1	Cabbage	*B. oleracea*	ssp. *capitata*	C1176	CC	1541	2.2
13	Bo-2	Cabbage	*B. oleracea*	ssp. *capitata*	C1220	CC	1606.8	2.3

1–11 WGS of *B. rapa* accessions were kindly provided by Xiaowu Wang (Key Laboratory of Horticultural Crops Genetic Improvement of Ministry of Agriculture, Sino-Dutch Joint Lab of Horticultural Genomics Technology, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China). *12–13* WGS of *B. oleracea* was generated with support of a grant from the *Golden Seed Project* (*Center for Horticultural Seed Developmenet,* ***No. 213003-04-3-SB430***), *Ministry of Agriculture, Food and Rural Affairs* (*MAFRA*)

mutations, influence recombination rates, protect chromosomes through telomerase-independent fashion, and maintain centromeres, which play a significant role in chromosome segregation (Pardue and DeBaryshe 2003; Wolfgruber et al. 2009; Biémont 2010; Sarilar et al. 2011; Goodier et al. 2012; Sampath et al. 2013). In *Brassica*, MITE transposons preferentially accumulate near or inside of genic regions indicating these likely play roles in gene evolution (Sarilar et al. 2011; Sampath et al. 2013, 2014).

Most plant centromeric DNA is composed of 150–180 bp tandem repeats and centromere-specific retrotransposons (CR; Jiang et al. 2003; Lim et al. 2007; Talbert and Henikoff 2010; Neumann et al. 2011; Jiang 2013). The centromeric tandem repeat arrays can extend to several megabases and are often interrupted by CRs, which can also insert into other CRs, forming a complex nested pattern, and play a significant role in centromere function and evolution (Jiang et al. 2003; Lim et al. 2007; Wei et al. 2013). Association of these tandem repeats and CRs with modified histone H3 (CENH3), the hallmark of active centromeres, further indicates their active role in centromere function (Neumann et al. 2011; Jiang 2013).

Some evidence has been presented to help explain the rapid evolution of centromeric tandem arrays across different centromeres within a species. Unequal crossover, gene conversion, and repeat transposition have been invoked as key players in the homogenization and spread of repeats intra-chromosomally, between sister chromatids, between homologous chromosomes, and between non-homologous chromosomes (Walsh 1987; Charlesworth et al. 1994; Cohen et al. 2003; Hall et al. 2005). Unequal crossovers

usually result in higher-order repeat units consisting of more than one type of element and variation in lengths of arrays (Hall et al. 2005; Talbert and Henikoff 2010). Other mechanisms such as gene conversion and repeat transposition may amplify satellite arrays and cause their spread into nonhomologous chromosomes (Hall et al. 2005).

In *Brassica*, CentBr and CRB are major components of the centromere (Lim et al. 2007). The CRB is a common centromeric component of the A, B, and C genomes. However, the absence of CentBr hybridization in *B. nigra* (B genome) indicates that the B genome diverged from the A and C genomes earlier, supporting the 9 MYA divergence time for the B genome (Fig. 7.3i; Lim et al. 2007; Koo et al. 2011). This was further supported by the FISH results with the subtelomeric repeat STR, which also showed genome-specific evolution. The BnSTR tandem repeat from *B. nigra* (B genome) did not hybridize to either the A or C genome, and BrSTR from the A genome did not hybridize to either the B or C genome, although BoSTR from the C genome hybridized to both the A and C genomes (Koo et al. 2011). However, those tandem repeats (CentB and STR) show high sequence similarity between species (Lim et al. 2005, 2007; Koo et al. 2011), suggesting that the tandem repeats subsequently diverged in the A, B, and C genomes after speciation even though they shared a single origin in the ancient genome.

The pericentromeric retrotransposon PCRBr showed A-genome specificity (Fig. 7.3i). PCRBr is a gypsy type retrotransposon and is accumulated in several chromosomes of *B. rapa* suggesting that these retrotransposons were rapidly amplified in the A genome after divergence from the C genome during the last 4.6 MYA (Wang et al. 2011; Liu et al. 2014). Additionally, CentBr1 and CentBr2 have diverged in sequence and chromosomal distribution in *B. rapa* and *B. oleracea*. CentBr2 has both *Hin*dIII (AAGCTT) and *Sau*3AI (GATC) restriction sites while CentBr1 has lost the *Sau*3AI site (Koo et al. 2011). This phenomenon was also observed for maize CentC and Cen4 (Kato et al. 2004). Collectively, these results highlight the dynamic nature of the genomes in the genus *Brassica* and present examples of lineage- and genome-specific rapid evolution of centromeric components (Koo et al. 2011).

7.4 Conclusion and Perspectives

As exemplified by Macas et al. (2007) in *Pisum sativum*, survey of plant genomes using low-coverage NGS data proved to be an excellent tool for capturing the highly repetitive genomic sequences that are mostly left out during assembly. Our application of this technique to *Brassica* species further corroborated the usefulness of this approach. Characterizing the genomic abundance and distribution of these repetitive sequences is further facilitated when 1x WGS genomic survey is coupled with molecular cytogenetic techniques such as FISH.

Using this approach, independent analysis of repetitive elements from genome assembly data can provide huge amount of information regarding genome structure and evolution when comparative analyses are performed with closely and distantly related species. This approach may also promote our knowledge of plants with huge genomes such as *Allium* (Jakse et al. 2008). Repetitive sequences can be analyzed using low-coverage WGS before completion of genome sequencing and can provide guidance for complete elucidation of the genome structure of the target plant. This combined genome survey and cytogenetic approach will also be useful for evolutionary genomics analysis of plant families lacking available genome sequences by allowing comparison of the repetitive yet highly informative portions of their genomes, as exemplified by our work in ginseng (*Panax ginseng*; Choi et al. 2014).

Acknowledgments This research was carried out with the support by Golden Seed Project (Center for Horticultural Seed Development, No. 213003-04-3-SB430), Ministry of Agriculture, Food and Rural Affairs (MAFRA), Ministry of Oceans and Fisheries (MOF), Rural Development Administration (RDA) and Korea Forest Service (KFS), Republic of Korea.

References

Biémont C (2010) A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics 186(4):1085–1093

Biémont C, Vieira C (2006) Genetics: junk DNA as an evolutionary force. Nature 443(7111):521–524

Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R et al (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491(7426):705–710

Cabo S, Carvalho A, Martin A, Lima-Brito J (2014) Structural rearrangements detected in newly-formed hexaploid tritordeum after three sequential FISH experiments with repetitive DNA sequences. J Genet 93(1):183–188

Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371(6494):215–220

Choi HI, Waminal NE, Park HM, Kim NH, Choi BS et al (2014) Major repeat components covering one-third of the ginseng (*Panax ginseng* C.A. Meyer) genome and evidence for allotetraploidy. Plant J 77(6):906–916

Cohen S, Yacobi K, Segal D (2003) Extrachromosomal circular DNA of tandemly repeated genomic sequences in *Drosophila*. Genome Res 13(6A):1133–1145

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284(5757):601–603

Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. Plant Biotechnol J 8(1):2–9

Goodier JL, Cheung LE, Kazazian HH Jr (2012) MOV10 RNA Helicase is a potent inhibitor of retrotransposition in cells. PLoS Genet 8(10): e1002941

Hall SE, Luo S, Hall AE, Preuss D (2005) Differential rates of local and global homogenization in centromere satellites from *Arabidopsis* relatives. Genetics 170(4):1913–1927

Hardman N (1986) Structure and function of repetitive DNA in eukaryotes. Biochem J 234(1):1–11

Harrison GE, Heslop-Harrison JS (1995) Centromeric repetitive DNA sequences in the genus *Brassica*. Theor Appl Genet 90(2):157–165

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res 16(10):1252–1261

Hershkovitz MA, Zimmer EA (1996) Conservation patterns in angiosperm rDNA ITS2 sequences. Nucl Acids Res 24(15):2857–2867

Huang S, Li R, Zhang Z, Li L, Gu X et al (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41(12):1275–1281

Jakse J, Meyer JD, Suzuki G, McCallum J, Cheung F et al (2008) Pilot sequencing of onion genomic DNA reveals fragments of transposable elements, low gene densities, and significant gene enrichment after methyl filtration. Mol Genet Genome 280(4):287–292

Jiang J (2013) Centromere evolution. In: Jiang J, Birchler JA (eds) Plant centromere biology. Wiley, Oxford, pp 159–168

Jiang J, Birchler JA, Parrott WA, Kelly Dawe R (2003) A molecular view of plant centromeres. Trends Plant Sci 8(12):570–575

Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G et al (2005) Evolution of genome size in Brassicaceae. Ann Bot 95(1):229–235

Kato A, Lamb JC, Birchler JA (2004) Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. Proc Natl Acad Sci USA 101(37):13554–13559

Khrustaleva LI, Kik C (2001) Localization of single-copy T-DNA insertion in transgenic shallots (*Allium cepa*) by using ultra-sensitive FISH with tyramide signal amplification. Plant J 25(6):699–707

Kim K, Lee SC, Lee J, Lee HO, Choi BS, Joh JH, Kim NH, Park HS, Yang TJ (2015) Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within Panax ginseng species. Plos One (in press)

Koo DH, Plaha P, Lim YP, Hur Y, Bang JW (2004) A high-resolution karyotype of *Brassica rapa* ssp. *pekinensis* revealed by pachytene analysis and multicolor fluorescence in situ hybridization. Theor Appl Genet 109(7):1346–1352

Koo DH, Hong CP, Batley J, Chung YS, Edwards D et al (2011) Rapid divergence of repetitive DNAs in *Brassica* relatives. Genomics 97(3):173–185

Lamb JC, Danilova T, Bauer MJ, Meyer JM, Holland JJ et al (2007a) Single-gene detection and karyotyping using small-target fluorescence in situ hybridization on maize somatic chromosomes. Genetics 175(3):1047–1058

Lamb JC, Meyer JM, Corcoran B, Kato A, Han F et al (2007b) Distinct chromosomal distributions of highly repetitive sequences in maize. Chrom Res 15(1):33–49

Lee HR, Zhang W, Langdon T, Jin W, Yan H et al (2005) Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in *Oryza* species. Proc Natl Acad Sci USA 102(33):11793–11798

Lim KY, Kovarik A, Matyasek R, Bezdek M, Lichtenstein CP et al (2000) Gene conversion of ribosomal DNA in *Nicotiana tabacum* is associated with undermethylated, decondensed and probably active gene units. Chromosoma 109(3):161–172

Lim KB, de Jong H, Yang TJ, Park JY, Kwon SJ et al (2005) Characterization of rDNAs and tandem repeats in the heterochromatin of *Brassicarapa*. Mol Cells 19 (3):436–444

Lim KB, Yang TJ, Hwang YJ, Kim JS, Park JY et al (2007) Characterization of the centromere and peri-centromere retrotransposons in *Brassicarapa* and their distribution in related *Brassica* species. Plant J 49(2):173–183

Liu S, Liu Y, Yang X, Tong C, Edwards D et al (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun 5. doi:10.1038/ncomms4930

Macas J, Neumann P, Navratilova A (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. BMC Genome 8:427

Martins C, Wasko AP (2004) Organization and evolution of 5S ribosomal DNA in the fish genome. In: Williams CR (ed) Focus in genome research. Nova Science Publishers, Hauppauge, pp 335–363

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11(1):31–46

Michael TP, Jackson S (2013) The first 50 plant genomes. Plant Genome 6(2)

Neumann P, Navratilova A, Koblizkova A, Kejnovsky E, Hribova E et al (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mob DNA 2(1):4

Nowak R (1994) Mining treasures from 'junk DNA'. Science 263(5147):608–610

Pagel M, Johnstone RA (1992) Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. Proc Roy Soc Lond Sr B: Biol Sci 249(1325):119–124

Pardue ML, DeBaryshe PG (2003) Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. Annu Rev Genet 37:485–511

Park HM, JeonEJ WaminalNE, ShinKS KweonSJ et al (2010) Detectioin of transgenes in three genetically modified rice lines by fluorescence in situ hybridization. Genes Genomics 32:527–531

Roa F, Guerra M (2012) Distribution of 45S rDNA sites in chromosomes of plants: structural and evolutionary implications. BMC Evol Biol 12(1):225

Sampath P, Lee SC, Lee J, Izzah NK, Choi BS et al (2013) Characterization of a new high copy Stowaway family MITE, BRAMI-1 in *Brassica* genome. BMC Plant Biol 13:56

Sampath P, Murukarthick J, Izzah NK, Lee J, Choi HI et al (2014) Genome-wide comparative analysis of 20 miniature inverted-repeat transposable element families in *Brassica rapa* and *B. oleracea*. PLoS ONE 9 (4):e94499

Santos AP, Wegel E, Allen GC, Thompson WF, Stoger E et al (2006) In situ methods to localize transgenes and transcripts in interphase nuclei: a tool for transgenic plant research. Plant Methods 2:18

Sarilar V, Marmagne A, Brabant P, Joets J, Alix K (2011) BraSto, a Stowaway MITE from *Brassica*: recently active copies preferentially accumulate in the gene space. Plant Mol Biol 77(1–2):59–75

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19(R2): R227–R240

Schatz MC, Witkowski J, McCombie WR (2012) Current challenges in de novo plant genome sequencing and assembly. Genome Biol 13(4):243

Shapiro JA, von Sternberg R (2005) Why repetitive DNA is essential to genome function. Biol Rev Camb Philos Soc 80(2):227–250

Suzuki G, Ogaki Y, Hokimoto N, Xiao L, Kikuchi-Taura A et al (2012) Random BAC FISH of monocot plants reveals differential distribution of repetitive DNA elements in small and large chromosome species. Plant Cell Rep 31(4):621–628

Swaminathan K, Varala K, Hudson ME (2007) Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. BMC Genomics 8:132

Talbert PB, Henikoff S (2010) Centromeres convert but don't cross. PLoS Biol 8(3):e1000326

van der Knaap E, Sanyal A, Jackson SA, Tanksley SD (2004) High-resolution fine mapping and fluorescence in situ hybridization analysis of sun, a locus controlling tomato fruit shape, reveals a region of the tomato genome prone to DNA rearrangements. Genetics 168 (4):2127–2140

Walsh JB (1987) Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics 115(3):553–567

Waminal N, Park HM, Ryu KB, Kim JH, Yang TJ et al (2012) Karyotype analysis of *Panax ginseng* C.A. Meyer, 1843 (Araliaceae) based on rDNA loci and DAPI band distribution. Comp Cytogenet 6(4):425–441

Waminal NE, Ryu KB, Park BR, Kim HH (2014) Phylogeny of cucurbitaceae species in korea based on 5S rDNA non-transcribed spacer. Genes Genomics 36(1):57–64

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, Huang S, Li X, Hua W, Wang J, Wang X, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park BS, Weisshaar B, Liu B, Li B, Liu B, Tong C, Song C, Duran C, Peng C, Geng C, Koh C, Lin C, Edwards D, Mu D, Shen D, Soumpourou E, Li F, Fraser F, Conant G, Lassalle G, King GJ, Bonnema G, Tang H, Wang H, Belcram H, Zhou H, Hirakawa H, Abe H, Guo H, Wang H, Jin H, Parkin IA, Batley J, Kim JS, Just J, Li J, Xu J, Deng J, Kim JA, Li J, Yu J, Meng J, Wang J, Min J, Poulain J, Wang J, Hatakeyama K, Wu K, Wang L, Fang L, Trick M, Links MG, Zhao M, Jin M, Ramchiary N, Drou N, Berkman PJ, Cai Q, Huang Q, Li R, Tabata S, Cheng S, Zhang S, Zhang S, Huang S, Sato S, Sun S, Kwon SJ, Choi SR, Lee TH, Fan W, Zhao X, Tan X, Xu X, Wang Y,

Qiu Y, Yin Y, Li Y, Du Y, Liao Y, Lim Y, Narusaka Y, Wang Y, Wang Z, Li Z, Wang Z, Xiong Z, Zhang Z, Brassica rapa C, Genome Sequencing Project (2011) The genome of the mesopolyploid crop species Brassica rapa. Nat Genet 43 (10):1035–1039

Wei L, Xiao M, An Z, Ma B, Mason AS, Qian W, Li J, Fu D (2013) New insights into nested long terminal repeat retrotransposons in Brassica species. Mol Plant 6(2):470–482

Wolfgruber T K, Sharma A, Schneider KL, Albert PS, Koo D-H, Shi J, Gao Z, Han F, Lee H, Xu R, Allison J, Birchler JA, Jiang J, Dawe RK, Presting GG (2009) Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. PLoS Genet 5(11):e1000743

Xiong ZY, Pires JC (2011) Karyotype and identification of all homoeologous chromosomes of allopolyploid Brassica napus and Its diploid progenitors. Genetics 187(1):37–49

Yu W, Lamb JC, Han F, Birchler JA (2007) Cytological visualization of DNA transposons and their transposition pattern in somatic cells of maize. Genetics 175 (1):31–39

The Common Ancestral Genome of the *Brassica* Species

8

Feng Cheng, Martin A. Lysak, Terezie Mandáková and Xiaowu Wang

Abstract

All *Brassica* species are derived from a common hexaploid ancestor, and this hexaploid ancestor has been further deduced to origin from a diploid species through a whole genome triplication event. The diploid ancestor has 7 chromosomes and resembles the karyotype of tPCK (translocation Proto-Calepineae Karyotype). The confirming evidences for the *Brassica*s' tPCK ancestor are from below three aspects: (1) The reconstructed genomic segments of the three subgenomes of all *Brassica* species keep the genomic structure of tPCK; (2) The locations of extant centromeres and the traces of paleocentromeres on the genomes of *Brassica*s support its ancestral diploid genome as tPCK; (3) The phylogeny tree and evolution analysis based on the whole genome sequences of several sequenced Brassicaceae species find that the *Brassicas* are evolved from a tPCK genome, such as the tPCK species *S. parvula*. The determination of the shared diploid ancestor for all *Brassica* species lays an important foundation for the genetic studies of *Brassica* crops.

8.1 Introduction

The genus *Brassica* contains 39 species and numerous varieties (http://www.theplantlist.org/). Many Brassicas are important crops or weeds. Among them, six species comprising the U's Triangle (Nagaharu 1935) are of economic importance because of their cultivation as vegetables, condiments, and source of oilseed. All these species are closely related and their genomes share very good synteny relationships (Panjabi et al. 2008). "Diploid" Brassicas

F. Cheng · X. Wang (✉)
Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China
e-mail: wangxiaowu@caas.cn

M.A. Lysak · T. Mandáková
CEITEC—Central European Institute of Technology, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_8

originated through one or more whole-genome triplication (WGT) events involving three highly similar ancestral genomes (Lagercrantz and Lydiate 1996; Lysak et al. 2005, 2007; Parkin et al. 2005; Wang et al. 2011b; Sharma et al. 2014).

The ancestral karyotype of *Brassica* has been under debate for more than half a century before its recent elucidation after the genome sequencing of *Brassica rapa* (Wang et al. 2011a). Many ancestral genomes with chromosome numbers ranging from $x = 3$ to 7 were proposed (e.g., Röbbelen 1960; Truco et al. 1996; Parkin et al. 2005; reviewed by Prakash and Hinata 1980). Truco et al. (1996) suggested that the diploid *Brassica* ancestor had five, six or seven chromosomes, based on the comparison of linkage maps among *B. rapa*, *Brassica oleracea,* and *Brassica nigra*. These authors favored an ancestral genome with six chromosomes (W1–W6), though the existence of a seventh chromosome (W7) could not be ruled out. Seven ancestral chromosomes were also purported by Mandáková and Lysak (2008) when they compared seven chromosomes of the consensual ancestral genome called Proto-Calepineae Karyotype (PCK), specific for $x = 7$ tribes of the crucifer Lineage II, with the structure of the *B. rapa* (AA) subgenome in *Brassica napus* (Parkin et al. 2005).

The whole genome sequencing of *B. rapa* opened new possibilities to investigate ancestral genomes of *Brassica* species (Cheng et al. 2011). In *B. rapa* ($n = 10$), genomic synteny comparison with *Arabidopsis thaliana* ($n = 5$), based on the 24 ancestral genomic blocks (GBs) of the Ancestral Crucifer Karyotype (ACK, $n = 8$; Schranz et al. 2006), identified three syntenic copies for each GB (Cheng et al. 2011). Alignment of these syntenic genomic fragments to the genome of *A. thaliana* revealed that the three syntenic subgenomes differ in the gene density and rate of gene loss (fractionation), and were classified as the least fractionated (LF), the medium fractionated (MF1), and the most fractionated (MF2) subgenomes (Wang et al. 2011b; Cheng et al. 2012b).

Based on the aforementioned results and the comparison of the three *B. rapa* subgenomes with genome sequences or genetic maps of other crucifer species, including *A. thaliana*, *Arabidopsis lyrata*, *Brassica oleracea*, *Brassica nigra*, *Caulanthus amplexicaulis*, *Raphanus sativus*, *Sinapis alba*, *Schrenkiella parvula*, and *Thellungiella salsuginea*, as well as with previously proposed Brassicaceae ancestral genomes (Schranz et al. 2006; Mandáková and Lysak 2008; Panjabi et al. 2008; Burrell et al. 2011; Dassanayake et al. 2011; Hu et al. 2011; Li et al. 2011; Nelson et al. 2011; Shirasawa et al. 2011; Wu et al. 2012), Cheng et al. (2013) draw the conclusion that the *B. rapa* genome arose via a WGT event involving three $n = 7$ genomes structurally resembling the ancestral translocation Proto-Calepineae Karyotype (tPCK). tPCK genome is an evolutionary younger variant of the PCK genome from which it differs by a reciprocal translocation (Mandáková and Lysak 2008; Cheng et al. 2013). In this chapter, we aim to summarize the evidence for the origin of *Brassica* ancestral genomes considering the following main aspects: (1) conserved associations of genomic blocks, (2) traces of ancestral centromeres, (3) phylogenetic relationships, and (4) comparisons of paleogenome evolution in other Brassicaceae taxa.

8.2 The Ancestral Diploid Karyotype of *B. rapa* Has the tPCK Genome Structure

Cheng et al. (2012b) identified the distribution of GBs in the ten chromosomes of *B. rapa*. Firstly, syntenic genes were determined between *B. rapa* and *A. thaliana* using the SynOrths tool (Cheng et al. 2012a). Then, genomic synteny relationships of 24 ancestral GBs between *A. thaliana* and *B. rapa* were obtained based on the identified syntenic genes. Finally, 71 of the purported 72 GBs (3×24 GBs) corresponding to three ancestral genomes involved in the WGT event were identified in the *B. rapa* genome; one copy

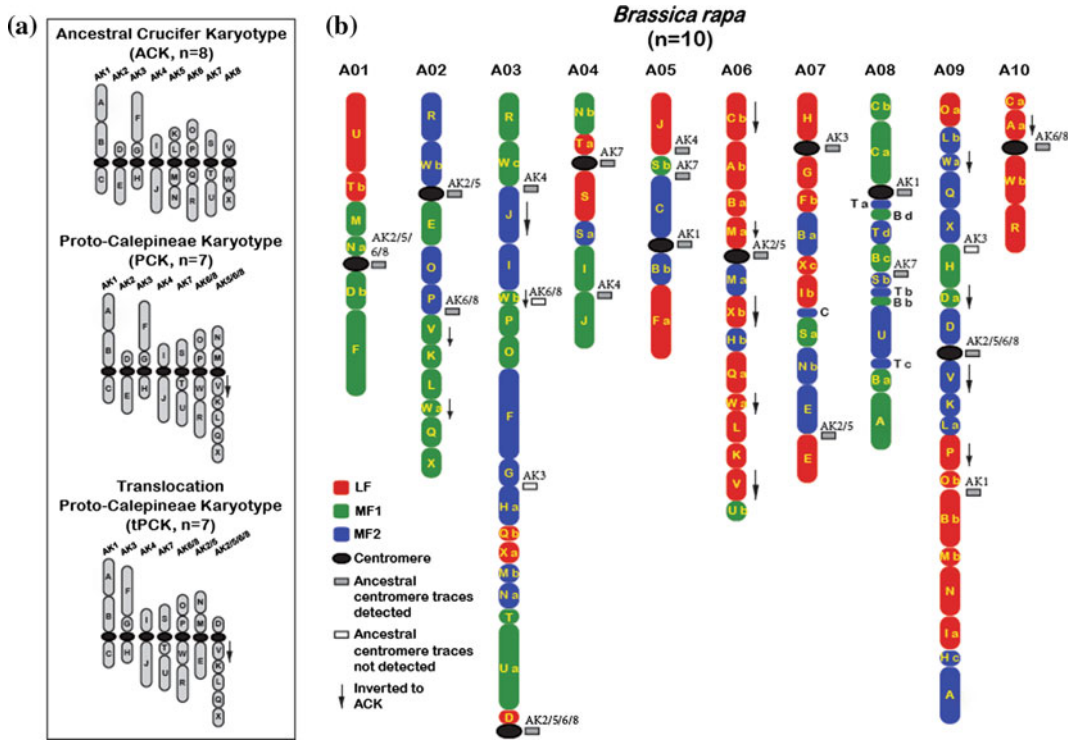

Fig. 8.1 Three diploid ancestral Brassicaceae genomes and the distribution of genomic blocks (GBs) within the 10 chromosomes of *B. rapa*. **a** Ancestral Crucifer Karyotype (ACK), Proto-Calepineae Karyotype (PCK), and translocation PCK (tPCK). Each genome comprises 24 GBs (*A–X*) (Schranz et al. 2006; Mandáková and Lysak 2008). The chromosome labels in PCK and tPCK reflect their presumed origin from more ancestral eight chromosomes of ACK. **b** The 10 *B. rapa* chromosomes (*A01–A10*) consisting of three sets of 24 GBs (71 blocks in total; one GB is missing) assigned to subgenomes LF (*red*), MF1 (*green*), and MF2 (*blue*). *Downward-pointing arrows* indicate GBs that are inverted relative to blocks with the ancestral (ACK) orientation within a single chromosome. Ancestral centromeres of three tPCK-like genomes detected or not detected in the *B. rapa* sequence data are labeled with *small rectangles*

of block G was presumably lost during genomic fractionation following WGT.

Ancestral GB associations were compared between three ancestral crucifer genomes, i.e. ACK, PCK, and tPCK (Fig. 8.1a), and the three *B. rapa* subgenomes (Fig. 8.1b). Using methods similar to playing a jigsaw puzzle, different variants of three tentative subgenomes of *B. rapa* were reconstructed by listing the syntenic genomic fragments of the *B. rapa* genome along the eight chromosomes of ACK and the seven chromosomes of PCK/tPCK. The genomic continuity was examined through checking the status of any two adjacent ancestral GBs defined as a GB association. GB associations within the three ancestral genomes were then compared with the structure of the three *B. rapa* subgenomes. Results showed that 10, 15, and 16 GB associations out of the 16, 17, and 17 ancestral GB associations within the ACK, PCK, and tPCK, respectively, were present in the *B. rapa* genome. In *B. rapa*, among 10 ACK-specific associations, four were retained in three genomic copies, and six in two copies. Of the 15 PCK-specific associations, eight were found in three copies, six in two, and one as a single copy. Lastly, of the 16 tPCK-specific block associations, eight were present in three copies, six in two copies, and two as a single genomic copy (Cheng et al. 2013). Moreover, both PCK- and tPCK-specific GB associations V/K/L/Wa/Q/X and O/P/W/R were identified in all three *B. rapa* subgenomes

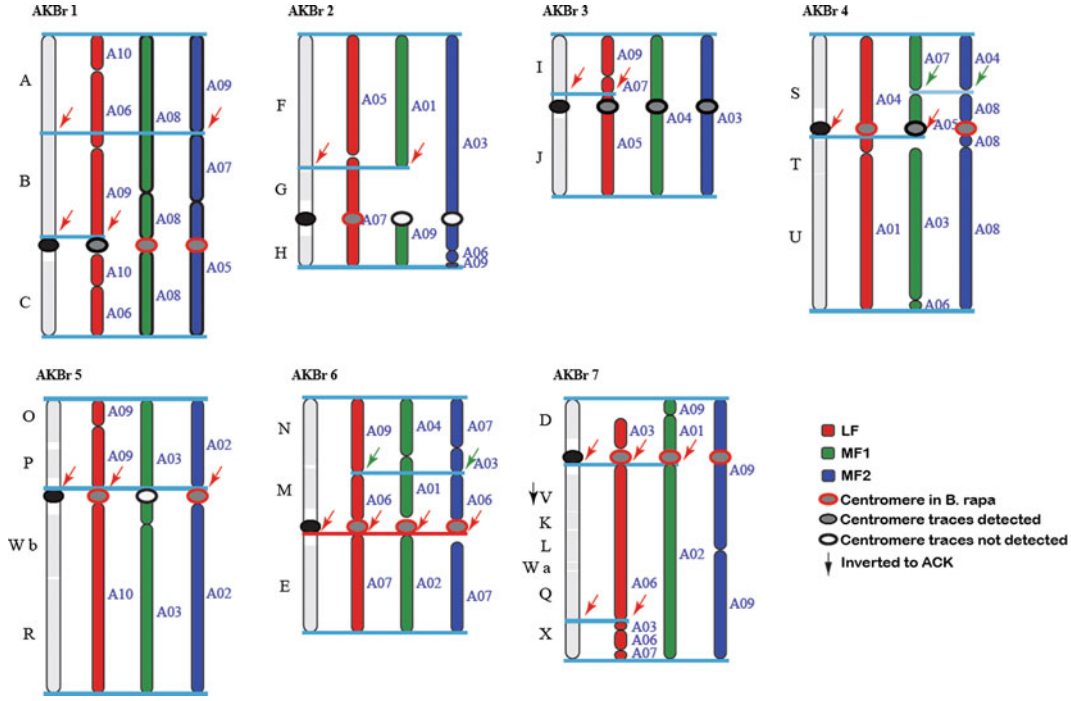

Fig. 8.2 Reconstruction of the three ancestral *B. rapa* subgenomes based on the structure of the tPCK. A tPCK-like genome was reconstructed for each of the three *B. rapa* subgenomes (LF, MF1, and MF2). Subgenome block breakpoints in the region of genomic block (GB) boundaries in tPCK are indicated by *red arrows*; *green arrows* highlight breakpoints shared by any two *B. rapa* subgenomes and not corresponding to ancestral GB boundaries. A *red horizontal line* links breakpoints shared among all three subgenomes, whereas *light blue lines* link shared breakpoints at the boundaries of two associated GBs. Centromeres and paleocentromeres traces, and theoretically purported but not detected paleocentromeres are shown as *ellipses of different colors*. AKBr is short for ancestral karyotype of Brassicas, which is the same to tPCK

(Fig. 8.1b). The three copies of the V/K/L/Wa/Q/X association are located on chromosomes A02 (MF1 genome), A06 (LF), and A09 (MF2), and further rearrangements of this association were found on chromosomes A06 and A09. Two copies of the O/P/W/R association were rearranged and located on chromosomes A02 (MF2) and A03 (MF1), while the third copy was split and located on chromosomes A09 (blocks O/P) and A10 (W/R). Altogether these data suggested that the three subgenomes of *B. rapa* have originated from either PCK or tPCK and not from ACK.

To further specify the origin of *B. rapa* subgenomes, Cheng et al. (2013) searched the *B. rapa* genome for signatures of the whole-arm reciprocal translocation differentiating tPCK genome from a more ancestral PCK genome. tPCK harbours GB associations D/V and M/E, which are absent in the PCK genome (Fig. 8.1a; Mandáková and Lysak 2008). In the MF2 subgenome, the association D/V is located on chromosome A09, whereas in the LF subgenome the same association is split between chromosomes A03 and A06 due to an inter-subgenomic translocation between block D (LF) and a part of block U (MF1) (Fig. 8.1b). The association M/E cannot be observed directly in the *B. rapa* genome. However, block associations $M_{(LF)}/M_{(MF2)}$ (A06) and $E_{(LF)}/E_{(MF2)}$ (A07) should be products of a purported inter-subgenomic translocation between two M/E associations. Evidence that the MF1 subgenome also descended from tPCK comes from the comparison between *B. rapa* and *B. nigra* genomes (Panjabi et al. 2008). The GB association R/W/M/O/P/E on chromosome B2 in

B. nigra is syntenic to the association R/W/$E_{(MF1)}$/O/P on chromosome A02 in *B. rapa*. Although the M/E association on B2 was interrupted by blocks O and P, the association of blocks M and E with blocks O, P, R, and W suggests its origin from the tPCK-specific associations O/P/W/R and M/E.

The more numerous GB associations shared between *B. rapa* and the PCK/tPCK than those between *B. rapa* and the ACK, as well as the tPCK-specific GB associations observed in the genome of *B. rapa*, suggest that the three *B. rapa* subgenomes had a close structural resemblance to the seven chromosomes of the ancestral tPCK genome (Fig. 8.2).

8.3 Paleocentromere Traces in the *B. rapa* Genome Provide Additional Support for Its Origin from the tPCK

Comparison between centromere positions in the tPCK genome and the extant genome of *B. rapa* indicated that all the ten *B. rapa* centromeres were inherited from 21 paleocentromeres of the hexaploid ancestral genome (Cheng et al. 2013). Using the *B. rapa* genome sequence, the fate of the 21 paleocentromeres corresponding to the three ancestral tPCK-like genomes has been analyzed in detail. The centromere-specific sequences were aligned to the *B. rapa* genome to determine the candidate regions corresponding to paleocentromeres. In total, 18 paleocentromere regions were detected, while three ($AK3_{(MF1)}$, $AK3_{(MF2)}$, and $AK6/8_{(MF1)}$) cannot be identified (Fig. 8.1b).

Among the 18 paleocentromere regions, 10 correspond to the 10 extant *B. rapa* centromeres. Centromeres on chromosomes A04, A05, A07, and A09 correspond to ancestral centromeres of AK7, AK1, AK3, and AK2/5/6/8, respectively. The centromeres of A01, A03, and A10 were derived via a translocation or pericentric inversion event from paleocentromeres of chromosomes AK2/5/6/8, AK2/5/6/8, and AK6/8, respectively. The origin of centromeres on A02, A06, and A08 is linked to the reduction of chromosome number in *B. rapa*. Here translocation events with breakpoints within the (peri) centromeric regions of two different ancestral chromosomes resulted in the loss of one of the two paleocentromeres. Of the eight remaining paleocentromeres, six were disrupted by translocations, while the last two belonged to entirely conserved ancestral chromosomes within *B. rapa* chromosomes A03 and A04 whose centromeres were inactivated or deleted (Figs. 8.1b and 8.2).

The distribution pattern of transposable elements (TEs) in the *B. rapa* genome also supports its origin from the tPCK genome. It is well documented that TEs are often enriched around centromere regions, which has been observed in genomes of many species such as *A. thaliana*, maize, and soybean (Initiative 2000; Schnable et al. 2009; Schmutz et al. 2010), and is true for the 21 tPCK-derived paleocentromeres in the present-day genome of *B. rapa*. Taking the three reconstructed subgenomes as a basis (Figs. 8.1a and 8.2), it is easy to calculate and plot TE densities for the 21 ancestral subgenomes. The results revealed that TE sequences continue to show a relatively high density around positions of the 21 paleocentromeres in *B. rapa*. This variation pattern is reflecting not only the position of the 10 extant *B. rapa* centromeres, but also the location of 11 lost paleocentromeres (Fig. 8.3).

8.4 Phylogeny Suggests a Shared Ancestry of the *Brassica* Mesohexaploid Genome and Extant tPCK-like Genomes

The sequenced genome of *S. parvula* ($n = 7$), which has the tPCK-like structure provides a chance to compare *B. rapa* sequences with an extant tPCK-like genome. K_s values (synonymous mutation rate per synonymous mutation site) represent the mutation frequency of nuclear sites that are selection-free. As the mutation frequency of nuclear sites under neutral selection (selection-free) is relatively stable and has a

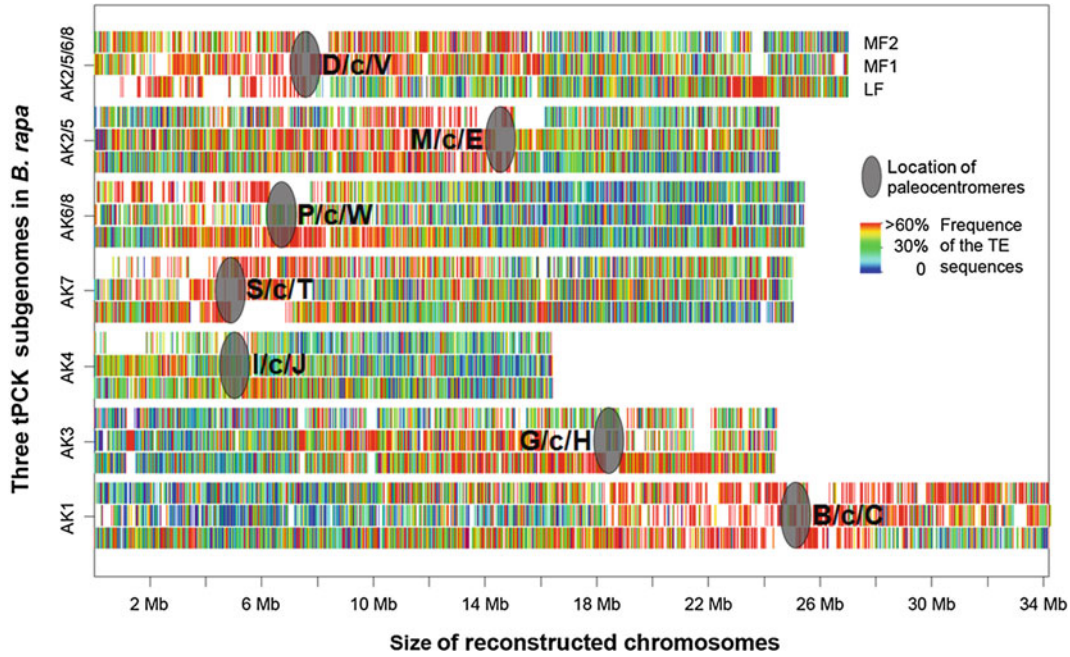

Fig. 8.3 Distribution of transposable elements (TE) supports the positions of the 21 paleocentromeres of the three tPCK-like subgenomes in the *B. rapa* genome. The color of each bin represents the ratio of TE sequences to the total sequences of the flanking region of a given gene used to reconstruct the tPCK subgenomes. Each paleocentromere is displayed as an association of two centromere-flanking genomic blocks (c = centromeric region). Modified after Cheng et al. (2014)

positive relationship to elapsed evolution time, K_s values are used for divergence time estimates. Because syntenic genes are homeologues inherited from a common ancestor, K_s values of syntenic genes serve as a proxy for divergence time of two species from their common ancestor.

By calculating K_s values among syntenic genes of *B. rapa, A. thaliana*, *A. lyrata*, *S. parvula*, and *Thellungiella salsuginea* (Initiative 2000; Dassanayake et al. 2011; Hu et al. 2011; Wang et al. 2011b; Wu et al. 2012), the evolutionary relationships among these species were elucidated (Cheng et al. 2013). Results showed that *B. rapa* is phylogenetically closer to *S. parvula* than to other analyzed genomes. K_s for *B. rapa* relative to *S. parvula* was the smallest one ($\sim$0.29–0.31; $\sim$10 million years of divergence; Koch et al. 2000), while *B. rapa* and *T. salsuginea* diverged $\sim$11.7 million years ago (Mya) (K_s $\sim$0.34–0.36), *B. rapa* and *A. lyrata* $\sim$14.3 Mya (K_s $\sim$0.41–0.45), and *B. rapa* and *A. thaliana* diverged $\sim$14.5 Mya (K_s $\sim$0.42–0.45). Furthermore, K_s analysis also revealed that the divergence times between *S. parvula* and each of the three *B. rapa* subgenomes are very close to the divergence time between the *B. rapa* subgenomes (i.e., the occurence of the WGT event). K_s values between each of the three *B. rapa* subgenomes and *S. parvula* range from 0.29 to 0.31, while those for each pair of *B. rapa* subgenomes range from 0.29 to 0.33 (i.e., $\sim$10.3 Mya). The phylogenetic tree built using the K_s loci of the above five species showed that *Brassica* and *Arabidopsis* diverged prior to the *Brassica* (Brassiceae)-*Thellungiella* split. Furthermore, the Brassiceae-specific WGT occurred near the divergence time between Brassiceae progenitors and the ancestor of *Schrenkiella*.

8.5 All Brassica Crops and Some Other Brassicaceae Species Have Evolved from the tPCK-like Genome

Brassica crops and other species from the tribe Brassiceae are close relatives forming a monophyletic group sharing a common hexaploid ancestor (Lysak et al. 2005). Once the diploid ancestors of *B. rapa* were determined, it is relatively straight-forward to analyze whether these ancestral genomes were shared also by other *Brassica* crop species and other Brassiceae/Brassicaceae species. The simplest way how to achieve this goal is to compare tPCK-specific signatures, GB associations V/K/L/Wa/Q/X, O/P/W/R, M/E and D/V, with available genome sequences or genetic maps of other species. Following Cheng et al. (2013), we provide evidence supporting the tPCK genome as an ancestral genome of several other diploid and mesopolyploid Brassicaceae species, such as *B. oleracea*, *B. nigra*, *R. sativus*, *S. alba*, *C. amplexicaulis*, *S. parvula* and *T. salsuginea*.

In the genome of *B. oleracea* (CC), three genomic copies of the block associations V/K/L/Wa/Q/X and O/P/W/R, and one copy of the D/V association were observed (Liu et al. 2014). Similar to the *B. rapa* genome, one copy of the association M/E is present in *B. oleracea*, two other M/E copies have been reshuffled and cannot be observed directly within the *B. oleracea* genome. For the *B. nigra* genome (BB), the genetic map of the allotetraploid *B. juncea* was used for genomic structure analyses. Two copies of V/K/L/(Wa)/Q/X were found on chromosomes B1 and B6 in *B. nigra*, with the third copy being split between B4 and B8. Two copies of O/P/W/R were found on B2 and B3, while the third copy was fractionated on B8 (Panjabi et al. 2008). Two copies of the D/V association are located on B1 and B6, while one copy of M/E was rearranged with O/P/W/R on B2 (Panjabi et al. 2008). These analyses together with the detail analysis of the *B. rapa* genome support the conclusion that three "diploid" *Brassica* genomes (AA, BB and CC) have evolved from a common or structurally very similar hexaploid ancestor(s) harboring three tPCK-like ancestral genomes.

An EST linkage map of radish (*R. sativus*) was used (Shirasawa et al. 2011) to dissect the ancestral structures of this genome (Cheng et al. 2013). Three copies of the two tPCK-specific block associations V/K/L/(Wa)/Q/X and O/P/W/R were observed on linkage groups of radish: the first block association was located on LG2, LG4, and LG6, while the latter one was located on LG5, LG6, and LG8, respectively. Furthermore, two copies of tPCK-specific association D/V were found to be linked to the association V/K/L/(Wa)/Q/X on LG2 and LG6, and one copy of O/P/W/R was fractionated with another tPCK-specific association (M/E) on LG5. Altogether these data suggest that three tPCK-like genomes were most likely involved in the origin of the hexaploid ancestor of the radish genome.

Comparative genomic analysis in white mustard (*S. alba*; (Nelson et al. 2011) and *C. amplexicaulis* (Burrell et al. 2011) based on genetic linkage mapping also suggested that these genomes descended from tPCK-like ancestral genomes. In the *S. alba* genome, the V/K/L/V/Q association was fractionated on chromosomes S10 and S11, while the O/P/W/R was fractionated on S09. The M/E association is located on S02, while the D/V association was not found (Nelson et al. 2011). In *Caulanthus*, the GB association D/V/K/L/(Wa)/Q/X was fragmented on LG3. The O/P/W/R association was observed on LG2 (with blocks P and R fragmented), but the M/E association was not observed (Burrell et al. 2011).

In species that still maintain the tPCK genome structure, such as the halophytes *S. parvula* and *T. salsuginea* (Dassanayake et al. 2011; Wu et al. 2012), their origin from a tPCK ancestral genome is self-evident.

These analyses show that tPCK and its more ancestral variant PCK are important ancestral genomes of several Brassicaceae clades, particularly from so-called Lineage II including economically important tribe Brassiceae (Franzke et al. 2010).

8.6 Conclusion

Based on the analyses of the *B. rapa* genome structure, multiple genome comparisons, signatures of paleocentromeres, as well as the Ks-based phylogenies, the diploid ancestors of *B. rapa* before the WGT were shown to have seven chromosomes and to structurally resemble the ancestral tPCK genome. Furthermore, the tPCK-like ancestor was found to be presumably a common to all *Brassica* species and some other Brassicaceae groups. The determination of their ancestors resolves the long-lasting dispute on the origin of mesopolyploid *Brassica* genomes, and paved the way toward reconstructing the evolutionary processes shaping genomes of all *Brassica* and Brassiceae crops.

References

Burrell AM, Taylor KG, Williams RJ, Cantrell RT, Menz MA, Pepper AE (2011) A comparative genomic map for *Caulanthus amplexicaulis* and related species (Brassicaceae). Mol Ecol 20:784–798

Cheng F, Liu S, Wu J, Fang L, Sun S et al (2011) BRAD, the genetics and genomics database for Brassica plants. BMC Plant Biol 11:136

Cheng F, Wu J, Fang L, Wang X (2012a) Syntenic gene analysis between *Brassica rapa* and other Brassicaceae species. Front Plant Sci 3:198

Cheng F, Wu J, Fang L, Sun S, Liu B et al (2012b) Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. PLoS One 7:e36442

Cheng F, Mandáková T, Wu J, Xie Q, Lysak MA, Wang X (2013) Deciphering the diploid ancestral genome of the Mesohexaploid *Brassica rapa*. Plant Cell 25:1541–1554

Cheng F, Wu J, Wang X (2014) Genome triplication drove the diversification of Brassica plants. Hortic Res 1:14024

Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H et al (2011) The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet 43:913–918

Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, Mummenhoff K (2010) Cabbage family affairs: the evolutionary history of Brassicaceae. Trends Plant Sci 16:108–116

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF et al (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet 43:476–481

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

Koch MA, Haubold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). Mol Biol Evol 17:1483–1498

Lagercrantz U, Lydiate DJ (1996) Comparative genome mapping in Brassica. Genetics 144:1903–1910

Li F, Hasegawa Y, Saito M, Shirasawa S, Fukushima A et al (2011) Extensive chromosome homoeology among Brassiceae species were revealed by comparative genetic mapping with high-density EST-based SNP markers in radish (*Raphanus sativus* L.). DNA Res 18:401–411

Liu S, Liu Y, Yang X, Tong C, Edwards D et al (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun (in press)

Lysak MA, Koch MA, Pecinka A, Schubert I (2005) Chromosome triplication found across the tribe Brassiceae. Genome Res 15:516–525

Lysak MA, Cheung K, Kitschke M, Bures P (2007) Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. Plant Physiol 145:402–410

Mandáková T, Lysak MA (2008) Chromosomal phylogeny and karyotype evolution in x = 7 crucifer species (Brassicaceae). Plant Cell 20:2559–2570

Nagaharu U (1935) Genome analysis in Brassica with special reference to the experimental formation of *B. napus* and peculiar mode of fertilication. Jpn J Bot 7:389–452

Nelson MN, Parkin IA, Lydiate DJ (2011) The mosaic of ancestral karyotype blocks in the *Sinapis alba* L. genome. Genome 54:33–41

Panjabi P, Jagannath A, Bisht NC, Padmaja KL, Sharma S et al (2008) Comparative mapping of *Brassica juncea* and *Arabidopsis thaliana* using Intron Polymorphism (IP) markers: homoeologous relationships, diversification and evolution of the A, B and C Brassica genomes. BMC Genom 9:113

Parkin IA, Gulden SM, Sharpe AG, Lukens L, Trick M et al (2005) Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. Genetics 171:765–781

Prakash S, Hinata K (1980) Taxonomy, cytogenetics and origin of crop Brassicas, a review. Oper Bot 55:1–57

Röbbelen G (1960) Beitra ge zur Analyse des Brassica-Genoms. Chromosoma 11:205–228

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Schranz ME, Lysak MA, Mitchell-Olds T (2006) The ABC's of comparative genomics in the Brassicaceae:

building blocks of crucifer genomes. Trends Plant Sci 11:535–542

Sharma S, Padmaja KL, Gupta V, Paritosh K, Pradhan AK et al (2014) Two plastid DNA lineages–Rapa/Oleracea and Nigra–within the tribe Brassiceae can be best explained by reciprocal crosses at hexaploidy: evidence from divergence times of the plastid genomes and R-block genes of the A and B genomes of *Brassica juncea*. PLoS One 9:e93260

Shirasawa K, Oyama M, Hirakawa H, Sato S, Tabata S et al (2011) An EST-SSR linkage map of *Raphanus sativus* and comparative genomics of the Brassicaceae. DNA Res 18:221–232

Truco MJ, Hu J, Sadowski J, Quiros CF (1996) Inter- and infra-genomic homology of the Brassica genomes: implications for their origin and evolution. Theor Appl Genet 93:1225–1233

Wang X, Wang H, Wang J, Sun R, Wu J et al (2011a) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Wang X, Wang H, Wang J, Sun R, Wu J et al (2011b) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Wu HJ, Zhang Z, Wang JY, Oh DH, Dassanayake M et al (2012) Insights into salt tolerance from the genome of *Thellungiella salsuginea*. Proc Natl Acad Sci USA 109:12219–12224

Genome Evolution after Whole Genome Triplication: the Subgenome Dominance in *Brassica rapa*

9

Feng Cheng, Jian Wu, Bo Liu and Xiaowu Wang

Abstract

Subgenome dominance is widely existed in plant species that experienced allopolyploidization. Subgenome dominance represents a series of biased pehnomenons as that one subgenome retains more genes, more dominantly expressed genes, less functional mutations etc., over the other subgenomes. *Brassica rapa*, which experienced a whole genome triplication (WGT) event ~11 million years ago, exhibits significant subgenome dominance, with the LF (the least fractionated) subgenome retains ~1.5 times more genes in average than the other two subgenomes MF1 and MF2 (more fractionated 1 and 2). Furthermore, paralogous genes in LF are always expressed to higher levels and accumulated less functional mutations than that located at MF1 and MF2. Further research found that small RNA mediated methylation of transposons that distributed at genes' flanking regions plays an important role in the formation of subgenome dominance. Finally, based on these findings, a two-step process was proposed to illustrate the WGT event in *B. rapa*.

9.1 Introduction

Whole genome duplication (WGD) or polyploidization is a common feature that is widely spread among the plant species, and many even experience several rounds of WGD. WGD occurred frequently in the evolutionary history of plants; it improved the tolerance of plants to mutations and provided abundant genetic material for the evolution of new features. This allowed plants to survive challenging external factors, such as environmental fluctuations and/or habitat changes. *Brassica rapa* is a crop species that evolved through rounds of polyploidization, such as the γ triplication, and α and β duplications. These three polyploidization events are shared by *B. rapa* and the model plant *Arabidopsis thaliana*, and most of the species in Brassicaceae. In addition to the three rounds of

F. Cheng · J. Wu · B. Liu · X. Wang (✉)
Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences,
Beijing 100081, China
e-mail: wangxiaowu@caas.cn

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_9

polyploidization, *B. rapa* also experienced an extra whole genome triplication (WGT) event that is shared by all *Brassica* crops.

The WGT of *B. rapa* serves as a good model to study the subgenome's evolution in polyploids. Polyploidization plays important roles in the evolution of plant species. However, the mechanism of subsequent evolution following polyploidization has not been sufficiently investigated. The lack of whole genome sequences for appropriate polyploid species has restrained related progress. The whole genome sequencing of *B. rapa* offered a good chance to investigate genome evolution following polyploidization. The WGT event in *B. rapa* occurred approximately 11 million years ago, which is long enough for the three subgenomes generated from the WGT to become well fractionated and differentiated. However, the event is young enough that the three subgenomes still can be separated unambiguously. The genomic differentiation and reshuffling has not fractionated the three subgenomes too heavily to be recognized, and most genes are clearly identifiable in the outgroup, *A. thaliana*. Owing to the genome sequences of *B. rapa*, the subgenome dominance phenomenon was observed and the mechanism regulating the phenomenon investigated. Based on this information, a two-step theory of polyploidization was hypothesized to explain the process of WGT that occurred in *B. rapa*.

This chapter will introduce studies of the structural and functional evolution of the *B. rapa* genome following WGT, mainly focusing on the subgenome dominance phenomenon. Subgenome dominance is believed to occur in allotetraploids (Garsmeur et al. 2014). The phenomenon is an integration of a set of bias features among subgenomes produced by a polyploidization event: (1) biased fractionation: one subgenome retains more genes, while others retain less; (2) dominant expression: one subgenome has more genes expressed at higher levels than their paralogs in other subgenomes; (3) biased mutation: genes from one subgenome tend to be more resistant to functional mutations and accumulate fewer such mutations than genes from other subgenomes. There is evidence that transposable elements (TEs) play an important role in the formation of subgenome dominance.

9.2 Reconstruction of the Three Subgenomes in *B. rapa*

To investigate the subgenome evolution in *B. rapa*, the accurate partitioning of the three subgenomes is the first and also the most important step. By multiple genome comparisons, researchers found that genomes of species in Brassicaceae can be divided into 24 genomic blocks (GBs; labeled from A to X) (Parkin et al. 2005; Schranz et al. 2006; Cheng et al. 2013). The genome that has eight chromosomes and one set of 24 GBs was suggested to be the ancestral common karyotype (ACK) of family Brassicaceae. The examples for extant species that keep the ACK genome structure are *Arabidopsis lyrata* and *Capsella rubella* (Hu et al. 2011; Slotte et al. 2013). It is the rearrangement sometimes accompanied with WGDs that gave birth to all the Brassicaceae species. The genomic fragments corresponding to these 24 GBs were defined in the genome of *A. thaliana*, using two genes to denote the boundaries of each GB. It is a useful resource for comparative genomic analyses in *Brassicaceae*.

The genome of *A. thaliana* was used as the reference to determine the distributions of GBs in the genome of *B. rapa*. Since *B. rapa* experienced an extra round of WGT compared with the diploid species in Brassicaceae as ACK, it should have three copies of 24 GBs, i.e., 72 GBs in total. Since the triplication event that occurred in the early stage of *Brassica* origin is much more recent than that of *Carica papaya* or *Vitis vinifera* ($\sim$80 Mya) or the most recent tetraploidy (α duplication) for *A. thaliana* lineage, the syntenic genes in *A. thaliana* and *B. rapa* can be determined at a high accuracy (Cheng et al. 2012a). In total, 7813, 5439, and 1675 genes were determined to have 1, 2, and 3 syntenic copies, respectively, of *B. rapa* genes in *A. thaliana*. The GB information in *A. thaliana* can be easily transferred to *B. rapa* based on these syntenic

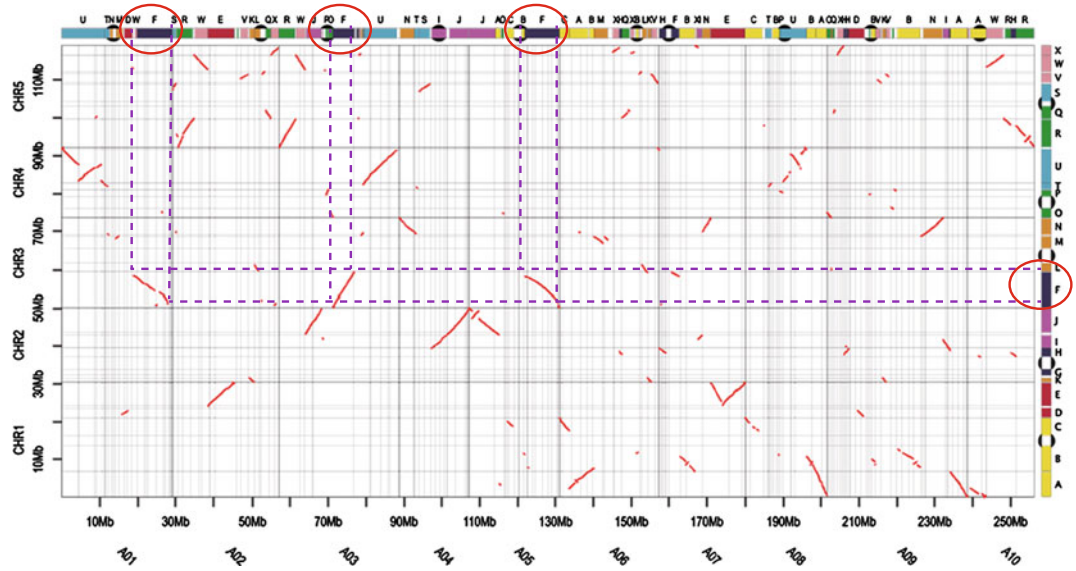

Fig. 9.1 Syntenic gene pairs in the genomes of *Arabidopsis thaliana* and *Brassica rapa* were used as anchors to transfer the genomic blocks information from the genome of *A. thaliana* to *B. rapa*. Block F was used as an example (*purple dashed lines*) to show how the positions of the three copies of F block were determined in the genome of *B. rapa*. Revised from figure in Wang et al. (2011)

gene pairs (Fig. 9.1), thus determining the distribution, as well as the fractionation information for the 72 GBs (actually 71 GBs detected, one copy of G block completely lost) across 10 chromosomes of *B. rapa*.

The three subgenomes can be accurately separated by comparing syntenic fragments of *B. rapa* with the diploid ancestral genomes. The continuous syntenic fragments of *B. rapa* to *A. thaliana* can be selected and ordered along the seven chromosomes of the diploid ancestor translocation Proto-Calepineae Karyotype (tPCK) (Cheng et al. 2013) or any other diploid genome. The three copies of syntenic fragments in *B. rapa* have different breakpoints compared with their ancestral tPCK genome; thus it is quite straight forward to place them along the chromosomes of tPCK, just like placing jigsaw puzzle pieces (Supplementary Fig. 1 in Cheng et al. 2014). After all these syntenic fragments were correctly placed, for each of the seven ancestral chromosomes in tPCK, we should obtain three copies corresponding to the three subgenomes in *B. rapa*.

9.3 Biased Gene Fractionation Among the Three Subgenomes of *B. rapa*

After WGD, subgenomes that coexist in one nucleus are always differentiated. This differentiation is easily detected by comparing the gene density among subgenomes generated from WGD (Thomas et al. 2006; Schnable et al. 2011). This bias in gene densities among subgenomes is called as biased gene fractionation. It has been observed in *A. thaliana* and maize (Thomas et al. 2006; Schnable et al. 2009; Woodhouse et al. 2010), and may be a common feature in species with ancient polyploid genomes (Sankoff et al. 2010).

Biased gene fractionation was observed among the three subgenomes of *B. rapa*. After the reconstruction of the three copies of chromosomes for all seven chromosomes of the ancestral genome tPCK, the gene loads for each chromosome copy could be easily investigated. For each of the seven chromosomes of tPCK,

Table 9.1 Numbers of dominantly expressed genes in the subgenomes LF, MF1, and MF2 of *Brassica rapa*, revised from table of Cheng et al. (2012b)

Organism/Accession	Twofold dominance		
	LF	MF1 (LF/MF1)	MF2 (LF/MF2)
Leaf	393	262 (1.50)	233 (1.69)
Stem	362	258 (1.40)	228 (1.59)
Root	356	273 (1.30)	221 (1.61)
Chiifu	363	253 (1.43)	216 (1.68)
L58	355	229 (1.55)	194 (1.83)

these is always one copy of a reconstructed chromosome from the *B. rapa* genome that has significantly more genes than the other two copies (~1.5-fold higher in gene density). While the other two copies of chromosomes with fewer genes contained slightly different numbers of genes. The significant difference in gene content cannot occur accidently at the chromosome level; thus it represents the real gene variation among the three subgenomes. Therefore, one set of seven chromosomes with the highest gene density for each were grouped and called the LF subgenome. Another set of seven chromosomes that had moderate numbers of genes were grouped to be the more fractionated subgenome 1 (MF1), while the remaining copies of the seven chromosomes that had the least genes represented the more fractionated subgenome 2 (MF2).

9.4 Dominant Gene Expression Among Subgenomes of *B. rapa*

Besides the variation of gene fractionation, paralogous genes in different subgenomes also show variant levels of expression. The subgenome that retains more genes is the one that has more genes expressed at higher levels than their paralogs in the other subgenomes. This phenomenon is called genome dominance. It has been observed in many genomes with recent polyploidization, such as maize (Schnable et al. 2011), the allotetraploids of *A. thaliana* and *A. arenosa* (Wang et al. 2006), and the natural allotetraploid *Tragopogon miscellus* (Buggs et al. 2010), as well as an allotetraploid cotton species (Senchina et al. 2003).

Genome dominance was observed in *B. rapa* across many accessions and by means of different statistical methods. Several whole genome transcriptome sequencing datasets were used to confirm the genome dominance phenomenon among subgenomes of *B. rapa*, such as mRNA-Seq data for different organs and different accessions or varieties of *B. rapa* (Wang et al. 2011, 2012; Cheng et al. 2012b). Two methods was used to evaluate the dominance expression among paralogs extracted from a confident fully retained homologs set (Cheng et al. 2012b). The first one is the twofold rule, in which only when a gene is expressed at least twofold higher than the other two homologs, is it considered to be dominantly expressed. Using this method, genome dominance was observed from all the available expression datasets. The subgenome LF, which has the most retained genes, always has more dominantly expressed genes (1.30–1.83 times) than that of subgenomes MF1 and MF2 (Table 9.1). Subgenome MF1 has slightly more dominantly expressed genes than subgenome MF2. The second method is the horse race experiment; a gene winning by any fraction of expression was considered the dominantly expressed one. Similar patterns of genome dominance were observed under the horse race experiment compared with that of the twofold rule.

The more strictly the method was applied, the more significant the observed genome dominance was among subgenomes of *B. rapa*. When determining the dominantly expressed genes using a more than twofold higher rule, such as a

three or fivefold rule, more significant genome dominance was observed. For instance, the subgenome LF will have relatively more dominant genes than that of MF1 and MF2. Furthermore, the median difference in syntenic gene pairs in which LF expressed at a higher level was marginally higher than the median difference for the pairs in which either MF1 or MF2 expressed at a higher level. These results supported the hypothesis that subgenome dominance effects should be stable and are controlled by factors in the genome of *B. rapa*.

9.5 Biased Distribution of Functional Mutations Among Subgenomes of *B. rapa*

Another difference among subgenomes of polyploids is the variation of functional mutations. The dominant subgenome, which retains more genes and has more genes expressed at higher levels, is also the subgenome whose genes have better defenses against functional mutations. This phenomenon was observed in maize (Schnable et al. 2011) and is also true in the genome of *B. rapa* (Cheng et al. 2012b).

Genes in subgenome LF accumulated fewer functional mutations than those of subgenomes MF1 and MF2. Biased distributions of functional mutations among subgenomes of *B. rapa* could also be considered as ongoing biased gene fractionation because many functional mutations resulted in pseudogenes. Using the resequencing data from different accessions of *B. rapa*, it was found that genes in the subgenome LF accumulate significantly fewer nonsynonymous single nucleotide polymorphisms (SNPs) and frameshift InDels than the two MF subgenomes. Take *B. rapa* strains L144, a rapid cycling laboratory accession, and VT117, a vegetable turnip accession, as examples (Cheng et al. 2012b), the former contains 561,367 SNPs and 45,995 InDels, and the latter contains 562,935 SNPs and 60,003 InDels in the 23,716 confident genes determined by their resequencing data. Among them, genes in subgenome LF always had fewer nonsynonymous SNPs (4574 and 4103 for L144 and VT117, respectively) and frameshift InDel

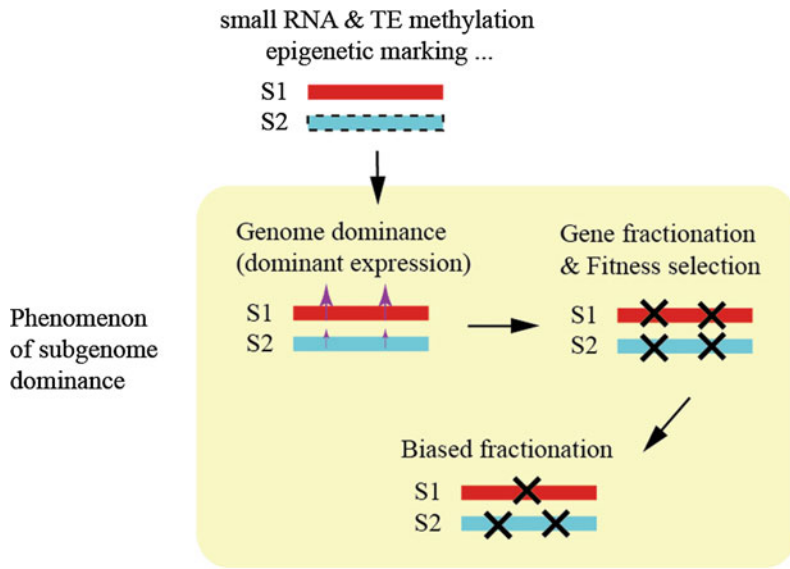

Fig. 9.2 Proposed formation of subgenome dominance in the genome of *Brassica rapa*. Phenomenon of dominant expression, ongoing biased functional mutation, and biased gene fractionation observed in *B. rapa* were united by the aim of improving plant fitness. Meanwhile, the gene dominant expression is likely to be regulated by the small RNA mediated methylation of transposable elements

(52 and 60 for L144 and VT117, respectively) mutations than the genes in the MFs (4834 and 4866 SNPs in the MF1 of L144 and turnip, respectively, and 4620 and 4530 SNPs in the MF2 of L144 and VT117, respectively; 72 and 77 InDels in the MF1 of L144 and VT117, respectively, and 73 and 83 InDels in the MF2 of L144 and VT117, respectively) of both L144 and VT117.

The three features of subgenome dominance —biased gene fractionation, dominant expression of genes and biased functional mutations— in *B. rapa* can be united in improving the fitness of plants. In this explanation system (Fig. 9.2), genes that are expressed to higher levels than their paralogs are more important for the biological activity of the plant. Thus, functional mutations of these genes would be more significant in reducing the plant's fitness than mutations of their syntenic paralogs. Therefore, natural selection conserves these dominantly expressed genes against functional mutations, whereas their paralogs accumulate more mutations and eventually fractionated, resulting in a higher gene density in the dominant subgenome and lower gene density in the dominated subgenomes. This explanation was first suggested in the maize genome and subsequently in the genome of *B. rapa* (Schnable et al. 2011, 2012; Cheng et al. 2012b). However, this explanation of the subgenome dominance phenomenon still leaves unanswered questions: what element controls the biased distribution of dominantly expressed genes among subgenomes?

9.6 Small RNA-Mediated Methylation of TEs Regulates Genome Dominance

Epigenetic modifications of TEs play important roles in regulating gene expression. In *A. thaliana*, Hollister and Gaut (2009), Hollister et al. (2011) found that the methylated TEs could suppress the expression of nearby genes. They suggested that there was a dynamic balance between gene expression and the activity of neighboring TEs in plants (Hollister and Gaut 2009). The activation of TEs reduces the stability of the plant genome, which is harmful, while the methylation of TEs will inactive them, which is beneficial to the plant. However, the methylated TEs will also suppress nearby gene expression, leading to the reduced fitness of the host. Thus, there should be a trade-off between the methylation of TEs and gene expression. By analyzing the small RNA data of *B. rapa*, it was revealed that small RNAs also play an important role in the formation of subgenome dominance in *B. rapa*.

Small RNAs regulate the dominance expression among the subgenomes of *B. rapa* through the methylation of TEs in the flanking regions of genes (Woodhouse et al. 2014). Based on the analysis of small RNA-Seq data, it was discovered that 24 bp small RNAs were mapped primarily to TE sequences located at the 5′ and 3′ regions of genes and more small RNAs mapped to genes located in the MF1 and MF2 subgenomes than in the LF. The biased targeting of small RNAs to TE sequences was much more significant when comparing dominantly expressed genes with their paralogs. These data suggest that small RNA-targeted TEs play an important role in the formation of subgenome dominance (Fig. 9.2). It is likely that the 24 bp small RNA-mediated TE methylation suppressed the expression of nearby genes, and its biased distribution in the subgenomes of *B. rapa* then led to subsequent subgenome dominance.

9.7 Theory of Two-Step Polyploidization and Its Relationship to Subgenome Dominance

A two-step theory was hypothesized to explain the polyploidization process of WGT that occurred in the ancestor of *B. rapa* (Wang et al. 2011; Cheng et al. 2012b; Tang et al. 2012). In the first step, the two tPCK genomes (precursors of MF1 and MF2) merged. A first round of genomic reshuffling and gene fractionation gave

birth to a new diploid (consisting of subgenomes MF1 and MF2). Since there is no significant genome dominance detected between MF1 and MF2 in the current genome of *B. rapa*, autotetraploidization cannot be excluded as a possible process for the first tetraploidization. However, considering that more small deletions occurred recently in the MF1subgenome than MF2 (Tang et al. 2012), allotetraploidization is favored as the first duplication event. In the second step, another tPCK genome (LF) was added to the fractionated diploid genome (MF1 and MF2). Then, a second round of genomic shuffling and gene fractionation resulted in the mesohexaploid ancestor of *B. rapa*. The "two" merged genomes (LF and MFs) are different, indicating that the second step was a process of alloploidization, resulting in the biased gene fractionation and dominant gene expression phenomenon.

References

Buggs RJ, Chamala S, Wu W, Gao L, May GD et al (2010) Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. Mol Ecol 19(Suppl1):132–146

Cheng F, Wu J, Fang L, Wang X (2012a) Syntenic gene analysis between Brassica rapa and other Brassicaceae species. Front Plant Sci 3:198

Cheng F, Wu J, Fang L, Sun S, Liu B et al (2012b) Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. PLoS One 7:e36442

Cheng F, Mandakova T, Wu J, Xie Q, Lysak MA et al (2013) Deciphering the diploid ancestral genome of the mesohexaploid *Brassica rapa*. Plant Cell 25:1541–1554

Cheng F, Wu J, Wang X (2014) Genome triplication drove the diversification of Brassica plants. Hortic Res 1:14024

Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A et al (2014) Two evolutionarily distinct classes of paleopolyploidy. Mol Biol Evol 31:448–454

Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19:1419–1428

Hollister JD, Smith LM, Guo YL, Ott F, Weigel D et al (2011) Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc Natl Acad Sci USA 108:2322–2327

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF et al (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet 43:476–481

Parkin IA, Gulden SM, Sharpe AG, Lukens L, Trick M et al (2005) Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. Genetics 171:765–781

Sankoff D, Zheng C, Zhu Q (2010) The collapse of gene complement following whole genome duplication. BMC Genom 11:313

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci USA 108:4069–4074

Schnable JC, Wang X, Pires JC, Freeling M (2012) Escape from preferential retention following repeated whole genome duplications in plants. Front Plant Sci 3:94

Schranz ME, Lysak MA, Mitchell-Olds T (2006) The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci 11:535–542

Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J et al (2003) Rate variation among nuclear genes and the age of polyploidy in Gossypium. Mol Biol Evol 20:633–643

Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F et al (2013) The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat Genet 45:831–835

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS et al (2012) Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. Genetics 190:1563–1574

Thomas BC, Pedersen B, Freeling M (2006) Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res 16:934–946

Wang J, Tian L, Lee HS, Wei NE, Jiang H et al (2006) Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics 172:507–517

Wang X, Wang H, Wang J, Sun R, Wu J et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Wang F, Li L, Li H, Liu L et al (2012) Transcriptome analysis of rosette and folding leaves in Chinese cabbage using high-throughput RNA sequencing. Genomics 99:299–307

Woodhouse MR, Schnable JC, Pedersen BS, LyonsE, Lisch D et al (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. PLoS Biol 8: e1000409

Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M et al (2014) Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. Proc Natl Acad Sci USA 111:5283–5288. doi: 10.1073/pnas.1402475111

Genome Triplication Drove the Diversification of *Brassica* Plants

10

Feng Cheng, Jian Wu, Jianli Liang and Xiaowu Wang

Abstract

Brassica species are significant in diversity. First, it has many close but different species, such as *Brassica rapa*, *Brassica oleracea*, *Brassica nigra*, *Brassica napus*, *Brassica juncea*, etc, many of which are important crops. Second, for each *Brassica* species, it is rich in morphotypes, they have distinctive and impressive traits, such as the heading leaves and enlarged roots in *B. rapa* (Chinese cabbage and turnip) or *B. oleracea* (cabbage and kohlrabi), and the enlarged inflorescences in *B. oleracea*, i.e. broccoli, cauliflower. All these *Brassica*s are evolved from a common hexaploidy ancestor that experienced a whole genome triplication (WGT) event. Studies show that WGT drove the diversification of *Brassica* plants in both the speciation and booming of morphotypes. Following WGT, the extensive block reshuffling and chromosome reduction of the triplicated diploid ancestor through rediploidization process as well as hybridization promoted the *Brassica* speciation. The biased gene retention and subgenome dominance effect further promoted function evolution of multi-copy genes, and finally lead to the expansion of rich morphotypes in *Brassica*s. Conclusively, the WGT event plays important role in driving the diversification of *Brassica*s by initiating the genome and gene-level evolution.

10.1 Diversification of *Brassica* Species

The rich diversity in phenotypes or morphotypes is the most distinct feature for *Brassica* plants. There are two levels of diversification for *Brassica* (Fig. 10.1): (1) Many *Brassica* species evolved from a common ancestor, and they are

F. Cheng · J. Wu · J. Liang · X. Wang (✉)
Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China
e-mail: wangxiaowu@caas.cn

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_10

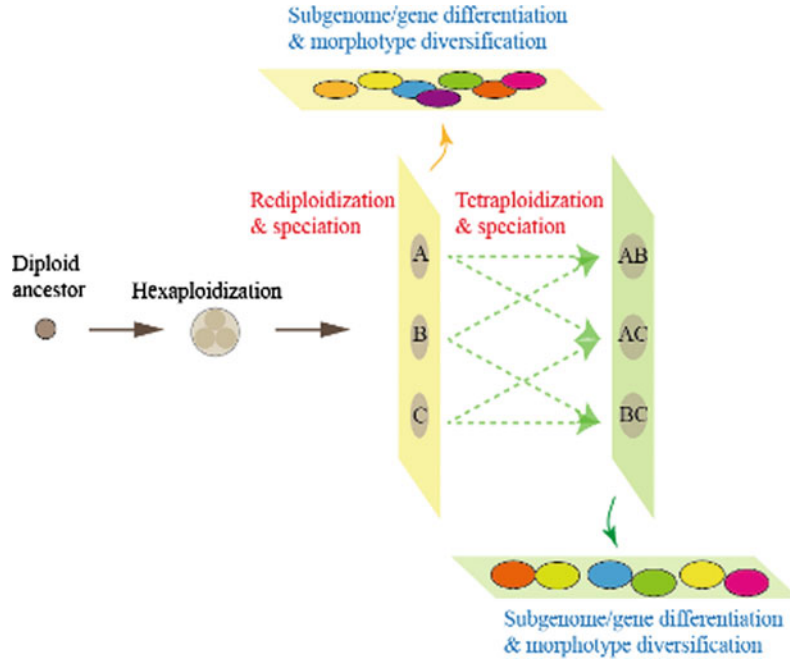

Fig. 1 Two levels of diversification for *Brassica* species following the WGT event. Level 1 is the speciation that occurred during the process of rediploidization and tetraploidization after WGT. Level 2 is the morphotype boom after the speciation, which is likely to have been driven by the biased fractionation of subgenomes, as well as the functional differentiation of multicopy genes generated from WGT

close relatives to each other. (2) There are abundant morphotypes for each *Brassica* species. First, the six main *Brassica* crops, referred to above, whose relationships were describes by U-triangle (Nagaharu 1935) represent the rich speciation of *Brassica*. Second, each of these *Brassica* species evolved into rich varieties with diversified phenotypes, including leafy heads, enlarged roots, other enlarged organs of stems or inflorescences, oilseeds, sarsons, and even ornamental features. In *Brassica rapa*, heading Chinese cabbage and pak choi are consumed as leafy vegetables. Chinese cabbage is specific for its large leafy head, whereas pak choi has relatively smaller leaves and does not form heading leaves. Turnip has an enlarged root that is edible or occasionally used as fodder. Caixin and purple caitai bolt rapidly and generate long, tender stems used as food. Morphotypes of oilseed *B. rapa* produce large, full seeds for oil extraction, and sarsons grow seed pods that are eaten in India. Some morphotypes of *B. rapa* develop beautiful leaf patterns or colors, and are thus used as ornamental plants. *Brassica oleracea* also has rich morphotypes. Heading *B. oleracea* is also used as a leaf vegetable, whereas oilseed *B. oleracea* produces edible oil. Cauliflower and broccoli, special morphotypes of *B. oleracea*, develop enlarged inflorescences that are eaten as vegetables. Other *Brassica* crops, such as *Brassica juncea*, have even greater numbers of rich morphotypes than *B. rapa* and *B. oleracea*. In addition to these cultivated crops, there are many wild relatives of the species in U-triangle that have greatly diversified phenotypes, further extending the diversity of *Brassica*. Additionally, many morphotypes or phenotypes shared among *Brassica* developed independently and in parallel, such as the heading leaves in *B. rapa* and *B. oleracea*, and enlarged roots in *B. rapa* and *B. juncea*.

The whole-genome triplication (WGT) that occurred in the common ancestor of *Brassica* crops played an important role in the diversification of *Brassica* plants. It was observed that new plant species always evolve after polyploidization. There are many such events in the evolutionary history of the plant kingdom, such as the diversification of early core eudicots (Jiao et al. 2102) (Lysak et al. 2005; Van de Peer et al. 2009). This process of diversification is also related to the events after whole-genome

duplication (WGD), including migrations, the fluctuations in environments, and/or human cultivation/selection (Proost et al. 2011), but polyploidization provides plants with the ability to be diversified and respond to changed habitats. Furthermore, from the aspect of genes, the multicopy genes generated from WGD could develop new functions (gene subfunctionalization or neofunctionalization); thus new traits or morphotypes of plants could evolve. *Brassica* species shared an additional common feature that they all experienced an extra WGT event, which occurred approximately 9–15 million years ago (MYA) (Beilstein et al. 2010; Wang et al. 2011b) or even approximately 28 MYA (Lukens et al. 2004; Lysak et al. 2005; Arias et al. 2014). The WGT event is important for the speciation and the morphotype expansion of genus *Brassica*. The subsequent genomic rearrangements and gene evolution initiated by WGT contributed to the booming of a variety of *Brassica* plants.

10.2 Chromosome Evolution after WGT and the *Brassica* Speciation

Brassica shared a common WGT event, which has been confirmed by different aspects, such as a comparative genomic analysis (Cheng et al. 2013). Genomic synteny between *B. rapa* and *Arabidopsis thaliana* clearly revealed the WGT event experienced by *B. rapa* (Wang et al. 2011b). Most genes inherited from their nearest common ancestor were shared by *B. rapa* and *A. thaliana* (80.2 and 73.8 % for *B. rapa* or *A. thaliana*, respectively) (Cheng et al. 2012a, b). Although big genomic fragments were fractionated during the rediploidization of *B. rapa*, the local gene order was conserved and syntenic fragments can be identified between *B. rapa* and *A. thaliana.* For each genomic fragment of *A. thaliana*, three syntenic copies were found in *B. rapa*. These three genome copies were generated from the WGT event (Wang et al. 2011b; Cheng et al. 2012b). Furthermore, a synteny analysis between *B. oleracea* and *A. thaliana* also showed that *B. oleracea* had good genomic collinearity with *A. thaliana*, and *B. oleracea* shared the same WGT event as that of *B. rapa* (Liu et al. 2014). Meanwhile, comparative studies of *B. juncea* and other *Brassica* using information from linkage maps showed that *B. nigra* also shared the same WGT event (Panjabi et al. 2008). Furthermore, previous works, including genomic structure analyses, paleocentromere evolution, and phylogenic studies, among multiple genomes of Brassicaceae evidenced that the diploid ancestor of *B. rapa* resembled the block arrangement of translocation Proto-Calepineae Karyotype (tPCK), which has seven chromosomes (Cheng et al. 2013). Finally, based on this information, we established that all *Brassica* crops referred to in U-triangle evolved from a common diploid tPCK genome that experienced a WGT event.

Chromosomal reduction and rearrangement, accompanied with paleocentromere descended from the hexaploid ancestor (tPCK × 3, $n = 21$), were important for the speciation of *Brassica* plants. After WGT, extensive chromosome reshuffling during rediploidization lead to the origin of closely related species in *Brassica*. In polyploids, it is understandable that having more than two copies of homologous chromosomes at the synapsis stage of meiosis will result in abnormal synaptonemal complexes, thereby decreasing the fertility of gametes. Logically, natural selection drives the rediploidization process with chromosomal rearrangement that all eliminate the extra homologous chromosomes. Further rounds of genomic reshuffling in the rediploid ancestor at different evolutionary time points then created different species of *Brassica* (Fig. 10.1). In the genome of *B. rapa*, chromosomes and paleocentromeres were reduced from 21 to 10 through multichromosome translocation, fusion, and inter/intrachromosomal recombination. These genomic reshuffling events should also have occurred in the origin of other *Brassica* species.

10.3 Gene Evolution after WGT and the Evolution of Rich Morphotypes in *Brassica*

Biased gene retention after WGT may promote the diversification of *Brassica* plants. Phytohormones, especially auxin, play important roles in the morphogenesis of plants (Santner and Estelle 2009). The genes that are involved in phytohormone signaling are thus important for the formation of diversified morphotypes (Gazzarrini and McCourt 2003; Santner and Estelle 2009). By comparing gene categories between *B. rapa* and other genomes, such as *A. thaliana*, *Carica papaya* or *Vitis vinifera*, it was found that auxin-related genes were expanded in *B. rapa* (Wang et al. 2011b). Furthermore, analysis on gene categories that retained only one or multiple copies found that genes involved in the response to almost all kinds of phytohormone signaling were significantly over-retained in the genomes of *B. rapa* (Wang et al. 2011b), as well as in *B. oleracea* (Liu et al. 2014). These redundant phytohormone related genes should contribute to the morphotype diversification of *Brassica* plants.

Subgenome dominance was observed among the three subgenomes of *B. rapa* (Cheng et al. 2012b; Tang et al. 2012). The subgenome dominance effect drove the differentiation of paralogous genes. The following differences related to subgenome dominance were found. (1) One subgenome retained more genes than the other two. It is clearly observed by counting the number of genes within the three reconstructed tPCK subgenomes, since one subgenome has approximately 1.5-times more genes than the other two subgenomes (Wang et al. 2011b; Cheng et al. 2012b). (2) There are more genes located in the over retained subgenome expressed at higher levels than in their paralogs. Using mRNA-Seq data generated for different organs of *B. rapa*, a comparison of paralogous gene pairs showed that a greater number of genes located in the over retained subgenome are expressed at higher levels than their paralogs in the other two subgenomes (Cheng et al. 2012b). (3) Genes in the dominant subgenome accumulated fewer functional mutations than those of the other subgenomes (Cheng et al. 2012b). The resequencing of different *B. rapa* morphotypes showed that genes located in the dominant subgenome accumulated fewer functional mutations than those located in the other two subgenomes, which was also considered as the ongoing biased fractionation in *B. rapa* (Cheng et al. 2012b).

WGT provided redundant genes as materials, or a buffer pool, for multicopy genes to evolve new functions, and the subgenome dominance effect facilitated the process by differentiating the multicopy genes that were located in different subgenomes of *B. rapa*. The newly evolved functions may promote the evolution of different morphotypes of *Brassica* species. In *A. thaliana*, many duplicated genes were subfunctionalized or neofunctionalized after several rounds of whole genome polyploidization, known as the α, β, and γ duplications. For example, genes from extra duplications have become subfunctionalized compared with those in *C. papaya*, such as the enzymes *CYP79A* and *CYP79B* who catalyze the first step of glucosinolate synthesis (Bekaert et al. 2012). Meanwhile, some genes in *A. thaliana* have become neofunctionalized to develop extra biosynthetic pathways for indole and methionine-derived aliphatic glucosinolates, which do not exist in *C. papaya*. Glucosinolate genes in *B. rapa* showed strong over retention after WGT (Wang et al. 2011a). These redundant genes could be subfunctionalization or neofunctionalization to develop new functions of glucosinolate metabolism in *B. rapa*, as in *A. thaliana*. It is expected that there are many more overretained genes in *B. rapa*. The subgenome dominance effect may promote the evolutionary process of these over retained genes by conserving one copy of them and differentiating the other copies to develop new functions. Finally, these differentiated genes will contribute to the evolution of rich *B. rapa* varieties and similar processes can also occur in other *Brassica* species.

10.4 Conclusions

A WGT event promoted the diversification of *Brassica* from two levels represented by speciation and the expansion of rich morphotypes. First, WGT drove the rediploidization process to stabilize the hexaploid genome. Genomic reshuffling and chromosome reduction gave rise to the speciation of diploid *Brassica* plants as *B. rapa*, *B. nigra*, and *B. oleracea*. The genomic differentiation of the three basic genomes in the U-triangle then generated the stable allotetraploid species such as *B. carinata*, *B. napus*, and *B. juncea*. Second, subgenome differentiation, biased gene retention, and gene subfunctionalization and/or neofunctionalization, after WGT promoted the parallel evolution of many special morphotypes for each *Brassica* species. Therefore, WGT initialized the genome- and gene-level evolution that further drove the *Brassica* speciation and generated a bulk of morphotypes for *Brassica*.

In the near future, further research should be conducted to investigate the evolution of each morphotype of *Brassica*. Previous studies of whole-genome sequences determined the genome- and gene-level evolution of only one accession of *B. rapa*. After that, additional *B. rapa* accessions or other *Brassica* species should be under extensive study to address the following questions: (1) What are the origins of different morphotypes in *Brassicas*? (2) What are the mechanisms of parallel evolution of certain morphotypes that developed independently in different *Brassica* species? (3) Which genes were involved in the development of each morphotype and in the regulation of important agronomic traits in *Brassica* crops? Answering these questions will increase our knowledge on the diversification of *Brassica* morphotypes and transfer the benefits of genomic studies to the application of genetic improvement for *Brassica* crops.

References

Arias T, Beilstein MA, Tang M, McKain MR, Pires JC (2014) Diversification times among *Brassica* (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence. Am J Bot 101:86–91

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S (2010) Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. Proc Natl Acad Sci USA 107:18724–18728

Bekaert M, Edger PP, Hudson CM, Pires JC, Conant GC (2012) Metabolic and evolutionary costs of herbivory defense: systems biology of glucosinolate synthesis. New Phytol 196:596–605

Cheng F, Wu J, Fang L, Wang X (2012a) Syntenic gene analysis between *Brassica rapa* and other Brassicaceae species. Front Plant Sci 3:198

Cheng F, Wu J, Fang L, Sun S, Liu B et al. (2012b) Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. PLoS One **7**:e36442

Cheng F, Mandakova T, Wu J, Xie Q, Lysak MA, Wang X (2013) Deciphering the diploid ancestral genome of the Mesohexaploid *Brassica rapa*. Plant Cell 25:1541–1554

Gazzarrini S, McCourt P (2003) Cross-talk in plant hormone signalling: what Arabidopsis mutants are telling us. Ann Bot 91:605–612

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR et al. (2102) A genome triplication associated with early diversification of the core eudicots. Genome Biol 13:R3

Liu S, Liu Y, Yang X, Tong C, Edwards D et al. (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Comun (in press)

Lukens LN, Quijada PA, Udall J, Pires JC, Schranz ME et al (2004) Genome redundancy and plasticity within ancient and recent *Brassica* crop species. Biol J Linn Soc 82:665–674

Lysak MA, Koch MA, Pecinka A, Schubert I (2005) Chromosome triplication found across the tribe Brassiceae. Genome Res 15:516–525

Nagaharu U (1935) Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilication. Jap J Bot 7:389–452

Panjabi P, Jagannath A, Bisht NC, Padmaja KL, Sharma S et al (2008) Comparative mapping of *Brassica juncea* and *Arabidopsis thaliana* using Intron Polymorphism (IP) markers: homoeologous relationships, diversification and evolution of the A, B and C Brassica genomes. BMC Genom 9:113

Proost S, Pattyn P, Gerats T, Van de Peer Y (2011) Journey through the past: 150 million years of plant genome evolution. Plant J 66:58–65

Santner A, Estelle M (2009) Recent advances and emerging trends in plant hormone signalling. Nature 459:1071–1078

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS et al. (2012) Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. Genetics 190:1563–1574

Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K (2009) The flowering world: a tale of duplications. Trends Plant Sci 14:680–688

Wang H, Wu J, Sun S, Liu B, Cheng F et al. (2011a) Glucosinolate biosynthetic genes in *Brassica rapa*. Gene 487:135–142

Wang X, Wang H, Wang J, Sun R, Wu J et al (2011b) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Comparative Analysis of Gene Conversion Between Duplicated Regions in *Brassica rapa* and *B. oleracea* Genomes

11

Jinpeng Wang, Hui Guo, Dianchuan Jin, Xiyin Wang and Andrew H. Paterson

Abstract

Plant genomes contain many duplicated genes, some of which were produced by recursive polyploidizations. These duplicated genes may evolve interactively and even concertedly through homoeologous recombination. Here, we explored likely gene conversion in *Brassica rapa* and *Brassica oleracea*. By checking gene colinearity, we detected 4296 duplicated genes existing in both the species, which were produced by whole-genome triplication from their common ancestor. Incongruities of homologous gene tree topologies indicated that 8 % of these duplicated genes were converted by one another after the divergence of *B. rapa* and *B. oleracea*. These converted genes are more often from larger duplicated chromosomal blocks, indicating that illegitimate recombination is more likely to occur between larger homoeologous chromosomal regions. This research contributed to understanding genome stability and gene evolution after polyploidization.

J. Wang · D. Jin · X. Wang
Center for Genomics and Computational Biology, North China University of Science of Technology, Tangshan 063000, Hebei, China

J. Wang · H. Guo · X. Wang
College of Life Sciences, North China University of Science and Technology, Tangshan 063000, Hebei, China

X. Wang (✉) · A.H. Paterson
Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602, USA
e-mail: wang.xiyin@gmail.com

H. Guo · A.H. Paterson
Department of Plant Biology, University of Georgia, Athens, GA 30602, USA
e-mail: paterson@uga.edu

D. Jin
College of Sciences, North China University of Science and Technology, Tangshan 063000, Hebei, China

A.H. Paterson
Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

A.H. Paterson
Department of Crop and Soil Science, University of Georgia, Athens, GA 30602, USA

A.H. Paterson (✉)
Department of Genetics, University of Georgia, Athens, GA 30602, USA
e-mail: paterson@uga.edu

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_11

11.1 Introduction

Plant genomes have been widely affected by recursive polyploidizations, which repeatedly double or triple the genome information in a cell over-night (Bowers et al. 2003, 2005; Jaillon et al. 2007; Soltis et al. 2008; Soltis and Soltis 2009; Abrouk et al. 2010; Tang et al. 2010; Jiao et al. 2011, 2012). Though wide-spread gene losses and DNA rearrangements often follow, mostly leading to restoration of diploid heredity, hundreds of duplicated genes are often preserved in colinearity on homoeologous chromosomes or chromosomal segments, retaining valuable traces of these abrupt evolutionary events (Wang et al. 2005; Gaeta et al. 2007; Paterson 2008; Paterson et al. 2009, 2012; Proost et al. 2011; Schnable et al. 2011; Freeling et al. 2012).

Recent research into illegitimate recombination between duplicated genes revealed that many duplicated genes might have been affected by gene conversion, with one copy of a pair of duplicates being converted to the DNA sequence by the other by a unidirectional recombination-like mechanism (Xu et al. 2008; Gaeta and Chris Pires 2009; Wang and Paterson 2011). A comparative analysis between rice and sorghum genomes showed that 12 % of rice duplicated genes and 14 % of sorghum duplicated genes were affected by conversion after the divergence of these lineages (Wang et al. 2007, 2009). Among those converted genes, 40 % were affected to their full gene length and the others in only partial sequence. These conversion events may have occurred tens of millions years ago.

A comparison between rice subspecies *indica* and *japonica* found evidence of more recent gene conversion, showing that ~8 % of rice genes may have been converted after the split of the two subspecies about 400,000 years ago (Zhu et al. 2007). One pair of grass chromosomes, e.g., rice chromosomes 11 and 12, their sorghum orthologous chromosomes 5 and 8, and corresponding chromosomes in other grasses, have been affected by prominent conversion (Wang et al. 2011a, b, c). After the split of rice and sorghum, nearly 60 % of rice and sorghum duplicated genes have been converted by their duplicated copies. Evidence from sequence similarity analysis, and independent analysis of *Oryza* species (including rice) indicated that near the termini of the short arms of *Oryza* chromosomes 11 and 12, gene conversion may be still ongoing, 70 million years or after the origination of these duplicated genes (Jacquemin et al. 2009; Wang et al. 2011a, b, c).

Analysis of eudicot genomes found more evidence of homo(eo)logous gene conversion. In a tetraploid cotton, *Acala Maxxa*, 40 % of paralogous genes from its two subgenomes At and Dt differ in sequence from their diploid progenitors. The vast majority of these mutations are convergent, with At genes converted to the Dt state at more than twice the rate (25 %) as the reciprocal (10.6 %) (Paterson et al. 2012). As to conversion between homologous chromosomes, sequencing 40 *Arabidopsis* F_2 plants and their parents showed that small gene conversion tracts, often biased, represented over 90–99 % of all recombination events. Moreover, the rate of alteration of protein sequence caused by gene conversion is reported to be more than 600-times that caused by mutation (Yang et al. 2012).

11.2 Comparative Inference of Gene Conversion in *B. rapa* and *B. oleracea*

The existence of large homoeologous blocks provides a chance for homoeologous (ectopic) DNA recombination, which may result in concerted evolution of duplicated genes as inferred previously in grasses (Wang et al. 2009, 2011a, b, c).

11.2.1 Rationale to Infer Gene Conversion

Annotated genes from *Brassica rapa* and *Brassica oleracea* were from sequencing project websites (Wang et al. 2011a, b, c; Liu et al. 2014). To find colinear homologs within a plant or between two plants, we run BLASTP to find

homologous genes. Homologs with E-values smaller than 1e-10 were taken as input for ColinearScan, which was adopted to infer DNA blocks containing 10 or more colinear genes. By checking chromosome numbers, it was not difficult to define orthologs between *B. rapa* and *B. oleracea*. By using an approach described previously, we defined three subgenomes A, B, and C, and found paralogs in each plant. If there was no gene loss, at corresponding locations there would be three colinear genes in each plant produced by the genome triplication, namely, Br-A, Br-B, and Br-C in *B. rapa*, and their respective orthologs, Bo-A, Bo-B, and Bo-C in *B. oleracea*, forming homologous gene sextet. However, due to wide-spread gene loss after the genome triplication, often we could not find sextets of homoeologs.

To infer gene conversion, based on sextets or incomplete groups, we defined homologous gene quartets, two paralogs in a plant and their respective orthologs in the other plant. Then we inferred synonymous nucleotide substitution rates (Ks) between them. We anticipated that orthologs were more similar than paralogs, in that speciation was after genome triplication. However, if paralogs in a genotype were more similar than orthologs, we considered that the paralogs might have been affected by gene conversion. Bootstrapping tests were repeated 100 times. To estimate Ks, we first aligned proteins of a homologous quartet with CLUSTALW, and after removing gaps, the protein alignment was then translated into cDNA alignment in codons. Ks were estimated by using the Nei-Gojobori approach implemented by BioPerl.

11.2.2 Characterization of Gene Conversion

By using ColinearScan to find gene colinearity and by checking sequence similarity between chromosomal regions, we inferred paralogous genes within *B. rapa* and *B. oleracea* genomes, respectively, and inferred orthologous genes between them. Here, we checked triplicated genes that were preserved in both *Brassica* species, which form homologous gene sextets. For genes in each sextet, we checked each quartet of homologs within them. We compared gene similarity or tree topology. We anticipated that the paralogs (duplicated genes) were more diverged than their respective orthologs. If not, we inferred that the paralogs might have been affected by gene conversion. We removed possible redundancy when counting converted gene pairs.

We detected 4296 homoeologous pairs of genes, involving 8592 (20.6 % of) *B. rapa* genes and 8592 (24.7 % of) *B. oleracea* genes. Most of these reside in 23 large duplicated blocks in *B. rapa* (Fig. 11.1a) and 19 large duplicated blocks in *B. oleracea* (Fig. 11.1b), distributed throughout the chromosomes. In total, we found that ~8 % of duplicates (368 and 343) in *B. rapa* and *B. oleracea* have been affected by gene conversion (Table 11.1). The conversion tracts vary in size, ranging from a few base pairs to full gene lengths.

11.2.3 Unbalanced Gene Conversion Among Chromosomes

Different chromosomes have been unequally affected by gene conversion (Fig. 11.1c, d). In *B. rapa*, the most affected chromosomes are Br01, Br04, and Br05, with >10 % of paralogs affected, whereas in *B. oleracea*, the most affected chromosomes are Bo01 and Bo06, with >10 % of paralogs affected. In contrast, no paralogous pair from between Br09 and Br01, Bo08 or Bo09, Bo04 and Bo09 has been affected. Genes residing in bigger chromosomes with more colinear homoeologs are more likely to be affected by conversion (Fig. 11.2). This means larger duplicated regions on these chromosomes may facilitate the occurrence of homoeologous recombination due to preserving more DNA homology.

11.2.4 Gene Conversion Occurs Correspondingly in Two *Brassica* Species

Gene conversion often occurs in both *Brassica* species in a corresponding manner, that is, if a duplicated gene pair were affected by gene

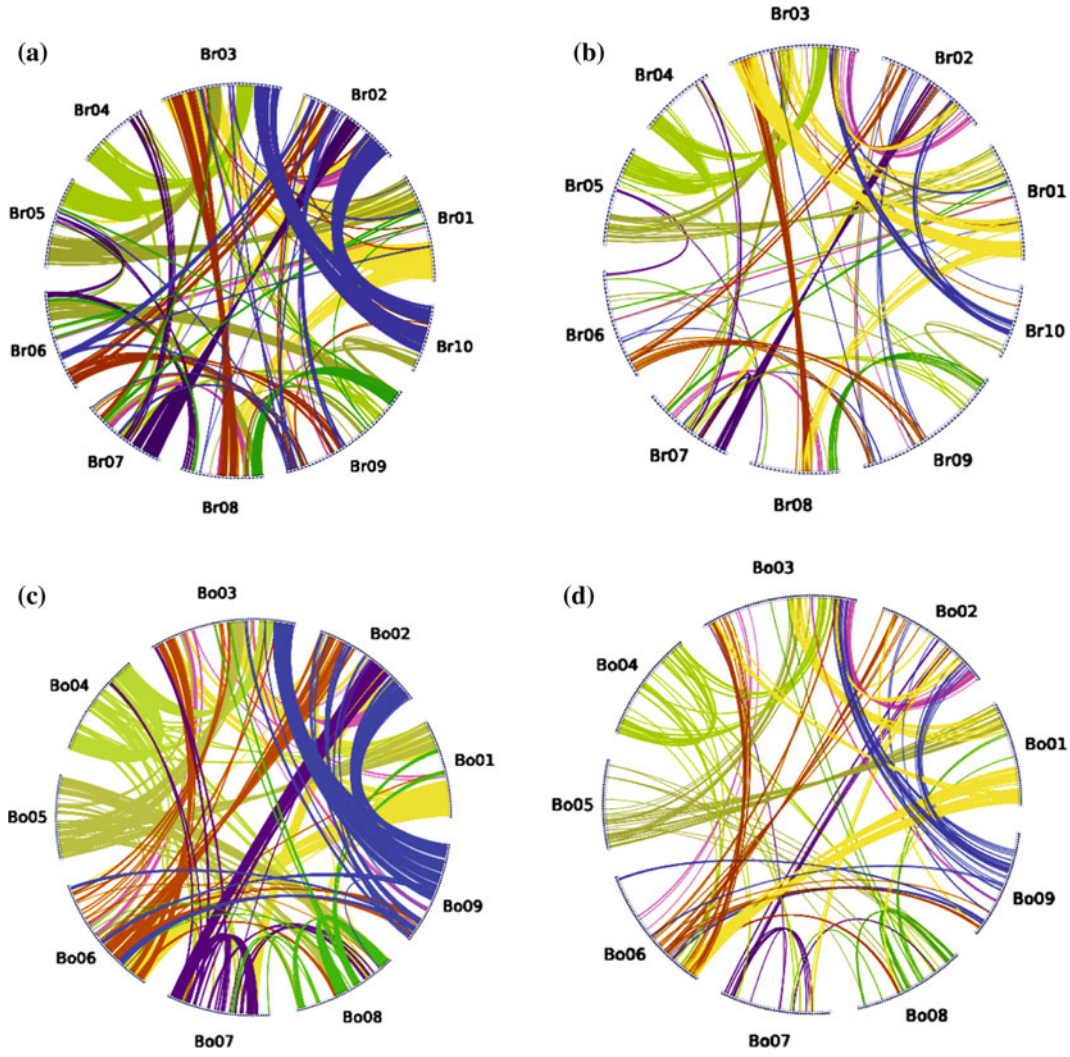

Fig. 11.1 Whole-genome triplication and gene conversion. Distributions of duplicated genes (**a**, **c**) and those converted (**b**, **d**) in *B. rapa* and *B. oleracea*, respectively

conversion in one species, so were their counterparts in the other species. Most homoeologous quartets (~92 %) were found to be converted in both species. Only 53, or about one-sixth, of homoeologous gene quartets showed evidence of independent concerted evolution, i.e. were inferred to have experienced independent conversion events in *B. rapa* or *B. oleracea*. That is, it is likely that 5/6 the events are likely to have occurred shortly after the triplication but before the lineages diverged, or co-occurred independently in each lineage.

11.2.5 Biased Gene Conversion Among Different Subgenomes

Previous publication (Wang et al. 2011a, b, c) revealed three subgenomes that formed the present genomes of *B. rapa* and *B. oleracea*, and here we characterize gene conversion between different subgenomes. As to the analysis, there is an occurrence bias of gene conversion among subgenomes. About 40–44 % of conversion events involved paralogs on subgenomes A and B in both species, substantially more than

Table 11.1 Gene conversion on chromosomes in *B. rapa* and *B. oleracea*

Chromosomes	Paralogs	Converted genes	Conversion rates
Br01	1029	107	0.104
Br02	1000	65	0.065
Br03	1914	175	0.091
Br04	555	62	0.112
Br05	930	97	0.104
Br06	499	33	0.066
Br07	735	65	0.088
Br08	689	56	0.081
Br09	648	48	0.074
Br10	577	28	0.049
Summary	8576	736	0.083
Bo01	1018	104	0.102
Bo02	957	59	0.062
Bo03	1619	129	0.080
Bo04	1099	84	0.076
Bo05	651	52	0.080
Bo06	1055	111	0.105
Bo07	641	34	0.053
Bo08	734	57	0.078
Bo09	802	56	0.070
Summary	8576	686	0.078

between other subgenome combinations (Table 11.2). However, this increase parallels gene numbers in the respective subgenomes, with the percentages of converted paralogs from any two subgenomes being similar. This suggests that gene conversion is related to homologous gene density, which determines the likelihood of illegitimate recombination to occur.

11.3 Gene Conversion and Genome Stability

11.3.1 Increased Genome Stability and Complexity After Polyploidization

Wide-spread and recursive polyploidizations have affected all flowering and seed plants, and may have been an important driving force of their evolution, especially the likely rapid divergence and speciation of lineages to form large groups of related species (Bowers et al. 2003; Ziolkowski et al. 2006; Soltis and Soltis 2009; Jiao et al. 2012). This should be a direct result of genome instability after whole-genome duplication/triplication (Marfil et al. 2006; Mazowita et al. 2006). These large-scale genome addition events duplicated/triplicated DNA content overnight, adding much genomic complexity and increasing interactions between chromosomes. Such interactions may include physically by DNA binding, knotting, splitting and breaking; and genetically by pairing, clustering, recombining, and segregating. A drive to recover diploid heredity may be the paramount source of force. Anyway, the majority of land plants favors diploid heredity and are adapted to finish a cycle of meiosis each year. Increased complexity will lead a lot of outcomes genetically, and the first among them is genomic instability.

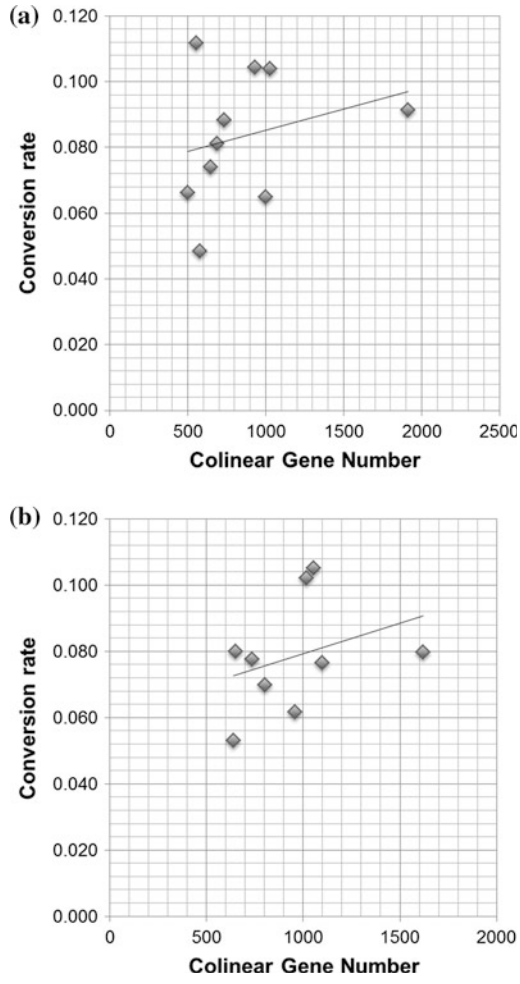

Fig. 11.2 Correlation of gene conversion and duplicated block size. **a** *B. rapa*. **b** *B. oleracea*

Table 11.2 Gene conversion in subgenomes in *B. rapa* and *B. oleracea*

Between subgenomes	Quartet #	*B. rapa*			*B. oleracea*		
		Converted #	Fraction of all converted	Percent. of quartet	Converted #	Fraction of all converted	Percent. of quartet
A–B	1790	149	0.40	0.08	150	0.44	0.08
A–C	1327	109	0.30	0.08	92	0.27	0.07
B–C	1171	110	0.30	0.09	101	0.29	0.09

Genomic instability is often accompanied by wide-spread gene losses, chromosomal rearrangement, and recombination between homo (eo)logous chromosomes or chromosomal segments (Wang et al. 2005; Feldman et al. 2012). If a polyploid came to recover diploid heredity, with one-to-one pairing of homologous chromosomes rather than pairing among multiple homo(eo)logous chromosomes, it may eventually regain much of its genomic stability. However, small scale chromosomal rearrangement may still continue to occur. The analysis of grass genomes indicated that the majority of genomic changes occurred before the divergence of major grass clades. For example, after the divergence of rice and sorghum, only ~2–3 % of genes were lost, resulting in minimal erosion of gene colinearity along orthologous chromosomes, in contrast to the loss of at least 65 % of genes duplicated in their common ancestor (Wang et al. 2005; Paterson et al. 2009). For another example, the majority of chromosomal rearrangement occurred before their divergence, and only a few such rearrangements can be identified in the sorghum lineage after its split with rice (Murat et al. 2010).

11.3.2 Homoeologous Recombination Is a Driving Force for Genomic Evolution

Homoeologous recombination is also a phenomenon of genomic instability, and can last much longer than other changes discussed above. As a result of this kind of illegitimate recombination, gene conversion transfers genetic information in a unidirectional manner. As gene conversion mechanisms proposed, it would increase DNA substitution rates, and therefore may play a role as a driving force of evolution (Chen et al. 2007; Wang and Paterson 2011). This has been attested to by comparative analysis of grass genes (Wang et al. 2009; Wang and Paterson 2011). After the ease of major genomic changes, homoeologous recombination and gene conversion can still occur millions of years after ancestral polyploidization (Wang et al. 2011a, b, c; Paterson et al. 2012). This has been evidenced from the analysis of both monocot to dicot plants. A particularly striking finding involves genes at the very end of rice chromosomes 11 and 12 and their counterparts in other grasses (Wang et al. 2007; Jacquemin et al. 2009; Paterson et al. 2012).

11.3.3 Gene Conversion and Homoeologous Block Length

Here, we revealed a correlation of longer lengths of duplicated blocks (or larger numbers of genes) with higher conversion rates, which agrees with previous findings in grasses. More colinear genes often mean higher DNA similarity between duplicated regions, which would increase the likelihood of homoeologous pairing. The chance of pairing is definitely much less between homoeologous than homologous chromosomes. Once it occurs, it would have some genetic outcomes, such as relatively low-level genomic instability, DNA mutations, and conversion.

11.4 Conclusion

Here, by performing comparative genomic analysis, we characterized gene conversion in *B. rapa* and *B. oleracea*. Gene conversion as a result of homoeologous recombination is a long lasting driving force of plant evolution. Widespread and recursive polyploidizations have played a pivotal role in the evolution, divergence and speciation of land plants. After the ease of genome shock (McClintock 1984) often in the early days after polyploidization, characterized by wide-spread gene losses and chromosomal rearrangements, genomes may recover much stability and return to diploid heredity. Though occurring at lower levels in later stages than early days after polyploidization, homoeologous recombination and gene conversion may last for a very long time, continuing to play a driving force in genomic evolution and genetic innovation.

References

Abrouk M, Murat F, Pont C, Messing J, Jackson S et al (2010) Palaeogenomics of plants: synteny-based modelling of extinct ancestors. Trends Plant Sci 15:479–487

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438

Bowers JE, Arias MA, Asher R, Avise JA, Ball RT, et al (2005) Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. Proc Natl Acad Sci USA 102:13206–13211

Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP (2007) Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet 8:762–775

Feldman M, Levy AA, Fahima T, Korol A (2012) Genomic asymmetry in allopolyploid plants: wheat as a model. J Exp Bot 63:5045–5059

Freeling M, Woodhouse MR, Subramaniam S, Turco G, Lisch D, Schnable JC (2012) Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr Opin Plant Biol 15:131–139

Gaeta RT, Chris Pires J (2009) Homoeologous recombination in allopolyploids: the polyploid ratchet. New Phytol 186:18–28

Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC (2007) Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. Plant Cell 19:3403–3417

Jacquemin J, Laudie M, Cooke R (2009) A recent duplication revisited: phylogenetic analysis reveals an ancestral duplication highly-conserved throughout the Oryza genus and beyond. BMC Plant Biol 9:146

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, et al (2011) Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR et al (2012) A genome triplication associated with early diversification of the core eudicots. Genome Biol 13:R3

Liu S, Liu Y, Yang X, Tong C, Edwards D, et al (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun 5:3930

Marfil CF, Masuelli RW, Davison J, Comai L (2006) Genomic instability in *Solanum tuberosum* × *Solanum kurtzianum* interspecific hybrids. Genome 49:104–113

Mazowita M, Haque L, Sankoff D (2006) Stability of rearrangement measures in the comparison of genome sequences. J Comput Biol 13:554–566

McClintock B (1984) The significance of responses of the genome to challenge. Science 226:792–801

Murat F, Xu JH, Tannier E, Abrouk M, Guilhot N et al (2010) Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. Genome Res 20:1545–1557

Paterson AH (2008) Paleopolyploidy and its impact on the structure and function of modern plant genomes. Genome Dyn 4:1–12

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457:551–556

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, et al (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492:423–427

Proost S, Pattyn P, Gerats T, Van de Peer Y (2011) Journey through the past: 150 million years of plant genome evolution. Plant J 66:58–65

Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci USA 108:4069–4074

Soltis PS, Soltis DE (2009) The role of hybridization in plant speciation. Annu Rev Plant Biol 60:561–588

Soltis DE, Bell CD, Kim S, Soltis PS (2008) Origin and early evolution of angiosperms. Ann N Y Acad Sci 1133:3–25

Tang H, Bowers JE, Wang X, Paterson AH (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci USA 107:472–477

Wang XY, Paterson AH (2011) Gene conversion in angiosperm genomes with an emphasis on genes duplicated by polyploidization. Genes (Basel) 2:1–20

Wang X, Shi X, Hao B, Ge S, Luo J (2005) Duplication and DNA segmental loss in the rice genome: implications for diploidization. New Phytol 165:937–946

Wang X, Tang H, Bowers JE, Feltus FA, Paterson AH (2007) Extensive concerted evolution of rice paralogs and the road to regaining independence. Genetics 177:1753–1763

Wang X, Tang H, Bowers JE, Paterson AH (2009) Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. Genome Res 19:1026–1032

Wang X, Tang H, Paterson AH (2011a) Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. Plant Cell 23:27–37

Wang X, Wang H, Wang J, Sun R, Wu J et al (2011b) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Wang XY, Tang HB, Paterson AH (2011c) Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major poaceae lineages. Plant Cell 23:27–37

Xu S, Clark T, Zheng H, Vang S, Li R et al (2008) Gene conversion in the rice genome. BMC Genom 9:93

Yang S, Yuan Y, Wang L, Li J, Wang W, et al (2012) Great majority of recombination events in Arabidopsis are gene conversion events. Proc Natl Acad Sci USA 109:20992–20997

Zhu Q, Zheng X, Luo J, Gaut BS, Ge S (2007) Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. Mol Biol Evol 24:875–888

Ziolkowski PA, Kaczmarek M, Babula D, Sadowski J (2006) Genome evolution in Arabidopsis/Brassica: conservation and divergence of ancient rearranged segments and their breakpoints. Plant J 47:63–74

Molecular Mapping and Cloning of Genes and QTLs in *Brassica rapa*

12

Guusje Bonnema

Abstract

In this chapter an overview is given of QTL studies performed in the species *Brassica rapa*. First we provide an overview of the types of molecular markers that have been used in time, and the genetic maps that have been constructed from a broad range of populations, both in terms of population type and morphotypes used for the crosses. Since the publication of the *B. rapa* genome sequence, numbers of molecular markers have increased exponentially. Second, an overview is given of QTL studies published in the last few years, and thereafter the QTL studies that resulted in cloning of the causal genes are presented. A brief overview is given of association mapping studies performed in *B. rapa*, and the chapter is concluded with proposed strategies to identify genes underlying both qualitative and quantitative traits.

12.1 Introduction

The species *Brassica rapa* encompasses many important crop species including heading Chinese cabbages, non-heading leafy vegetables like pakchoi and the Japanese mizuna's, mibuna's and neep greens, vegetable and fodder turnips and both annual and biannual oilcrops (Zhao et al. 2005; Bonnema et al. 2011). The diversity both between and within these diverse crop types is enormous, and a valuable resource for crop improvement (Zhao et al. 2005; Pino Del Carpio et al. 2011a, b; Lee et al. 2013).

Classical genetic analysis has been used to unravel the inheritance of various qualitative and quantitative traits. Identification of these trait loci can be exploited for the development of genetic markers that can be used for marker-assisted selection of the traits of interest. This information can also lead to the identification and isolation of causative genes that increase our understanding of the trait variation. For these genetic analyses, a wide range of bi-parental mapping populations including among others F_2, F_3, backcross (BC), recombinant inbred line (RIL) or doubled haploid (DH) have been developed, that segregate for the traits of interest (see also Table 12.1).

G. Bonnema (✉)
Wageningen UR Plant Breeding, Wageningen, The Netherlands
e-mail: Guusje.Bonnema@wur.nl

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_12

Table 12.1 Overview of QTL studies in *B. rapa* since 2011 until February 2014

Article title	Traits analysed	QTL	Candidate genes	Population/markers	Authors, year
Genetic dissection of leaf development in *Brassica rapa* using a "genetical genomics" approach	Leaf size and shape	Leaf trait QTL, flowering time QTL and eQTL	Yes, major regulators	90 DH lines 1328 cM 509 markers AFLP, SSR, SNPs	Xiao et al. (2014)
Quantitative trait loci x environment interactions for plant morphology vary over ontogeny in *Brassica rapa*	Vegetative traits, stem elongation, onset reproduction	QTL x environment x ontogeny interaction		R5009IMB211 RILs 227 RFLP and SSR markers in Iniguez-Luy et al. (2009)	Dechaine and Brock (2014)
Identification and mapping of a novel dominant resistance gene, *TuRB07* to turnip mosaic virus in *Brassica rapa.*	Turnip mosaic virus	Single dominant locus*TuRB07*	CC-NBS-LRR, *Bra018862* and *Bra018863*	DH, F_2 and two BC_1 populations SSRs	Jin et al. (2014)
RNA-seq based SNPs in some agronomically important oleiferous lines of *Brassica rapa* and their use for genome-wide linkage mapping and specific-region fine mapping	Tetralocular ovary (*Tet-o* locus)	Fine-map of *tet-o* locus		F_7-RIL population derived from a Chiifu x Tetra (ssp. trilocularis, Yellow sarson) cross 594 SNP and 138 IP and SSR markers in single copy genes	Paritosh et al. (2013)
Quantitative trait loci mapping in *Brassica rapa* revealed the structural and functional conservation of genetic loci governing morphological and yield component traits in the A, B, and C subgenomes of *Brassica* species	Morphology and yield traits	Many QTLs for morphological traits and yield, set of candidate genes	Next-generation seq data, *A. thaliana* and comparison QTL clusters in *B. napus* and *B. juncea* to identify candidate genes. Expression profiling in parental leaf tissue	190 F_2 from a cross between Chinese cabbage Chiifu and rapid cycling. BAC-SSRs, EST-SSRs and Intron Polymorphisms (IP)	Li et al. (2013a)
Fine-mapping of the clubroot resistance gene *CRb* and development of a useful selectable marker in *Brassica rapa*	The clubroot resistance (CR) gene *CRb*		140-kb genomic region with candidate resistance genes.	2032 F_2 plants generated by selfing *B. rapa* 'CR Shinki.'	Kato et al. (2013)

(continued)

Table 12.1 (continued)

Article title	Traits analysed	QTL	Candidate genes	Population/markers	Authors, year
Mapping quantitative trait loci for yield-related traits in Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*)	Yield-related traits, heading traits	QTL-by-environment interactions		192 DH lines (Chinese cabbage x Chinese cabbage cross) 43 SSRs 190 sequence-related amplified polymorphism (SRAP)	Liu et al. (2013b)
A naturally occurring long insertion in the first intron in the *Brassica rapa FLC2* gene causes delayed bolting	Bolting	Bolting	*BrFLC2, BrFLC3* and *BrFLC3'*	F_2 population (385 plants) derived from the Tsukena No. 2 9 "Early" 110 SSR, FLC based	Kitamoto et al. (2013)
Fine-mapping and identification of candidate *Br-or* gene controlling orange head of Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*)	Orange head Chinese cabbage accumulating significant amounts of carotenoids	Single recessive gene, *Br-or*	*ORF1* encoding carotenoid isomerase, involved in isomerization of carotenoids	F_2S_4 mapping population (1724 plants). SCAR, SSR and InDel based on reference genome	Zhang et al. (2013)
Genetic analysis of health-related secondary metabolites in a *Brassica rapa* recombinant inbred line population.	Seed tocopherol and seedling metabolite concentrations.	Wide range of QTL	*BrVTE1* gene om A03 for tocopheroles	RIL population, 160 F_7 100 SNPs, 130 AFLP, 27 InDel, and 13 SSR	Bagheri et al. (2013a, b)
Identification of seed-related QTL in *Brassica rapa*	Seed traits, oil content, siliques	Wide range of QTL		F2 97 AFLPs and 21 SSRs	Bagheri and Pino-Del-Carpio (2013)
Identification and characterization of *Crr1a*, a gene for resistance to clubroot disease (*Plasmodiophorabrassicae* Woronin) in *Brassica rapa* L	Clubroot disease	Two gene loci, *Crr1a* and *Crr1b*	$Crr1a^{G004}$(TIR-NB-LRR), Complementatio tests	3700 F_2 Plants	Hatakeyama et al. (2013)

(continued)

Table 12.1 (continued)

Article title	Traits analysed	QTL	Candidate genes	Population/markers	Authors, year
Fine-mapping of the clubroot resistance gene *CRb* and development of a useful selectable marker in *Brassica rapa*	Clubroot resistance gene	Single locus *Crb*	140 kb region with several candidate genes, among others TIR-NBS-LRR	2032 F2 plants generated by selfing *B. rapa* 'CR Shinki	Kato et al. (2013)
Nucleotide sequence variation of GLABRA1 contributing to phenotypic variation of leaf hairiness in Brassicaceae vegetables	Leaf hairiness	Single locus	*GLABRA1 (GL1)*		Li et al. (2013b)
miR319a-targeted BrpTCP genes modulate head shape in *Brassica rapa* by differential cell division arrest in leaf regions	Chinese cabbage head shape		miR319a targeted *BrpTCP4*	150 RILs	Mao et al. (2013)
The *Brassica rapa FLC* homologue *FLC2* is a key regulator of flowering time, identified through transcriptional co-expression networks	Flowering time	QTL, eQTL, co-expression networks	*BrFLC2* major regulator	90 DH lines 278 AFLPs, 50 SSRs, 125 Flowering time (SNP) markers, 2 IBP (intron-based polymorphism), and 1 CAPS	Xiao et al. (2013)
QTL Mapping of Leafy Heads by Genome Resequencing in the RIL Population of *Brassica rapa*	Six head traits	18 QTLs	*BrpGL1* candidate gene for trichomes; AP2 domain TF *BrpESR1*—no petiole phenotype, leaf serration—*BrpSAW1*	150 recombinant inbred lines (RILs) 2209 SNP markers	Yu et al. (2013)
Genetic analysis of morphological traits in a new, versatile, rapid-cycling *Brassica rapa* recombinant inbred line population	Plant architecture and seed characteristics			160 F_7 RILs 100 SNPs, 130 AFLPs, 27 InDel, and 13 SSR markers	Bagheri et al. (2012)

(continued)

Table 12.1 (continued)

Article title	Traits analysed	QTL	Candidate genes	Population/markers	Authors, year
Genetic mapping and localization of quantitative trait loci for chlorophyll content in Chinese cabbage (*Brassica rapa* ssp. *pekinensis*)	Chlorophyll a and b content	QTLs	Additive and dominant QTL	$F_{2:3}$ mapping population; 238 marker loci (SSR)	Ge et al. (2012)
Construction of genetic linkage map and mapping of QTL for seed color in *Brassica rapa*	Seed color	QTLs on A09		RILs from cross Yellow seeded cv x brown seeded line	Kebede et al. (2012)
A naturally occurring InDel variation in *BraA.FLC.b* (*BrFLC2*) associated with flowering time variation in *Brassica rapa*	Flowering time	Association mapping QTL	*BrFLC2*	159 *B. rapa* accessions	Wu et al. (2012)
A large insertion in bHLH transcription factor BrTT8 resulting in yellow seed coat in *Brassica rapa*	Seed color	Single recessive maternal gene	BrTransparant testa 8 (*BrTT8*)	BC5 black-seeded parent as a donor to Yellow sarson	Li et al. (2012)
Mapping quantitative trait loci for leaf and heading-related traits in Chinese cabbage (*Brassica rapa* L. ssp. *pekinesis*)	Seven leaf and head-related traits	17 QTL	No candidates, no reference genome	139 F3 lines from a cross between two Chinese Cabbages	Ge et al. (2011)
The genetic architecture of ecophysiological and circadian traits in *Brassica rapa*	QTL association between circadian and ecophysiological traits	Several QTLs; correaltion and colocation analysis		150 RILs *B. rapa*, yellow sarson R500 x rapid cycling IMB211; 224 RFLP and microsatellite markers (Iniguez-Luy et al. 2009)	Edwards et al. (2011)

(continued)

Table 12.1 (continued)

Article title	Traits analysed	QTL	Candidate genes	Population/markers	Authors, year
Regulatory hotspots are associated with plant gene expression under varying soil phosphorus supply in *Brassica rapa*	Cis- and trans-eQTL and their environmental response to low phosphorus availability; P use efficiency-related QTL	Trans-eQTL hotspots on A06 and A01	A06 hotspot enriched with P metabolism-related Gene Ontology terms; A01 hotspot with chloroplast- and photosynthesis- related terms (A01)	67 selected informative RILs*; B. rapa*, yellow sarson R500 x rapid cycling IMB211 (Iniguez-Luy et al. 2009) 125 gene Expression Markers (GEMs) based on leaf transcript profiles on Agilent Brassica 95 k 60-mer arrays (Trick et al. 2009a	Hammond et al. (2011)
Genetic architecture of the circadian clock and flowering time in *Brassica rapa*	QTL for circadian rythm and flowering time	All the flowering-time QTL partially overlap with circadian period QTL	Micro-synteny between *Arabidopsis* and *B. rapa* genomes to identify candidate genes for these QTL PRR7	50 *B. rapa* accessions, 8 wild populationd (california), 159 RILs from *B. rapa*, yellow sarson R500 x rapid cycling IMB211 (Iniguez-Luy et al. 2009)	Lou et al. (2011)
Sequence-characterized amplified region and simple sequence repeat markers for identifying the major quantitative trait locus responsible for seedling resistance to downy mildew in Chinese cabbage (*Brassica rapa* ssp. pekinensis)	QTL for seedling downy mildew resistance		Fine mapping by developing SSR markers from BAC clones, based on physical interval	100 DH lines from a cross between 91 and 112 and T12-19, susceptible and resistant resp. to downy mildew	Yu et al. (2011)

12.2 Genetic Markers and Genetic Maps

For *B. rapa*, over 20 genetic linkage maps have been constructed using a range of marker types. In the 1990s, restriction fragment length polymorphism (RFLP) markers were mainly used (Song et al. 1991; Hoenecke and Sernyk 1992; Teutonico and Osborn 1994; Lagercrantz and Lydiate 1996; Kole et al. 1997), then, from around 1995 to 2006, amplified fragment length polymorphism (AFLP), RFLP and random amplified polymorphic DNA (RAPD) markers were widely used (Kim et al. 2006), and from 2005 to the present, markers like simple sequence repeats (SSR; Choi et al. 2007; Kim et al. 2009; Wang et al. 2011c; Kebede et al. 2012; Suwabe et al. 2012; Yu et al. 2012), single nucleotide polymorphism (SNP; Li et al. 2011; Wang et al. 2011a; Chung et al. 2013; Paritosh et al. 2013) and insertion-deletions (InDels; Choi et al. 2007; Kim et al. 2009; Wang et al. 2011a; Suwabe et al. 2012; Yu et al. 2012; Liu et al. 2013a) gradually have become more favorable. *B. rapa* has 10 chromosomes, so genetic maps should consist of 10 corresponding linkage groups. In the early mapping studies, linkage groups were often not assigned to chromosome numbers, and if so, the orientation of the linkage groups with respect to chromosome orientation was not defined. This made it virtual impossible to compare mapping studies both within *B. rapa* and across *B. rapa, Brassica napus, Brassica juncea,* etc. The orientation of linkage groups A03 and A10 in two important published *B. rapa* reference genetic maps, was for example different; the orientation in the F_2 linkage map of Kim et al. (2006) was opposite to the orientation in the DH linkage map published by Choi et al. (2007). The orientation as published by Choi et al. (2007) followed the international standard, as this orientation is similar to that of the *B. napus* maps used for comparative mapping published by Parkin et al. (2005). When reading papers, one needs to make sure to interpret linkage group number and orientation correctly. The advantage of sequence tagged markers, like SSRs, InDels and SNPs, is that they are sequence based, are transferable across mapping populations, and thus can serve as anchors to link mapping populations, and additionally can anchor genetic maps to the *B. rapa* reference genome.

With the publication of the *B. rapa* genome sequence in 2011, the organization of the three subgenomes representing a genome triplication event that preceded the origin of the diploid Brassica species, *B. rapa, Brassica oleracea*, and *Brassica nigra*, became clear (Wang et al. 2011b). Studies on the ancestral genomes and the patterns of local gene losses and insertions, increased insight in this complex genome (Lysak et al. 2005; Town et al. 2006; Wang et al. 2011b; Cheng et al. 2012, 2013; Lou et al. 2012; Tang et al. 2012). Recent developments in sequencing technology, accompanied by dramatic cost reductions (Muers 2011) and improved data analysis techniques, have made it feasible to generate and analyze genomic data of large numbers of genotypes, which allows the discovery of sequence variants and generation of sequence-based markers including SNPs and InDels (Edwards et al. 2013). The abundance and distribution throughout the genome make these types of markers ideal for high-resolution genetic mapping and association mapping studies. In the last few years, several papers describing improved genetic maps of *B. rapa* have been published. This includes papers that describe the genetic and physical mapping of certain classes of genes, like the glucosinolate biosynthetic genes (Zang et al. 2009; Wang et al. 2011c). These papers describe identification of 102 putative GS genes in *B. rapa* as the orthologs of 52 GS genes in *Arabidopsis thaliana,* and the genetic map positions of all but one. High-density genetic maps were presented for diverse crossing combinations, between different crop types. Whole-genome resequencing of two parental lines (Chinese cabbage Z16 and rapid cycling L144) resulted in the development of genetic markers to construct a genetic map of the corresponding DH population with 415 InDel markers and 92 SSR markers. Using this linkage map, 152 scaffolds could be anchored to

chromosomes, encompassing more than 82.9 % of the *B. rapa* genome (Wang et al. 2011c). In a follow up study, 26,693 InDel markers markers were identified and a selection of 503 InDel markers was made to evaluate their use in other mapping populations (Liu et al. 2013a, b). Screening of seven *B. rapa* accessions resulted in 387 (77 %) polymorphic markers. The use of these markers was also evaluated in *B. napus* (the AC genome) and *B. juncea* (the AB genome). More than 90 % amplified a PCR product; 25 % showed polymorphism between the two *B. napus* accessions and 8 % between the two *B. juncea* accessions, making this set of novel PCR-based InDel markers a valuable resource for genetic studies and breeding programs in diverse *Brassica* species. In yet another study, a genetic map of a DH population from a cross between a Chinese Cabbage and a European turnip is described with 629 markers, including 112 SSRs, 129 InDels and 370 other marker types, and using the sequence based markers, this map could be anchored to other reference maps for *B. rapa* (Yu et al. 2011). A number of maps have also been published recently that consist of multiple marker types (AFLP, SSR, CAPs), enriched with gene-specific markers. The DH population from a cross between a Yellow Sarson and a Pakchoi was enriched with flowering time genes (Xiao et al. 2013) and homologs of *A. thaliana* genes involved in leaf development (Xiao et al. 2014).

A 60,000 (60 k) SNP Infinium genotyping array for *B. napus* (A and C genomes) was produced in 2012 by the International Brassica SNP Consortium in cooperation with Illumina, Inc., San Diego, CA, USA (Snowdon and Friedt 2004; Snowdon and Iniguez Luy 2012;. Edwards et al. 2013). This SNP array makes it possible to efficiently generate high-density, sequence-based, genome-wide polymorphism screens and genetic maps. This array has also been tested on both a *B. rapa* and a *B. oleracea* DH populations and a collection with many accessions. Most A genome markers worked well in *B. rapa*, while C genome markers worked in *B. oleracea*, with very few functional in both the genomes. In each DH population around 20 % of the markers were polymorphic, yielding high-density maps of around 7000 markers (add initial Bonnema personal information). In another study an ultra-dense genetic bin map (465 bins) with 9177 SRAP markers, 1737 integrated unique Solexa paired-end sequences and 46 SSR markers representing 10,960 independent genetic loci was assembled of a F_7 RIL of 92 lines from a cross between a Yellow Sarson and a Chinese cabbage DH line (Li et al. 2011). These marker platforms make it feasible not only to generate high- density genetic maps, but also to compare different genetic maps and QTL studies. Despite the possibility to generate high-density genetic maps, in many studies the numbers of genetic markers used are still relatively low. As however biparental population sizes are often not so large (100–200 genotypes), the numbers of recombinants are also limited and these high numbers of markers are not necessary.

12.3 QTL Studies in *B. rapa*

For Quantitative Trait Locus (QTL) analyses, a wide range of biparental mapping populations including among others F_2, F_3, BC, RIL or DH have been developed, that segregate for the traits of interest (see also Table 12.1). In this paragraph, an overview of recent genetic mapping studies in *B. rapa*, aiming to identify single gene loci and QTLs for a wide range of traits, is presented and it is discussed how these studies have been facilitated since the publication of the *B. rapa* Chiifu genome sequence in 2011.

To get an impression of the number of QTL studies and whether the publication of the *B. rapa* genome in 2011 resulted in an increase in the numbers of these studies, a search was done in Scopus (http://www.scopus.com/), as the largest abstract and citation database of peer-reviewed literature. The keywords Brassica AND QTL resulted in 309 documents and the number of publications clearly increased since 2011 (till 2005 between 5 and 10 yearly, with an increase to around 24 yearly for 2007–2009, followed by an increase to almost 40 in 2012 and 2013). A similar

trend is seen when only looking at *B. rapa* and QTL (total 111 documents, around 20 papers in each 2012 and 2013). A follow up question is, whether these more recent studies were based on genetic maps with increased marker density and whether the analysis continued beyond identification of QTL and single loci, for example by postulating candidate genes, based on genome sequence and gene expression analyses. To be able to answer these questions, Table 12.1 lists QTL mapping papers published in 2011, 2012, 2013 and January 2014.

From Table 12.1 it is clear, that QTL mapping studies in *B. rapa* are published for a wide range of traits, ranging from plant morphological traits and flowering time, to seed and silique traits, yield, diverse resistances (clubroot, turnip mosaic virus, downy mildew), health related secondary metabolites (tocopherols, glucosinolates, etc.) and even food processing traits. Many of those studies identify regions explaining variation for the traits under study, and many studies also phenotype the populations repeatedly (different years or different locations). Recently most studies analyze the genome sequence under the QTL region to predict candidate genes for the traits of interest. Depending on the trait under study, candidate genes are also predicted based on similar trait variation and candidate genes in *A. thaliana*. An example is the identification of the recessive Turnip Mosaic Virus (TuMV) resistance gene *retro02* (Qian et al. 2013). This gene was fine-mapped using InDel markers in an F_2 population of 239 individuals to a 0.9-cM interval between two markers, which limited the interval to two scaffolds on chromosome A04 of the *B. rapa* genome. A candidate gene Bra035393, encoding a eukaryotic initiation factor eIF (iso) 4E protein, was predicted within the mapped resistance locus. The parental alleles were sequenced and a polymorphism (A/G) was found in exon 3 resulting in a (Gly/Asp) amino acid substitution, correlating with resistance/susceptibility. This correlation was also evident in four resistant and three susceptible lines.

Only very few studies combine QTL mapping with gene expression profiling over the individuals of the segregating population, in order to predict candidate genes. Colocation of trait QTL with gene expression QTL (e-QTL) and gene co-expression analyses can increase our understanding of the putative roles of candidate genes in trait variation. This approach was used in the genetic analysis of flowering time variation with *BrFLC2* identified as a key regulator of flowering time (Xiao et al. 2013). A similar approach was used to define a number of candidate genes for leaf developmental and plant architecture traits in *B. rapa* (Xiao et al. 2014). Genetic analysis illustrated that often QTL for plant architecture, leaf development and flowering time co-localized, suggesting pleiotropic regulation of leaf development and plant architectural traits in *B. rapa*. Four of these colocalizing phenotypic QTLs mapped at the positions of flowering time- and leaf trait candidate genes with their cis-eQTL, and cis- or trans-eQTL for homologs of genes playing a role in leaf development in *A. thaliana*. The leaf genes *B. rapa KIP-RELATED PROTEIN 2* (*BrKRP2*_A03) colocalized with QTL for leaf shape and plant height; *B. rapa ERECTA* (*BrER*_A09) colocalized with QTL for leaf color and leaf shape; *B. rapa LONGIFOLIA1* (*BrLNG1*_A10) colocalized with QTL for leaf size, leaf color, plant branching and flowering time; while the major flowering time gene, *B. rapa FLOWERING LOCUS C* (*BrFLC2*_A02) colocalized with QTL explaining variation in flowering time, plant architectural traits and leaf size. Complementation studies are needed to study the role of these genes in the phenotypic variation studied. Using the same population as the one used to study the genetic variation for flowering time and leaf development, also secondary metabolites, with special focus on glucosinolates, were studied. Combining metabolomics with transcriptomics, several candidate genes regulating glucosinolate composition in *B. rapa* leaves are predicted (Pino-Del-Carpio et al. 2014). In yet another study regulatory hotspots associated with plant gene expression under varying soil phosphorous supplies in *B. rapa* are identified and candidate genes for phosphor use efficiency mapping at these hotspots are listed (Hammond et al. 2011).

12.4 Cloning of Genes Underlying Quantitative Trait Loci

Despite the availability of the *B. rapa* genome sequence, only very few studies resulted in the actual cloning of the genes responsible for the trait, followed by functional complementation of the gene in either *A. thaliana* or in *B. rapa*. The map-based cloning and characterization of the *Crr1a* gene conferring resistance to clubroot in *B. rapa* was recently reported (Hatakeyama et al. 2013). This gene was cloned after fine-mapping using 1920 F_2 plants and identifying bacterial artificial chromosome (BAC) clones spanning the resistance locus. Final proof of the role of *BrCrr1a* in resistance to clubroot was by complementation both susceptible *A. thaliana* and *B. rapa* with the resistance allele of *Crr1a*. *Crr1a* encoded a Toll-Interleukin-1 receptor/ nucleotide-binding site/leucine-rich repeat (TIR-NB-LRR) protein. The gene for seed coat color in *B. rapa* was identified as a bHLH transcription factor *B. rapa Transparent Testa* 8 (BrTT8); a large insertion resulted in the yellow seed color. Functional complementation tests exhibited a phenotype reversion from yellow to brown seeds in the *A. thaliana* tt8-1 mutant. The TT8 gene regulates the accumulation of proanthocyanidins (PAs) in the seed coat. In another study, the *miR3191*-targeted genes that modulate head shape by differential cell division arrest in specific leaf regions of Chinese cabbage were identified (Mao et al. 2013). This is a nice example where QTL studies are combined with candidate genes based on *A. thaliana* gene function, gene expression analysis and complementation studies to identify genes underlying the heading trait. In this study 150 RILs from a cross between a Chinese cabbage and a non-heading Pakchoi accession were phenotyped for their head shape; 58 RILs produced compact heads, and 92 RILs formed small heads, or failed to form any head. The RILs with compact heads could be divided into four groups based on head shape (round, conical, cylindrical and oblong), and rosette leaf morphology correlated to each head shape. The cylindrical head shape was associated with wavy leaf margins and authors mentioned that this was reminiscent of *cin* mutants of *Antirrhinum majus* and *jaw-D* mutants of Arabidopsis. These genes are targeted by miR319, which led the authors to search miR319-targeted *BrpTCP* genes that may affect leaf curvature and head shape. At the 14th nucleotide of the miRNA binding site in *BrpTCP4-3*, a C to T substitution was identified, increasing its sequence complementary to miR319. Expression levels of *BrpTCP4-1* gene varied greatly across the 150 RILs, and also within a leaf. Transgenic plants expressing p35S::*Brp-MIR319a2* resulting in silencing of *BrpTCP* genes caused the transition of leafy heads from the round to the cylindrical shape. They concluded that the cylindrical shape of leafy head is associated with the decreased expression of *BrpTCP4*. To the best of our knowledge no other QTLs have been cloned so far (by end 2013) in *B. rapa*, while few genes underlying monogenic traits have been cloned (*BrTT8* and *BrCrr1a*). From these examples one can conclude that when appropriate segregating populations are available, and the traits under investigation can be phenotyped reliably, the *B. rapa* genome sequence has made it feasible to clone the genes underlying the traits under study. This is based on the availability of large numbers of genetic markers with physical positions since the availability of the genome sequence, the relative ease to analyze resequenced genomes with respect to the reference genome, and the possibility to inspect the genome sequence underlying the QTL. The reality is that most QTL studies do not proceed to cloning the underlying genes. The underlying reasons are not clear. On the one hand, the aim of many studies is the identification of QTL that can be exploited for crop improvement, and not ultimate cloning of the gene. On the other hand, many groups simply do not continue their efforts after identification of QTLs towards cloning the underlying genes, as number of recombinants are too less, and investments in increased population sizes is the limiting factor.

12.5 Association Mapping Studies

In Table 12.1, only QTL studies using biparental mapping populations are described. However, association mapping studies, using collections of *B. rapa* accessions/genotypes representing a wide range of genetic variation, offer additional opportunities to identify genetic markers explaining trait variation. The advantage of association mapping studies is that one can sample allelic variation present in the collection under study, and that generally recombination between loci is high. The disadvantage is that population structure can cause false-positive associations, while correction for population structure often leads to false-negative associations. In addition, variation caused by rare alleles cannot be identified, as statistical methods do not allow this. The role of both the *BrFLC2* and *BrFLC1* genes in regulation of flowering time were illustrated using association mapping studies (Yuan et al. 2009; Wu et al. 2012). Two other papers presented core collections of *B. rapa* accessions for association mapping studies and presented marker trait associations for flowering time, leaf variation and phytate content (Zhao et al. 2007, 2010). This same *B. rapa* core collection was used in a study to find association between metabolites (tocopherols, carotenoids, chlorophylls and folate) in *B. rapa* leaves and molecular markers (Pino-Del-Carpio et al. 2011a). These studies used limited numbers of markers (SSRs and AFLPs) and would greatly benefit from high-density marker platforms, like the SNP arrays described above or resequencing strategies, possibly including genome complexity reduction steps. In a recent publication, the term associative transcriptomics was launched (Harper et al. 2012). This new method uses transcriptome sequencing to identify and score molecular markers representing variation in both gene sequences and gene expression, which can then be associated with trait variation. In the allopolyploid *B. napus*, the authors identified genomic deletions in *B. napus* orthologues of the transcription factor HAG1 that underlie two quantitative trait loci for glucosinolate content of seeds.

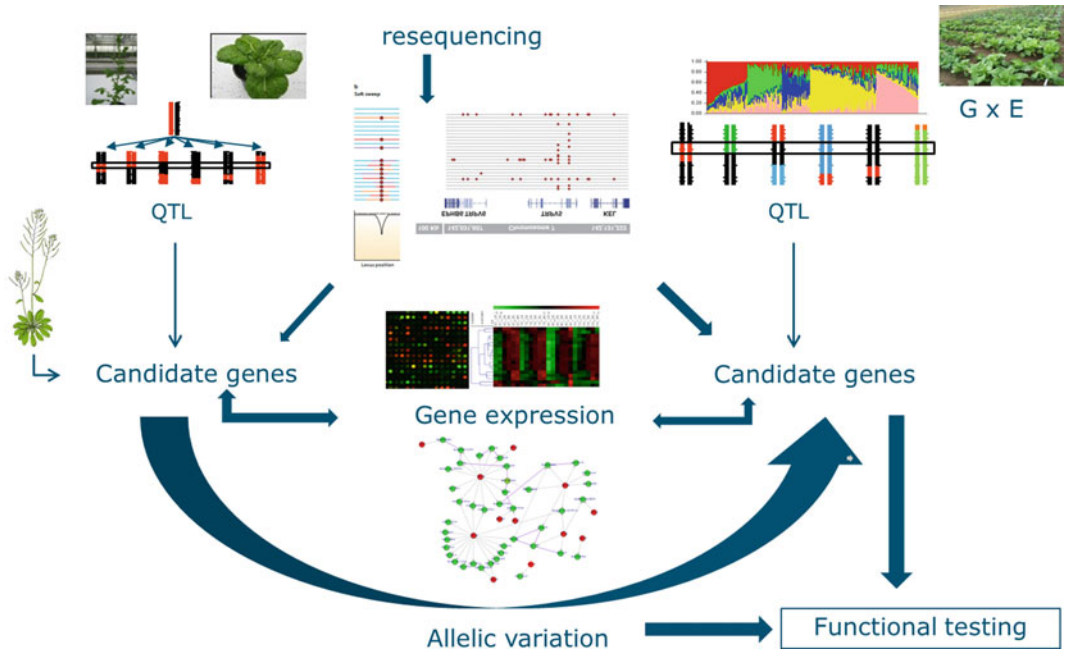

Fig. 12.1 Proposed strategy to clone and functional test genes underlying quantitative trait variation. *G* genotye, *E* environment

12.6 Concluding Remarks

The *B. rapa* genome sequence, the recently published *B. oleracea* genome sequence (Liu et al. 2014; Parkin et al. 2014), the *B. napus* genome sequence (Chalhoub et al. 2014) and the soon to be published *B. nigra* and *B. juncea* genome sequences create great opportunities to exploit published QTL studies aiming at cloning underlying genes. These genome sequences will increase our insight into genome synteny among *A. thaliana*, and the A, B and C genomes. This insight, combined with the increased use of sequence-based (high-density) molecular marker maps, makes it possible to in silico integrate published QTL studies, even across species (*B. rapa, B. oleracea, B. napus, B. juncea* and even radish). This can lead to the identification of QTL hotspots and identification of candidate genes underlying these QTLs for further studies. The same approach was already followed in a recent publication where the authors integrated in silico 1960 QTLs associated with 13 seed yield and yield-related traits from 15 *B. napus* mapping experiments over the last decade (Zhou et al. 2013). They identified conserved QTLs and multifunctional loci and predicted 146 genes underlying the QTLs for flowering time and other yield-related traits by comparative mapping with the Arabidopsis genome, which may advance fine-mapping of genes.

We propose a strategy, which is visualized in Fig. 12.1, in which both biparental QTL studies and association mapping studies are combined to identify candidate genes explaining trait variation. Resequencing genotypes under study will generate molecular markers and allelic variation of the candidate gene. Gene co-expression studies can add information about the role of the candidate genes in trait variation. As a very last step, functional testing of the genes to proof their function in *A. thaliana* and *B. rapa* needs optimization of transformation protocols for *B. rapa.*

References

Bagheri H, El-Sod M, van Oorschot I, Hanhart C, Bonnema G et al (2012) Genetic analysis of morphological traits in a new, versatile, rapid-cycling *Brassica rapa* recombinant inbred line population. Front Plant Sci 3:183

Bagheri H, El-Soda M, Kim HK, Fritsche S, Jung C et al (2013a) Genetic analysis of health-related secondary metabolites in a *Brassica rapa* recombinant inbred line population. Int J Mol Sci 14:15561–15577

Bagheri H, Pino-del-Carpio D, Hanhart C, Bonnema G, Keurentjes J, Aarts M (2013b) Identification of seed-related QTL in *Brassica rapa*. Spanish J Agric Res 11(4):1085–1093

Bonnema G, Pino-Del-Carpio D, Zhao J (2011) Diversity analysis and molecular taxonomy of Brassica vegetable crops, in: Sadowski J. (Ed), Brassica vegetables. Series: Genetics, genomics and breeding of crop plants, Kole C. (Ed.), Science publishers, Jersey, British Isles, pp 47–72

Chalhoub B, Denoeud F, Liu SY, Parkin IAP et al (2014) Early allopolyploid evolution in the post-neolithic *Brassica napus* oilseed genome. Science 345(6199): 950–953

Cheng F, Wu J, Fang L, Sun S et al (2012) Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa* (S-H Shiu, Ed). PLoS ONE 7: e36442

Cheng F, Mandáková T, Wu J, Xie Q, Lysak MA, Wang X (2013) Deciphering the diploid ancestral genome of the Mesohexaploid *Brassica rapa*. Plant Cell 25:1541–1554. doi:10.1105/tpc.113.110486

Choi SR, Teakle GR, Plaha P, Kim JH, Allender CJ et al (2007) The reference genetic linkage map for the multinational *Brassica rapa* genome sequencing project. Theor Appl Genet 115:777–792

Chung H, Jeong YM, Mun JH, Lee SS, Chung WH, Yu HJ (2013)Construction of a genetic map based on high-throughput SNP genotyping and genetic mapping of a TuMV resistance locus in *Brassica rapa*. Mol Gen Genom 289:149–160

Dechaine J, Brock M (2014) Quantitative trait loci× environment interactions for plant morphology vary over ontogeny in *Brassica rapa*. New Phytol 201:657–669

Edwards CE, Ewers BE, Williams DG, Xie Q, Lou P, Xu X, McClung CR, Weinig C (2011) The genetic architecture of ecophysiological and circadian traits in *Brassica rapa*. Genetics 189:375–390

Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet 126:1–11

Ge Y, Ramchiary N, Wang T, Liang C, Wang N et al (2011) Mapping quantitative trait loci for leaf and

heading-related traits in Chinese cabbage (*Brassica rapa* L. ssp. pekinesis). Hort Environ Biotech 52:494–501

Ge Y, Wang T, Wang N, Wang Z, Liang C et al (2012) Genetic mapping and localization of quantitative trait loci for chlorophyll content in Chinese cabbage (*Brassica rapa* ssp. pekinensis). Scientia Hortic 147:42–48

Hammond JP, Mayes S, Bowen HC, Graham NS, Hayden RM et al (2011) Regulatory hotspots are associated with plant gene expression under varying soil phosphorus supply in *Brassica rapa*. Plant Phys 156:1230–1241

Harper AL, Trick M, Higgins J, Fraser F, Clissold L et al (2012) Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. Nat Biotech 30:798–802

Hatakeyama K, Suwabe K, Tomita RN, Kato T, Nunome T et al (2013) Identification and characterization of Crr1a, a gene for resistance to clubroot disease (Plasmodiophora brassicae Woronin) in *Brassica rapa* L. PLoS ONE 8:e54745

Hoenecke ME, Sernyk JL (1992) A genetic linkage map of restriction fragment length polymorphism loci for *Brassica rapa* (syn *campestris*). Genome 35:746–757

Iniguez-Luy FL, Lukens L, Farnham MW, Amasino RM, Osborn TC (2009) Development of public immortal mapping populations, molecular markers and linkage maps for rapid cycling *Brassica rapa* and *B. oleracea*. Theor Appl Genet 120:31–43

Jin M, Lee S-S, Ke L, Kim JS, Seo M-S, Sohn S-H, Park B-S, Bonnema G (2014) Identification and mapping of a novel dominant resistance gene, TuRB07 to Turnip mosaic virus in *Brassica rapa*. Theor Appl Genet 127 (2):509–519

Kato T, Hatakeyama K, Fukino N, Matsumoto S (2013) Fine mapping of the clubroot resistance gene CRb and development of a useful selectable marker in *Brassica rapa*. Breed Sci 63:116–124

Kebede B, Cheema K, Greenshields DL, Li C, Selvaraj G et al (2012) Construction of genetic linkage map and mapping of QTL for seed color in *Brassica rapa*. Genome 55:813–823

Kim JS, Chung TY, King GJ, Jin M, Yang TJ et al (2006) A sequence-tagged linkage map of *Brassica rapa*. Genetics 174:29–39

Kim H, Choi SR, Bae J, Hong CP, Lee SY et al. (2009) Sequenced BAC anchored reference genetic map that reconciles the ten individual chromosomes of *Brassica rapa*. BMC Genomics 10:432

Kitamoto N, Yui S, Nishikawa K, Takahata Y, Yokoi S (2013) A naturally occurring long insertion in the first intron in the *Brassica rapa* FLC2 gene causes delayed bolting. Euphytica 196:213–223

Kole, P. Kole, R. Vogelzang, TC Osborn (1997) Genetic linkage map of a *Brassica rapa* recombinant inbred population. J Heredity 88:553–557

Lagercrantz U, Lydiate DJ (1996) Comparative genome mapping in Brassica. Genetics 144:1903–1910

Lee J, Lim Y-P, Han C-T, Nou I-S, Hur Y (2013) Genome-wide expression profiles of contrasting inbred lines of Chinese cabbage, Chiifu and Kenshin, under temperature stress. Genes & Genomics 35 (3):273–288

Li W, Zhang J, Mou Y, Geng J, McVetty PBE et al (2011) Integration of Solexa sequences on an ultradense genetic map in *Brassica rapa* L. BMC Genomics 12:249

Li X, Chen L, Hong M, Zhang Y, Zu F et al (2012) A large insertion in bHLH transcription factor BrTT8 resulting in yellow seed coat in *Brassica rapa*. PLoS ONE 7:e44145

Li X, Ramchiary N, Dhandapani V, Choi SR, Hur Y et al (2013a) Quantitative trait loci mapping in *Brassica rapa* revealed the structural and functional conservation of genetic loci governing morphological and yield component traits in the A, B, and C subgenomes of Brassica species. DNA Res 20:1–16

Li F, Zou Z, Yong HY, Kitashiba H, Nishio T (2013b) Nucleotide sequence variation of GLABRA1 contributing to phenotypic variation of leaf hairiness in Brassicaceae vegetables. Theor Appl Genet 126:1227–1236

Liu B, Wang Y, Zhai W, Deng J, Wang H et al (2013a) Development of InDel markers for Brassica rapa based on whole-genome re-sequencing.Theor Appl Genet 126:231–239

Liu Y, Zhang Y, Xing J, Liu Z, Feng H (2013b) Mapping quantitative trait loci for yield-related traits in Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*). Euphytica 193: 221–234

Liu S, Liu Y, Yang X, Tong C, Edwards D et al (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nature Comm 5:3930

Lou P, Xie Q, Xu X, Edwards CE, Brock MT et al (2011) Genetic architecture of the circadian clock and flowering time in *Brassica rapa*. Theor Appl Genet 123:397–409

Lou P, Wu J, Cheng F, Cressman LG, Wang X, McClung CR (2012) Preferential retention of circadian clock genes during diploidization following whole genome triplication in *Brassica rapa*. The Plant Cell Online 24(6):2415–2426

Lysak MA, Koch MA, Pecinka A, Schubert I (2005) Chromosome triplication found across the tribe *Brassiceae*. Genome research 15(4):516–525

Mao Y, Wu F, Yu X, Bai J, He Y (2013) miR319a-targeted BrpTCP genes modulate head shape in *Brassica rapa* by differential cell division arrest in leaf regions. Plant Physiol 164:710–720

Muers M (2011) Technology: getting Moore from DNA sequencing. Nat Rev Genet 12:586

Paritosh K, Yadava SK, Gupta V, Panjabi-Massand P, Sodhi YS et al (2013) RNA-seq based SNPs in some agronomically important oleiferous lines of *Brassica rapa* and their use for genome-wide linkage mapping and specific-region fine mapping. BMC genomics 14:463

Parkin IAP, Gulden SM, Sharpe AG, Lukens L, Trick M, et al. (2005) Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana.* Genetics 171:765–781

Parkin IA, Koh C, Tang H, Robinson SJ, Kagale S et al (2014) Transcriptome and methylome profiling reveals

relics of genome dominance in the mesopolyploid *Brassica oleracea*. Genome Biol 15:R77

Pino Del Carpio D, Basnet RK, De Vos RCH, Maliepaard C, Paulo MJ et al (2011a) Comparative methods for association studies: a case study on metabolite variation in a *Brassica rapa* core collection (PK Ingvarsson, Ed.). PLoS ONE 6:10

Pino Del Carpio D, Basnet RK, De Vos RCH, Maliepaard C, Visser R, Bonnema G (2011b) The patterns of population differentiation in a Brassica rapa core collection. Theor Appl Genet 122:1105–1118

Pino-Del-Carpio D, Basnet RK, Arends D, Lin K, DeVos RCH et al (2014) Regulatory network of secondary metabolism in *Brassica rapa*: Insight into the glucosinolate pathway. PLoS ONE 9(9):e107123. doi:10.1371/journal.pone.0107123

Qian W, Zhang S, Zhang S, Li F, Zhang He tal (2013) Mapping and candidate-gene screening of the novel Turnip mosaic virus resistance gene retr02 in Chinese cabbage (*Brassica rapa* L.). Theor Appl Genet 126:179–188

Snowdon RJ, Friedt W (2004) Review: molecular markers in Brassica oilseed breeding: current status and future possibilities. Plant Breed 8:1–9

Snowdon RJ, Iniguez Luy FL (2012) Potential to improve oilseed rape and canola breeding in the genomics era. Plant Breed 131:351–360

Song, K M, suzuki JY, Williams MKSH, Osborn TC (1991) A linkage map of *Brassica rapa* (syn . campestris) based on restriction fragment length polymorphism loci. Theor Appl Genet 82:296–304

Suwabe K, Suzuki G, Nunome T, Hatakeyama K, Mukai Y et al (2012) Microstructure of a *Brassica rapa* genome segment homoeologous to the resistance gene cluster on Arabidopsis chromosome 4. Breed Science 62:170–177

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS et al (2012) Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. Genetics 190:1563–1574

Teutonico RA, Osborn TC (1994) Mapping of RFLP and qualitative trait loci in *Brassica rapa* and comparison to the linkage maps of B. napus, *B. oleracea*, and *Arabidopsis thaliana*. Theor Appl Genet 89:885–894

Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, Hine EE, Althoff R, Arbogast TS, Tallon LJ (2006) Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. The Plant Cell 18(6):1348–1359

Trick M, Cheung F, Drou N, Fraser F, Lobenhofer EK, Hurban P, Magusin A, Town CD, Bancroft I (2009) A newly-developed community microarray resource for transcriptome profiling in *Brassica* species enables the confirmation of *Brassica*-specific expressed sequences. BMC Plant Biology 9(1):50

Wang Y, Sun S, Liu B, Wang H, Deng J et al (2011a) A sequence-based genetic linkage map as a reference for *Brassica rapa* pseudochromosome assembly. BMC Genomics 12:239

Wang X, Wang H, Wang J, Sun R, Wu J et al (2011b) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Wang H, Wu J, Sun S, Liu B, Cheng F, et al. (2011c) Glucosinolate biosynthetic genes in *Brassica rapa*. Gene 487:135–142

Wu J, Wei K, Cheng F, Li S, Wang Q et al (2012) A naturally occurring InDel variation in BraA.FLC.b (BrFLC2) associated with flowering time variation in *Brassica rapa*. BMC Plant Biol 12:151

Xiao D, Zhao JJ, Hou XL, Basnet RK, Carpio DPD et al (2013) The *Brassica rapa* FLC homologue FLC2 is a key regulator of flowering time, identified through transcriptional co-expression networks. J Exp Bot 64:4503–4516

Xiao D, Wang HG, Basnet RK, Zhao JJ, Lin K et al (2014) Genetic dissection of leaf development in *Brassica rapa* using a "genetical genomics" approach. Plant Physiol 164:1309–1325

Yu S, Zhang F, Zhao X, Yu Y, Zhang D (2011) Sequence-characterized amplified region and simple sequence repeat markers for identifying the major quantitative trait locus responsible for seedling resistance to downy mildew in Chinese cabbage (*Brassica rapa* ssp. pekinensis). Plant Breed 130:580–583

Yu S, Zhang F, Zhao X, Yu Y, Zhang D et al (2012) An improved *Brassica rapa* genetic linkage map and locus-specific variations in a doubled haploid population. Plant Mol Biol, Reporter 31:558–568

Yu X, Wang H, Zhong W, Bai J, Liu P, He Y (2013) QTL Mapping of leafy heads by genome resequencing in the RIL population of *Brassica rapa* (R Wu, Ed.). PLoS ONE 8:e76059

Yuan YX, Wu J, Sun RF, Zhang XW, Xu DH et al (2009) A naturally occurring splicing site mutation in the *Brassica rapa* FLC1 gene is associated with variation in flowering time. J Exp Bot 60:1299–1308

Zang YX, Kim HU, Kim JA, Lim MH, Jin M et al (2009) Genome-wide identification of glucosinolate synthesis genes in *Brassica rapa*. The FEBS J 276:3559–3574

Zhang J, Li H, Zhang M, Hui M, Wang Q et al (2013) Fine mapping and identification of candidate Br-or gene controlling orange head of Chinese cabbage (*Brassica rapa* L. ssp. pekinensis). Mol Breed 32:799–805

Zhao J, Wang X, Deng B, Lou P, Wu J et al (2005) Genetic relationships within *Brassica rapa* as inferred from AFLP fingerprints. Theor Appl Genet 110:1301–1314

Zhao J, Paulo M-J, Jamar D, Lou P, Van Eeuwijk F et al (2007) Association mapping of leaf traits, flowering time, and phytate content in *Brassica rapa*. Genome 50:963–973

Zhao J, Artemyeva A, Del Carpio DP, Basnet RK, Zhang N et al (2010) Design of a *Brassica rapa* core collection for association mapping studies. Genome**53**:884–898

Zhou QH, Fu DH, Mason AS, Zeng YJ, Zhao CX et al (2013) In silico integration of quantitative trait loci for seed yield and yield-related traits in *Brassica napus*. Mol Breed 33:881–894

Impact Molecular Marker and Genomics-Led Technologies on *Brassica* Breeding

13

Jianjun Zhao

Abstract

At the early generations of plant breeding, landrace cultivars were developed by selection from favorable variations in traits of interest including yield, and resistance to diseases, and other traits. Now new technologies including hybridization and recently developed molecular tools have been developed, which are speeding up the modern commercial plant breeding program. Herein, we present an review on DNA marker development and its wide utility in marker-assisted breeding (MAS) and genomic selection. The review of marker-assisted selection in *Brassica rapa* is summarized and discussed.

13.1 Features of *Brassica* Breeding

The *Brassica* genus comprises a number of economically important species among which the three elementary diploid species are *B. rapa* ($2n$ = 20; genome composition AA), *B. nigra* ($2n$ = 16; genome composition BB) and *B. oleracea* ($2n$ = 18; genome composition CC), and the three amphidiploids *B. juncea* ($2n$ = 36; genome composition AABB), *B. napus* ($2n$ = 38; genome composition AACC) and *B. carinata* ($2n$ = 34; genome composition BBCC). The latter three species originated through interspecific hybridization between any two of the three diploid species. The relationship between the different genomes is clearly outlined in the well-known "Triangle of U" by U in 1935 (UN 1935).

In *Brassica* breeding systems, strong heterosis has long been shown in these *Brassica* crops. Oil seed crops such as *B. napus* and *B. juncea* show a high percent of heterosis in different crosses (Fu et al. 1990; Jain et al. 1994). Vegetables such as Chinese cabbage of *B. rapa* and cabbage of *B. oleracea* also exhibit high heterosis. The paleohexaploid crop *B. rapa* displays extreme morphological diversity, and includes leafy vegetables, turnips, and oil types that all differ based on which organs are consumed (Zhao et al. 2005; Bonnema et al. 2011). In general, F_1 hybrid breeding is useful for all *Brassica* crops.

Doubled haploid (DH) technology has been widely applied to generate inbred lines and self-incompatibility (SI) has been used in

J. Zhao (✉)
College of Horticulture, Hebei Agricultural University, Lekai Southern Street 2596, 071001 Baoding, China
e-mail: jjz1971@aliyun.com

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_13

vegetable *Brassicas* successfully to produce F_1 hybrids. However, self-incompatibility is not always stable, and may be suppressed by high temperature or drought. Cytoplasmic male sterility (CMS) is another technology applicable to hybrid production, which is stable and applicable to all *Brassica* crops. CMS has been covered in several excellent reviews (Yamagishi and Bhat 2014). Studies suggest that CMS in lines of different origins may have common molecular mechanisms. A similar resemblance may also be found among nuclear fertility restorer (Rf) genes, and with multiple ways of fertility restoration through evolution of different Rf genes.

In general, breeding of *Brasscia* aims at increasing yield, improving agronomic characteristics and improving quality. In oil types, one important task in breeding programs is to increase seed oil content and seed yield, although it is difficult to achieve these two simultaneously. In vegetables *Brassica* breeding programs have different objectives and priorities since each vegetable type is characterized by its own characteristics. The market demands are considered by breeders in designing the most desirable ideotype, like Chinese cabbage with ovate or cylindrical heads favored in different geographical regions. Bolting resistance is an important breeding aim to enable year-round heading Chinese cabbage production. Disease resistant varieties are also very much needed. Clubroot, caused by *Plasmodiophora brassicae*, is one of the most damaging diseases in *Brassica* crops because the majority of commercial *B. rapa* cultivars is very susceptible, which implies that breeding for resistance has a high priority. DNA marker technology can speed up the traditional breeding programs, which is using in practice breeding.

13.2 Marker-Assisted Breeding

Molecular markers are widely used in plant breeding and genetic research, offering great promise for plant breeding. By using of DNA markers in plant breeding is called marker-assisted selection.

13.2.1 Types of DNA Markers

Commonly used DNA markers are restriction fragment length polymorphism (RFLP), single sequence repeat (SSR), rapid amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), insertion/deletion polymorphism (InDel) and single nucleotide polymorphism (SNP). These makers can help plant breeders select more efficiency for desirable crop traits. However each kind of marker is not always advantages for specific purposes, one should be careful to make benefits analysis relative to conventional breeding method. The most widely used markers are SSRs or microsatellites, with highly reliable, co-dominant, simple and cheap to use and generally highly polymorphic. And it can easily be amplified by polymerase chain reaction (PCR) using primers designed from flanking sequences of the SSR motifs. SNP markers derived from specific DNA sequences of markers that are linked to a gene or quantitative trait locus (QTL) are also extremely useful for marker-assisted breeding (MAS) (Sanchez et al. 2000), with the majority being biallelic, and can be tightly linked to or are the actual cause of allelic differences in traits (Ashrafi et al. 2012).

What type of markers will be used in MAS? Several considerations including reliability, quantity and quality of DNA required, technical procedure for marker assay, level of polymorphism, and cost could be taken into account (Mohler and Singrun 2004). Genetic distance between markers and target loci is preferred less than 5 cM, in which flanking markers can predict phenotype and will increase the reliability of the markers. In practice breeding, large amounts and high quality of DNA is necessary, which will add to the cost of the procedures. In techniques, the level of simplicity and the time required are critical considerations, which should meet the request of crops transplanting. Furthermore, highly polymorphic markers in elite breeding

material should be developed. Of course, the cost of marker assay must be reasonable in order for MAS to be feasible.

13.2.2 Marker Development

Normally two strategies of bulked segregant analysis (BSA) and Map/QTL based are used in markers development.

In BSA marker development, traits are known as dominant or recessive single genes. In a segregated population, individuals with extreme trait (e.g., resistant and susceptible individuals) are pooled and genotyped. Recently new method of MutMap was developed and applied to the identification of genes responsible for agronomically important traits (Abe et al. 2012). The MutMap is based on selecting mutants of interest at M3–M5 generations and crossing them to the parental line, followed by evaluation of phenotypes in segregating F_2 progeny. Fekih et al. developed MutMap$^+$ (Fekih et al. 2013), a versatile extension of MutMap that is based on selfing of heterozygous plants showing wild-type phenotype and identified in M2 progeny segregating for wild-type and a mutant phenotype of interest that is recessive homozygous.

In Map/QTL marker development, polygenic traits are considered. The detection of genes or QTLs controlling traits is possible due to genetic linkage analysis, which is based on the principle of genetic recombination during meiosis (Tanksley 1993). A number of well phenotyped individuals are fingerprinted with hundred(s) of molecular markers. Association of markers will be established and traits can be positioned on the constructed genetic linkage map. Marker development pipeline includes several steps reviewed by Collard and Mackill (2008). Firstly, population development and good phenotyping are important and will take the most time. Segregating populations such as temporary populations (F_2, F_3 or backcross BC) and permanent populations (recombinant inbreds and doubled haploids) are frequently used. Secondly, linkage map is constructed and then QTL mapping is performed. Markers associated with QTLs can be directly useful in MAS. Thirdly, QTL validation/confirmation has become widely accepted in recent years. This step involves confirmation of position and effect of QTLs, verification of QTLs in independent populations and testing in different genetic backgrounds, and/or fine mapping. Fourthly, marker validation steps involve testing of markers in important breeding material, identifying of 'toolbox' of polymorphic markers within a 10 cM 'window' spanning and flanking a QTL, and converting markers into a form of easy detection. Benefits of using markers in MAS compared with conventional phenotypic selection, advantages of early selection at seedling stage, simpler screening, and single plant selection can be exploited by breeders to accelerate the breeding process. In practice, breeder should also think the limitations of MAS such as startup expenses and labor costs, false positives because of recombination between the marker and the gene of interest, reliability and easier to use.

In *B. rapa*, the UniGenes database (http://www.ncbi.nlm.nih.gov/unigene/) is used often to develop gene target expressed sequence tags (EST) and SSRs marker from genes with unique identity and position in the genome. The publication of the genome sequence of *B. rapa* (http://brassicadb.org/brad/index.php) has further facilitated the mining of UniGene-derived microsatellite and SNP markers that could serve as unique, locus-specific and functional markers. One example of developing genetic markers for genes involved in flowering time regulation is published recently (Xiao et al. 2013), in which a total of 190 gene/paralog-specific primer pairs designed to amplify *B. rapa* genes amplified polymorphic bands between the parents of the one DH population.

13.2.3 Application *of MAS in Breeding*

Collard and Mackill described the main uses of DNA markers in plant breeding including five broad areas (Collard and Mackill 2008): marker-assisted evaluation of breeding material, marker-assisted backcrossing, gene pyramiding,

early generation selection, and combined MAS. In practice, marker-assisted backcrossing for a single gene and for multiple genes is widely used in hybrid breeding.

13.2.3.1 Marker-Assisted Backcrossing for a Single Gene

Backcrossing has been a widely used technique in plant breeding for almost a century. The use of DNA markers in backcrossing (marker-assisted backcrossing, or MAB) can greatly increase the efficiency of selection when applied in a proper way. MAB is the simplest form of MAS, in which the goal is to incorporate a major gene (targeted gene) from an agronomically inferior source (the donor parent, DP) into an elite cultivar or breeding line (the recurrent parent, RP). In most cases, the RP used for backcrossing has a large number of desirable attributes but is deficient in only a few characteristics. The desired outcome is a line containing only the major gene from the donor parent, with the recurrent parent genotype present everywhere else in the genome.

Three general levels (foreground, recombinant, and background selection) of MAB can be described (Holland 2004), which is same as the two types (foreground and background selection) recognized by Hospital (2003). The first level is referred to as 'foreground selection', in which the breeder selects plants having the marker allele of the donor parent at the target locus. The best plants with reproductive-stage traits can be identified in the seedling stage for backcrossing. Furthermore, recessive alleles can be selected, which is difficult to do using conventional methods. In the broad sense, the second and third levels belong together to the 'background selection', referring to the use of tightly linked flanking markers for recombinant selection and unlinked markers to select for the RP.

The second level refers to as 'recombinant selection' in a strict sense, involving selecting backcrossing individual with the target gene and recombination events between the target locus and linked flanking markers. It is important to reduce the size of the donor chromosome segment containing the target locus. The rate of decrease of the donor fragment is slower than for unlinked regions and many deleterious genes may be linked to the target gene from the donor parent (linkage drag), which can be eliminated by using markers linked to a target gene instead of using conventional selection. The second level selection is usually performed using at least two backcross generations.

The third level of MAB refers to as 'background selection' in the strict sense, involving selecting backcross progeny with the greatest proportion of RP genome. Background markers used in selection are random markers that are unlinked to the target gene/QTL on all other chromosomes, in which the RP recovery can be greatly accelerated and can be achieved by BC4, BC3 or even BC2.

In practice, both foreground and background selections are often conducted in the same backcross program, either simultaneously or sequentially. The efficiency of MAB depends on a number of factors, including the population size of each backcross generation, distance of markers from the target locus, and number of background markers used. In a typical MAB program (Hospital 2003), in each backcross generation heterozygotes were selected at the target locus. Recurrent parent alleles were selected at markers flanking the target locus (2 cM on either side) and at three markers on each nontarget chromosome. Results show faster recovery of the recurrent parent genome with MAS as compared to conventional backcrossing when foreground and background selection are combined. The recurrent parent genome is recovered more slowly on the chromosome carrying the target locus than on other chromosomes because of the difficulty in breaking linkage with the target donor allele. Methods for optimizing sample sizes and selection strategies in MAS are discussed by Bonnett et al. (2005), Frisch and Melchinger (2001), and Frisch et al. (1999a, b).

In *B. rapa*, many genetic maps have been constructed using a range of marker types and different mapping populations, used for QTL studies of root morphology (Lu et al. 2008), downy mildew (Yu et al. 2009), flowering time and leaf morphology (Lou et al. 2007; Li et al. 2009), hairiness and seed coat color

(Zhang et al. 2009), glucosinolate traits (Lou et al. 2008), and head-forming traits (Kubo et al. 2010). Recently an integrated map was constructed containing 540 markers (226 UniGene-derived microsatellite, 309 SSRs, and five additional markers) that covered a total length of 1086.6 cM (Wang et al. 2014). Single-copy markers were further identified for future applications in MAS of important economic traits. However, there is limited information about MAS in practical breeding. More markers linked to clubroot resistance (CR) in *B. rapa* are developed and have a potential to be applied in MAS. In a recent publication (Zhang et al. 2014), the *CRb* gene conferring resistance to clubroot disease is fine-mapped in *B. rapa* and the development of CR markers is summarized. In total eight CR loci have been identified and designated as the series *Crr1–Crr4*, *CRa–CRc* and *CRk*. These CR genes have been mapped on the chromosomes of *B. rapa*: *CRa*, *CRb*, *CRk* and *Crr3* reside on chromosome A03; *Crr1*, *Crr2*, *Crr4*, and *CRc* lie on chromosomes A08, A01, A06, and A02, respectively. Some markers linked to these loci have been successfully used in MAS (Piao et al. 2010) and in pyramiding of three CR genes (Matsumoto et al. 2012) in Chinese cabbage breeding. In 2012, the group of Zhongyun Piao reported that the clubroot resistance gene *CRb* from donor parent 'CR Shinki DH line' was introgressed into Chinese cabbage inbred line '91-12B' as recurrent parent by advanced backcross program and four-stage of marker-assisted selection. In this research, 7 near isogenic lines (NILs) carrying the *CRb* gene were selected (Zhang et al. 2012).

13.2.3.2 Marker-Assisted Backcrossing for Multiple Genes

A trait with complex variation is controlled by several genes, which are influenced by several genetic and environmental factors. Despite the success in polygene mapping, using markers to select for multiple genes is not straightforward, and less proven, than selection for a single gene.

Flint-Garcia et al. (2003) compared the efficiency of phenotypic recurrent selection versus MAS for multiple QTL markers (2nd generation European corn borer 2-ECB and rind penetrometer RPR resistance) in maize. Results indicated that MAS was effective in selecting for both resistance and susceptibility to 2-ECB, but not always as effective as phenotypic selection. In some cases, MAS was effective in moving the population in one direction, but not in the other. These results demonstrated that MAS can be an effective selection tool for both RPR and 2-ECB resistance.

In MAS, pyramiding is the process of combining several genes together into a single genotype. MAS-QTL pyramiding approach is based on a strategy to efficiently accumulate beneficial QTLs in a single line. The most widespread application for pyramiding has been for combining multiple disease resistance genes. Castro et al. (2003) provided an example of the combination of quantitative resistance by pyramiding of a single stripe rust gene and two QTLs in barley. Preliminary results indicated combining qualitative and quantitative resistance genes improved resistance levels in the presence of a virulent race of the pathogen.

The value of alleles from wild relatives was demonstrated in a MAS study for blackmold resistance in tomato (Robert et al. 2001). Five QTL alleles for resistance, previously detected in wild *Lycopersicon cheesmanii*, were backcrossed into a cultivated tomato background and the backcross progenies were evaluated. Three of the five alleles were effective in reducing disease severity; however, only one of the effective alleles was not associated with negative horticultural traits. The authors proposed fine mapping studies to determine if markers could be used to separate resistance from the undesirable traits.

Zong et al. (2012) proposed a novel QTL pyramid breeding scheme with marker assisted and phenotype selections (MAPS) in rice, which could significantly reduce the workload and improve the efficiency of conventional phenotype selection. This scheme allowed pyramiding of as many as 24 QTLs at a single hybridization without massive cross work. In this research, QTLs were validated the effectiveness by using the chromosome segment substitution lines

(CSSLs). On the purpose of breeding high-yielding rice varieties, a QTL pyramid breeding strategy was employed. The MAPS crop breeding scheme was further proposed, in which three procedures were involved. Phase I represented selection of parental lines using conventional phenotypic selection. Phase II depicted the process of QTL identification for target traits by setting up the genetic linkage map using several hundred F_2 plants. After QTL identification, PCR-based markers, which would represent polymorphism closely linked to the selected QTL peaks, would be prepared for the following MAS procedure. Phase III represented the process of QTL pyramiding. Several thousand F_2 seedlings would be planted for genotype selection using PCR-based markers mentioned above. After genotype selection, several hundred positive-effect-QTL-containing lines, which could be heterozygous or homozygous, would be then transferred into the field for phenotype selection. After phenotype selection, there might be only less than 100 lines matching breeder's requirements. Seeds of each selected line would constitute the F_3 population for further genotype and phenotype selection. The same processes of genotype and phenotype selection described above would be repeated for subsequent generations until genotypes of all the selected positive-effect QTLs became homozygous. Then genotype selection would be stopped and the phenotype selection should be continued until the new elite lines with better traits no longer showed segregation.

In Chinese cabbage (*Brassica rapa* ssp. *pekinensis*), the research of Matsumoto et al. (2012) proved that CR can be reinforced through the accumulation of varied resistance genes. Five DH CR lines with an individual CR locus were crossed with each other. A subsequent selection for resistance was performed using sequence characterized amplified region markers in segregating generations. Finally, four homozygous lines for three resistance genes (*CRa*, *CRk*, and *Cite*) and the F_1 hybrids between them were developed. CR homozygous pyramiding lines for three CR genes exhibited exceedingly high resistance against all of 6 field isolates.

13.2.3.3 Marker-Assisted Evaluation of Breeding Materials

Prior to crossing hybridization and breeding line development, there are several applications in which DNA marker data may be useful for breeding, such as cultivar identity, assessment of genetic diversity and parent selection, and confirmation of hybrids.

Most of the commercial varieties are hybrids, in their seeds production DNA markers have been used to define heterotic groups that can be used to exploit heterosis. Heterotic groups and patterns are of fundamental importance in hybrid breeding. In 2013, Reif et al. (2003) evaluated the usefulness of SSR markers for defining heterotic groups and patterns in subtropical germplasm, and examined applications of SSR markers for broadening heterotic groups by systematic introgression of other germplasm. For intermediate and early maturity subtropical germplasm, two heterotic groups could be suggested consisting of a flint and dent composite. The relationships between the populations obtained by SSR analyses are in excellent agreement with pedigree information. However, it is not yet possible to predict the exact level of heterosis based on DNA marker data.

In practice breeding, the maintenance of high levels of genetic purity is essential in hybrid production. The development of inbred lines for use in producing superior hybrids is a very time-consuming and expensive procedure. DNA Markers can be used to confirm the true identity of individual plants. In hybrid rice, Yashitola et al. (2002) used SSR and STS markers were used to confirm purity, which was considerably simple and feasible.

Genetic relationship and genetic variation within breeding material can greatly help the breeder in identifying a superior genotype that can be released as a new cultivar to farmers for commercial production. Broadening the genetic base of core breeding material requires the identification of diverse strains for hybridization with elite cultivars (Reif et al. 2005). Accurate assessment of the levels and patterns of genetic diversity can be invaluable in crop breeding for diverse applications including analysis of genetic

variability in cultivars, identifying diverse parental combinations to create segregating progenies with maximum genetic variability for further selection, and introgressing desirable genes from diverse germplasm into the available genetic base (Mohammadi and Prasanna 2003). In breeding, core selection increasingly uses molecular marker-based dissimilarity and clustering methods, under the implicit assumption that markers and genes of interest are genetically correlated. In practice, low marker densities mean that genome-wide correlations are mainly caused by genetic differentiation, rather than by physical linkage (Heerwaarden et al. 2013).

Besides the method of phenotypical and agronomical data, molecular markers are usually the most efficient way of estimating genetic relationships and genetic variation. A number of methods are currently available for analysis of genetic diversity in germplasm accessions, breeding lines, and populations, including similarity matrix, genetic distance analysis (GDA), principal component analysis (PCA) and principal coordinate analysis (PCoA), clustering methods and dendrogram construction. The benefits and possible applications in your breeding program depends on the objective(s) of the experiment, the level of resolution required, the resources and technological infrastructure available, and the operational and time constraints (Mohammadi and Prasanna 2003).

In *B. rapa*, the genetic relationships among 161 accessions representing all different morphotypes from geographical locations worldwide were studied using AFLP fingerprinting. Cluster analysis revealed groups that often corresponded to cultivar groups, and most interestingly indicated that different morphotypes are often more related to other morphotypes from the same region than to similar morphotypes from different regions (Bonnema et al. 2011). The same data set was imported into the program STRUCTURE that uses a Bayesian approach to cluster accessions; four subpopulations of different sizes were revealed. The grouping of all accessions in these four subpopulations were used as a reference in a later study (Zhao et al. 2010), five subpopulations were further identified for 239 accessions, in which the oil accessions from VIR (Vavilov Research Institute of Plant Industry in Russia) formed a fifth group together Japanese turnips and Pakistani winter oils.

13.3 Genomic Selection Breeding

Rapid developments in next generation sequencing (NGS) technologies over the last decade have opened up many new opportunities to explore the relationship between genotype and phenotype with greater resolution than ever before. As the cost of sequencing has decreased, breeders have begun to utilize NGS with increasing regularity to sequence large populations of plants, increasing the resolution of gene and QTL discovery and providing the basis for modeling complex genotype-phenotype relationships at the whole-genome level (Varshney et al. 2014). Genomic selection (GS), as a new application of MAS, is based on the simultaneous estimation of effects on the phenotype of all available loci, haplotypes, and markers without a previous selection of markers with effects on the phenotype (Jannink et al. 2010). Rather than seeking to identify individual loci significantly associated with a trait, GS uses all marker data as predictors of performance and consequently delivers more accurate predictions. It means that genome-wide marker genotyping is used for GS rather than selected markers showing significant associations with the traits of interest. This is a very useful tool, which can be used for fast purification of parent lines or for the selection of recombinants.

In practice, GS is applied in a population that is different from the reference population in which the marker effects were estimated. Genomic selection uses two types of datasets: a training set and a validation set. The training set is the reference population in which the marker effects were estimated. The validation set contains the selection candidates (derived from the reference population) that have been genotyped (but not phenotyped) and selected based on marker effects estimated in the training set. Since marker

technology is continuously reducing the cost per data point and increasing the number of available markers. GS offers the opportunity to increase the selection gains per unit of time. Selection can be based on GS predictions, potentially leading to more rapid and lower cost gains from breeding. Several groups have recently started exploring the GWS approach in both self- and cross-pollinated crops with some modifications for both types of crops (Bernardo 2010).

Association mapping including candidate gene association and genome-wide association provide a better resolution of the genetic maps thanks to the availability of markers distributed throughout of the genome (Zhu et al. 2008). Advantages of association mapping compared with QTL mapping in designed biparental segregating populations are that typically more variation can be observed, there is no need to develop segregating populations, and generally the mapping resolution is higher compared with that in populations from controlled crosses. In *B. rapa*, Zhao et al. (2007, 2010) applied association mapping for identification of genetic markers associated with leaf traits, flowering time and phosphate levels, and to compare the outcome of association mapping with QTL detected in DH populations that we developed for this purpose.

13.4 Summary and Future Outlook

DNA markers have been integrated in conventional schemes or used to substitute for conventional phenotypic selection. Currently, new high-throughput marker genotyping platforms have been developed, for example a number of assay types are commonly used in SNP genotyping (Kumar et al. 2012), which will speed MAS to be used in practice breeding. Since 2006, there have been a few success stories about the development of varieties using SNPs in publications derived from academic research (Mammadov et al. 2012). Although the private sector does not normally release details of its breeding methodologies to the public, several papers published by Monsanto (Eathington et al. 2007; Rosso et al. 2011), Pioneer Hi-bred (Zheng et al. 2008), Syngenta (Ribaut and Ragot 2007), and Dow AgroSciences (Ren et al. 2011) indicate that commercial organizations are the main drivers in the application of SNP markers in MAS (Ragot et al. 2007).

MAS has already proven valuable for backcrossing of major genes into elite parents, using both foreground and background selection. Knowledge of actual gene sequences will make MAS more powerful and informative across a range of genetic backgrounds. The use of MAS for multiple QTLs for complex traits is exploiting. With improvements in statistical methodologies and experimental design, integration of MAS for QTLs with phenotypic selection seems a reasonable approach. For certain traits time and cost savings will be a major driver of the application of MAS.

With the released genomic information and the increasing amount of molecular markers that become more available for *Brassica* crops, large amounts of polymorphic markers can do for the elucidation of complex traits. Genome or transcriptome sequences give breeders access to genes, their genomic position and function, as well as to large collections of markers that can be used for obtaining high density genetic maps, or for MAS (Tester and Langridge 2010). New analysis tools are also becoming available and friendly-use, breeders have to use these toolboxes well and need to have a clear strategy of breeding program. Reduced costs and optimized strategies for integrating MAS with phenotypic selection are needed before the technology can reach its full potential. If the effectiveness of the new designed strategies is validated, the MAS should be used more often in practice *Brassica* breeding program.

References

Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H et al (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. Nat Biotechnol 30:174–178

Ashrafi H, Hill T, Stoffel K, Kozik A, Yao J et al (2012) De novo assembly of the pepper transcriptome (*Capsicum annuum*): a benchmark for in discovery

of SNPs, SSRs and candidate silico genes. BMC Genom 13:571
Bernardo R (2010) Genome wide selection with minimal crossing in self-pollinated crops. Crop Sci 50:624–627
Bonnema G, Carpio DPD, Zhao JJ (2011) Diversity analysis and molecular taxonomy of *Brassica* vegetable crops. In: Kole C, Sadowski J (eds) Genetics, genomics and breeding of crop plants., Vegetable *Brassicas*Science Publishers, Enfield, pp 81–124
Bonnett DG, Rebetzke GJ, Spielmeyer W (2005) Strategies for efficient implementation of molecular markers in wheat breeding. Mol Breed 15:75–85
Castro AJ, Capettini F, Corey AE, Filichkina T, Hayes PM et al (2003) Mapping and pyramiding of qualitative and quantitative resistance to stripe rust in barley. Theor Appl Genet 107:922–930
Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Phil Trans Roy Soc Sr B 363 (1491):557–572
Eathington SR, Crosbie TM, Edwards MD, Reiter RS, Bull JK (2007) Molecular markers in a commercial breeding program. Crop Sci 47:S154–S163
Fekih R, Takagi H, Tamiru M, Abe A, Natsume S et al (2013) MutMap: genetic mapping and mutant identification without crossing in rice. PLoS ONE 8(7): e68529
Flint-Garcia SA, Darrah LL, McMullen MD, Hibbard BE (2003) Phenotypic versus marker-assisted selection for stalk strength and second-generation European corn borer resistance in maize. Theor Appl Genet 107:1331–1336
Frisch M, Melchinger AE (2001) Marker-assisted backcrossing for introgression of a recessive gene. Crop Sci 41:1485–1494
Frisch M, Bohn M, Melchinger AE (1999a) Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. Crop Sci 39:967–975
Frisch M, Bohn M, Melchinger AE (1999b) Comparison of selection strategies for marker assisted backcrossing of a gene. Crop Sci 39:1295–1301
Fu T, Yang G, Yang X (1990) Studies on "three line" Polimacytoplasmic male sterility developed in *Brassica napus*. Plant Breed 104:115–120
Heerwaarden J, Odong TL, Eeuwijk FA (2013) Maximizing genetic differentiation in core collections by PCA-based clustering of molecular marker data. Theor Appl Genet 126:763–772
Holland JB (2004) Implementation of molecular markers for quantitative traits in breeding programs—challenges and opportunities. In: Proceedings for the 4th international crop science congress, Brisbane, Australia, 2004; 26 September-1 October. Published on CDROM. Web site www.cropscience.org.au
Hospital F (2003) Marker-assisted breeding. In: Newbury HJ (ed) Plant molecular breeding. Blackwell Publishing, Oxford, pp 30–59
Jain A, Bhatia S, Banga SS, Prakash S, Lakshmikumaran M (1994) Potential use of random amplified polymorphic DNA (RAPD) to study the genetic diversity in Indian mustard (*Brassicajuncea* (L) Czern and Coss) and its relationship with heterosis. Theor Appl Genet 88:116–122
Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genom 9:166–177
Kubo N, Saito M, Tsukazaki H, Kondo T, Matsumoto S, Hirai M (2010) Detection of quantitative trait loci controlling morphological traits in *Brassica rapa* L. Breed Sci 60:164–171
Kumar S, Banks TW, Cloutier S (2012) SNP Discovery through next-generation sequencing and its applications. Int J Plant Genom 2012:15. doi:10.1155/2012/831460
Li F, Kitashiba H, Inaba K, Nishio T (2009) A *Brassica rapa* linkage map of EST-based SNP markers for identification of candidate genes controlling flowering time and leaf morphological traits. DNA Res 16:311–323
Lou P, Zhao JJ, Kim JS, Shen S, Dunia PDC et al (2007) Quantitative trait loci for flowering time and morphological traits in multiple populations of *Brassica rapa*. J Exp Bot 58:4005–4016
Lou P, Zhao J, He H, Hanhart C, Dunia PDC et al (2008) Quantitative trait loci for glucosinolate accumulation in *Brassica rapa* leaves. New Phytol 179:1017–1032
Lu G, Cao JS, Yu XL, Xiang X, Chen H (2008) Mapping QTLs for root morphological traits in *Brassica rapa* L. based on AFLP and RAPD markers. J Appl Genet 49:23–31
Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S (2012) SNP markers and their impact on plant breeding. Int J Plant Genom 2012:11. doi:10.1155/2012/728398
Matsumoto E, Ueno H, Aruga D, Sakamoto K, Hayashida N (2012) Accumulation of three clubroot resistance gene through marker-assisted selection in Chinese cabbage (*Brassicarapa* ssp. pekinensis). J Jpn Soc Hort Sci 81(2):184–190
Mohammadi SA, Prasanna BM (2003) Analysis of genetic diversity in crop plants—salient statistical tools and cconsiderations. Crop Sci 43:1235–1248
Mohler V, Singrun C (2004) General considerations: marker-assisted selection. In: Lorz H, Wenzel G (eds) Biotechnology in agriculture and forestry., Molecular marker systemsSpringer, Berlin, pp 305–317
Piao ZY, Wu D, Wang M, Zhang T (2010) Marker-assisted selection of near isogenic lines for clutroot resistant gene in Chinese cabbage. Acta Hort Sin 37(8):1264–1272
Ragot M, Lee M, Guimaraes E, et al (2007) Marker-assisted selection in maize: current status, potential, limitations and perspectives from the private and public sectors. Marker-assisted selection, current status and future perspectives in crops, Livestock, Forestry and Fish, pp 117–150
Reif JC, Melchinger AE, Xia XC, Warburton ML, Hoisington DA et al (2003) Use of SSRs for establishing heterotic groups in subtropical maize. Theor Appl Genet 107:947–957

Reif JC, Hamrit S, Heckenberger M, Schipprack W, Maurer HP et al (2005) Trends in genetic diversity among European maize cultivars and their parental components during the past 50 years. Theor Appl Genet 111:838–845

Ren R, Nagel BA, Kumpatla SP, et al (2011) Maize cytoplasmic male sterility (Cms) C-type restorer Rf4 gene. Molecular markers and their use. Google Patents

Ribaut JM, Ragot M (2007) Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. Exp Bot 58(2):351–360

Robert VJM, West MAL, Inai S, Caines A, Arntzen L et al (2001) Marker-assisted introgression of black-mold resistance QTL alleles from wild *Lycopersicon cheesmanii* to cultivated tomato (*L. esculentum*) and evaluation of QTL phenotypic effects. Mol Breed 8:217–233

Rosso ML, Burleson SA, Maupin LM, Rainey KM (2011) Development of breeder-friendly markers for selection of MIPS1 mutations in soybean. Mol Breed 28 (1):127–132

Sanchez AC, Brar DS, Huang N, Li Z, Khush GS (2000) Sequence tagged site marker-assisted selection for three bacterial blight resistance genes in rice. Crop Sci 40:792–797

Tanksley S (1993) Mapping polygenes. Annu Rev Genet 27:205–233

Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science 327:818–822

UN (1935) Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. Jpn J Bot 7:389–452

Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. PLOS Biol 12(6):e1001883

Wang Z, Ge Y, Jing J, Han X, Piao ZY (2014) Integrated genetic linkage map based on UGMS and gSSR markers in *Brassica rapa*. Sci Hort 179:293–300

Xiao D, Zhao JJ, Hou XL, Basnet RK, Carpio DP et al (2013) The *Brassica rapa* FLC homologue FLC2 is a key regulator of flowering time, identified through transcriptional co-expression networks. J Exp Bot 64:4503–4516

Yamagishi H, Bhat SR (2014) Cytoplasmic male sterility in *Brassicaceae* crops. Breed Sci 64:38–47

Yashitola J, Thirumurugan T, Sundaram RM, Naseerullah MK, Ramesha MS et al (2002) Assessment of purity of rice hybrids using microsatellite and STS markers. Crop Sci 42:1369–1373

Yu SC, Zhang FL, Yu RB, Zou YM, Qi JN et al (2009) Genetic mapping and localization of a major QTL for seedling resistance to downy mildew in Chinese cabbage (*Brassica rapa* ssp. *pekinensis*). Mol Breed 23:573–590

Zhang JF, Lu Y, Yuan YI, Zhang XW, Geng JF et al (2009) Map-based cloning and characterization of a gene controlling hairiness and seed coat color traits in *Brassica rapa*. Plant Mol Biol 69:553–563

Zhang T, Wu D, Zhao Z, Wang Z, Piao ZY (2012) Development of near isogenic lines for clubroot resistance in chinese cabbage and their assessment. Mol Plant Breed 10(6):722–730 (in Chinese)

Zhang T, Zhao Z, Zhang CY, Pang WX, Choi SR et al (2014) Fine genetic and physical mapping of the *CRb* gene conferring resistance to clubroot disease in *Brassica rapa*. Mol Breed 34:1173–1183

Zhao J, Wang X, Deng B, Lou P, Wu J et al (2005) Genetic relationships within *Brassica rapa* as inferred from AFLP fingerprints. Theor Appl Genet 110:1301–1314

Zhao J, Paulo MJ, Jamar D, Lou P, van Eeuwijk F et al (2007) Association mapping of leaf traits, flowering time, and phytate content in *Brassica rapa*. Genome 50(10):963–973

Zhao J, Artemyeva A, Carpio DPD, Basnet RK, Zhang NW et al (2010) Design of a *Brassica rapa* core collection for association mapping studies. Genome 53:884–898

Zheng P, Allen WB, Roesler K et al (2008) A phenylalanine in DGAT is a key determinant of oil content and composition in maize. Nat Genet 40(3):367–372

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. Plant Genome 1:5–20

Zong G, Wang A, Wang L, Liang G, Gu M et al (2012) A pyramid breeding of eight grain-yield related quantitative trait loci based on marker-assistant and phenotype selection in rice (*Oryza sativa* L.). J Genet Genom 39:335–350

The Database for *Brassica* Genome Studies—BRAD

14

Feng Cheng, Xiaobo Wang, Jian Wu and Xiaowu Wang

Abstract

More and more *Brassica* species and relative species from Brassicaceae have been sequenced along the technology improvement of seuquencing and genome assembly. Now, how to apply these bulk genomic datasets to assist the scientific research and breeding work becomes an urgent issue that needs to be addressed. To build functional and user-friendly databases that provide both the basic genome sequences and the elaborately analyzed data, as well as some frequently used tools is one of the solutions. A useful genome database for *Brassica* crops should have below features. (1) Free access to all the basic datasets, such as the genome and gene annotation files; (2) The common BLAST tool to compare all sequences available for *Brassica* species; (3) Genome visualization tool to show all kinds of genomic elements in one frame; (4) Well functional annotation of predicted genes for the newly sequenced *Brassica* genome, it's much better if links of orthologs are made between genes of *Brassica* and the model plant *Arabidopsis thaliana*; (5) Information of molecular markers, genetic maps, and population etc. of the *Brassica* species. In this chapter, we take the database BRAD (http://brassicadb.org) as an example to introduce the databases in the research field of *Brassica*.

14.1 Introduction

To assist researchers and breeders to better understand and use these genomic datasets from *Brassica* species, many *Brassica* databases have been built. For genomic studies in *Brassica*, useful *Brassica* databases are necessary to integrate or visualize these bulk datasets to assist the

F. Cheng · X. Wang · J. Wu · X. Wang (✉)
Institute of Vegetables and Flowers,
Chinese Academy of Agricultural Sciences,
Beijing 100081, China
e-mail: wangxiaowu@caas.cn

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_14

research and breeding work of related users. These databases include the *Brassica* database BRAD (http://brassicadb.org), Brassica.info (http://www.brassica.info/), BrassEnsembl (http://www.brassica.info/BrassEnsembl/index.html), BrassicaDB (http://brassica.nbi.ac.uk/BrassicaDB/), CropStoreDB (http://www.cropstoredb.org/), and BolBase (http://www.ocri-genomics.org/bolbase/index.html). Each of these databases has a different emphasis, Brassica.info serves as a platform to integrate genomic resources and release news of projects or activities on *Brassica* studies, and it also provides downloading services for some genomic data. BrassEnsembl visualizes different sets of *Brassica* genomic data under a single frame. CropStoreDB provides a practical approach to managing crop genetic data, while BolBase focuses on genomic structure comparisons in the genome of *Brassica oleracea*. Among them, BRAD is the database that aims at building a bridge between the genomes of *Arabidopsis thaliana* and those of the *Brassica* species (Cheng et al. 2011), transferring the research information of genomic studies and the bulk gene functional studies from the model species *A. thaliana* to the newly sequenced *Brassica* species.

In this chapter, we will focus on BRAD and present an overview of the major functions of this *Brassica* database, especially the most important and useful aspect of BRAD—the multiple genomes and gene synteny analyses among *Brassica* and other Brassicaceae species, such as *A. thaliana*. The introduction of BRAD here aims at informing users about what kind of data can be retrieved from the *Brassica* databases, and what kind of analysis can be performed using them. Initially, the BRAD database was built for the community to release and share bulk *Brassica rapa* genomic data. With continuous updating and research progress, BRAD evolved into an important repository for whole genome scale genomic data from all *Brassica* species and relative species in Brassicaceae. It now provides datasets of 12 genomes, such as the *Brassica* A, C, and AC genomes (Wang et al. 2011a; Liu et al. 2014), as well as other Brassicaceae species *A. thaliana*, *Arabidopsis lyrata*, *Schrenkiella parvula*, *Thellungiella halophila*, *Thellungiella salsuginea*, *Aethionema arabicum*, *Capsella rubella*, *Leavenworthia alabamica*, and *Sisymbrium irio* (Initiative 2000; Dassanayake et al. 2011; Hu et al. 2011; Wu et al. 2012; Haudry et al. 2013; Slotte et al. 2013; Yang et al. 2013), including their de novo assembled genome sequences and predicted gene models, associated annotations (InterPro, KEGG2, and SwissProt) and syntenic relationships. BRAD was also designed as an initial access point for other *Brassica* web pages and resources.

Fig. 14.1 Homepage and the navigation of BRAD. Five sections shown as the five navigation menus at the top of the website (*red box*): Browse, Search, Tools, Download, and Links

As shown by the navigation menu on the top of the homepage, there are five sections in BRAD (Fig. 14.1): Browse, Search, Tools, Downloading, and Links. The first three are the major sections, with "Browse" providing marker information for genetic maps and listing genes of important families. The "Search" section provides search functions for users to retrieve information on their interested genes, including the annotations, coding sequences, syntenic gene relationships, which is an important function of BRAD. The "Tools" section provides BLAST services, GBrowse, and its syntenic module to visualize genomic datasets. The last two sections are smaller, with "Downloading" providing accessions for bulk dataset downloads from BRAD, and "Links", which offers websites of other *Brassica* related databases. Generally, based on the information or function affiliation, the contents of these five sections can be separated into four parts: (1) orthologous genes; (2) genomic visualization; (3) molecular tools; and (4) featured data browsing. Below is an introduction to the four grouped tools in BRAD.

14.2 Orthologous Genes Among *Brassicas* and Other Brassicaceae Species

Syntenic and nonsyntenic genes are useful, especially for predicting genes in newly-sequenced genomes. BRAD helps users to share gene information from the well-studied model plant *A. thaliana* and apply it to new genomes.

14.2.1 Syntenic Paralogs and Orthologs

Syntenic gene analysis among multiple genomes in Brassicaceae is the core function and featured valuable resource in BRAD. Syntenic genes are homologs in two or more genomes that are inherited from the most recent common ancestor, and thus, these genes have maintained both the sequence's homology and the linear relationship on the chromosomes among these species. The tool SynOrths was applied to determine accurate syntenic orthologs between two genomes (Cheng et al. 2012a). SynOrths determines if two genes of two genomes are a syntenic pair using both their sequence similarity and the colinearity—the homology of their flanking genes (Fig. 14.2). In detail, SynOrths uses one genome, for example, *B. rapa* as the query genome and the others as the subject genomes. Under default parameters (m = 20, n = 100, and r = 0.2), it first identifies homologous genes between two genomes using BLASTP (E-value < 1×10^{-20} or best hits). The 20 closest genes (m = 20), flanking either side of the gene in the query *B. rapa* genome, are then compared with the 100 closest genes (n = 100) flanking either side of the gene in the subject genome. If at least 20 % (r = 0.2) of the best hits for the 40 genes (20 × 2 for both sides) in the query *B. rapa* genome are found within the 200 genes (100 × 2 for both sides) in the subject genome, then the original pair of are designated as a syntenic ortholog candidate. Syntenic genes among the 12 genomes in Brassicaceae were determined using this tool accompanied with further analysis, and they were integrated into

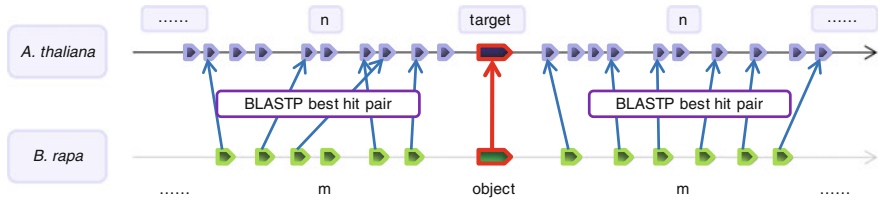

Fig. 14.2 Algorithm of syntenic gene determination in SynOrths. Both the sequence homology of the gene pair and their flanking genes are considered to determine whether they are under synteny

one sheet and embedded in BRAD as a searching function of syntenic genes. Again, using *B. rapa* as an example, 30,773 syntenic pairs between *B. rapa* and *A. thaliana* were obtained, and there were 9293, 6683, and 2346 *A. thaliana* genes, having one, two, and three paralogs, respectively, in the three subgenomes LF, MF1, and MF2 of *B. rapa* (the three subgenomes originated from a genome triplication). LF, MF1, and MF2 are abbreviations for less fractionized, more fractionized 1 and more fractionized 2, respectively, denoting subgenomes with more or fewer genes retained (Wang et al. 2011a; Cheng et al. 2012b). Now BRAD provides the gene set that shows conserved synteny between the three subgenomes of *Brassica* and *A. thaliana*, as well as other Brassicaceae species. These syntenic genes are re-sorted and listed according to the genes' order in the seven chromosomes of the *Brassica* ancestor (tPCK genome structure) (Cheng et al. 2012b). By typing any gene ID from the 12 Brassicaceae genomes, their syntenic genes in the other genomes and the homologous relationships of flanking genes can be easily obtained (Fig. 14.3). In addition, a small dialog window, which is linked to each gene ID in the output

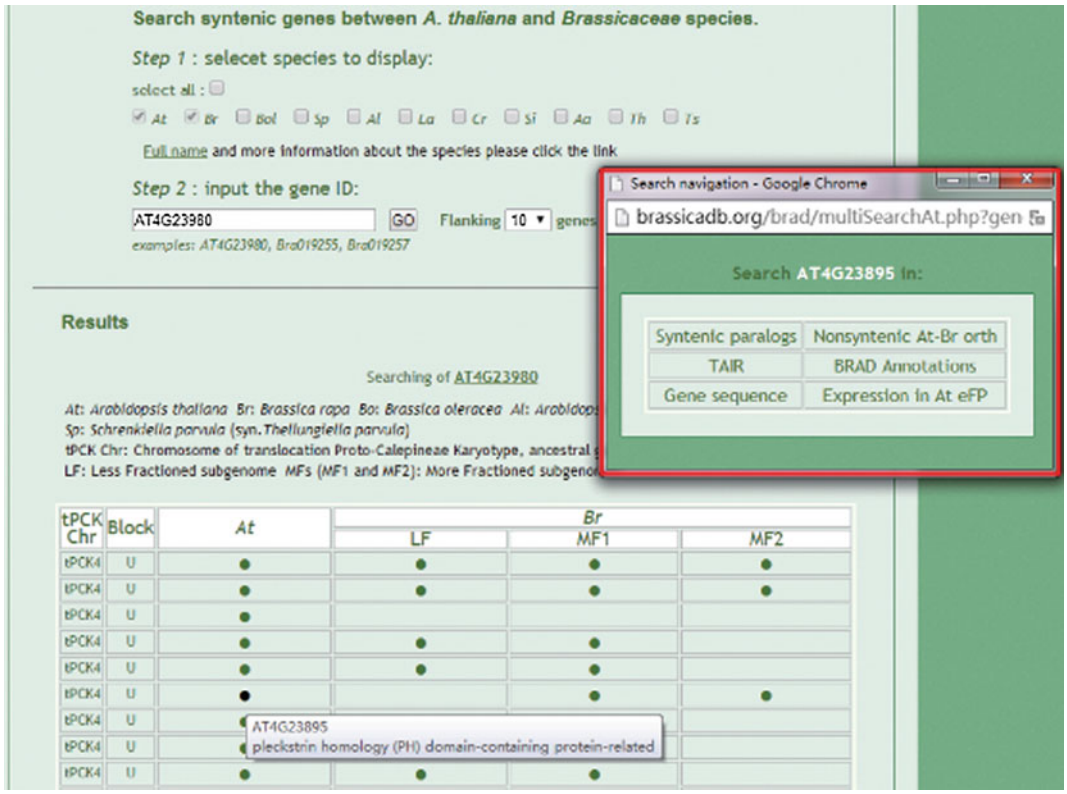

Fig. 14.3 Syntenic genes among *Brassica* and other Brassicaceae species. Taking the *A. thaliana* gene AT4G23895 as an example, search results are shown in a table. Each *solid circle* in the table represents a gene. The first column lists the chromosomes in the *Brassica* ancestral genome tPCK, followed by genomic blocks and *A. thaliana* gene IDs. The next three columns show genes from the subgenomes LF, MF1, and MF2 of *Brassica rapa*. In each *row*, listed genes are in a syntenic relationship. Moving the cursor over an *A. thaliana* gene gives a floating box containing the gene's annotation, while moving the cursor over a *B. rapa* gene causes the supporting information on its syntenic relationship to the *A. thaliana* gene to appear in a pop-up window. By clicking a gene in the table, users will open a small dialog window (*red box*) that offers links to its annotation, the best BLASTX hit to *A. thaliana*, and the function and Gene Ontology (GO) of the matching gene, as shown in Fig. 14.2

table, gives more information on the according gene (Fig. 14.3), which will help users to retrieve more useful data for their study. The dialog window will let users access information, not only in BRAD, but also in other databases, such as TAIR, for further detailed information. This dialog window allows simple and easy navigation, integrating all available information on a certain gene into one menu for users to access by clicking. This dialog function is used commonly in most pages of BRAD.

14.2.2 Nonsyntenic Orthologs

The function of searching nonsyntenic genes is a supplement of syntenic gene analyzing, which helps users to obtain homologous genes that originated through transposition events. Nonsyntenic genes in two genomes were determined under the following two rules: (1) the BLASTP alignment should be satisfied by: identity >70 %, coverage of both genes >75 %, and 2) the two genes should not be a syntenic pair. Using *B. rapa* and *A. thaliana* as examples, a total of 17,159 such nonsyntenic orthologs were determined.

14.3 Visualization of Genomic Features and Genomic Synteny

A user-friendly visualization page of genomic data will help users to understand and apply the data to easily assist their research. Here, two visualized pages were built in BRAD. However, the visualization aspect still needs improvement.

14.3.1 Genome Browse (GBrowse)

The Genome Browser tool developed by the Generic Model Organism Database Project (Donlin 2009), (http://gmod.org) was adopted by BRAD to visualize the Brassicaceae genomes. Three major levels are displayed: chromosome where the search target is located, genome segment with flanking regions of the search target, and the exact target. BRAD GBrowse now provides available information, including predicted gene models, transposons, different types of RNA sets, and genetic markers, for these species.

14.3.2 Synteny Blocks

A synteny module of GBrowse, SynBrowse, was used to visualize the relationship of genomic synteny between the 12 Brassicaceae genomes. Information on syntenic blocks was identified based on the relationship of synteny genes determined by SynOrths. The genomic fragments in which the syntenic genes retain the linear relationship between the two genomes have been merged to synteny blocks. These conserved and pairwise blocks were then visualized by SynBrowse. The genes that show in the synteny blocks were provided with links to the synteny gene searching section, which then leads users to the one-to-one gene synteny information and further resources through the dialog window.

14.4 Useful Tools for Molecular Studies

BRAD developed or adopted some useful tools and functions for molecular studies, including gene function annotations, local functional elements searching, and BLAST service.

14.4.1 Searching Genes by Keywords of Function Annotations

Six kinds of annotation datasets were provided here: SwissProt, Trembl, KEGG, InterPro domain, Gene Ontology, and the BLASTX (best hit) of *Brassica* to *A. thaliana*. These datasets are used to annotate different aspects of newly predicted gene models, such as nucleotide sequences, proteins, and domains. SwissProt and Trembl annotations are generated by BLASTP best hit

(cutoff E-value: 1e-5) based on the predicted proteins in the Swiss-Prot and TrEMBL databases. Predicted genes are mapped to KEGG pathways based on the best hit from the Swiss-Prot database. InterPro is used to annotate motifs and domains of predicted genes by comparison to public databases, including Pfam, PRINTS, PROSITE, ProDom, and SMART by using applications, such as hmmpfam, fprintscan, ScanRegExp profilescan, blastprodom, and hmmsmart. Gene Ontology information is extracted from the InterPro results. In total, in the gene annotation pages, for all the predicted genes of the 12 genomes listed above, we collected 244,836 gene ontology records, 275,161 InterPro domain annotation records, 84,568 KEGG records, 71,076 SwissProt records, 37,220 Trembl records, and 14,790 BLASTX best hits to *A. thaliana*. This bulk information will help users to have a better understanding of genes in newly sequenced genomes. The most up-to-date information can be checked through the address: http://brassicadb.org/brad/searchAll.php. Orthologs between new genomes and the model plant *A. thaliana* are also used to annotate these new genes.

14.4.2 Screening for Elements in Local Genomic Regions

This section was developed to help users to locate genomic elements that are colocated with molecular markers, other elements, or flanking the region of interest. Users can easily perform the search by inputting a physical position, a gene ID, or genetic marker, accompanied with the size of the flanking regions to be searched. All of the genomic features, such as genes, transposons, and RNAs (miRNA, tRNA, rRNA, and snRNA) that are located in the searched region will be collected and displayed in a table. A link to GBrowse provides an option to visualize the search region on the background of the complete chromosome. It is a useful function for certain studies, such as the fine mapping of QTLs. Once QTLs are obtained, existing markers from BRAD can be used directly for searching, while new markers should be aligned to the genome sequence with the BLAST tool in BRAD to locate their physical positions. The flanking regions of these markers can then be checked using this tool to locate candidate genomic elements, such as genes or small RNAs that might be the causal factors of the QTLs. As the research progresses, BRAD can further enable the searching of flanking regions by adding more datasets, making it an integrative and valuable resource pool for molecular geneticists and breeders.

14.4.3 Bulk Resources for BLAST Services

Standard wwwblast modules were adopted by BRAD to help users to perform homologous sequence alignment and searching. The BLAST databases collect and provide genomes, genes and proteins sequences, or EST sequences available for the 12 Brassicaceae species. With this tool and resources, users can screen for homologous relationships between their studied sequence and the Brassicaceae databases in BRAD easily and efficiently.

14.5 Resources of Browsing and Bulk Data Downloading

In this part, BRAD provides browsable pages of gene lists from important gene families and genetic maps. BRAD also offers access to all the datasets that are employed in BRAD for users to download.

14.5.1 Browse of Linkage Maps and Gene Families

Gene lists for important gene families, such as auxin, glucosinolate, flowering, transcription factor, and resistant genes, as well as another 182 gene families and their orthologous relationship to *A. thaliana* are provided for browsing. For

each gene ID that appears on these pages, a link to a small dialog window is provided.

Besides the gene family information, BRAD collected 1160 genetic markers, including 758 SSR and 402 InDel markers, covering all 10 chromosomes (Choi et al. 2007; Kim et al. 2009) from three population lines of *B. rapa*: RCZ16_DH, JWF3P, and VCS_DH. RCZ16_DH is a population developed from a cross between a rapid cycling line, L144, and a summer type Chinese cabbage doubled haploid (DH) line Z16 (Wang et al. 2011b). Markers of RCZ16_DH were developed based on the resequencing data of their parents L144 and Z16. The other two maps, JWF3P and VCS_DH, were integrated from public database http://www.brassica-rapa.org. This information can be browsed by clicking on links on the pages, the information is collected and displayed in a hierarchical structure, with more clicks providing users with more detailed information.

Gene family information on other Brassicaceae species and the genetic markers of other populations or *Brassica* species will be easily added to BRAD when the data is available.

14.5.2 Download and External Links

BRAD provides accessions for bulk data downloads, including genome, gene, and protein sequences, gene annotations, and other predicted genomic elements. In addition, BRAD also collects and provides numerous community resources either as data or external website links, including websites of laboratories focusing on Brassicaceae, or *Brassica* breeding.

14.6 General Guidelines for Using BRAD

14.6.1 Browse Molecular Markers and Genetic Maps

For each marker in the "Browse" section, BRAD presents its genetic and physical positions, primer information, as well as its parental populations. Users can access these data in the following order: chromosome selection → population specification → detailed marker information → click marker ID for primer information.

14.6.2 Search Using Annotations and Syntenic Genes

In the annotation search section, users can find genes with interesting functions by typing a keyword, such as flower or growth, and then relevant records will be compared from the six annotation datasets as described above. Clicking on the selected records will then lead users to genes with annotations related to the keyword. A further click of the gene ID will show users with more detailed gene information. Syntenic genes can only be searched by using gene IDs of either species stored by BRAD. In the web of syntenic paralogs, the pull-down 'flanking' menu has two options (10 or 20), which means it will extend 10 or 20 genes up- and downstream of the searched gene. In the tabulated output, such as in Fig. 14.3, the targeted gene is in the middle of its flanking genes. Each *A. thaliana* gene corresponds to 1, 2, or 3 genes in the three subgenomes of *B. rapa*. "No circle" indicates that there is no gene identified. Moving the cursor over the ID of a gene expands the functional annotations of *A. thaliana* genes and the detailed supporting information of synteny relationships of *B. rapa* genes to that of *A. thaliana*.

14.6.3 Search Navigation

As referred to above, the dialog window is a useful and easy to follow tool in BRAD. BRAD embedded this JavaScript-driven small dialog window as the navigation tool for each gene ID in any of the output tables, which help users to quickly access all the information on an interested gene in BRAD. By combining the access points for many datasets into one window, the navigation can lead users to different resources on the target genes, which facilitates the use of BRAD. The navigation window now integrates resources

such as gene annotations, syntenic or nonsyntenic orthologs, gene sequence, functional elements in the gene's flanking regions and data visualization in GBrowse, links to information of TAIR databases, and gene expression of *A. thaliana*.

14.7 Conclusion and Perspective

BRAD, a database focused on gene function illustration and multiple genome comparisons among *Brassica* and other Brassicaceae species, especially the model plant *A. thaliana*, has been built in time because the genome studies on *Brassica* species are increasing. Compared with other databases on *Brassica* plants, BRAD keeps its specific core function and advantages, especially its deep mining of syntenic genes and genomic blocks in the newly assembled *Brassica* genomes, as well as the use of the bulk research information from the model plant *A. thaliana*. By the continuous improvement of applications and the integration of more available datasets in the future, BRAD will help scientists and breeders to fully and efficiently use the information on genomic and genetic datasets of *Brassica* plants. BRAD is a valuable resource for the scientists of comparative genomics, plant evolution, and molecular biology, and the breeders of *Brassica*.

In the future, BRAD should integrate datasets of all *Brassica* species available. In addition, to meet the demands of the genetic and genomic studies on *Brassica* crops, the following data types and applications should be updated on BRAD. (1) Resequencing data of different lines in each *Brassica* species; SNP, InDel, or SV loci and their frequency in populations, as well as the haplotypes (derived from SNPs variation) of *Brassica* germplasm collections that are generated from resequencing data. (2) Transcriptome sequencing or mRNA-Seq data, gene expression data in different organs, and different accessions of each *Brassica* species. (3) Whole genome methylation and small RNA-Seq data for *Brassica* species. (4) Visualization and interactive technologies should be adopted to improve BRAD to be a much more user-friendly database. Solutions to visualize the syntenic relationships of multiple genomes from the chromosome scale, to genomic segments, to genes or tens of nucleotides level should be employed, thus helping users to have a better understanding of genome and gene evolution at different scales among different *Brassica* species and their relation to other Brassicaceae species.

References

Cheng F, Liu S, Wu J, Fang L, Sun S et al (2011) BRAD, the genetics and genomics database for Brassica plants. BMC Plant Biol 11:136

Cheng F, Wu J, Fang L, Wang X (2012a) Syntenic gene analysis between Brassica rapa and other Brassicaceae species. Front Plant Sci 3:198–380

Cheng F, Wu J, Fang L, Sun S, Liu B et al (2012b) Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. PLoS One **7**:e36442

Choi SR, Teakle GR, Plaha P, Kim JH, Allender CJ et al (2007) The reference genetic linkage map for the multinational Brassica rapa genome sequencing project. Theor Appl Genet 115:777–792

Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H et al (2011) The genome of the extremophile crucifer Thellungiella parvula. Nat Genet 43:913–918

Donlin MJ (2009) Using the generic genome browser (GBrowse). Curr Protoc Bioinform Chapter9: Unit 9 9

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M et al (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet 45:891–898

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF et al (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet 43:476–481

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

Kim H, Choi SR, Bae J, Hong CP, Lee SY et al (2009) Sequenced BAC anchored reference genetic map that reconciles the ten individual chromosomes of Brassica rapa. BMC Genom 10:432

Liu S, Liu Y, Yang X, Tong C, Edwards D et al (2014) The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nature Comunication In press

Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F et al (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Genet 45:831–835

Wang X, Wang H, Wang J, Sun R, Wu J et al (2011a) The genome of the mesopolyploid crop species Brassica rapa. Nat Genet 43:1035–1039

Wang Y, Sun S, Liu B, Wang H, Deng J et al (2011b) A sequence-based genetic linkage map as a reference for Brassica rapa pseudochromosome assembly. BMC Genom 12:239

Wu HJ, Zhang Z, Wang JY, Oh DH, Dassanayake M et al. (2012) Insights into salt tolerance from the genome of Thellungiella salsuginea. Proc Natl Acad Sci USA 109:12219–12224

Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J et al (2013) The Reference Genome of the Halophytic Plant Eutrema salsugineum. Front Plant Sci 4:46

Future Prospects

15

Xiaowu Wang and Feng Cheng

Abstract

Releasing of the *Brassica rapa* var. Chiifu genome provided a reference for genome evolution research, gene discovery and breeding of *Brassica* crops. However, because of the limitation of the current technology, there is still a great space for the genome to be improved. These include increasing of the assembled repeat sequences, anchoring of the assemblies, accuracy of the predicted gene models and annotation of functional genome elements. More genomes of relative species were released. Comparative genomics genomes should be conducted to understand the *B. rapa* genome under a wider background. The reference genome provided possibilities to explore the genetic variation in *B. rapa* by GWAS. But large scale genome wide SNPs have to be generated. Extensive application of the genome data needs also improvement of the genome database.

The release of the *Brassica rapa* var. Chiifu genome (Wang et al. 2011) is not only of importance for genome evolution research but also facilitates gene discovery and breeding of *Brassica* crops. We can now rebuild the evolutionary route of the mesopolyploid *Brassica* genome (Cheng et al. 2013) and bridge the rich knowledge obtained from *Arabidopsis* to this cultivated crop species. However, this is just a starting point for applying genomics to improve *Brassica* crops. Sequencing of the *B. rapa* genome is far from finished. The quality of the genome assembly has room for improvement owing to advances in sequencing technologies. The accuracy of the annotation can be increased with the accumulation of more RNA-seq data. The genome should be better understood by being studied from different angles. More tools and resources need to be established to successfully transfer the knowledge from *Arabidopsis* to the *Brassica* crops and help the breeders to increase the rate of breeding.

Owing to the limitations of sequencing technology and in the power of the assembly pipeline, the present version of the *B. rapa* genome is still a draft, and a large part of the genome has

X. Wang (✉) · F. Cheng
Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China
e-mail: wangxiaowu@caas.cn

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_15

not been assembled yet. The present assembly of the *B. rapa* genome was assembled using mostly paired-end reads of 75 bp. Based on theoretical estimations, the genome size is 485 Mb; however, only 285 Mb were assembled, which is 58.7 % of the predicted total genome. It was determined that the missing 200 Mb is largely repetitive sequences, which are also very important for understanding the genome (Wang et al. 2011). Even within the assembled sequences, there are a large number of gaps waiting to be filled.

The sequencing technologies are developing rapidly, and the assembling pipelines are continuously improving. The read lengths of the most widely used Illumina sequencing method has now reached 250 bp and are increasing rapidly. The assembling pipelines, such as ALLPATHS-LG (Gnerre et al. 2011), MaSuRCA (Zimin et al. 2013), SOAPdenovo2 (Luo et al. 2012), and Velvet (Zerbino and Birney 2008), can create contigs by using longer k-mers over 100 bp or even dynamic k-mers (Platanus) (Kajitani et al. 2014). This improvement can greatly increase the quality of the assembly, especially the lengths of repetitive sequences that can be assembled. The third-generation sequencing technologies, such as Pacific bio and nanopore (Branton et al. 2008; Gupta 2008; Mardis 2008; Metzker 2009; Ku and Roukos 2013), which can produce very long reads over 10 kb, are emerging and maturing. Although the accuracy of the reads is not as high as second-generation sequencing methods, the read length can improve significantly the lengths of the contigs and the assemblage quality of repetitive sequences. There is still plenty of room to improve the recent *B. rapa* genome assembly (version 1.5) by adopting new sequencing and assembling technologies.

The scaffolds anchored to the linkage groups can also be increased by using high-density maps constructed with markers produced by high-throughput sequencing. The chromosome pseudomolecular version 1.5 of the present *B. rapa* assembly was created using linkage maps constructed with a total of 1673 traditional markers (mostly SSRs and InDels). Approximately 90 % of the 285 Mb assembly was anchored. Recently, (Yu et al. 2013) reported a high-density linkage map generated by resequencing 150 recombinant inbred lines (RILs) derived from the cross between heading and nonheading Chinese cabbage. The map contained 2209 bin markers produced from more than 1 million single nucleotide polymorphisms (SNPs). Such a high quality map, and similar maps generated in the future, should greatly facilitate the anchoring of scaffolds, especially when more sequence is assembled for the next version of the *B. rapa* genome.

Improving the annotation of a reference genome is continuous work. Using gene expression data to support gene model predictions is the most important way to improve the annotation. The current versions of the gene models were mostly predicted based on the limited expressed sequence tag (EST) data generated by traditional EST sequencing. Recently a large number of high-throughput RNA-seq data have been generated (Cheng et al. 2012; Mun et al. 2012; Wang et al. 2012; Paritosh et al. 2013; Song et al. 2013; Tong et al. 2013), which provided rich resources for increasing the accuracy of gene model predictions. The abundance of RNA-seq data also provided opportunities to detect alternative transcript splicing of genes. Furthermore, technologies to isolate RNA from specific tissues or single cells are now combinable with high-throughput sequencing. This allows the detection of tissue- or cell-specific and lowly expressed mRNA, which will further improve the gene model predictions.

Repetitive sequences compose most of the *B. rapa* genome; however, they are still poorly studied. Repetitive sequences of the *B. rapa* genome have been neither well assembled nor well characterized. This part of the genome is largely hidden. The application of new sequencing technology and sequence assemblers will improve the assembling of repetitive sequences, which will allow us to better understand the structure of these sequences in the genome.

In human genomics, there is an ENCyclopedia Of DNA Elements (ENCODE) project (ENCODE 2004), which generates a variety of

whole-genome datasets that, in total, describe the sequence, alternative transcript splicing, DNA methylation, regulatory protein occupancy (CHiP-seq), small and long noncoding RNA levels, points of chromosomal contact (Hi–C), and a variety of chromatin/nucleosome occupancy characteristics. It is obvious that the human ENCODE project is extremely valuable in advancing human health. The value of ENCODE-like data for crop improvement can also be expected.

A number of Brassicaceae species, including *Arabidopsis lyrata* (Hu et al. 2011), *Capsella rubella* (Slotte et al. 2013), *Schrenkiella parvula* (syn. *Thellungiella parvula*) (Dassanayake et al. 2011), *Leavenworthia alabamica* (Haudry et al. 2013), *Sisymbrium irio* (Haudry et al. 2013), and *Aethionema arabicum* (Haudry et al. 2013) were sequenced recently. The very close relatives of *B. rapa*, *Brassica oleracea* (Liu et al. 2014) and *Brassica nigra*, and its allotetraploid relatives, *Brassica napus* and *Brassica juncea*, have also been sequenced and will be publically available soon. Tools that can visually present the comparative results and integrate functional annotations from different species will now be of vital importance. This is not only because these comparative tools can be used for transferring knowledge from species to species, but also because these tools can help investigators who are not working in the field of genomics to apply genomics in their own fields.

B. rapa has a genome that is still rapidly changing. A better understanding of the *B. rapa* species cannot be achieved using only a reference genome. There are at least two factors that drive change in the *B. rapa* genome. First, as a species with relatively recent whole-genome triplication, the *B. rapa* genome is still experiencing gene fractionation. Second, the transposons in *B. rapa* are very active. Both of these factors create large numbers of variations within the species. The concepts of pan and core genomes were adopted to describe the genome of *B. rapa* by (Lin et al. 2014). They defined a core and a pan genome in *B. rapa* after comparing the reference genome with the resequencing data of a rapid-cycling line and a turnip accession. However, this was not nearly enough information for defining an accurate core and pan genome. Increasing high-throughput sequencing methods and price reductions for sequencing data have now allowed the resequencing of a large number of accessions. Resequencing of a collection of accessions representing different *B. rapa* morphotypes and geographic origins will produce information capable of more accurately defining the core and pan genomes of *B. rapa*. However, using the Chiifu genome as a reference, we can only look at the genes present in that genome. Many genes or genetically active features are accession-specific and cannot been seen in the Chiifu reference genome. The de novo assembly of some representative accessions of different morphotypes is necessary to define more comprehensive core and pan genomes for *B. rapa*.

B. rapa is morphologically very polymorphic. Under artificial selection during domestication, it has evolved many different morphotypes with extreme morphological characteristics, such as an enlarged hypocotyl stem in turnip, enlarged leafy heads in heading Chinese cabbage, and strong axillary branching described by Mizuna et al. Detecting the genes involved in extreme traits will not only help to illustrate how artificial selection impacts a genome and shapes it into a crop, but will also be of importance in improving the agronomic traits. Large scale resequencing can detect selection signals, such as selection sweeps and F_{ST}, which can be used to pinpoint selected genes and further unravel the genetic mechanisms behind the morphological plasticity of *B. rapa.* The domestication and spreading history of *B. rapa* crops are still mysteries. With the accumulation of whole genome sequence information from a large number of *B. rapa* accessions with different origins, genomic evidence can be collected to solve these mysteries. However, unlike *B. oleracea*, in which many wild species were found in the Mediterranean region, no native wild *B. rapa* has been found. This makes the investigation of *B. rapa*'s domestication more challenging.

Association mapping, often in the form of genome-wide association study, is a tool used to

map quantitative traits based on linkage disequilibrium (LD), which was first developed in human genetics. It has the advantage of mapping quantitative traits with high resolution in a way that is statistically very powerful. To perform association mapping, the entire genome needs to be scanned for significant associations between a panel of SNPs and a particular phenotype, which requires an extensive knowledge of the SNPs within the organism of interest's genome. To take advantage of association mapping in locating loci for important traits in *B. rapa*, we have to develop tools that can detect SNPs in a high enough density to locate associating genome blocks. As a species with the feature of outcrossing and an evolutionary history of millions of years, we can expect that the LD of the species is not large, which will make it difficult to use DNA chip technology to detect enough SNPs for association mapping. To perform association mapping in *B. rapa*, high-throughput resequencing should be used to generate enough SNPs.

The *Brassica* database (BRAD) (Cheng et al. 2011) is a portal for *Brassica* genome information. BRAD has recently begun focusing on providing services for mining the genomic data of *B. rapa*, including a genome browser, synteny block information, and gene annotation data. It also hosts the genomic data of *B. oleracea* and other Brassicaceae species with published genomes. It is predicted that the genome sequences of other *Brassica*, such as *B. nigra*, *B. napus*, *B. juncea* and *B. carinata*, and species closely related to *Brassica*, such as the Raphanus species, will soon be published. Integrating their genome data into BRAD will enable the comparison of the *B. rapa* genome with the genomes of its closely related species and aid in unraveling the evolution of the complicated genomes of the *Brassica* species. There are also large-scale resequencing projects of *B. rapa* underway. Databases and tools for exploring resequencing data will be designed and included in BRAD to facilitate gene mining and variety breeding of *B. rapa* crops. The concept of a diversity-fixed foundation set has been proposed for *B. rapa*, and we believe that with the creation of such a set of genetic materials, genomics, transcriptomics, metabolomics, and even phenomics data will be produced and made publically available. Platforms should be established to integrate, "view", and explore the omics data.

References

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA et al (2008) The potential and challenges of nanopore sequencing. Nat Biotechnol 26:1146–1153

Cheng F, Liu S, Wu J, Fang L, Sun S et al (2011) BRAD, the genetics and genomics database for *Brassica* plants. BMC Plant Biol 11

Cheng F, Wu J, Fang L, Sun S, Liu B et al (2012) Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. PLoS One **7**:e36442

Cheng F, Mandakova T, Wu J, Xie Q, Lysak MA et al (2013) Deciphering the diploid ancestral genome of the Mesohexaploid Brassica rapa. Plant Cell 25:1541–1554

Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H et al (2011) The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet 43:913–918

Encode C (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. Science 306:636–640

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108:1513–1518

Gupta PK (2008) Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol 26:602–611

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M et al (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet 45:891–898

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF et al (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet 43:476–481

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y et al (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res

Ku CS, Roukos DH (2013) From next-generation sequencing to nanopore sequencing technology: paving the way to personalized genomic medicine. Expert Rev Med Devices 10:1–6

Lin K, Zhang N, Severing EI, Nijveen H, Cheng F et al (2014) Beyond genomic variation—comparison and

functional annotation of three *Brassica rapa* genomes: a turnip, a rapid cycling and a Chinese cabbage. BMC Genomics 15

Liu S, Liu Y, Yang X, Tong C, Edwards D et al (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun (in press)

Luo R, Liu B, Xie Y, Li Z, Huang W et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402

Metzker ML (2009) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46

Mun JH, Yu HJ, Shin JY, Oh M, Hwang HJ et al (2012) Auxin response factor gene family in *Brassica rapa*: genomic organization, divergence, expression, and evolution. Mol Genet Genomics 287:765–784

Paritosh K, Yadava SK, Gupta V, Panjabi-Massand P, Sodhi YS et al (2013) RNA-seq based SNPs in some agronomically important oleiferous lines of *Brassica rapa* and their use for genome-wide linkage mapping and specific-region fine mapping. BMC Genomics 14:463

Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F et al (2013) The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat Genet 45:831–835

Song XM, Liu TK, Duan WK, Ma QH, Ren J et al (2013) Genome-wide analysis of the GRAS gene family in Chinese cabbage (*Brassica rapa* ssp. pekinensis). Genomics 103:135–146

Tong C, Wang X, Yu J, Wu J, Li W et al (2013) Comprehensive analysis of RNA-seq data reveals the complexity of the transcriptome in *Brassica rapa*. BMC Genomics 14:689

Wang X, Wang H, Wang J, Sun R, Wu J et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Wang F, Li L, Li H, Liu L, Zhang Y et al (2012) Transcriptome analysis of rosette and folding leaves in Chinese cabbage using high-throughput RNA sequencing. Genomics 99:299–307

Yu X, Wang H, Zhong W, Bai J, Liu P et al (2013) QTL mapping of leafy heads by genome resequencing in the RIL population of *Brassica rapa*. PLoS ONE 8: e76059

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL et al (2013) The MaSuRCA genome assembler. Bioinformatics 29:2669–2677