Brassica juncea Genome Sequencing: Structural and Functional Insights

12

Thakku R. Ramkumar, Sagar S. Arya, Divyani D. Kumari, and Sangram K. Lenka

Abstract

Indian mustard (*Brassica juncea*) is one of the oldest cultivated amphidiploid crops by human civilization. Natural interspecific crossing between diploid progenitors *B. nigra* (BB) and *B. rapa* (AA) resulted in evolution of this allotetraploid plant species with a total genome size of 1068 Mb. For genetic improvement of the desired traits, it is a prerequisite to unravel the whole genome sequence of this allotetraploid crop and its diploid progenitors. There are several genome and transcriptome sequencing initiatives conducted in this regard to unravel structure and functional annotation of genes in the genome. Similarly, this genomic information was used to obtain species specific molecular insights into the important agronomic traits such as fatty acid biosynthesis, anti-nutritional factors, resilience to climatic perturbations and pathogen resistance. The nuclear and organellar genome sequencing efforts in *B. juncea*, therefore, helped in improving our understanding of the complex allotetraploid architecture and building a foundation to utilize the information for translational genomics and precession breeding in the future.

12.1 Introduction

Brassica juncea, commonly known as Indian mustard, is an allopolyploid (AABB) formed by the cross between *B. rapa* (AA) and *B. nigra* (BB). This *B. juncea* ($2n = 36$) allopolyploidy was discovered by Prakash and Chopra in 1991 by artificial resynthesizing (Prakash and Chopra 1991). Cytological investigations using flow cytometry revealed the genome size to be 922 Mb (Yang et al. 2016). Previously, the *B. juncea* genome size was believed to be 1068 Mb, of which, 529 Mb and 632 Mb has been acquired from *B. rapa* and *B. nigra*, respectively. The genome of *Brassica* spp. has been used to examine gene function divergence and genome evolution as a result of polyploidy, widespread duplication, and hybridization. These investigations and characterization are performed under the title of "Multinational Brassica Genome Project (MBGP)", established in 2002 (https://www.brassica.info/home/about_mbgp.html); the project's main goal was to distribute and organize Brassica genomic resources for the benefit of the global Brassica research community.

T. R. Ramkumar
Department of Microbiology and Cell Sciences, IFAS, University of Florida, Gainesville, FL 32611, USA

S. S. Arya · D. D. Kumari · S. K. Lenka (✉)
The Energy and Resources Institute, TERI-Deakin Nano Biotechnology Centre, Gurugram, Haryana 122003, India
e-mail: sagar.arya@teri.res.in

C. Kole and T. Mohapatra (eds.), *The Brassica juncea Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-030-91507-0_12

Table 12.1 Difference in the diploid and amphiploid genetic makeup of the *Brassica* species

Species	Genome type	n	IC	GS
B. rapa	AA	10	0.539	529
B. nigra	BB	8	0.647	632
B. oleracea	CC	9	0.71	696
B. juncea	AABB	18	1.092	1068
B. napus	AACC	19	1.154	1132
B. carinata	BBCC	17	1.308	1284

1C: refers to the size of haploid genome, GS: genome size in Mbps, n: number of homologous pairs of chromosomes

This helped greatly to study the qualitative and quantitative information of the diversity within *Brassica* spp. It allowed to generate a connection between phenotypic character and genomic loci responsible and to characterize the genes, and detail the substantial numbers of functional proteins within the organism. The database powered by these services has also enabled the identification and characterization of molecular genetic markers and the construction of high-resolution genetic maps. From this, gathering information of plant evolution complexity and its translation to the model plant has be come easy.

Genomic sequencing only provides information of one allelic form of each gene, but to find all variations in the heritable phenotype, three approaches are essential; gene expression information, comparison with known gene sequences, and genetic association studies. Further, functional genomics provides insights of the potential role of a gene by analyzing the pattern of gene expression. Since few genes are highly conserved, their core function can be predicted by comparing with annotated genes or available databases. Identifying the genetic markers can make it easy to pick certain alleles in the genome which are associated with important function. Genome sequence becomes applicable for crop improvement through marker-assisted breeding. Marker-assisted selection is likely to connect with a novel agronomic characteristic when an alternative allele of a gene is sequenced across different genotypes. These findings permit the identification of closely related genes with improved quality and quantities which target the improvement of biological pathways. The sequence of the *B. juncea* genome will identify orthologous genes, aim for germplasm enhancement. Table 12.1 provides with the basic understanding of the difference in the diploid and amphiploid genetic makeup of the *Brassica* species. The model plant *Arabidopsis* has high structural similarity with *Brassica* genes especially in the coding regions (~85% nucleotide conservation) (Cavell et al. 1998). Therefore, genome sequencing will serve as the foundation for translational genomics of this crop based on the functional insights of genes obtained from *Arabidopsis*. Here, we will contextualize the structural and functional details of *B. juncea* genome sequencing initiatives using various next-generation sequencing platforms.

12.2 Technological Advancements in Plant Genome Sequencing

There are several sequencing technologies that can be applied for genome sequencing in plants. These are primarily divided into two approaches, one of which is whole genome shotgun sequencing. In the shotgun sequencing approach, the complete genome is broken down into fragments ranging between 5 and 20 kb. Sequencing each fragment can be done by any suitable sequencing method and the sequence assembly can be done with computational tools to construct the whole genome sequence. Another approach is the application of bacterial artificial chromosomes (BACs) alongside genetic mapping, fingerprinting, and end sequencing. Till date, no technology exists that can read DNA

from one end to the other of even moderately sized chromosome.

The initial microbial genome sequencing was primarily conducted by whole genome shotgun technology. In this approach genomic DNA is sheared to smaller insert libraries of 0.4–1.2 kb and 5–20 kb library. The 0.4–1.2 kb library is used for sequencing and the 5–20 kb library acted as template in assembly. This technology could not meet the demands of eukaryotic genome sequencing, which contains long range repeats. Several technologies such as BAC, yeast artificial chromosome (YAC), plasmic-based conventional clones with large-insert (PBC) and bacteriophage P1-derived artificial chromosome (PAC) were developed by several groups. Though the linear YAC vectors, with yeast centromere and arms, can carry inserts of 1000 kb, the limitations such as instability of insert, chimerism, less amenable in extraction and purification alongside similarly sized yeast chromosomes, led researchers to look for alternative technologies too (Anderson 1993). Meanwhile, in the early 1990s, the BAC library technology emerged as backbone technology to sequence the large complex eukaryotic genomes, while the microbial genome sequencing was dominated with whole genome shotgun technology (Zhang and Wu 2001). The BACs are vector derivatives of *E. coli* F factor with low copy number and autonomous replicative capability (Shizuya et al. 1992), and can host inserts of 120–200 kb efficiently, and even of 300 kb (Luo and Wing 2003). With larger insert sizes, the BAC-based sequencing gave an advantage during assembly over shotgun sequencing. The first BAC library for plants was developed for *Sorghum bicolor* (Woo et al. 1994), followed with rice (Tao et al. 1994), *Arabidopsis* (Choi et al. 1995), and other plants. Physical map of the candidate genomes was constructed using the corresponding species' library, which in turn were used to assist whole genome sequence assembly.

The chromosome mapping, often termed as physical mapping, is the sequential arrangement of several overlapping contigs into putative chromosome, on which DNA markers such as genes of agronomic traits, repeats, transposons, serve as landmarks. This gives the landscape of DNA markers with estimated distance, across the assembled chromosome. Genome assembly is the process of arranging nucleotide sequences into the correct manner and represent model for the actual genome. Physical mapping of chromosomes/genome is required to precisely place the assembled scaffolds at the right place of the genome, thus increasing the contiguity precision of genome construction. For instance, the 135 Mb *Arabidopsis* genome is comprised of five chromosomes. Hauge et al. (1991) utilized cosmid library of 17,000 cosmid clones to construct the physical map of *Arabidopsis thaliana* genome. These 17,000 cosmid clones were assembled into 750 contigs, spanning 90–95% of the entire *Arabidopsis* genome, however, lacking the landscape and DNA marker information. Later Schmidt et al. (1995) utilized YAC clones to construct the physical map of *Arabidopsis* chromosome 4 with much more resolution. This study resulted in the construction of four contigs with information about the distribution of 112 DNA markers, 20 unmapped genes, transposable elements and other random genomic DNA fragments, and covered 90–95% of 21 Mb chromosome 4 of Arabidopsis (Schmidt et al. 1995). Subsequently the physical maps of chromosome 5 (Schmidt et al. 1997; Kotani et al. 1997), chromosome 2 (Zachgo et al. 1996), and chromosome 3 (Camilleri et al. 1998) were also constructed using available YAC libraries. Following this, a complete *A. thaliana* physical map with BAC was constructed (Mozo et al. 1999). This model genome was further enriched with mutant information and gene expression data for functional annotation which subsequently became the cornerstone for Brassica genome annotation.

12.3 Sequencing Platforms and Data Generation for Brassica Genome

Multiple platforms were employed in *B. juncea* genome projects for sequencing data generation. Yang et al. (2016) used the Illumina HiSeq 2000,

Illumina HiSeq 2500 platforms for 13 paired-end (PE) and mate-paired (MP) Illumina libraries (175.8 ×) and PacBio RSII sequencing platform for 1 single-molecule reads library (12.03 ×), and 222 × of BioNano genome mapping data for *B. juncea* var. Tumida *T.* genome sequencing (Yang et al. 2016). Following this, Paritosh et al. (2021) sequenced *Brassica juncea* var. Varuna with dual library system, employing Illumina HiSeq 1000 sequencer platform for llumina PE libraries (200–350 bp) and PacBio RSII platform for PacBio long range (30–50 kb) DNA (PB) libraries. The combination of PE and PB gave higher resolution and better contiguity (Paritosh et al. 2021). The progenitor species with A genome, the *B. rapa* was sequenced with Illumina GA II platform with shotgun sequencing libraries ranging three different insert sizes as long (∼2 kb, 5 kb and 10 kb), medium (∼500 bp), and short (∼200 bp), further supported with Sanger sequencing to fill the gaps in assembly (Wang et al. 2011). This reference genome was reannonated with sequencing data from Illumina 55 Gb (114 ×) PE, 8.7 Gb (18 ×) MP libraries with Illumina Hiseq2500 platform and 6.5 Gb (13.4 ×) reads of single-molecule sequencing PacBio PB libraries data with an average length of 12 Kb with PacBio Sequel II platform (Cai et al. 2017). The genome information is further improved with 19.40 Gb data generated obtained from PacBio library-based PacBio Sequel II platform (Zhang et al. 2018). Zhang et al. also utilized Illumina HiSeq 4000 platform, for Hi-C mapping with 2 × 125 bp reads, for improved chromosome. For the sequencing of the B genome progenitor *B. nigra*, Yang et al. (2016) generated 10 PE and MP Illumina libraries (95.99 ×) for *B. nigra* doubled haploid line (YZ12151). Perumal et al. (2020) used C2 and Ni100 cultivars. A total of 82 Gb (137 ×) for C2 cultivar and 115 Gb (192 ×) for Ni100 cultivar were generated with Illumina HiSeq 2500 and Oxford NanoporeMinION and GridION platforms were used for further assembly. Paritosh et al. (2020, 2021) constructed genome sequence for *B. nigra* var. Sangam with three different sized, PE libraries (200–350, 300–450, and 400–550 bp) and MP libraries (2–3, 4–6, and 10 kb) and Illumina HiSeq 1000 and Illumina MiSeq platforms.

12.4 k-mer Analysis and Error Correction

k-mers are short sequence substrings obtained from the insert reads. These are used to estimate the genome size, by dividing the total number of the k-mers with the depth of the major peaks. k-mers are also used to remove the errors during sequence assembly. A multiple appearance of specific k-mer in the read data reflects errorless sequencing, and error sequences were omitted with lower frequencies. K-mers of 16nt (Wang et al. 2011), 17nt (Yang et al. 2016; Perumal et al. 2020), 21nt (Paritosh et al. 2020), 24nt (Perumal et al. 2020) for short reads and even k-mers of 91 bp (Cai et al. 2017), and 100 bp (Paritosh et al. 2021) were used to cover the complex region with PacBio reads.

12.5 Tools and Technologies Used in *B. Juncea* and Its Progenitor Genomes Assembly

The de novo assembly with short reads of *B. juncea* var. Tumida reference genome was done by ALLPATHS-LG and the gaps were filled with Pacbio RS II reads (Yang et al. 2016). The Irys View RefAligner utility was utilized to draft assemble the genomic scaffolds, and the scaffolds were anchored to the chromosomal map obtained from BioNano data (Yang et al. 2016). The availability of *B. juncea* var. Tumida as reference made the *B. juncea* var. Varuna genome assembly easier. The assembly was carried out using Canu assembler (V1.4) with PacBio sequence reads and the resolution of the sequences was cross examined by mapping with Illumina short read sequences using BWA-MEM. This allowed the *B. juncea* var. Varuna genome to be more precise (Paritosh et al. 2021). The *B. rapa* reference genome (Wang et al. 2011) was de novo assembled using SOAPdenovo and was validated using NUCmer. The

constructed chromosome assemblies were porous with gaps, in particular with the repetitive regions (Wang et al. 2011). The reference genome was reassembled to contigs by Cai et al. (2017) with new reads, as well as with the reads of Wang et al. (2011) using SOAPDenovo2. These assembled contigs were then linked into scaffolds using SSPACE (Cai et al. 2017). These scaffolds were then examined with BAC library sequences by BLAST as well as with Illumina short reads and anchored into super-scaffolds using GapCloser. The PacBio sequences were also used for gap filling using PBjelly_V15.2.20. The reassembled genome was then further validated with CEGMA analysis (Cai et al. 2017). Zhang et al. (2018) again constructed *B. rapa* genome by de novo assembly aided with Hi-C mapping data for better assembly. Using Canu (v1.5), the PacBio reads were assembled and the resolution was enriched by aligning with Illumina short reads using BWA (v0.7.15) and the assembly was then validated with Pilon (v1.22). The aligned and assembled scaffolds were further anchored onto chromosomes using BioNano optical mapping data with Lachesis. The Lachesis was also used to construct the spatial genome to visualize proximal interacting sequences with Hi-C data and plotting with ggplot2 package (Zhang et al. 2018). For *B. nigra* reference genome de novo assembly, the Illumina reads were assembled using ALLPATHS-LG and the gaps were then filled using GapCloser. The generated scaffolds were then aligned by BLAST with publicly available BAC library sequences of *B. nigra*. These scaffolds were then assembled into pseudo-chromosomes using available genetic map. The assembled genome sequence was further validated with CEGMA v.2.3 (Yang et al. 2016). Perumal et al. (2020) independently sequenced genomes of two other varieties of *B. nigra*. The generated Nanopore reads were assembled using CANU 1.6 and crosschecked with three other assemblers; SMARTDenovo, wtdbg, Miniasm (Perumal et al. 2020). The assembled contigs were iterated with high-quality Illumina reads using PILON. These iterated contigs were then scaffolded using Dovetail's HiRise pipeline and were anchored into pseudo-molecules using available genetic maps as references (Perumal et al. 2020). Paritosh et al. (2020, 2021) sequenced genome of *B. nigra* var. Sangam. The genome was assembled with Nanopore reads using CANU 1.6 and the contigs were mapped with high-quality Illumina reads using BWA-MEM (v0.7.12) for higher resolution and iteration with PILON (v1.23) (Paritosh et al. 2020). In parallel, the group also assembled the genome of *B. nigra* var. Sangam with Illumina reads using MaSuRcA, and the BWA-MEM was used to align and position these contigs with PacBio reads. The scaffolding was done with SSPACE-LongRead.pl script and SSPACE-STANDARD-3.0.pl script (Paritosh et al. 2021). The assembled contigs were anchored onto chromosomes using BioNano optical map data and the generated genetic data was polished with PILON (Paritosh et al. 2021).

12.6 Sequencing of Progenitor Diploid Genomes

Population analysis-based genetic linkage maps were earlier constructed for *B. oleracea* (Slocum et al. 1990), *B. rapa* (Song et al. 1991; Chyi et al. 1992), *B. napus* (Ferreira et al. 1994), *B. nigra* (Lagercrantz and Lydiate 1995), and other *Brassica* species as well (Thormann et al. 1994). *B. rapa* (AA) and *B. nigra* (BB) are the genomic progenitors of *B. juncea* (AABB). RNA-seq based single nucleotide polymorphism (SNP) marker mapping revealed collinearity between *B. nigra* genome and *B. juncea* B genome, and the *B. juncea* A genome with genome of *B. rapa* (Paritosh et al. 2014). The estimations by several groups put *B. rapa* (AA) genome to be 450–550 Mb, and confirmed to be of 442.9 MB with BioNano data (Zhang et al. 2018). The *B. rapa* genome underwent genome triplication, and localized rearrangements such as deletions, insertions, inversions, and substitutions (Hong et al. 2008). Wu et al. (2000) constructed binary vector-based library for *B. rapa* for cytoplasmic male sterility locus studies. Subsequently, a genome-wide BAC library was constructed for *B. rapa* jointly by multiple groups (Mun et al.

2008; Hong et al. 2006), and 12,017 BAC-end sequences were analyzed for TEs, SSRs, centromeric satellite repeats and genes, that could also act as markers. The coverage of genome with TEs (14%), 43,000 genes (16.8%) and 1392 different SSRs, with estimated 110,000 SSRs (~one SSR/~4.8 kb), were mapped with this and BAC-end sequence analyses (Hong et al. 2006, 2008). Meanwhile, multiple groups developed different *B. rapa* genetic linkage maps marked with different genetic markers (Kim et al. 2006; Choi et al. 2007). Wang et al. (2011) genome sequenced *B. rapa* by whole genome shotgun sequencing technology with short (~200 bp), medium (~500 bp), and long (~2, 5 and 10 kb) Illumina GA II libraries, and constructed an annotated draft genome. The paired short read sequences upon assembly covered 222 Mb deriving unique scaffolds. The long-read MP sequences ($\geq$2 kb) were not initially used to avoid chimeric reads and incorrect sequences that are common to such libraries. The obtained unique scaffolds were then matched with paired-end sequences to assemble the unique contigs and was matched and corrected with the available BAC libraries (Mun et al. 2008; Hong et al. 2006), and previously assembled chromosome sequences by BAC technology (Mun et al. 2010). Thus, the sequence assembly covered 283.8 Mb, with an estimated cover of >98% of the gene space (Wang et al. 2011). The gaps in draft genome were predominantly with repeat and centromeric regions. Search analyses with 214,425 expressed sequence tags (ESTs) and 52,712 unigenes from NCBI and other sources identified 41,174 putative coding genes. Analysis of *B. rapa* gene family revealed 1003 (5.9%) families out of 16,917 to be lineage specific (Wang et al. 2011), while the rest shared features closely with related plants such as *Arabidopsis, Carica papaya*, and *Vitis vinifera*. The genome data is further upgraded with more short reads and long PacBio reads with reannotation (Cai et al. 2017). The gaps in the previous versions are predominantly in the repeat regions and centromeric regions are due to usage of short read sequences. Zhang et al. (2018) overcome this drawback and further enriched the *B. rapa* genome sequence and developed a reference genome employing third generation sequencing technologies such as optical mapping, Hi-C, and single-molecule sequencing. This latest study with integrated approach put the *B. rapa* genome size to 442.9 Mb, shrinking the previous estimations of 450–550 Mb (Zhang et al. 2018). Optical mapping of *B. rapa* genome was performed alongside PacBio long-read sequencing with a focus on centromeric and repeat regions to fill the gaps from previous genome versions. Hi-C libraries were also generated to visualize the proximal and physically interacting DNA elements. The resequencing and reannotation provide the *B. rapa* genome with 45,985 protein-coding genes with 45,411 genes (98.75%) annotated on chromosomes, leaving behind 574 genes (1.25%) on scaffolds. The *B. nigra* (BB) genome has an estimated size of 570–607.8 Mb, while studies by Paritosh et al. (2020) suggested the *B. nigra* genome to 522 Mb. The *B. nigra* underwent whole genome duplication, and subsequent rearrangements, making it a complex genome for sequencing and annotation. Yang et al. (2016) genome sequenced the *B. nigra* genome, covering 396.9 Mb of the genome, and annotated 49,826 genes. Perumal et al. (2020) assembled *B. nigra* genome with whole genome shotgun Illumina PE (300–700 bp insert) sequencing, Roche/454and Illumina MP libraries (3–45 kb insert) and long reads from Nanopore sequencing. The short reads were assembled with the help of long reads and the genome was annotated with RNA seq data (Perumal et al. 2020). Perumal et al. (2020) predicted 59,877 and 67,030 genes in two different genotypes, of which 55,022 (92.0%) and 59,780 (89.2%) genes were annotated with Uniprot database. Following this, Paritosh et al. (2020) utilized Optical Mapping and Nanopore long-read sequencing to sequence and assemble the *B. nigra* genome with higher contiguity. The study identified 57,249 putative genes of which 42,444 genes were matched with transcriptome data. Table 12.2 enlists the sequencing and annotations efforts of *B. juncea* progenitors.

Table 12.2 Sequencing and annotation of *B. juncea* progenitors

NCBI accession	Organism	Description
PRJNA558855	*B. rapa*	Single-molecule real-time sequencing technology; Gene expression; multi-isolate
PRJNA445393	*B. rapa Subspecies Chinese*	Next-generation sequencing; transcriptome; multi-isolate
PRJNA244166	*B. rapa*	Transcriptome or gene expression
PRJNA298858	*B. rapa*	Global transcriptome profiling, transcriptome or gene expression; multi-isolate
PRJNA249065	*B. rapa*	Annotation; refSeqgenome; mono-isolate
PRJNA244166	*B. rapa*	Deep transcriptome sequencing; transcriptome or gene expression; mono-isolate
PRJNA636608	*B. nigra*	Nanopore sequencing; raw sequence reads; multispecies
PRJNA642332	*B. nigra* var. Sangam	Oxford nanoporeMinION; representative genome
PRJNA285130	*B. nigra* cultivar *inbred line YZ12151*	Illumina HiSeq; chromosome
PRJNA615316	*B. juncea*	Genome sequencing and assembly; multispecies
PRJNA516907	*B. nigra*	hybrid Illumina/Roche 454 sequencing; multispecies
PRJNA327666	*B. nigra*	RefSeqgenome; mono-isolate
PRJNA324621	*B. nigra* var. Sangam	Genome sequencing, assembly, raw sequence reads, transcriptome; mono-isolate
PRJNA320480	*B. nigra*	Genome sequencing, assembly, raw sequence reads, gene expression; multispecies
PRJNA311781	*B. nigra*	RefSeqgenome; mono-isolate

The data is deposited to NCBI database (date of access 20th Aug 2021)

12.7 Genome Sequencing of Cultivated *B. Juncea*

Sequencing of *B. juncea* has served to resolve the complicated allopolyploid genome using the tools and technologies described earlier. *B. juncea* carries an allopolyploid genome (AABB). The draft genome sequence for *B. juncea* was constructed (Yang et al. 2016) with multiple sequencing technologies primarily with PE and MP Illumina libraries, and certain portions with long-read Bionano and PacBio sequencing. Flow cytometric analysis suggested the genome size as 922 Mb. The first draft genome of *B. juncea* var. Tumida was done using shotgun reads by de novo assembly, single-molecule long reads (PacBio sequencing), genetic mapping and genomic (optical) mapping (BioNano sequencing). The assembled *B. Juncea* var. Tumida genome resolved the complication of allopolyploid genome, and it's constructed by de novo assembly. Subsequently, it was also demonstrated that the homeolog expression dominance in *B. juncea* to explore the A and B subgenomes for their transcriptional behavior. The draft genome covered 85% (784 Mb) of the genome and annotated with 80,050 genes, 21lncRNAs, 2,638 tRNAs, 511 rRNAs, 3,725 small RNAs, 1,402 microRNAs, and 15,418 small nuclear RNAs. The *B. juncea* genome appears to have lost identified 562 and 545 genes in A and B subgenomes (Yang et al. 2016). Genome-wide RNA-seq analysis of the homoeologous genes from different developmental stage, tissues and two newly resynthesized *B. juncea* revealed, in all the samples, on average 16.2% of genes displayed homeolog expression dominance, whereas only 8.2% of genes showed expression dominance toward to BjuB over BjuA excluding

B. juncea resynthesized lines. On the basis of this observation, its inferred that no significant global dominance for the subgenomes further strengthening recent polyplodization of *B. juncea* (Yang et al. 2016). The sequencing and assembly of the *B. juncea* genome facilitate the agricultural trait improvements in this important crop.

The constructed *B. juncea* genome was further enriched for improved contiguity using long read sequencing technologies (Paritosh et al. 2020). The chromosome-scale genome assembly of *B juncea* generates a complete architecture of the A and B genome. They subjected *B. juncea* var. Varuna to SMRT sequencing on the PacBio RSII platform. This study provided marked improvement in B genome assembly of Varuna variety, unlike earlier sequenced Tumida variety, exposing extensive gene block fragmentations and gene block associations than reported earlier. The quality of genome sequence was improved to the extent that, the longest contig covered 35.8 Mb and the largest scaffold covered 72.1 Mb. RNA-seq guided gene annotations revealed 101 959 putative genes, in A (46,381) and B (55,578) subgenomes, of which 93% genes were validated with uniprot entries (Paritosh et al. 2021). Transposable elements covered 385 Mb (45.8% of *B. juncea* genome) of the genome, with higher representation in B subgenome covering 259 Mb of 51% of the subgenome than the A subgenome, that covered 113 Mb of 33.9% of A subgenome (Paritosh et al. 2021). Further, we have also discussed organellar genome sequencing in details. The details of genome sequencing efforts in *B. juncea* are summarized in Table 12.3.

12.8 Genome Annotation

Genome annotation is a method of genome study to identify functional elements along the sequence. It is the description of individual genes and their proteins or predicting the meaning of the genome sequence. Such prediction can be structural annotation (including location of coding regions, open reading frames (ORFs), and their corresponding regulatory motifs, functional annotation (integration of additional data such as spatio-temporal expression profile at transcriptomic and proteomic level, interaction network, physiological and biochemical functions), and essential components (CDS, mRNAs, pseudogenes, promoters, and poly-A signals)). Annotation of *B. juncea* var. Varuna genome was carried out for transposable elements, centromeres, and genes. For transposable element *de-novo* prediction approach, centromeric region is identified by correlation plot and RNA-seq analysis is performed for gene annotation. There are 1590 consensus repeats generated when merged with *Arabidopsis thaliana* transposable element database, repeats were classified into retrotransposons, DNA transposons, and other repeats. Comparing to the A chromosome [(113 Mb (∼33.9%)], the B chromosome [∼259 Mb (∼51%)] has more repeats. A genome of *B. juncea* has a centromeric region already identified in *B. rapa* (A genome) and *B. oleracea* (C genome), i.e., CentBr1 and CentBr2 but absent in the B genome centromeric region. However, seven new B genome-specific repeats along with three new A genome-specific centromeric repeats were identified from *B. juncea*. In gene annotation, a total of 40,208 full-length high-quality sequences were obtained and a total of 105,354 genes were predicted of which 48,270 in the A genome and 57,084 in the B genome (validated by the non-redundant proteins in the UniProt plant database). Other studies have found a total of 82,008 genes–38,232 (∼79%), 43,776 (∼76%) within the A and B genome, respectively (Paritosh et al. 2014; Yang et al. 2016). *B. juncea* var. Varuna, B genome has shown high similarity with *B. nigra* genome. Orthologous study has also identified a total of 19,404 orthologous groups of which 39,383 genes of A had orthologs in B, and 42,727 genes of B had orthologs in the A genome. Similarly, with using non-redundant nucleotide and NCBI protein sequences, SWISS-PROT, Gene ontology (GO), Cluster of Orthologous Groups (COG), and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases, the Tumida variety was also annotated for 80,050 protein-coding genes. Additionally, Yang et al.(2016) identified 511

Table 12.3 Publicly available *B. juncea* genome sequences, assembly and associated initiatives useful for genome annotation

NCBI accession	Project data type	Scope	Institute	Remark
PRJNA719608	Transcriptome	Mono-isolate	Guizhou Academy of Agricultural Sciences, China	Transcriptome sequencings of *B. juncea* var. G170 (drought tolerant and sensitive) were performed for identifying differentially expressed genes related to drought stress. The identified genes were considered as potential candidates for enhancing drought tolerance in *B. juncea*
PRJEB41123	Genome assembly (Chromosome scale) of *B. juncea* (AABB) to compare the architecture of the A and B genomes	Mono-isolate	University of Delhi, India	A and B genome harbors high number of LTR/Gypsy retrotransposons, followed by centromeric repeats, and *B. nigra* specific gene clusters, that segregates the collinearity between A and B genomes
PRJNA673122	Raw sequence reads	Mono-isolate	University of York, UK	NA
PRJNA672814	Raw sequence reads	Multispecies	Huazhong Agricultural University, China	Anthocyanin studies of *B. juncea*
PRJNA670607	Transcriptome	Multispecies	Biozeron Shenzhen Inc., China	Gene expression studies of *B. juncea*
PRJNA662836	Raw sequence data	Multi-isolate	Punjab Agricultural University, India	*B. fruticulosa* is a wild species belonging to Brassicaceae, which carry resistance to various biotic stresses. Introgression lines were developed to transfer desirable variation from *B. fruticulosa* into *B. juncea*
PRJNA663021	Raw sequence data	Multispecies	Punjab Agricultural University, India	Molecular-genetic analysis of *B. juncea*: Development of *B. fruticulose* introgression lines to transfer desired variations into *B. juncea*
PRJNA639209	Raw sequence data from different inbred lines	Multi-isolate	Punjab Agricultural University, India	The project aims to SNP genotype of land races, advanced breeding materials, determinate mustard, resynthesized

(continued)

Table 12.3 (continued)

NCBI accession	Project data type	Scope	Institute	Remark
				genotypes and introgression lines of *B. juncea*. The data will be used to establish diversity conduits and association panels
PRJNA615316	Genome sequencing, assembly	Multispecies	Hunan Agricultural University, China	De novo assembly of a chromosome-scale genome of *B. juncea* landrace Sichuan Huangzi and resequencing of 480 *B. juncea* accessions
PRJNA562990	Transcriptome—raw sequence reads	Multispecies	Huazhong Agricultural University, China	Transcriptome profile revealed that cytokinin promoted callus sprouting in *B. juncea*
PRJNA550308	Genome sequencing and assembly	Mono-isolate	The Centre for Genetic Manipulation of Crop Plants, University of Delhi, India	Genome assembly (highly contiguous chromosome-scale) of an oleiferous type of *B. juncea* var. Varuna, with PacBio SMRT reads
PRJNA544908	Transcriptome—raw sequence reads	Multispecies	Xinyang Normal University, China	Transcriptome elucidation in the purple leaf of *B. juncea* and identified differentially expressed genes involved in anthocyanin biosynthesis
PRJNA507350	Transcriptome—raw sequence reads	Mono-isolate	University of York, UK	Transcriptomic validation in *B. juncea* by dissecting the genetic architecture of quality and agronomic traits
PRJNA499035	Raw sequence reads	Multispecies	Institute for Plant Breeding	Resequencing data for the parents of a population of interspecific *Brassica* hybrids. The parents are *B. napus* (Ag-Spectrum), *B. juncea* (Purple leaf mustard) and *B. carinata* (C2). The resulting population was allohexaploid
PRJNA497315	Mitochondrial genome and transcriptome analysis	Multispecies	Huazhong Agriculture University, China	Mitochondrial genome assembly of five alloplasmiccytoplasmic male sterile (CMS) lines in *B. juncea* and transcriptome analysis with the RNA editing profile

(continued)

Table 12.3 (continued)

NCBI accession	Project data type	Scope	Institute	Remark
PRJNA479948	Transcriptome—raw sequence reads	Multispecies	Central University of Punjab, India	Expression analysis to identify vital genes and pathways involved in the regulation of various metabolic and biological processes, particularly related to cadmium stress. A transcriptomic investigation to understand the response of *B. juncea* to cadmium stress could provide insights into improving its phytoremediation efficiency
PRJNA477240	Transcriptomics	Mono-isolate	Sichuan Agricultural University, China	Transcriptomicsto provides insight into stem development in *B. juncea* var. Tumida
PRJEB26751	SNP data	Mono-isolate	Sher-e-Kashmir University of Agricultural Sciences and Technology—Jammu, India	The project involves ddRAD-sequencing followed by identification of SNP identification
PRJNA471033	Raw sequence reads	Multi-isolate	University of York, UK	*B. juncea* (Varuna x Heera) developed from an DH population VHDH
PRJNA448707	Transcriptome or gene expression	Multi-isolate	University of Minnesota Duluth, Minnesota, US	RNA-Seq study of expression in *B. juncea* nectaries and leaves from different lines
PRJNA431509	Epigenomic data	Multispecies	ICAR—Institute of vegetable science, India	High through-put sequencing of small RNA
PRJNA430791	Raw sequence reads	Multispecies	School of Life Science, Lanzhou University, China	Investigated the molecular mechanisms of anthocyanin accumulation in *B. juncea* leaves
PRJNA395472	Transcriptome	Multi-isolate	Zhejiang A & F University, China	Transcriptional changes in *B. juncea* leaves after armyworm chewing
PRJNA383771	Raw sequence reads, assembly, transcriptome	Mono-isolate	ICAR-Indian Institute of Agricultural Biotechnology, India	RNA-seq of *B. juncea* under various Nitrate treatments
PRJNA339019	Transcriptome	Mono-isolate	Southeast Chongqing Academy of Agricultural Sciences, Chongqing, China	Transcriptomic profiling of arthrobacterium–induced systemic resistance in *B. juncea* var. Tumida infected by *Plasmodiophorabrassicae*

(continued)

Table 12.3 (continued)

NCBI accession	Project data type	Scope	Institute	Remark
PRJNA323808	Genome sequencing and assembly	Mono-isolate	SRM University, India	Whole genome sequencing and analysis of *B. juncea*
PRJNA321670	Transcriptome or gene expression	Multi-isolate	Department of Plant Molecular Biology, University of Delhi, India	The transgenic line BnCRY2aOE over-expressing a blue light photoreceptor (BnCRY2a) flowered earlier than wild-type plants. To identify the downstream candidate genes involved in regulating the early flowering transgenic phenotype, a genome-wide microarray analysis of the transgenic vs. wild type plants was performed. The microarray analysis unraveled the differential up-regulation of genes involved in flower development, cell differentiation and growth as well as hormone biosynthesis/signaling in comparison to wild-type
PRJNA319668	Transcriptome or gene expression	Multispecies	Zhejiang University, China	Analyzing and evaluating reversibility and heritability of DNA methylation resulting due to in vitro grafting between *B. oleracea* and *B. juncea*
PRJNA312980	Transcriptome or gene expression	Mono-isolate	Sichuan Normal University, China	Identified *B. juncea* transcripts using de novo assembly of the Illumina HiSeq 4000 sequencing
PRJNA301284	RefSeq genome	Mono-isolate	SRM University, India	*B. juncea* genome reference project
PRJNA298501	Raw sequence reads	Multispecies	Huazhong Agricultural University, China	Investigated the anthocyanin formation in five *Brassica* species, followed by transcriptome analysis between purple and green leaves. The response of potential key genes was examined for pigmentation as well as the physiological roles of anthyocyaninsin *Brassica* plant development

(continued)

Table 12.3 (continued)

NCBI accession	Project data type	Scope	Institute	Remark
PRJNA296365	Transcriptome or gene expression	Multi-isolate	Department of Botany, University of Delhi, India	Expression analysis of cold-stressed siliques in *B. juncea*
PRJNA290942	Transcriptome or gene expression	Mono-isolate	ICAR—National Research centre on plant biotechnology	Comparative transcriptome analysis between CMS (*Moricandiaarvensis*) *B. juncea* var. Pusa bold and its fertility restorer line
PRJEB9362	Transcriptome	Mono-isolate	Fondazione Edmund Mach, Italy	Identification of novel and conserved miRNAs involved in redox regulation of salt stress *B. juncea*
PRJNA289188	Transcriptome profiling	Mono-isolate	Zhejiang university, China	To obtain ESTs from *B. juncea* var. Tumida
PRJNA285130	Genome sequencing, assembly	Multispecies	Zhejiang university, China	*B. juncea* var. timuda genome sequencing and assembly
PRJNA277020	Transcriptome profiling	Multi-isolate	Banaras Hindu University, India	Microarray analysis of Indian mustard plants subjected to arsenate stress
PRJNA276704	Transcriptome or gene expression	Multi-isolate	Jawaharlal Nehru University, India	Assembly (de novo) and stress related transcriptomic profiling of a salinity-tolerant *B. juncea* var. CS52
PRJNA271638	Transcriptome or gene expression	Multi-isolate	ICAR-Directorate of Rapeseed Mustard Research, India	Elucidating salt (NaCl)-induced changes in whole transcriptome of *B. juncea* var. CS-52
PRJNA271633	Transcriptome or gene expression	_	ICAR-Directorate of Rapeseed Mustard Research, India	Salt stress induced changes in transcriptome of *B. juncea*
PRJNA270523	Transcriptome or gene expression	Multi-isolate	Plant Genomics and Stress Biology, Department of Botany, Delhi University, India	RNA-Seq of *B. juncea* in response to abiotic stresses
PRJNA245462	Transcriptome or gene expression	Multi-isolate	Centre for Genetic Manipulation of Crop Plants, Delhi University, India	Transcriptome profiling to identify differently expressed genes in various parts of *B. juncea* var. Heera
PRJNA244493	Transcriptome or gene expression	Mono-isolate	Bose Institute, India and Genotypic Technology Pvt. Ltd., India	Differentially expressed transcripts of susceptible *B. juncea* challenged with *Alternariabrassicicola*

(continued)

Table 12.3 (continued)

NCBI accession	Project data type	Scope	Institute	Remark
PRJNA231241	Transcriptome or gene expression	Multi-isolate	Plant Genomics and Stress Biology, Department of Botany, Delhi University, India	Genome-wide perspective of miRNAome in *B. juncea,* induced by abiotic stresses
PRJNA195539	Transcriptome or gene expression	Mono-isolate	Hunan Agricultural University, China	*B. juncea* transcriptome analysis
PRJNA185138	Transcriptome	Multi-isolate	Zhejiang University, China	Profiling of miRNA in *B. juncea* CMS and maintainer fertile lines by deep sequencing
PRJNA174360	Genetic map	Mono-isolate	_	Doubled haploid mapping population derived from a cross between Varuna and Heera was used to analyze yield influencing traits in *B. juncea.* SSR and RFLP markers were mapped onto an already existing AFLP map for the QTL analysis
PRJNA173786	Cheung_97 genetic map	Mono-isolate	_	Cheung_97 genetic map of *B. juncea* is based on RFLP detected by anonymous cDNA markers from *B. napus*
PRJNA173691	Panjabi_08 genetic map	Mono-isolate	Centre for Genetic Manipulation of Crop Plants, University of Delhi, India	Panjabi_08 genetic map of *B. juncea* was developed with PCR-based Intron Polymorphism markers to study segmental structure of the A and B genomes of *Brassica*
PRJNA169393	Transcriptome or gene expression	Multi-isolate	Agriculture and Food Systems, The University of Melbourne, Australia	Understanding the molecular mechanisms of *B. juncea* underpinning physiological tolerances to salinity and alkaline salinity
PRJNA136273	Transcriptome or gene expression	Multi-isolate	Prof. J.P. Khurana, Department of Plant Molecular Biology, University of Delhi, India	Elucidating the role of Cryptochrome 1, which modulates the development in *Brassica* by regulating the gene expression (involved in light, stress and phytohormone response)
PRJNA74717	RefSeq genome	Mono-isolate	Nanjing Agricultural University, State Key Laboratory of Crop Genetics and Germplasm Enhancement, P.R. China	Mitochondrial genome of *B. juncea*

(continued)

Table 12.3 (continued)

NCBI accession	Project data type	Scope	Institute	Remark
PRJNA72397	Transcriptome and gene expression	Multi-isolate	College of Bioinformation, Chongqing University of Posts and Telecommunications, China	Transcriptome analysis of the stem tumor mustard *B. juncea* var. Tumida
PRJNA43723	Pradhan_03 genetic map	Mono-isolate	University of Delhi, India	Pradhan_03 genetic map of *B. juncea* is derived by crossing of Varuna and Heera parental lines
PRJNA615316	RefSeq genome	–	Hunan Agricultural University, China	Representative genome of *B. juncea* var Sichuan Huangzi

Source NCBI-https://www.ncbi.nlm.nih.gov/bioproject/?term=brassica+juncea

rRNAs, 2,638 tRNAs,3,725 small RNAs, 21 long noncoding RNAs, 15,418 small nuclear RNAs and 1402 microRNAs, from the *B. juncea* genome and even extracted syntenic ortholog gene pairs (28,228 and 28,917) from the *B. juncea* subgenomes. Gene loss has been observed during speciation, a total of 562 (A subgenome) and 545 (B subgenome) genes were considered as lost, referring to their common ancestral genomes. Likewise, many structural and functional annotations have been performed to investigate the structure, function, and regulation of the genome which are enlisted in Table 12.4.

12.9 Organellar Genome Sequencing

12.9.1 Mitochondrial Genome Insights

Mitochondrial genomes in angiosperms are more complex compared to other organisms (Chang et al. 2011). Analyses of different angiosperm mitochondrial genome sequences have revealed several common characteristics; however, a number of diverse mitotypes have evolved within each genus/species. Sequencing and comparative investigations of six *Brassica* mitotypes indicates a mechanism for mitochondrial genome development in *Brassica*, which includes certain events of evolution like mutation, genome compaction, duplication, and rearrangements (Chang et al. 2011). Further, the analysis of the evolutionary relatedness between *Brassica* mitotypes provides information of mitochondrial genome of *B. juncea*. The size of whole single circular mitochondrial genome of *B. juncea* is 219,766 bp [GenBank: JF920288], which is reported to be similar to that of *B. rapa*, i.e., 219,747 bp. Whereas, the mitochondrial genome sizes of *B. carinata* and *B. oleracea* are greater than both *B. juncea* and *B. rapa*, i.e., 232,241 bp and 360,271 bp, respectively. The report suggests that the *B. juncea* mitotype is an evolutionary outcome of an inherited mitotype of (*B. rapa*) cam with certain modifications (Chang et al. 2011). The percentage of the total functional genes in the mitochondrial genome of *B. juncea* are almost the same compared to *B. rapa, B. oleracea* and *B. napus* mitotypes, expect *B. carinata*, where the similarity is 27.98% (Chang et al. 2011).

One of the main aspects of mitochondria genomes is producing cytoplasmic male sterility (CMS), which is essential to explore the heterosis in crop plants and is routinely utilized as a model to investigate the interactions between nuclear and cytoplasmic. These interactions are either spontaneous or formed through interspecific

Table 12.4 Structural and functional annotations performed in *B. juncea*

Annotation	Description	References
Nucleotide-binding site leucine-rich repeat genes	Identified 289 NLR genes with a ratio of 0.61:1 of toll/interleukin-1 receptor-NLR to non-toll/interleukin-1 receptor-NLR genes	(Inturrisi et al. 2020)
Heat shock transcription factor	A total of 60 HSF transcription factors were identified; establish a phylogenetic relationship	(Li et al. 2020a)
Superoxide dismutase proteins (SOD), belongs to the family of metalloenzyme	Identified 29 genes; and cis-regulatory elements in the promoter region, of which 10 SOD genes were abiotic stress-responsive	(Verma et al. 2019)
Receptor-like protein (RLP) and Receptor-like kinase (RLK) genes	A total of 228 RLPS and 493 RLKS were identified	(Yang et al. 2021)
Jasmonate ZIM domain proteins	38 genes were identified in Tumida variety	(Cai et al. 2020)
Auxin response factors	65 *B. Juncea* genes that encode ARF proteins were identified, further promoter *cis*-element also demonstrated in all the 65 genes	(Li et al. 2020b)
Teosinte branched1/*Cycloidea*/ proliferating cell factors	Identified 62 genes from the *B. Juncea* var. Tumida	(He et al. 2020)
Chitinase gene family	Identified 47 genes	(Mir et al. 2020)
Nonexpressor of pathogenesis-related genes	Identified 19 genes, which cluster into six separate groups in the genome of *B. juncea* var. Tumida	(Wang et al. 2021)
GRAS transcription factors	A total of 88 GRAS genes were identified	(Li et al. 2019)
Regulatory roles of lncRNAs	Identified 7613 lncRNAs, of which 1614 lncRNAs are involved in heat and drought related stress response	(Bhatia et al. 2020)

crossing, which results into an alloplasmic condition/cytoplasmic substitution (Gaikwad et al. 2006). CMS characterized, particularly by non-Mendelian inheritance and the suppression of the production of viable pollen. The CMS could result due to mutations, recombination, or rearrangements in the genome of mitochondria. Knowledge of the *B. juncea* mitochondrial genome is crucial to develop superior phenotypes. In this quest, numerous attempts are being undertaken to sequence and assemble the mitochondrial genomes of *B. juncea* lines. For instance, a study performed on five alloplasmic *B. juncea* lines revealed that the mitochondrial genomes from 221 to 256 kb (Wu et al. 2019), which is somewhat greater compared to the typical *B. juncea* size reported earlier (Chang et al. 2011). Another comparative study between the mitochondrial genomes of the *hau* CMS line and its iso-nuclear maintainer line in *B. juncea* showed a difference in the genome sizes and GC content (Heng et al. 2014). Mitochondrial genome of *B. junceahau* CMS is reported to be of 247,903 bp in size with a GC content of 45.08%, whereas that of another normal line (J163-4) and a maintainer line are 219,863 bp in size with a GC content of 45.23%. Further, the mitochondrial genome has numerous genes, of which, 35 are protein encoding, 25 are tRNA, 3 are rRNAs, and 29 are ORFs of unknown function. Whereas, genes in maintainer line are 36 for protein encoding, 22 for tRNA, 3 for rRNAs, and 31 are unidentified ORFs. In addition, sub-stoichiometrical coexistence of distinct

mitotypes is confirmed in *hau* CMS lines as well as its maintainer lines in *B. juncea* (Heng et al. 2014). Further, it was demonstrated that a cytotoxic protein ORF288 associated with male sterility in *B. juncea* causes aborted pollen development (Jing et al. 2012). The toxicity generating region of ORF288 is reported to be located near the N-terminus, which repressed the growth when heterologously expressed in *E. coli* (Heng et al. 2018). However, heterologous expression of ORF288 portions indicates that the region which induces CMS is present between amino acids 73 and 288, whose heterologous expression did not inhibit *E. coli* growth. It is reported that the transcript levels of *orf288* are associated with altered nuclear gene expression and the *hau* CMS system. Apart from *orf288*, it also reported that *orf220* gene causes male sterility in *B. juncea* (Yang et al. 2010). Both positive and negative correlations have been reported between the occurrence of CMS and CMS-associated orfs, when CMS-associated orfs were targeted and expressed in mitochondria. Certain orfs can cause male- or semi-sterility, while some do not. Understanding of nuclear-mitochondrial compatibility is crucial as it can result in differential expression of mitochondrial genes, e.g., mitochondrial *aptα* gene (Gaikwad et al. 2006). Furthermore, it is essential to identify the molecular basis of mitochondrial recombination in the male sterile cytoplasmic hybrids to understand the underlying mechanism as well as the environmental factors affecting fertility reversion. For instance, analysis of floral bud transcriptome under both pollination and non-pollination state showed a variability in the expression of Muts HOMOLOG1 (*MSH1*) (a nuclear gene which regulates illegitimate recombination in plant mitochondria) in response to different sugars, which shows that physiological changes are involved in the pollination signaling and fertility reversion in CMS plants (Zhao et al. 2021). This is indicative that the gynodioecy is a reproductive plan of action that might incorporate ecologically responsive genes such as *MSH1* as a switch for fertility-sterility transition under reproductive isolation. It was found that the mitochondrial genome in revertant lines of *B. juncea* CMS cytoplasm undergoes substoichiometric shift to suppress *orf220* copy number, whereas *MSH1-RNAi* with increased *orf220* copy number are male sterile (Zhao et al. 2016). This provides a valuable insight into substoichiometric shift in CMS induction, fertility reversion as well as interplay of *MSH1*.

12.9.2 Chloroplast Genome Sequencing

Chloroplast is a vital organelle in plants, which harbors genetic and enzymatic resource required for photosynthesis, as well as various important biosynthesis pathways such as pigments, fatty acids, vitamins, and amino acids. Understanding of the chloroplast genome is therefore essential to enhance the qualitative and quantitative traits of *B. juncea*. As chloroplast genome engineering of *B. jucnea* through chloroplast can increase the oil content, improve oil quality, as well as make it resistant to most of the abiotic and biotic factors. Prabhudas et al. (2016) are the first to sequence the complete chloroplast genome of *B. juncea* (Indian mustard). The sequencing carried out on Illumina Hiseq 2500 platform generated 100 bp paired raw reads, which were qualitatively trimmed followed by read quality assessment.

The chloroplast genome size is reported to be 153,483 bp, with 36.36% of GC content (Prabhudas et al. 2016). The assembled chloroplast genome exhibited a quadripartite structure consisting of an 83,286 bp large single copy region and an 17,775 bp single copy region, which are separated by a pair of IRa and IRb (inverted repeats a and b) of 26,211 bp each. The GC content was 34.12, 29.20, and 42.34% for large single copy region, single copy region, and inverted repeats, respectively. In the entire chloroplast genome, a total number of 113 genes were annotated, of which, 79 genes are protein encoding, 30 are tRNA, and 4 are rRNA. Total 15 genes including 9 protein-encoding genes and 6 tRNA genes had either a single of couple of introns. Gene duplications in inverted repeat regions were extended to 4 rRNA, 7 tRNA, as well as 6 protein-encoding genes.

B. juncea share the chloroplast genome with one of the hybridization donor *B. rapa*, which fits the U model (Li et al. 2017). The phylogenetic analysis indicated that the branch length of *B. juncea* chloroplast genome is close to *B. oleracea*, *B. napus*, and *B. rapa* more distant from *B. nigra* and *B. carinata*, suggesting a divergence from the two Brassica's.

12.10 Conclusion

Rapid development in next-generation sequencing technologies has undoubtedly accelerated the sequencing programs of *B. juncea* genomes along with its progenitors to generate metadata at less cost and time. The *B. juncea* bio-projects have sequenced, assembled, analyzed and have annotated the genome to highlight the impact of recombination between progenitor-genomes of *Brassica* spp. Apart from providing fundamental knowledge about the evolution of *B. juncea*, the bio-projects also helped in the identification of candidate genes involved in various physiological and biochemical processes. Importantly, the information can be utilized form trait mapping to targeted tinkering of the genetic makeup of *B. juncea* and to further improve resistance to biotic and abiotic stresses, enhance climate resilience, increase yield as well as quality and nutritional traits. The availability of the genomic data will further advance the knowledge of cultivated *B. juncea* lines and help to produce better tool kits for molecular breeding and crop improvement. Considering the level of diversity and wealth of the species in the *Brassica* genus, there is a lot yet to explore about *B. juncea* to clarify the relationship between genotypes/phenotypes. Similarly, progress in proteome and transcriptome research has the capacity to reveal the structural and functional role of genes in *B. juncea*, altered in response to physiological and environmental processes and pathogen attacks. Overall, the information of nuclear and organellar genomes, as well as the increasing amount of "omics" resources will certainly contribute toward the effective use of *B. juncea*.

References

Anderson C (1993) Genome shortcut leads to problems. Science 259:1684–1688

Bhatia G, Singh A, Verma D, Sharma S, Singh K (2020) Genome-wide investigation of regulatory roles of lncRNAs in response to heat and drought stress in *Brassica juncea* (Indian mustard). Environ Exp Bot 171:103922

Cai C, Wang X, Liu B, Wu J, Liang J, Cui Y, Cheng F, Wang X (2017) *Brassica rapa* genome 2.0: a reference upgrade through sequence re-assembly and gene re-annotation. Mol Plant 10:649–651

Cai Z, Chen Y, Liao J, Wang D (2020) Genome-wide identification and expression analysis of jasmonate ZIM domain gene family in tuber mustard (*Brassica juncea* var. tumida). PLoS One 15:e0234738

Camilleri C, Lafleuriel J, Macadre C, Varoquaux F, Parmentier Y, Picard G, Caboche M, Bouchez D (1998) A YAC contig map of Arabidopsis thaliana chromosome 3. Plant J 14:633–642

Cavell A, Lydiate D, Parkin I, Dean C, Trick M (1998) A 30 centimorgan segment of Arabidopsis thaliana chromosome 4 has six collinear homologues within the *Brassica napus* genome. Genome 41:62–69

Chang S, Yang T, Du T, Huang Y, Chen J, Yan J, He J, Guan R (2011) Mitochondrial genome sequencing helps show the evolutionary mechanism of mitochondrial genome formation in Brassica. BMC Genomics 12:1–12

Choi S, Creelman RA, Mullet JE, Wing RA (1995) Construction and characterization of a bacterial artificial chromosome library of *Arabidopsis thaliana*. Plant Mol Biol Rep 13(2):124–128

Choi SR, Teakle GR, Plaha P, Kim JH, Allender CJ, Beynon E, Piao ZY, Soengas P, Han TH, King GJ (2007) The reference genetic linkage map for the multinational *Brassica rapa* genome sequencing project. Theor Appl Genet 115(6):777–792

Chyi Y-S, Hoenecke M, Sernyk J (1992) A genetic linkage map of restriction fragment length polymorphism loci for *Brassica rapa* (syn. campestris). Genome 35(5):746–757

Ferreira M, Williams P, Osborn T (1994) RFLP mapping of *Brassica napus* using doubled haploid lines. Theor Appl Genet 89(5):615–621

Gaikwad K, Baldev A, Kirti P, Mohapatra T, Bhat S, Chopra V, Prakash S (2006) Organization and expression of the mitochondrial genome in CMS (Moricandia) *Brassica juncea*: nuclear-mitochondrial incompatibility results in differential expression of the mitochondrial atpα gene. Plant Breed 125(6):623–628

Hauge BM, Giraudat J, Hanley S, Hwang I, Kohchi T, Goodman HM (1991) Physical mapping of the Arabidopsis genome and its applications. In: Plant Molecular Biology vol 2. Springer, pp 239–248

He J, He X, Chang P, Jiang H, Gong D, Sun Q (2020) Genome-wide identification and characterization of

TCP family genes in *Brassica juncea* var. tumida. Peer J 8:e9130

Heng S, Wei C, Jing B, Wan Z, Wen J, Yi B, Ma C, Tu J, Fu T, Shen J (2014) Comparative analysis of mitochondrial genomes between the hau cytoplasmic male sterility (CMS) line and its iso-nuclear maintainer line in *Brassica juncea* to reveal the origin of the CMS-associated gene orf288. BMC Genomics 15 (1):1–12

Heng S, Gao J, Wei C, Chen F, Li X, Wen J, Yi B, Ma C, Tu J, Fu T (2018) Transcript levels of orf288 are associated with the hau cytoplasmic male sterility system and altered nuclear gene expression in *Brassica juncea*. J Exp Bot 69(3):455–466

Hong CP, Plaha P, Koo D-H, Yang T-J, Choi SR, Lee YK, Uhm T, Bang J-W, Edwards D, Bancroft I (2006) A Survey of the *Brassica rapa* genome by BAC-end sequence analysis and comparison with Arabidopsis thaliana. Molecules Cells (Springer Science and Business Media BV) 22(3)

Hong CP, Kwon S-J, Kim JS, Yang T-J, Park B-S, Lim YP (2008) Progress in understanding and sequencing the genome of *Brassica rapa*. Int J Plant Genomics 2008

Inturrisi F, Bayer PE, Yang H, Tirnaz S, Edwards D, Batley J (2020) Genome-wide identification and comparative analysis of resistance genes in *Brassica juncea*. Mol Breed 40(8):1–14

Jing B, Heng S, Tong D, Wan Z, Fu T, Tu J, Ma C, Yi B, Wen J, Shen J (2012) A male sterility-associated cytotoxic protein ORF288 in *Brassica juncea* causes aborted pollen development. Jexp Bot 63(3):1285–1295

Kim JS, Chung TY, King GJ, Jin M, Yang T-J, Jin Y-M, Kim H-I, Park B-S (2006) A sequence-tagged linkage map of *Brassica rapa*. Genet 174(1):29–39

Kotani H, Nakamura Y, Sato S, Kaneko T, Asamizu E, Miyajima N, Tabata S (1997) Structural analysis of *Arabidopsis thaliana* chromosome 5. II. Sequence features of the regions of 1,044,062 bp covered by thirteen physically assigned P1 clones. DNA Res 4 (4):291–293

Lagercrantz U, Lydiate DJ (1995) RFLP mapping in *Brassica nigra* indicates differing recombination rates in male and female meioses. Genome 38(2):255–264

Li P, Zhang S, Li F, Zhang S, Zhang H, Wang X, Sun R, Bonnema G, Borm TJ (2017) Front Plant Sci 8:111

Li M, Sun B, Xie F, Gong R, Luo Y, Zhang F, Yan Z, Tang H (2019) Identification of the GRAS gene family in the *Brassica juncea* genome provides insight into its role in stem swelling in stem mustard. Peer J 7: e6682

Li M, Xie F, Li Y, Gong L, Luo Y, Zhang Y, Chen Q, Wang Y, Lin Y, Zhang Y (2020a) Genome-wide analysis of the heat shock transcription factor gene family in *Brassica juncea*: structure, evolution, and expression profiles. DNA Cell Biol 39(11):1990–2004

Li W, Chen F, Wang Y, Zheng H, Yi Q, Ren Y, Gao J (2020b) Genome-wide identification and functional analysis of ARF transcription factors in *Brassica juncea* var. tumida. PloS One 15(4):e0232039

Luo M, Wing RA (2003) An improved method for plant BAC library construction. In: Plant functional genomics. Springer, pp 3–19

Mir ZA, Ali S, Shivaraj S, Bhat JA, Singh A, Yadav P, Rawat S, Paplao PK, Grover A (2020) Genome-wide identification and characterization of Chitinase gene family in *Brassica juncea* and Camelina sativa in response to Alternaria brassicae. Genomics 112 (1):749–763

Mozo T, Dewar K, Dunn P, Ecker JR, Fischer S, Kloska S, Lehrach H, Marra M, Martienssen R, Meier-Ewert S (1999) A complete BAC-based physical map of the *Arabidopsis thaliana* genome. Nat Genet 22 (3):271–275

Mun J-H, Kwon S-J, Yang T-J, Kim H-S, Choi B-S, Baek S, Kim JS, Jin M, Kim JA, Lim M-H (2008) The first generation of a BAC-based physical map of *Brassica rapa*. BMC Genomics 9(1):1–11

Mun J-H, Kwon S-J, Seol Y-J, Kim JA, Jin M, Kim JS, Lim M-H, Lee S-I, Hong JK, Park T-H (2010) Sequence and structure of *Brassica rapa* chromosome A3. Genome Biol 11(9):1–12

Paritosh K, Pradhan AK, Pental D (2020) A highly contiguous genome assembly of *Brassica nigra* (BB) and revised nomenclature for the pseudochromosomes. BMC Genomics 21(1):1–12

Paritosh K, Yadava SK, Singh P, Bhayana L, Mukhopadhyay A, Gupta V, Bisht NC, Zhang J, Kudrna DA, Copetti D (2021) A chromosome-scale assembly of allotetraploid *Brassica juncea* (AABB) elucidates comparative architecture of the A and B genomes. Plant Biotechnol J 19(3):602

Paritosh K, Gupta V, Yadava SK, Singh P, Pradhan AK, Pental D (2014) RNA-seq based SNPs for mapping in *Brassica juncea* (AABB): synteny analysis between the two constituent genomes A (from *B. rapa*) and B (from *B. nigra*) shows highly divergent gene block arrangement and unique block fragmentation patterns. BMC Genomics 15(1):1–14

Perumal S, Koh CS, Jin L, Buchwaldt M, Higgins EE, Zheng C, Sankoff D, Robinson SJ, Kagale S, Navabi Z-K (2020) A high-contiguity *Brassica nigra* genome localizes active centromeres and defines the ancestral Brassica genome. Nat Plants 6(8):929–941

Prabhudas SK, Raju B, Kannan Thodi S, Parani M, Natarajan P (2016) The complete chloroplast genome sequence of Indian mustard (*Brassica juncea* L.). Mitochondrial DNA A 27(6):4622–4623

Prakash S, Chopra V (1991) Cytogenetics of crop Brassicas and their allies. In: Developments in Plant Genetics and Breeding, vol 2. Elsevier, pp 161–180

Schmidt R, West J, Love K, Lenehan Z, Lister C, Thompson H, Bouchez D, Dean C (1995) Physical map and organization of *Arabidopsis thaliana* chromosome 4. Science 270(5235):480–483

Schmidt R, Love K, West J, Lenehan Z, Dean C (1997) Description of 31 YAC contigs spanning the majority

of *Arabidopsis thaliana* chromosome 5. Plant J 11 (3):563–572

Shizuya H, Birren B, Kim U-J, Mancino V, Slepak T, Tachiiri Y, Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci 89(18):8794–8797

Slocum M, Figdore S, Kennard W, Suzuki J, Osborn T (1990) Linkage arrangement of restriction fragment length polymorphism loci in *Brassica oleracea*. TheorAppl Genet 80(1):57–64

Song K, Suzuki J, Slocum M, Williams P, Osborn T (1991) A linkage map of*Brassica rapa* (syn. campestris) based on restriction fragment length polymorphism loci. Theor Appl Genet 82(3):296–304

Tao Q, Zhao H, Qiu L, Hong G (1994) Construction of a full bacterial artificial chromosome (BAC) library of *Oryza sativa* genome. Cell Res 4(2):127–133

Thormann C, Ferreira M, Camargo L, Tivang J, Osborn T (1994) Comparison of RFLP and RAPD markers to estimating genetic relationships within and among cruciferous species. Theor Appl Genet 88(8):973–980

Verma D, Lakhanpal N, Singh K (2019) Genome-wide identification and characterization of abiotic-stress responsive SOD (superoxide dismutase) gene family in *Brassica juncea* and *B. rapa*. BMC Genomics 20 (1):1–18

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43(10):1035–1039

Wang P, Zhao Z, Zhang Z, Cai Z, Liao J, Tan Q, Xiang M, Chang L, Xu D, Tian Q (2021) Genome-wide identification and analysis of NPR family genes in *Brassica juncea* var. tumida. Gene 769:145210

Woo S-S, Jiang J, Gill BS, Paterson AH, Wing RA (1994) Construction and characterization of bacterial artificial chromosome library of Sorghum bicolor. Nucleic Acids Res 22(23):4922–4931

Wu Y, Tulsieram L, Tao Q, Zhang H-B, Rothstein SJ (2000) A binary vector-based large insert library for *Brassica napus* and identification of clones linked to a fertility restorer locus for Ogura cytoplasmic male sterility (CMS). Genome 43(1):102–109

Wu Z, Hu K, Yan M, Song L, Wen J, Ma C, Shen J, Fu T, Yi B, Tu J (2019) Mitochondrial genome and transcriptome analysis of five alloplasmic male-sterile lines in *Brassica juncea*. BMC Genomics 20 (1):1–15

Yang J, Liu X, Yang X, Zhang M (2010) Mitochondrially-targeted expression of a cytoplasmic male sterility-associated orf220 gene causes male sterility in *Brassica juncea*. BMC Plant Biol 10 (1):1–10

Yang J, Liu D, Wang X, Ji C, Cheng F, Liu B, Hu Z, Chen S, Pental D, Ju Y (2016) The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. Nat Genet 48(10):1225–1232

Yang H, Bayer PE, Tirnaz S, Edwards D, Batley J (2021) Genome-wide identification and evolution of receptor-like kinases (RLKs) and receptor like proteins (RLPs) in *Brassica juncea*. Biology 10(1):17

Zachgo EA, Wang ML, Dewdney J, Bouchez D, Camilleri C, Belmonte S, Huang L, Dolan M, Goodman HM (1996) A physical map of chromosome 2 of *Arabidopsis thaliana*. Genome Res 6(1):19–25

Zhang H-B, Wu C (2001) BAC as tools for genome sequencing. Plant Physiol Biochem 39(3–4):195–209

Zhang L, Cai X, Wu J, Liu M, Grob S, Cheng F, Liang J, Cai C, Liu Z, Liu B (2018) Improved *Brassica rapa* reference genome by single-molecule sequencing and chromosome conformation capture technologies. Hortic Res 5(1):1–11

Zhao N, Xu X, Wamboldt Y, Mackenzie SA, Yang X, Hu Z, Yang J, Zhang M (2016) MutS HOMOLOG1 silencing mediates ORF220 substoichiometric shifting and causes male sterility in *Brassica juncea*. J Exp Bot 67(1):435–444

Zhao N, Li Z, Zhang L, Yang X, Mackenzie SA, Hu Z, Zhang M, Yang J (2021) MutS HOMOLOG1 mediates fertility reversion from cytoplasmic male sterile *Brassica juncea* in response to environment. PlantCell Environ 44(1):234–246