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