Genes, Genomes and Genomics ©2007 Global Science Books

The *Brassica rapa* Genome Sequencing Initiative

Martin Trick[1] • Ian Bancroft[2] • Yong Pyo Lim[3*]

[1] Department of Computational and Systems Biology, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK
[2] Department of Crop Genetics, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK
[3] Department of Horticulture, College of Agriculture and Life Sciences, Chungnam National University, Daejeon 305-764, Korea

Corresponding author: * **yplim@cnu.ac.kr**

ABSTRACT

Brassica rapa ssp. *pekinensis* (Chinese cabbage), a member of the Brassicaceae, is both an economically important crop and a model plant for the study of polyploidization and phenotypic evolution. Its genome (the *Brassica* A-genome) is being sequenced by the Multinational *Brassica rapa* Genome Sequencing Project (M*Br*GSP), an international consortium of six countries (Korea, Canada, UK, China, USA, and Australia). This consortium has developed a number of genomic resources, including 2 genetic mapping populations, 2 BAC libraries, 22 cDNA libraries, 107,280 BAC-end sequences, and 104,914 ESTs and has constructed a genetic and physical map of *B. rapa* using them. As a strategy for sequencing the genespace, the project has adopted a BAC-by-BAC sequencing approach. In order to implement this, 629 seed BACs mapped genetically throughout the genome have been selected and sequenced. At this preliminary phase of the project, eight of the ten chromosomes have been allocated between the participants; Korea (R3 and R9), Canada (R2 and R10), UK/China (R1 and R8), USA (R6) and Australia (R7) and they are being separately sequenced. In this article, we assess the current status of the *B. rapa* genome sequencing project, including the available and developing computational and bioinformatics infrastructure.

Keywords: BAC-by-BAC sequencing approach, *Brassica rapa* ssp. *pekinensis*, genomic resources, genome structure, Multinational *Brassica rapa* Genome Sequencing Project (M*Br*GSP)

CONTENTS

INTRODUCTION

The genus *Brassica* is one of the core genera within the tribe *Brassiceae* (family Brassicaceae) and includes a number of crops with adaptation to a wide of variety of agroclimatic conditions. Economically, *Brassica* crops are widely used as important sources of vegetable oil, fresh and preserved vegetables and also condiments. The genetic relationships between six important *Brassica* species are classically represented in U's triangle (U 1935) as three different diploids (the A, B and C genomes) and the three amphidiploids formed between them by hybridization. *Brassica rapa* (AA, 2n=20, 529 Mbp/1C), *B. nigra* (BB, 2n=16, 632 Mbp/1C) and *B. oleracea* (CC, 2n=18, 696 Mbp/1C) are the monogenomic diploids (U 1935; Johnston *et al.* 2005). The remaining three species, *B. juncea* (AABB, 2n=36, 1,068 Mbp/1C), *B. napus* (AACC, 2n=38, 1,132 Mbp/1C) and *B. carinata* (BBCC, 2n=34, 1,284 Mbp/1C) are amphidiploids, which have evolved via hybridization between different monogenomic diploids (U 1935; Johnston *et al.* 2005).

Brassica species are characterized by remarkable morphological diversity with regard to inflorescences, leaves, stems, roots, and apical buds (Paterson *et al.* 2000). This phenomenon has been suggested to be causally linked to both recent and ancient changes in genomic ploidy levels (Lukens *et al.* 2004). In *Brassica* species, such polyploidization events have brought about genome-scale multiplication and rearrangements with subsequent genetic variation, including insertions, deletions, and substitutions. The results may have produced novel phenotypic variation for important traits (Paterson *et al.* 2000; Lukens *et al.* 2004). Thus, analysis of *Brassica* genomics may provide us with greater insight into the rapid phenotypic evolution of plants associated with polyploidization.

Due to the roles of *Brassica* species as important crops throughout the world, and their recognized potential as model plants for polyploidization studies, genome sequencing projects have been proposed at various times for each of the three basal diploid *Brassica* species. A cultivar of *B. rapa* ssp. *pekinensis* (Chinese cabbage) has been selected as the target for A genome sequencing in the Multinational

Received: 7 March, 2007. Accepted: 29 April, 2007.

Invited Mini-Review

Brassica rapa Genome Sequencing Project (M*Br*GSP), a component of the Multinational *Brassica* Genome Project (MBGP). The project is being undertaken by six countries (Korea, Canada, UK, USA, Australia and China), with the objective of determining the complete sequence of the genespace of this genome using a BAC-by-BAC sequencing approach (Lim *et al.* 2006).

STRUCTURAL FEATURES OF THE *B. rapa* GENOME

Genome triplication

Brassica species are phylogenetically related to *Arabidopsis thaliana*, having diverged 17-18 million years ago (MYA) from their common ancestor (Yang *et al.* 2006). *A. thaliana*, which has been completely sequenced, has a rather small genome (about 146 Mb), relatively little repetitive DNA and a high gene density (The Arabidopsis Genome Initiative 2000). Comparative genetic studies between diploid *Brassica* species and *A. thaliana* have revealed the presence of collinear chromosome segments, a conserved gene order, and a high degree of sequence conservation, but with some variation in gene contents occurring as the result of deletions or insertions (Lagercrantz 1998; O'Neill and Bancroft 2000; Rana *et al.* 2004; Park *et al.* 2005; Kim *et al.* 2006; Town *et al.* 2006; Yang *et al.* 2006). Most importantly, these studies have clearly demonstrated that diploid *Brassica* species contain extensively triplicated counterparts of the corresponding homologous segments of the *A. thaliana* genome, thereby suggesting that diploid *Brassica* species may have been derived from a hexaploid ancestor. In conjunction with this theory, Yang *et al.* (2006) recently reported that three paralogous subgenomes of diploid *Brassica* species triplicated 13 to 17 MYA, very soon after the *Arabidopsis* and *Brassica* divergence occurred, and that these have undergone a dynamic and ongoing diploidization process via chromosome rearrangements, substitutions, insertions, and deletions. The ancestral genome triplication in the various diploid *Brassica* species has resulted in a 3-5–fold inflation in extant genome sizes with respect to *A. thaliana*.

Karyotype of *B. rapa*

Recently, a molecular cytogenetic map of *B. rapa* was constructed on the basis of analysis of mitotic metaphase and pachytene chromosomes via fluorescence *in situ* hybridization (FISH) using three repetitive DNA sequence probes, 5S rDNA, 45S rDNA and C11-350H, a pericentromeric satellite repeat (Koo *et al.* 2004). The observed lengths of the mitotic metaphase chromosomes ranged from 1.46 μm to 3.30 μm, forming the basis of a karyotype (Chr 1-10) which can be reconciled with the nomenclature for genetic linkage groups (R1-10). The mean lengths of the ten chromosomes at pachytene ranged from 23.7 μm to 51.3 μm, a total of 385.3 μm, which is 17.5-fold longer than that of the mitotic metaphase chromosomes. Five 45S rDNA loci were identified by FISH on chromosomes 1, 2, 4, 5, and 7 and three 5S rDNA loci on chromosomes 2, 7 and 10. The C11-350H loci were located on all the mitotic metaphase chromosomes, except chromosomes 2 and 4. Based on these results, the total length of all heterochromatic regions was estimated to be 38.2 μm, which is approximately 10% of the total length of pachytene chromosomes. The pachytene karyotype was determined to consist of two metacentric chromosomes (chromosomes 1 and 6), five submetacentric chromosomes (chromosomes 3, 4, 5, 9 and 10), two subtelocentric chromosomes (chromosomes 7 and 8) and one acrocentric chromosome (chromosome 2).

Sequence composition of the genome

We surveyed the *B. rapa* genome via the analysis of 12,017 *Hin*dIII BAC-end sequences (Hong *et al.* 2006). From this sample we would estimate the *B. rapa* genome to contain 43,000-45,000 genes, a complement 1.6-1.7 times greater than that seen in the genome of *A. thaliana*. Genomic triplication events in *B. rapa* have been suggested to lead to an increase in gene number with subsequent gene loss (Bowers *et al.* 2003; Yang *et al.* 2006). Transposable elements (TEs) have been estimated to occupy approximately 14% of the genome (covering approximately 74 Mb), with a predominance of retrotransposons, 8.2 times greater than that observed previously in *A. thaliana* (Zhang and Wessler 2004). Their amplification in *B. rapa* may have played a crucial role in both evolution and genomic expansion. Simple sequence repeats (SSRs) have been estimated to occur with a frequency of approximately one per 5.5 kb within the *B. rapa* genome, as compared to approximately one per 3.2 kb within the *A. thaliana* genome.

THE *B. rapa* GENOME SEQUENCING PROJECT

Genomic resources

We have developed a doubled haploid (DH) population comprising 78 lines, which was derived from anther culture of the F_1 progeny between the Chinese cabbage inbred lines 'Chiifu' and 'Kenshin'. Using the same parents, we also developed a recombinant inbred population, which comprises over 251 lines, now at the F_7 generation. By isolating high molecular weight leaf DNA from the inbred line 'Chiifu', we have constructed two bacterial artificial chromosome (BAC) libraries using the restriction enzymes, *Hin*dIII and *Bam*HI. The *Hin*dIII and *Bam*HI libraries, which were designated as KBrH and KBrB, consisted of 56,592 and 50,688 clones, with an average insert size of 115 kb and 124 kb, respectively (Park *et al.* 2005; Lim *et al.* 2006). Recently, the ends of all of these BAC clones (107,280) have been sequenced by the M*Br*GSP groups. These BAC-end sequences will be employed for the selection of the minimal tiling paths for genome sequencing, in combination with the BAC fingerprinting data, and also as a resource for development of genetic markers. Twenty two cDNA libraries have also been constructed by Korean research groups from different plant tissues, including leaves, roots, cotyledons, ovules, siliquae, and anthers. The total number of the expressed sequence tags (ESTs) from these libraries is 104,914, with an average length of 575 bases. All these resources underpin the *Brassica* A-genome sequencing project.

Current status of genome sequencing

Using the 78 DH lines described above, a reference genetic linkage map has been constructed (Choi *et al.* unpublished). This map comprises a total of 556 markers, including 278 AFLP, 235 SSR, 25 RAPD and 18 ESTP markers. Ten linkage groups have been identified and designated as R1 to R10, through alignment and orientation using common SSR markers. The total length of the linkage map was 1,183 cM, with an average interval between adjacent loci of 2.68 cM. The linkage groups of the genetic map have been reconciled with the chromosomes of the cytogenetic map using linkage group-specific genetic markers.

Physical mapping has also facilitated the identification of seed BACs, for the BAC-by-BAC sequencing of individual chromosomes. All clones (56,592) in the *Hin*dIII BAC library have been fingerprinted for the construction of a physical map of *B. rapa* (Lim *et al.* unpublished). Among the fingerprinted BAC clones, 17,303 clones have been assorted into 3,854 contigs, representing 90.2% of the genome. The generated BAC contigs have been integrated with the reference genetic map via a search of BAC clones hybridized to genetic markers against a FingerPrinted Contigs (FPC) database. Furthermore, in order to construct a deep-coverage BAC physical map of *B. rapa*, 50,688 *Bam*HI BAC clones have been fingerprinted, using the SNaPshot® methodology (Luo *et al.* 2003).

For whole-genome sequencing, a BAC-by-BAC sequencing approach has been adopted. As the initial step in this process, a Korean *Brassica* genomics team, NIAB, has selected and sequenced 629 seed BACs genetically mapped throughout the whole genome. These seed BACs will be used as the primary scaffold for genome sequencing. At this preliminary stage of the work, eight of the ten chromosomes/linkage groups of *B. rapa* have been allocated to the participating countries as follows; Korea (R3 and R9), Canada (R2 and R10), UK/China (R1 and R8), USA (R6) and Australia (R7). At present, R4 and R5 remain unassigned.

Bioinformatics infrastructure

Bioinformatics and computational resources dedicated to *Brassica* genome research have evolved dramatically over recent years (see, for instance, Beckett *et al.* 2005). By leveraging information from the Arabidopsis genome sequence and by acquiring greater levels of coverage and resolution as the number of publicly available Brassica DNA sequences has risen (with now over a million accessions), there is much of the necessary infrastructure in place to extract added value from the genome sequencing initiative.

At the time of writing, around 212,000 Brassica EST sequences are publicly available – with a further 475,000 due to be released by June 2007. The size and inherent redundancy of this dataset means that a computational priority is the clustering and assembling of these ESTs into tentative consensus (TC) sequences. This, for example, should allow the construction of a fully representative Brassica GeneChip™ (to be undertaken by the Affymetrix company) and will also reduce both the computational overhead and the output complexity for conducting *in silico* alignments to the Arabidopsis sequence. Some work on assembling and aligning such Brassica TCs has already been done (Love *et al.* 2005), with the results available online at http://bioinformatics.pbcbasc.latrobe.edu.au/basc/cgi-bin/ESTDBSearchView.cgi. It is generally found that, when using BLASTN, around 90% of Brassica EST sequences identify Arabidopsis homologues at E-values of less than 10^{-10}. The results of such *in silico* alignments of the raw Brassica EST datasets are available through several different genome browser systems, the ATIDB database (http://atidb.org; Pan *et al.* 2003), AtEnsembl (http://atensembl.arabidopsis.info) and the Brassica-Arabidopsis Comparative Genome Viewer (http://brassica.agr.gc.ca/navigation/viewer_e.shtml).

Evidence for non-correspondences between Brassica ESTs and candidate homologues amongst the Arabidopsis gene models could form the core of a new phase of re-annotation of the Arabidopsis genome, exploiting the power of comparative genomics. At present, such an analysis is to some extent confounded by the conflation of orthologues and paralogues within these various *in silico* alignments. With the coming of the sequencing initiative, the availability of the cognate genomic sequences for the TCs, together with evidence for paralogous assemblies, will help to resolve such ambiguities. Alignment of Brassica ESTs to hitherto un-annotated regions of the Arabidopsis genome is another possibility afforded by *Brassica*-Arabidopsis comparative genomics. Any of the browsers cited above should reveal these rare, but informative, instances. The ATIDB database system also allows for a programmatic discovery of such "new" genes. A preliminary *in silico* experiment with ATIDB (Trick, unpublished results) has revealed 132 such alignments, suggesting the presence of un-annotated genes distributed over all 5 Arabidopsis chromosomes; many of these may correspond to conserved miRNA loci.

By far the largest component of the currently available Brassica sequences comes from single-pass reads of genomic DNA, or Genome Survey Sequences (GSSs). At present there are 808,039 of these, with nearly 600,000 coming from paired end-reads of a *Brassica oleracea* shotgun library constructed from mechanically sheared 2-3 kb fragments, contributed by collaboration between TIGR and the Cold Spring Harbor Laboratory in 2002. Using BLASTN, around 53% of these identify homologous Arabidopsis genome sequences with E-values less than 10^{-10}. A BLAST query interface and a text search facility against a preliminary annotation for this dataset is still available at http://www.tigr.org/tdb/e2k1/bog1/, although both the primary and meta-data are also now largely subsumed in the Arabidopsis-*Brassica* genome browser resources cited above. The primary data has proven to be an extremely useful resource, providing about 0.5x coverage with good read quality, and has been exploited to great effect for Arabidopsis gene discovery or re-annotation (Ayele *et al.* 2005), in a fashion complementary to the EST-based approach described above.

The remainder of the GSS sequences are derived from the paired BAC-end sequences of the *Brassica rapa* KBrH, -B and -S libraries described above, contributed by the various members of the sequencing initiative. These raw data can, in principle, be used simply like those from the *B. oleracea* GSSs, but their special provenance means that they also impart important synteny information and now they are central to the process of constructing the tiling path for the genome sequencing project. Hypothesising that primary sequence homologies for the majority of these paired reads should reflect an underlying microsynteny with the Arabidopsis genome sequence, algorithms can be devised to "virtually map" the Brassica BAC clones. One can programmatically search for paired end-reads that map *in silico* by sequence identity to loci on the same Arabidopsis pseudomolecule, that are separated by a physical distance adjusted for inter-genome inflation/deflation and whose strand matches are self-consistent. This approach has been adopted in ATIDB where, by satisfying these criteria, about 14% of the complete set of BAC clones has been mapped, with the results displayed graphically (http://atidb.org/cgi-perl/gbrowse/atibrowse/). Again, similar strategies have been adopted by both the AtEnsembl database and the BASC project of the PGG bioinformatics group where, in addition to integrating AtEnsembl, a text-based query facility for BAC-ends is also available (Love *et al.* 2006; http://hornbill.cspp.latrobe.edu.au/bacendmapper.html).

With the sequencing initiative coming on stream, we are now moving on to the handling and analysis of finished BAC sequences. A comprehensive website detailing the Korean groups' activities and the raw data from the sequencing of R3 and R9 is at http://www.brassica-rapa.org. The logistics of the project, now expanded to eight of the *B. rapa* chromosomes, will require a number of management tools, some of which are already in place. For instance, a BAC tracking database is now available at http://149.144.200.11/cgi-bin/bac_status/BACStatus.cgi. This aims to serve as a comprehensive reference point for all the collaborators in the *Br*GSP, and incorporates the data for the BACs being sequenced in Korea for linkage groups R3 and R9 that is also served from http://www.brassica-rapa.org.

The major deliverable of the genome sequencing project is to be the rapid availability of assembled, functionally annotated sequence for the *B. rapa* genespace. A "Brassica-centric" view of the emerging assembly, integrated with genetic maps and QTL data using the Ensembl system, will be available from http://ensembl.warwick.ac.uk. It has been agreed that the project will use a modified annotation pipeline from TIGR, using gene prediction programs trained on Brassica. In this way, all the data generated from the partners will be annotated to the same standard. As a first step, a prototype annotation pipeline has been developed, producing a first-pass annotation of 522 seed BACs (http://brassica.bbsrc.ac.uk/cgi-bin/GFF_contents.cgi?db=jic_brassica) which is rendered graphically through the GBrowse genome browser (Stein *et al.* 2002). The system incorporates a number of features, including scanning with sequenced genetic markers, scanning for microsatellite tracts with automated primer selection, scanning with Brassica ESTs

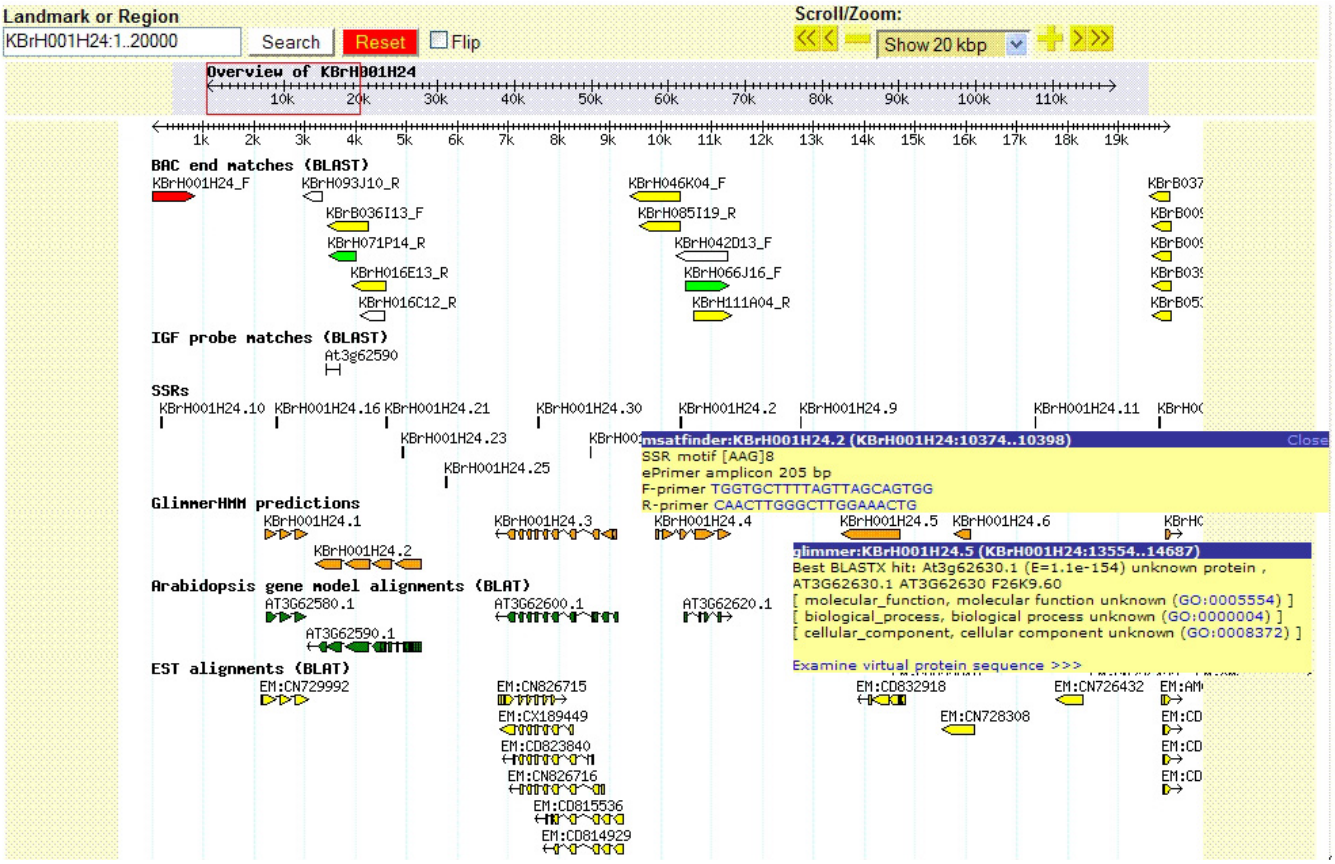

Fig. 1 Screenshot of a representative first-pass annotation of a *B. rapa* seed BAC, KBrH001H24, showing alignments with mapped probes and with ESTs, *in silico* discovery of SSRs and *ab initio* gene predictions. The BAC-end match for KBrH071P14_R, coloured green in the upper left part of the figure, signifies a pre-computed, likely overlap with another seed BAC. All these annotations are available from http://brassica.bbsrc.ac.uk.

with genomic alignment and *ab initio* gene prediction, initially using GlimmerHMM trained on *Arabidopsis*, with links to homologous Arabidopsis proteins and Gene Ontology terms. Recently the seed BACs have been re-scanned with all the BAC-end sequences, with the results rendered and marked up specifically to assist with the selection step for the next BAC in the path (**Fig. 1**). A "do-it-yourself" version of this pipeline is also available over the web at http://brassica.bbsrc.ac.uk/annotate.html. A similar facility is offered by the PGG bionformatics group at http://bioinformatics.pbcbasc.latrobe.edu.au/cgi-bin/AnnotatorView.cgi.

CONCLUSION

The genome sequence of *B. rapa* should not only provide us with novel insights into plant evolution, particularly the polyploidization-diploidization process, but will also provide excellent opportunities for efficient crop improvement in *Brassica* species, not only in *B. rapa*.

ACKNOWLEDGEMENTS

This research was supported by grants from the Rural Development Administration (BioGreen 21 Program) and the Korean Science and Engineering Foundation (R21-2004-000-10010-0), Republic of Korea. Preparatory work for the UK's contribution to the sequencing initiative was supported by BBSRC grant BBS/B07330 and the BBSRC's Competitive Support Grant to the John Innes Centre.

REFERENCES

Ayele M, Haas BJ, Kumar N, Wu H, Xiao Y, van Aken S, Utterback TR, Wortman JR, White OR, Town CD (2005) Whole genome shotgun sequencing of *Brassica oleracea* and its application to gene discovery and annotation in Arabidopsis. *Genome Research* **15**, 487-495

Beckett P, Bancroft I, Trick M (2005) Computational tools for *Brassica*-Arabidopsis comparative genomics. *Comparative and Functional Genomics* **6**, 147-152

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

Hong CP, Plaha P, Koo DH, Yang TJ, Choi SR, Lee YK, Uhm T, Bang JW, Edwards D, Bancroft I, Park BS, Lee J, Lim YP (2006) A Survey of the *Brassica rapa* genome through BAC-end sequence analysis and comparison with *Arabidopsis thaliana*. *Molecules and Cells* **22**, 300-307

Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, Lopez R, Price HJ (2005) Evolution of genome size in Brassicaceae. *Annals of Botany* **95**, 229-235

Kim JS, Chung TY, King GJ, Jin M, Yang TJ, Jin YM, Kim HI, Park BS (2006) A sequence-tagged linkage map of *Brassica rapa*. *Genetics* **174**, 29-39

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

Lagercrantz U (1998) Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. *Genetics* **150**, 1217-1228

Lim YP, Plaha P, Choi SR, Uhm T, Hong CP, Bang JW, Hur YK (2006) Towards unravelling the structure of *Brassica rapa* genome. *Physiologia Plantarum* **126**, 585-591

Love CG, Robinson AJ, Lim GAC, Hopkins CJ, Batley J, Barker G, Spangenberg GC, Edwards D (2005) *Brassica* ASTRA: an integrated database for *Brassica* genomic research. *Nucleic Acids Research* **33**, D656-D659

Love C, Logan E, Erwin T, Kaur J, Lim GAC, Hopkins C, Batley J, James N, May S, Spangenberg G, Edwards D (2006) Integrating and interrogating diverse *Brassica* data within an Ensembl structured database. *Acta Horticulturae* **706**, 77-82

Lukens LN, Quijada PA, Udall J, Pires JC, Schranz ME, Osborn TC (2004) Genome redundancy and plasticity within ancient and recent *Brassica* crop species. *Biological Journal of the Linnean Society* **82**, 665-674

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82**, 378-389

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

Pan X, Liu H, Clarke J, Jones J, Bevan M, Stein L (2003) ATIDB: *Arabidopsis thaliana* insertion database. *Nucleic Acids Research* **31**, 1245-1251

Park JY, Koo DH, Hong CP, Lee SJ, Jeon JW, Lee SH, Yun PY, Park BS, Kim HR, Bang JW, Plaha P, Bancroft I, Lim YP (2005) Physical mapping and microsynteny of *Brassica rapa* ssp. *pekinensis* genome corresponding to a 222 kb gene-rich region of *Arabidopsis* chromosome 4 and partially duplicated on chromosome 5. *Molecular Genetics and Genomics* **274**, 579-588

Paterson AH, Lan TH, Amasino R, Osborn TC, Quiros C (2001) *Brassica* genomics: a complement to, and early beneficiary of, the *Arabidopsis* sequence. *Genome Biology 2: Reviews* 1011.1-1011.4

Rana D, van den Boogaart T, O'Neill CM, Hynes L, Bent E, Macpherson L, Park JY, Lim YP, Bancroft I (2004) Conservation of the microstructure of genome segments in *Brassica napus* and its diploid relatives. *Plant Journal* **40**, 725-733

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S (2002) The generic genome browser: a building block for a model organism system database. *Genome Research* **12**, 1599-1610

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815

Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, Hine EE, Althoff R, Arbogast TS, Tallon LJ, Vigouroux M, Trick M, Bancroft I (2006) Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell* **18**, 1348-1359

U N (1935) Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Japanese Journal of Botany* **7**, 389-452

Yang TJ, Kim JS, Kwon SJ, Lim KB, Choi BS, Kim JA, Jin M, Park JY, Lim MH, Kim HI, Lim YP, Kang JJ, Hong JH, Kim CB, Bhak J, Bancroft I, Park BS (2006) Sequence-level analysis of the diploidization process in the triplicated *FLOWERING LOCUS C* region of *Brassica rapa*. *Plant Cell* **18**, 1339-1347

Zhang X, Wessler SR (2004) Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. *Proceedings of the National Academy of Sciences USA* **101**, 5589-5594