The Database for *Brassica* Genome Studies—BRAD

14

Feng Cheng, Xiaobo Wang, Jian Wu and Xiaowu Wang

Abstract

More and more *Brassica* species and relative species from Brassicaceae have been sequenced along the technology improvement of seuquencing and genome assembly. Now, how to apply these bulk genomic datasets to assist the scientific research and breeding work becomes an urgent issue that needs to be addressed. To build functional and user-friendly databases that provide both the basic genome sequences and the elaborately analyzed data, as well as some frequently used tools is one of the solutions. A useful genome database for *Brassica* crops should have below features. (1) Free access to all the basic datasets, such as the genome and gene annotation files; (2) The common BLAST tool to compare all sequences available for *Brassica* species; (3) Genome visualization tool to show all kinds of genomic elements in one frame; (4) Well functional annotation of predicted genes for the newly sequenced *Brassica* genome, it's much better if links of orthologs are made between genes of *Brassica* and the model plant *Arabidopsis thaliana*; (5) Information of molecular markers, genetic maps, and population etc. of the *Brassica* species. In this chapter, we take the database BRAD (http://brassicadb.org) as an example to introduce the databases in the research field of *Brassica*.

14.1 Introduction

To assist researchers and breeders to better understand and use these genomic datasets from *Brassica* species, many *Brassica* databases have been built. For genomic studies in *Brassica*, useful *Brassica* databases are necessary to integrate or visualize these bulk datasets to assist the

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

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_14

research and breeding work of related users. These databases include the *Brassica* database BRAD (http://brassicadb.org), Brassica.info (http://www.brassica.info/), BrassEnsembl (http://www.brassica.info/BrassEnsembl/index.html), BrassicaDB (http://brassica.nbi.ac.uk/BrassicaDB/), CropStoreDB (http://www.cropstoredb.org/), and BolBase (http://www.ocri-genomics.org/bolbase/index.html). Each of these databases has a different emphasis, Brassica.info serves as a platform to integrate genomic resources and release news of projects or activities on *Brassica* studies, and it also provides downloading services for some genomic data. BrassEnsembl visualizes different sets of *Brassica* genomic data under a single frame. CropStoreDB provides a practical approach to managing crop genetic data, while BolBase focuses on genomic structure comparisons in the genome of *Brassica oleracea*. Among them, BRAD is the database that aims at building a bridge between the genomes of *Arabidopsis thaliana* and those of the *Brassica* species (Cheng et al. 2011), transferring the research information of genomic studies and the bulk gene functional studies from the model species *A. thaliana* to the newly sequenced *Brassica* species.

In this chapter, we will focus on BRAD and present an overview of the major functions of this *Brassica* database, especially the most important and useful aspect of BRAD—the multiple genomes and gene synteny analyses among *Brassica* and other Brassicaceae species, such as *A. thaliana*. The introduction of BRAD here aims at informing users about what kind of data can be retrieved from the *Brassica* databases, and what kind of analysis can be performed using them. Initially, the BRAD database was built for the community to release and share bulk *Brassica rapa* genomic data. With continuous updating and research progress, BRAD evolved into an important repository for whole genome scale genomic data from all *Brassica* species and relative species in Brassicaceae. It now provides datasets of 12 genomes, such as the *Brassica* A, C, and AC genomes (Wang et al. 2011a; Liu et al. 2014), as well as other Brassicaceae species *A. thaliana*, *Arabidopsis lyrata*, *Schrenkiella parvula*, *Thellungiella halophila*, *Thellungiella salsuginea*, *Aethionema arabicum*, *Capsella rubella*, *Leavenworthia alabamica*, and *Sisymbrium irio* (Initiative 2000; Dassanayake et al. 2011; Hu et al. 2011; Wu et al. 2012; Haudry et al. 2013; Slotte et al. 2013; Yang et al. 2013), including their de novo assembled genome sequences and predicted gene models, associated annotations (InterPro, KEGG2, and SwissProt) and syntenic relationships. BRAD was also designed as an initial access point for other *Brassica* web pages and resources.

Fig. 14.1 Homepage and the navigation of BRAD. Five sections shown as the five navigation menus at the top of the website (*red box*): Browse, Search, Tools, Download, and Links

As shown by the navigation menu on the top of the homepage, there are five sections in BRAD (Fig. 14.1): Browse, Search, Tools, Downloading, and Links. The first three are the major sections, with "Browse" providing marker information for genetic maps and listing genes of important families. The "Search" section provides search functions for users to retrieve information on their interested genes, including the annotations, coding sequences, syntenic gene relationships, which is an important function of BRAD. The "Tools" section provides BLAST services, GBrowse, and its syntenic module to visualize genomic datasets. The last two sections are smaller, with "Downloading" providing accessions for bulk dataset downloads from BRAD, and "Links", which offers websites of other *Brassica* related databases. Generally, based on the information or function affiliation, the contents of these five sections can be separated into four parts: (1) orthologous genes; (2) genomic visualization; (3) molecular tools; and (4) featured data browsing. Below is an introduction to the four grouped tools in BRAD.

14.2 Orthologous Genes Among *Brassicas* and Other Brassicaceae Species

Syntenic and nonsyntenic genes are useful, especially for predicting genes in newly-sequenced genomes. BRAD helps users to share gene information from the well-studied model plant *A. thaliana* and apply it to new genomes.

14.2.1 Syntenic Paralogs and Orthologs

Syntenic gene analysis among multiple genomes in Brassicaceae is the core function and featured valuable resource in BRAD. Syntenic genes are homologs in two or more genomes that are inherited from the most recent common ancestor, and thus, these genes have maintained both the sequence's homology and the linear relationship on the chromosomes among these species. The tool SynOrths was applied to determine accurate syntenic orthologs between two genomes (Cheng et al. 2012a). SynOrths determines if two genes of two genomes are a syntenic pair using both their sequence similarity and the colinearity—the homology of their flanking genes (Fig. 14.2). In detail, SynOrths uses one genome, for example, *B. rapa* as the query genome and the others as the subject genomes. Under default parameters (m = 20, n = 100, and r = 0.2), it first identifies homologous genes between two genomes using BLASTP (E-value < 1×10^{-20} or best hits). The 20 closest genes (m = 20), flanking either side of the gene in the query *B. rapa* genome, are then compared with the 100 closest genes (n = 100) flanking either side of the gene in the subject genome. If at least 20 % (r = 0.2) of the best hits for the 40 genes (20 × 2 for both sides) in the query *B. rapa* genome are found within the 200 genes (100 × 2 for both sides) in the subject genome, then the original pair of are designated as a syntenic ortholog candidate. Syntenic genes among the 12 genomes in Brassicaceae were determined using this tool accompanied with further analysis, and they were integrated into

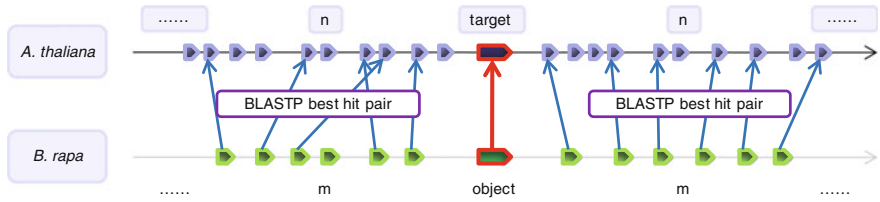

Fig. 14.2 Algorithm of syntenic gene determination in SynOrths. Both the sequence homology of the gene pair and their flanking genes are considered to determine whether they are under synteny

one sheet and embedded in BRAD as a searching function of syntenic genes. Again, using *B. rapa* as an example, 30,773 syntenic pairs between *B. rapa* and *A. thaliana* were obtained, and there were 9293, 6683, and 2346 *A. thaliana* genes, having one, two, and three paralogs, respectively, in the three subgenomes LF, MF1, and MF2 of *B. rapa* (the three subgenomes originated from a genome triplication). LF, MF1, and MF2 are abbreviations for less fractionized, more fractionized 1 and more fractionized 2, respectively, denoting subgenomes with more or fewer genes retained (Wang et al. 2011a; Cheng et al. 2012b). Now BRAD provides the gene set that shows conserved synteny between the three subgenomes of *Brassica* and *A. thaliana*, as well as other Brassicaceae species. These syntenic genes are re-sorted and listed according to the genes' order in the seven chromosomes of the *Brassica* ancestor (tPCK genome structure) (Cheng et al. 2012b). By typing any gene ID from the 12 Brassicaceae genomes, their syntenic genes in the other genomes and the homologous relationships of flanking genes can be easily obtained (Fig. 14.3). In addition, a small dialog window, which is linked to each gene ID in the output

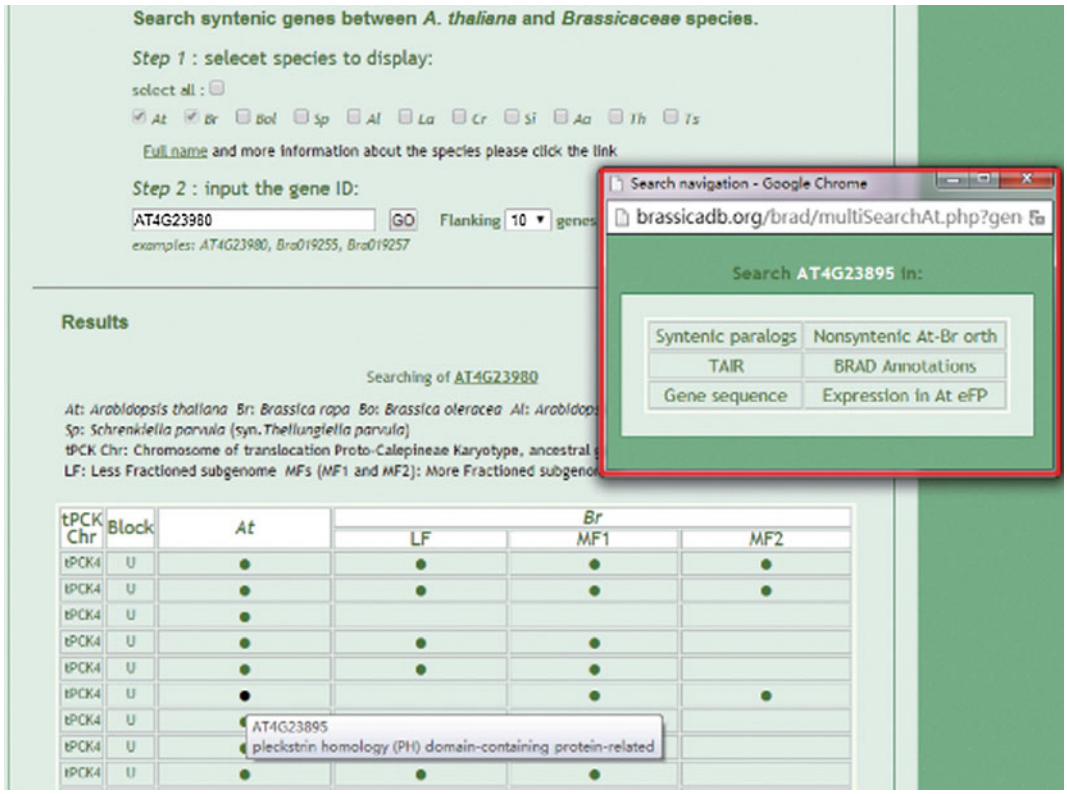

Fig. 14.3 Syntenic genes among *Brassica* and other Brassicaceae species. Taking the *A. thaliana* gene AT4G23895 as an example, search results are shown in a table. Each *solid circle* in the table represents a gene. The first column lists the chromosomes in the *Brassica* ancestral genome tPCK, followed by genomic blocks and *A. thaliana* gene IDs. The next three columns show genes from the subgenomes LF, MF1, and MF2 of *Brassica rapa*. In each *row*, listed genes are in a syntenic relationship. Moving the cursor over an *A. thaliana* gene gives a floating box containing the gene's annotation, while moving the cursor over a *B. rapa* gene causes the supporting information on its syntenic relationship to the *A. thaliana* gene to appear in a pop-up window. By clicking a gene in the table, users will open a small dialog window (*red box*) that offers links to its annotation, the best BLASTX hit to *A. thaliana*, and the function and Gene Ontology (GO) of the matching gene, as shown in Fig. 14.2

table, gives more information on the according gene (Fig. 14.3), which will help users to retrieve more useful data for their study. The dialog window will let users access information, not only in BRAD, but also in other databases, such as TAIR, for further detailed information. This dialog window allows simple and easy navigation, integrating all available information on a certain gene into one menu for users to access by clicking. This dialog function is used commonly in most pages of BRAD.

14.2.2 Nonsyntenic Orthologs

The function of searching nonsyntenic genes is a supplement of syntenic gene analyzing, which helps users to obtain homologous genes that originated through transposition events. Nonsyntenic genes in two genomes were determined under the following two rules: (1) the BLASTP alignment should be satisfied by: identity >70 %, coverage of both genes >75 %, and 2) the two genes should not be a syntenic pair. Using *B. rapa* and *A. thaliana* as examples, a total of 17,159 such nonsyntenic orthologs were determined.

14.3 Visualization of Genomic Features and Genomic Synteny

A user-friendly visualization page of genomic data will help users to understand and apply the data to easily assist their research. Here, two visualized pages were built in BRAD. However, the visualization aspect still needs improvement.

14.3.1 Genome Browse (GBrowse)

The Genome Browser tool developed by the Generic Model Organism Database Project (Donlin 2009), (http://gmod.org) was adopted by BRAD to visualize the Brassicaceae genomes. Three major levels are displayed: chromosome where the search target is located, genome segment with flanking regions of the search target, and the exact target. BRAD GBrowse now provides available information, including predicted gene models, transposons, different types of RNA sets, and genetic markers, for these species.

14.3.2 Synteny Blocks

A synteny module of GBrowse, SynBrowse, was used to visualize the relationship of genomic synteny between the 12 Brassicaceae genomes. Information on syntenic blocks was identified based on the relationship of synteny genes determined by SynOrths. The genomic fragments in which the syntenic genes retain the linear relationship between the two genomes have been merged to synteny blocks. These conserved and pairwise blocks were then visualized by SynBrowse. The genes that show in the synteny blocks were provided with links to the synteny gene searching section, which then leads users to the one-to-one gene synteny information and further resources through the dialog window.

14.4 Useful Tools for Molecular Studies

BRAD developed or adopted some useful tools and functions for molecular studies, including gene function annotations, local functional elements searching, and BLAST service.

14.4.1 Searching Genes by Keywords of Function Annotations

Six kinds of annotation datasets were provided here: SwissProt, Trembl, KEGG, InterPro domain, Gene Ontology, and the BLASTX (best hit) of *Brassica* to *A. thaliana*. These datasets are used to annotate different aspects of newly predicted gene models, such as nucleotide sequences, proteins, and domains. SwissProt and Trembl annotations are generated by BLASTP best hit

(cutoff E-value: 1e-5) based on the predicted proteins in the Swiss-Prot and TrEMBL databases. Predicted genes are mapped to KEGG pathways based on the best hit from the Swiss-Prot database. InterPro is used to annotate motifs and domains of predicted genes by comparison to public databases, including Pfam, PRINTS, PROSITE, ProDom, and SMART by using applications, such as hmmpfam, fprintscan, ScanRegExp profilescan, blastprodom, and hmmsmart. Gene Ontology information is extracted from the InterPro results. In total, in the gene annotation pages, for all the predicted genes of the 12 genomes listed above, we collected 244,836 gene ontology records, 275,161 InterPro domain annotation records, 84,568 KEGG records, 71,076 SwissProt records, 37,220 Trembl records, and 14,790 BLASTX best hits to *A. thaliana*. This bulk information will help users to have a better understanding of genes in newly sequenced genomes. The most up-to-date information can be checked through the address: http://brassicadb.org/brad/searchAll.php. Orthologs between new genomes and the model plant *A. thaliana* are also used to annotate these new genes.

14.4.2 Screening for Elements in Local Genomic Regions

This section was developed to help users to locate genomic elements that are colocated with molecular markers, other elements, or flanking the region of interest. Users can easily perform the search by inputting a physical position, a gene ID, or genetic marker, accompanied with the size of the flanking regions to be searched. All of the genomic features, such as genes, transposons, and RNAs (miRNA, tRNA, rRNA, and snRNA) that are located in the searched region will be collected and displayed in a table. A link to GBrowse provides an option to visualize the search region on the background of the complete chromosome. It is a useful function for certain studies, such as the fine mapping of QTLs. Once QTLs are obtained, existing markers from BRAD can be used directly for searching, while new markers should be aligned to the genome sequence with the BLAST tool in BRAD to locate their physical positions. The flanking regions of these markers can then be checked using this tool to locate candidate genomic elements, such as genes or small RNAs that might be the causal factors of the QTLs. As the research progresses, BRAD can further enable the searching of flanking regions by adding more datasets, making it an integrative and valuable resource pool for molecular geneticists and breeders.

14.4.3 Bulk Resources for BLAST Services

Standard wwwblast modules were adopted by BRAD to help users to perform homologous sequence alignment and searching. The BLAST databases collect and provide genomes, genes and proteins sequences, or EST sequences available for the 12 Brassicaceae species. With this tool and resources, users can screen for homologous relationships between their studied sequence and the Brassicaceae databases in BRAD easily and efficiently.

14.5 Resources of Browsing and Bulk Data Downloading

In this part, BRAD provides browsable pages of gene lists from important gene families and genetic maps. BRAD also offers access to all the datasets that are employed in BRAD for users to download.

14.5.1 Browse of Linkage Maps and Gene Families

Gene lists for important gene families, such as auxin, glucosinolate, flowering, transcription factor, and resistant genes, as well as another 182 gene families and their orthologous relationship to *A. thaliana* are provided for browsing. For

each gene ID that appears on these pages, a link to a small dialog window is provided.

Besides the gene family information, BRAD collected 1160 genetic markers, including 758 SSR and 402 InDel markers, covering all 10 chromosomes (Choi et al. 2007; Kim et al. 2009) from three population lines of *B. rapa*: RCZ16_DH, JWF3P, and VCS_DH. RCZ16_DH is a population developed from a cross between a rapid cycling line, L144, and a summer type Chinese cabbage doubled haploid (DH) line Z16 (Wang et al. 2011b). Markers of RCZ16_DH were developed based on the resequencing data of their parents L144 and Z16. The other two maps, JWF3P and VCS_DH, were integrated from public database http://www.brassica-rapa.org. This information can be browsed by clicking on links on the pages, the information is collected and displayed in a hierarchical structure, with more clicks providing users with more detailed information.

Gene family information on other Brassicaceae species and the genetic markers of other populations or *Brassica* species will be easily added to BRAD when the data is available.

14.5.2 Download and External Links

BRAD provides accessions for bulk data downloads, including genome, gene, and protein sequences, gene annotations, and other predicted genomic elements. In addition, BRAD also collects and provides numerous community resources either as data or external website links, including websites of laboratories focusing on Brassicaceae, or *Brassica* breeding.

14.6 General Guidelines for Using BRAD

14.6.1 Browse Molecular Markers and Genetic Maps

For each marker in the "Browse" section, BRAD presents its genetic and physical positions, primer information, as well as its parental populations. Users can access these data in the following order: chromosome selection → population specification → detailed marker information → click marker ID for primer information.

14.6.2 Search Using Annotations and Syntenic Genes

In the annotation search section, users can find genes with interesting functions by typing a keyword, such as flower or growth, and then relevant records will be compared from the six annotation datasets as described above. Clicking on the selected records will then lead users to genes with annotations related to the keyword. A further click of the gene ID will show users with more detailed gene information. Syntenic genes can only be searched by using gene IDs of either species stored by BRAD. In the web of syntenic paralogs, the pull-down 'flanking' menu has two options (10 or 20), which means it will extend 10 or 20 genes up- and downstream of the searched gene. In the tabulated output, such as in Fig. 14.3, the targeted gene is in the middle of its flanking genes. Each *A. thaliana* gene corresponds to 1, 2, or 3 genes in the three subgenomes of *B. rapa*. "No circle" indicates that there is no gene identified. Moving the cursor over the ID of a gene expands the functional annotations of *A. thaliana* genes and the detailed supporting information of synteny relationships of *B. rapa* genes to that of *A. thaliana*.

14.6.3 Search Navigation

As referred to above, the dialog window is a useful and easy to follow tool in BRAD. BRAD embedded this JavaScript-driven small dialog window as the navigation tool for each gene ID in any of the output tables, which help users to quickly access all the information on an interested gene in BRAD. By combining the access points for many datasets into one window, the navigation can lead users to different resources on the target genes, which facilitates the use of BRAD. The navigation window now integrates resources

such as gene annotations, syntenic or nonsyntenic orthologs, gene sequence, functional elements in the gene's flanking regions and data visualization in GBrowse, links to information of TAIR databases, and gene expression of *A. thaliana*.

14.7 Conclusion and Perspective

BRAD, a database focused on gene function illustration and multiple genome comparisons among *Brassica* and other Brassicaceae species, especially the model plant *A. thaliana*, has been built in time because the genome studies on *Brassica* species are increasing. Compared with other databases on *Brassica* plants, BRAD keeps its specific core function and advantages, especially its deep mining of syntenic genes and genomic blocks in the newly assembled *Brassica* genomes, as well as the use of the bulk research information from the model plant *A. thaliana*. By the continuous improvement of applications and the integration of more available datasets in the future, BRAD will help scientists and breeders to fully and efficiently use the information on genomic and genetic datasets of *Brassica* plants. BRAD is a valuable resource for the scientists of comparative genomics, plant evolution, and molecular biology, and the breeders of *Brassica*.

In the future, BRAD should integrate datasets of all *Brassica* species available. In addition, to meet the demands of the genetic and genomic studies on *Brassica* crops, the following data types and applications should be updated on BRAD. (1) Resequencing data of different lines in each *Brassica* species; SNP, InDel, or SV loci and their frequency in populations, as well as the haplotypes (derived from SNPs variation) of *Brassica* germplasm collections that are generated from resequencing data. (2) Transcriptome sequencing or mRNA-Seq data, gene expression data in different organs, and different accessions of each *Brassica* species. (3) Whole genome methylation and small RNA-Seq data for *Brassica* species. (4) Visualization and interactive technologies should be adopted to improve BRAD to be a much more user-friendly database. Solutions to visualize the syntenic relationships of multiple genomes from the chromosome scale, to genomic segments, to genes or tens of nucleotides level should be employed, thus helping users to have a better understanding of genome and gene evolution at different scales among different *Brassica* species and their relation to other Brassicaceae species.

References

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

Cheng F, Wu J, Fang L, Wang X (2012a) Syntenic gene analysis between Brassica rapa and other Brassicaceae species. Front Plant Sci 3:198–380

Cheng F, Wu J, Fang L, Sun S, Liu B et al (2012b) Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. PLoS One **7**:e36442

Choi SR, Teakle GR, Plaha P, Kim JH, Allender CJ et al (2007) The reference genetic linkage map for the multinational Brassica rapa genome sequencing project. Theor Appl Genet 115:777–792

Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H et al (2011) The genome of the extremophile crucifer Thellungiella parvula. Nat Genet 43:913–918

Donlin MJ (2009) Using the generic genome browser (GBrowse). Curr Protoc Bioinform Chapter9: Unit 9 9

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M et al (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet 45:891–898

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF et al (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet 43:476–481

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

Kim H, Choi SR, Bae J, Hong CP, Lee SY et al (2009) Sequenced BAC anchored reference genetic map that reconciles the ten individual chromosomes of Brassica rapa. BMC Genom 10:432

Liu S, Liu Y, Yang X, Tong C, Edwards D et al (2014) The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nature Comunication In press

Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F et al (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Genet 45:831–835

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

Wang Y, Sun S, Liu B, Wang H, Deng J et al (2011b) A sequence-based genetic linkage map as a reference for Brassica rapa pseudochromosome assembly. BMC Genom 12:239

Wu HJ, Zhang Z, Wang JY, Oh DH, Dassanayake M et al. (2012) Insights into salt tolerance from the genome of Thellungiella salsuginea. Proc Natl Acad Sci USA 109:12219–12224

Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J et al (2013) The Reference Genome of the Halophytic Plant Eutrema salsugineum. Front Plant Sci 4:46