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.

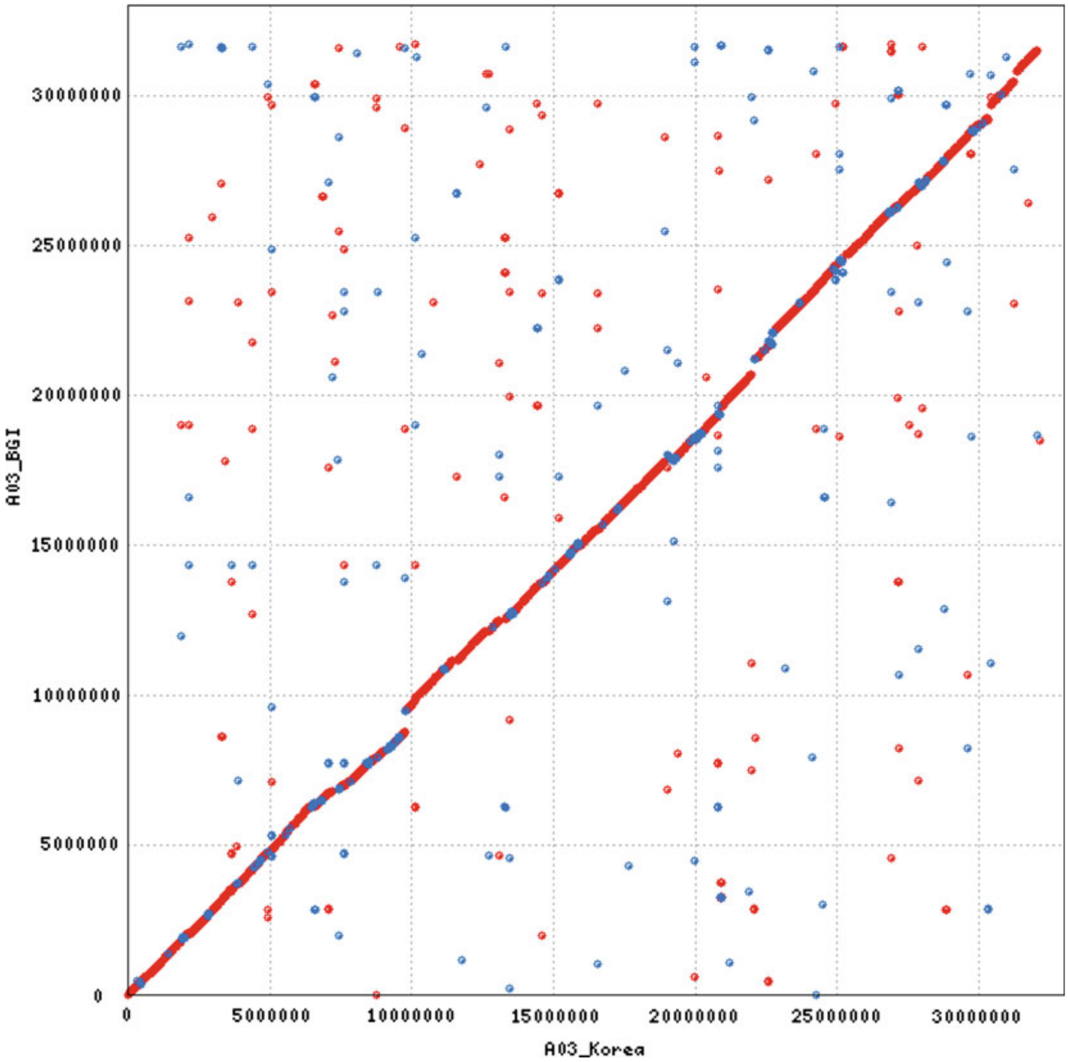

Fig. 2.1 Mummer plot of pseudochromosome A03 from Mun et al. versus that from the WGS assembly

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