The genome of the mesopolyploid crop species Brassica rapa

Wang, Xiaowu; Wang, Hanzhong; Wang, Jun; Sun, Rifei; Wu, Jian; Liu, Shengyi; Bai, Yinqi; Mun, Jeong-Hwan; Bancroft, Ian; Cheng, Feng; Huang, Sanwen; Li, Xixiang; Hua, Wei; Wang, Junyi; Wang, Xiyin; Freeling, Michael; Pires, J. Chris; Paterson, Andrew H.; Chalhoub, Boulos; Wang, Bo; Hayward, Alice; Sharpe, Andrew G.; Park, Beom-Seok; Weisshaar, Bernd; Liu, Binghang; Li, Bo; Liu, Bo; Tong, Chaobo; Song, Chi; Duran, Christopher; Peng, Chunfang; Geng, Chunyu; Koh, Chushin; Lin, Chuyu; Edwards, David; Mu, Desheng; Shen, Di; Soumpourou, Eleni; Li, Fei; Fraser, Fiona; Conant, Gavin; Lassalle, Gilles; King, Graham J.; Bonnema, Guusje; Tang, Haibao;

Publisher's version / Version de l'éditeur:
https://doi.org/10.1038/ng.919
Nature Genetics, 43, 10, 2011-08-28

The genome of the mesopolyploid crop species *Brassica rapa*

The *Brassica rapa* Genome Sequencing Project Consortium

Abstract:

The Brassicaceae family which includes *Arabidopsis thaliana*, is a natural priority for reaching beyond botanical models to more deeply sample angiosperm genomic and functional diversity. Here we report the draft genome sequence and its annoation of *Brassica rapa*, one of the two ancestral species of oilseed rape. We modeled 41,174 protein-coding genes in the *B. rapa* genome. *B. rapa* has experienced only the second genome triplication reported to date, with its close relationship to *A. thaliana* providing a useful outgroup for investigating many consequences of triplication for its structural and functional evolution. The extent of gene loss (fractionation) among triplicated genome segments varies, with one copy containing a greater proportion of genes expected to have been present in its ancestor (70%) than the remaining two (46% and 36%). Both a generally rapid evolutionary rate, and specific copy number amplifications of particular gene families, may contribute to the remarkable propensity of Brassica species for the development of new morphological variants. The *B. rapa* genome provides a new resource for comparative and evolutionary analysis of the Brassicaceae genomes and also a platform for genetic improvement of Brassica oil and vegetable crops.

Introduction

Botanical models such as *Arabidopsis thaliana* have been of enormous value in providing early insights into angiosperm (flowering plant) genome structure and function, but deeper sampling is essential toward understanding the botanical diversity that sustains humanity. For example, *A. thaliana* has experienced *two* genome duplications since its divergence from *Carica*, with rapid DNA sequence divergence, extensive gene loss, and fractionation of ancestral gene order eroding the resemblance of *A. thaliana* to ancestral Brassicales [1]. In the past few million years alone, *A. thaliana* has experienced a ~30% reduction in genome size [2] and 9-10 chromosomal rearrangements [3,4] that differentiate it from its sister species *A. lyrata* which retains near-perfect collinearity with representatives of a different genus, *Capsella rubella*.

The Brassicaceae family which includes *A. thaliana*, is a natural priority for more deeply sampling angiosperm diversity. With more than 300 additional genera, Brassica species alone includes many important vegetables that are widely used in the cuisine of many cultures (*B. rapa*: Chinese cabbage, Pak-choi and turnip; *B. oleracea*: broccoli, cabbage and cauliflower) as well as oilseed crops (*B. napus*, *B. rapa*, *B. juncea* and *B. carinata*) which provide collectively 12% of world edible vegetable oil production [5]. The six widely cultivated Brassica species are also a classical example of the importance of polyploidy in botanical evolution, described by the "U's triangle" [6], with the three diploid species *B. rapa* (A genome), *B. nigra* (B genome), and *B. oleracea* (C genome) having formed the amphidiploid species *B. juncea* (A and B genomes), *B. napus* (A and C

genomes), and *B. carinata* (B and C genomes) by hybridization. Comparative physical mapping studies have confirmed genome triplication in a common ancestor of *B. oleracea* [7] and *B. rapa* [8] since its divergence from the *A. thaliana* lineage 13-17 MYA [9,10] or more [11].

Genome sequence assembly and annotation

Using 72x coverage of paired short read sequences generated by Illumina GA II technology, and stringent assembly parameters, we assembled the genome of *B. rapa* ssp. *pekinensis* line Chiifu-401-42 to N50 sequence contig size of 27.3 kb and N50 scaffold size of over 339 kb (Supplementary Information S1 and S2). The assembled sequence covers 283.8 Mb, which was estimated covering over 98% of the gene space (Table S2.1T1) and slightly larger than the estimated size of the euchromatic space (220 Mb) [12]. These assemblies show excellent agreement with 647 BACs [12] and previously assembled chromosome A3 [13] sequenced by Sanger technology (Supplementary Information S2.2). Integration with 199,452 BAC-end sequences produced 159 super-scaffolds representing 90% of the assembled sequences, with N50 scaffold size greater than 1.97 Mb. Genetic mapping of 1,427 uniquely aligned markers in *B. rapa* enabled us to produce 10 pseudo-chromosomes, including 90% of the assembly (Supplementary Information 2.3).

The difference in physical size of the *A. thaliana* and *B. rapa* genomes is largely due to transposable elements. Within the *B. rapa* assembly, 39.47% of all sequences were annotated as putative LTR transposons (16.01%), DNA transposons (8.20%) or LINE

elements (5.63%) (Supplementary Table S3.1T3). Although widely dispersed throughout the genome (Figure 1), the transposon-related sequences were most abundant in the vicinity of centromeres, inversely related to the abundance of protein-coding genes. We estimated that transposon-related sequences occupy 39.51% of the genome, with the proportions of retrotransposons (with LTRs), DNA transposons and LINEs being 27.14%, 3.20% and 2.82%, respectively (Supplementary Table S3.1T4). As we could not map onto the assembly 14.4% of the reads (Table S1.2T1), which were likely from repetitive elements, the percentage of the repetitive elements might be under estimated.

We modeled 41,174 protein-coding genes in the *B. rapa* genome, distributed throughout the chromosome arms but with lower density near centromeres (Figure 1). Gene models have an average transcript length of 2,015 bp, coding length of 1,172 bp, and a mean of 5.03 exons per gene, all similar to that observed in *A. thaliana* [14] (Supplementary Information S3.3.1). A total of 95.8% of genes have a match in at least one of the public protein databases (SwissProt, TrEMBL, InterPro, GO terms and KEGG pathways) (Supplementary Information S3.3.2), and 99.3% were represented among the public EST collections or *de novo* Illumina mRNA-seq data (27.1M PE-reads) representing various tissues and developmental stages and stress treatments, supporting the accuracy of the *B. rapa* gene-predictions (Supplementary Figure S3.3.1F2).

Among the total of 16,917 *B. rapa* gene families, only 1,003 (5.9%) of identified gene families appear to be lineage specific, with 15,725 (93.0%) shared with *A. thaliana*, and 9,909 (58.6%) shared by all of *A. thaliana* [14], *C. papaya* [15] and *V. vinifera* [16] (Figure 2)

(Supplementary Information S3.5). Among the *B. rapa* specific gene families, no TE was identified (Table S3.3.3T1). However, 748 of them (74.58%) had no hit in any of the 4 databases (GO, InterPro, Swiss-Prot or TrEMBL), indicating that characterization of their functions may be important to explain the speciation of the Brassica genus.

Stabilizing the genome of a mesohexaploid

Whole genome duplication has been observed in all plant species sequence to date, including *A. thaliana* which has three paleo-polyploidy events [17]: a paleohexaploidy (γ) event shared with most dicots (astrids and rosids); and two paleotetraploidy events (β then α) shared with other members of the order Brassicales. *B. rapa* shares this complex history, with the addition of a whole-genome triplication thought to have occurred between 13 and 17 MYA [9,10], making 'mesohexaploidy' characteristic of the Brassiceae tribe of the Brassicaceae [18]. Using stringent criteria for the identification of collinear genome segments (at least 40 pairs of paralogous genes per Mb of sequence), we determined 54.7% of the *B. rapa* assembly to be at least duplicated, and present as numerous disjointed segments, as illustrated in Supplemental Figure S4.1F1a.

Unlike the only other genome triplication reported to date [16], its close relationship to *A. thaliana* provides a useful outgroup for investigating the adaptation of the Brassica lineage to the triplicated state. In total, 108.6 Mb (90.01 %) of the *A. thaliana* genome and 259.6 Mb (91.13%) of the *B. rapa* genome assembly was contained within collinear blocks. We confirmed almost complete triplication of the *B. rapa* genome relative to *A. thaliana* (Figure 3), and by inference to the postulated Brassicaceae ancestral genome (n=8). Bayesian molecular dating was adopted to estimate the neutral evolutionary rate

and WGT time using the program MULTIDIVTIME (http://statgen.ncsu.edu/thorne/multidivtime.html), the paralogous anchored in the trilpled segments (Figure S4.1F1a) and their orthologous (Table S4.1T1) dated the mesohexaploidy event between 5 and 9 MYA (Supplemental Figure S3.6F1), more recently than reported previously[11]. Relatively few rearrangements differentiate the organization of the *A. thaliana* genome from that of the ancestral genome (~7), whereas many more (>50) differentiate the chromosome structure of the *B. rapa* genome from that of the ancestral genome. In addition, there are numerous instances of rearrangements within collinearity blocks, for example, inverted segments were found within block J on A05, block A on A6, and block R on A10, and at all three occurrences of block B (on A7, A8 and A9) (Figure 3). As the latter occurs at the same position within all three versions of the *B. rapa* block, we can deduce the ancestral arrangement of that block was most likely represented in *B. rapa*, with the inversion having occurred in the *A. thaliana* lineage.

The Brassica mesohexaploidy offers a unique opportunity to study the retention of whole genome triplicates. Assuming an initial gene count similar to that of *A. thaliana* (*i.e.* around 30,000), the newly-formed hexaploid would have had about 90,000 genes. Our count of about 42,000, indicates substantial gene loss following hexaploidy, a process that is typical of post-polyploidy evolution in eukaryotes [19].

We identified each of the orthologous blocks in the *B. rapa* genome corresponding to the A to X ancestral blocks using syntenic orthologs between *B. rapa* and *A. thaliana* (Supplemental data S5.2) and found significant disparity in gene loss across the

duplicated blocks (Supplementary Figure S5.2F1). Among 21 syntenic regions, 20 showed significant deviations from equivalent gene frequencies ($P < 0.05$). To illustrate this variation, we concatenated the least fractionated blocks (LF), the medium fractionated blocks (MF1) and most fractionated blocks (MF2) and calculated the proportions of genes retained in each set of blocks (or sub-genomes), relative to *A. thaliana*. The LF sub-genome retains 70% of the genes found in *A. thaliana*, whereas MF1 and MF2 retain substantially lower proportions of retained genes (46% and 36%, respectively; Figure 4). Based on the analysis of synonymous base substitution rates (Ks values), as shown in supplementary Table S5.2T1, the three sub-genomes are indistinguishable.

The Brassica mesohexaploidy may have occurred in either one step (*e.g.* fusion of reduced and unreduced gametes of a diploid species, followed by chromosome doubling to produce the hexaploid), or two (*e.g.* initial genome doubling to form a tetraploid, followed some time later, by fusion of a gamete from this tetraploid with a gamete from a diploid, followed by chromosome doubling). Our observation of differentially fractionated sub-genomes is consistent with the latter hypothesis, where sub-genomes MF1 and MF2 had undergone substantial fractionation in a tetraploid nucleus before fractionation commenced in the LF genome in the more recently formed hexaploid. However, biased fractionation following tetraploidy (albeit less extreme than we observed) has been reported in *A. thaliana* [20] and recently in maize [21], where it was hypothesized to be the result of epigenetic marking of the genome of one parent (higher

DNA methylation and lower histone acetylation on the silenced, hence more fractionated, segments), and this hypothesis cannot be excluded.

The large homoeologous blocks resulting from Brassica mesohexaploidy are fertile ground for the occurrence of ectopic DNA recombination, which may result in concerted evolution of duplicated (triplicated) genes for tens of millions of years. By finding and comparing Brassica-Arabidopsis homologous gene quartets, including two alpha- or beta-duplicates in Brassica and their respective orthologs in Arabidopsis, we found that, respectively, 25% and 30% of Brassica and Arabidopsis duplicates are more similar to their intragenomic paralog than to their (temporally more closely related) intergenomic ortholog, suggesting appreciable gene conversion since divergence of these lineages (Supplementary S5.5). The sizes of affected regions varies from 10 bp to more than 2 Kbp. Remarkably, a majority (67% and 53, respectively) of conversion events co-occurred in parallel in both species, suggesting that intrinsic properties of specific genes such as sequence and functional conservativeness, may have contributed to the occurrence of conversion. The new homeologs produced by Brassica-specific triplication also show evidence of conversion. Genes proximal to telomeres tend to have smaller nucleotide substitution rates than distal genes (P-value=0.0004), likely a result of higher conversion rates in the former and consistent with prior findings from grasses [22,23].

Genes preferentially retained or families expanded following the Brassica paleohexaploidy.

The gene dosage hypothesis [24] predicts that gene functional categories encoding products that interact with one another or in networks, such as "ribosome protein," "transcription factor," and "proteasomal protein" should be over-retained, and genes with products that do not interact with other gene products should be under retained. Using transcription factor genes (TF) as an example, we found an approximate doubling of gene retention following the *A. thaliana* α tetraploidy event (preceding the *B. rapa* hexaploidy). The pre-grass tetraploidy expanded TFs about 6-fold [25] and the maize tribal tetraploidy expanded sorghum-maize orthologous TFs 4.3-fold [21]. *B. rapa* TFs with a detectable ortholog in *A. thaliana*, are significantly over-retained (Supplementary Table S6T1). Similar results were obtained for genes encoding known protein subunits of cytoplasmic ribosomes, and for genes known to be involved with the proteosome. We found clear under-retention of genes associated with DNA repair, nuclease activity, binding, and chloroplast associated, genes thought to encode products with few interactions (Supplementary Table S6.4T3) [25,26]. In general, some major categories of our homeolog retention data are precedented.

The GO annotation classes of over-retained genes suggests that genome triplication may have expanded gene families that underlie environmental adaptability as observed in other polyploid species [27]. Genes with GO terms associated with response to important environmental factors including salt, cold, osmotic stress, light, wounding, pathogen (broad spectrum) defense, and both cadmium and zinc ions, were over retained. Genes responding to plant hormones (jasmonic acid, auxin, salicylic acid, ethylene, brassinosteroid, cytokinin, and abscisic acid) were also over retained (Figure 5).

Characteristics of a crop genome

Under selection, Brassica species have a remarkable propensity for the development of new morphological variants [28]. This has led to domestication and selective breeding resulting in a range of different crop types, in *B. rapa* including enlarged overlapping leaves in heading Chinese cabbage, enlarged roots and hypocotyls in turnip, arrested inflorescences of brocoletto, highly branching shoots in Mizuna and flat-growing leaves in Wutacai. This morphological diversity makes *B. rapa* an excellent species for the study of plant morphological evolution as well as the process of domestication and directed selection.

One factor contributing to rapid morphological evolution may be a general acceleration of nucleotide substitution rates. Different Brassicales taxa have been evolving at very divergent rates, with polyploidization and different generation times potentially contributing to rate variations. For 2275 orthologous groups of genes in *B. rapa*, *A. thaliana*, papaya and grape (Table S5.6T1), nucleotide substitution rates in *B. rapa* were greater than all the others, with average Ks and Ka values 69% and 24% faster than papaya, and 1% and 7% faster than *A. thaliana* (Table S5.6T2). A much slower evolving rate in papaya may be explained by its longer generation time as a perennial than *B. rapa* and *A. thaliana* as annuals. Extra polyploidizations (two in *A. thaliana* and three in *B. rapa*) may have also contributed to rate elevation for providing pushing force of DNA mutation due to genomic instability and gene redundancy.

Another factor in the morphological plasticity in *B. rapa* may be expansion in the number of auxin related genes, and the morphological diversity of this species may be explained in part by continuing changes in gene content. The dynamic and differential distribution of the hormone auxin controls many plant growth and morphological developmental processes [29-31], and auxin is a true morphogen in female gametophytic development in *A. thaliana*[37]. We identified 347 *B. rapa* genes related to auxin synthesis, transportation, signal transduction and inactivation, in contrast to 187 such genes in *A. thaliana* (Supplemental Table S7.1.1T2). Gene families involved in auxin synthesis (5 members of *TAA* or *TAR*, 16 members of *YUC* in *B. rapa*; Supplemental Tables S7.1.1T1 and S7.1.1T2), transportation (9 members of *AUX1* and *PIN*; Supplemental Figures S7.1.1F3 and 7.1F4), and signal transduction (15 members of *TIR1*, 12 of *TPL*, 31 of *ARF* and 51 of *IAA*; Supplemental Figures S7.1.1F5-S7.1.1F8) have been expanded by genome triplication, and additional amplification by tandem duplication was observed for *GH3* and *SAUR* (45 and 143 members, respectively; Supplemental Figures S7.1.1F9 and S7.1.1F10). The possible role of multiple auxin-related gene networks in environmental adaptation is worth continued study.

B. rapa has also experienced striking amplification of the plant-specific TCP gene family, important in the evolution and specification of plant morphology [32]. *B. rapa* has 40 TCP genes, more than *A. thaliana* (24), grape (19), or papaya (21), and recursive polyploidization has contributed to the expansion (Fig S7.1.2F1). It is suspected that class I and II TCP transcription factors act antagonistically by competing for common targets or partners [33], making the relative numbers of the two classes interesting. There is a larger

class II TCP subfamily in *B. rapa* (Class I: 16; class II: 19) than previously reported in other plants, where class I proteins exceed class II by 1.2-2 fold [34].

The regulation of flowering, key to many Brassica morphologies, shows contrasting impacts of mesohexaploidy. Flowering Locus C (FLC), encoding a MADS protein that acts as a repressor of flowering in Arabidopsis [35], has 4 copies (2 WGD, 2 tandem) in Arabidopsis and 8 in *B. rapa* (Fig S7.2F1), with both WGT and tandem duplications contributing to this expansion. Likewise, 5 of 6 *B. rapa* VERNALIZATION1 (VRN1) genes [36] produced by the recent whole-genome triplication have been preserved (Fig S7.2F2), and the three *A. thaliana* CONSTANS-LIKE (COL) genes [37] were also further multiplied (Fig S7.2F3). However, GIGANTEA (GI) genes promoting flowering under long days [38] have been strictly limited to only one copy (Fig S7.2F4), with >80% protein similarity. Though there is a second GI homoeolog in both papaya and grape, they are highly diverged. Likewise, SHORT VEGETATIVE PHASE (SVP), a floral repressor in the thermosensory pathway [39], has only two Brassica orthologs, likely produced by the recent WGT, and low copy number in other plants with the few paralogs being highly diverged in sequence (e.g. <53% similarity in protein sequences) (Fig S7.2F5). LEAFY (LFY) and APETALA1 (AP1), both pivotal for the vegetative to reproductive transition [40], exemplify contrasting gene copy number evolution, with Brassica AP1 genes in fast expansion, following the trend in *A. thaliana* (Fig S7.2F6); while LFY is single copy in all analyzed plants except for two copies in *B. rapa*.

Synthesis

The comparison of *B. rapa* and *A. thaliana*, much like a prior comparison of the cereals sorghum and rice [41], illustrates how deeper sampling of the angiosperm family tree may shed new light on the evolution of both genome structure and gene function in plants of central importance to humanity. Further opportunities abound – particularly attractive is the closely related *Brassica oleracea*, which enjoys an even greater range of morphologies than *B. rapa*, and may be an excellent system in which to test some of the hypotheses offered herein about the genetic control of morphological evolution.

Deeper investigation of the angiosperms is also important to shedding light on the causes and consequences of genome *triplication* – while duplications abound, only two paleo-triplications have been reported to date. Virtually nothing is known about the differences in evolutionary challenges and opportunities resulting from these two types of events, although several major crops that are hexaploid might be facile systems for further study.

Methods

Genome sequencing and assembly. Approximately 72-fold shotgun coverage was generated using Illumina GA sequencing from small (~200 bp), medium (~500 bp), and long (~2 Kb, 5 Kb and 10 Kb) insert libraries (Supplementary information S1.). These paired reads were assembled into preliminary scaffolds using SOAP de novo [42]. Further assembly of scaffolds used sequence data from 199,452 BAC ends.

Integration of shotgun assembly with genetic maps. The scaffolds were anchored to the *B. rapa* genetic linkage map using 1,427 uniquely aligned markers from an integrated linkage map developed from four populations (Supplemental Information S2.3). In addition, 1,054 markers mapped to the *B. napus* A genome were used to verify and aid the alignment. Chromosomes were orientated by alignment to the reference A genome linkage groups in Parkin et al [43] (equivalent to N1-N10). Where genetic information was not available from Brassica maps, scaffolds order and/or orientation was inferred based on evidence of conserved collinearity with the *A. thaliana* gene order.

Protein-coding gene annotation. After pre-masking for transposable elements, genes were predicted using multiple gene prediction softwares and via homology based searches, all data was combined via GLEAN (Supplemental Information S3.3).

Inter- and intra-genomic alignments. Synteny within and between species was assessed using McScan and an all-against-all BLASTP comparison for pairwise gene clustering. Pairwise segments were extended by clustered genes from dynamic programming to build syntenic plots of *B. rapa vs A. thaliana*.

Acknowledgements

This work was primarily funded by the Chinese Ministry of Science and Technology, Ministry of Agriculture, Ministry of Finance, the National Natural Science Foundation of China. Other funding sources included: Core Research Budget of the Non-profit Governmental Research Institution; the European Union 7th Framework Project; fund from Shenzhen Municipal Government of China; the Danish Natural Science Research Council; Korean National Academy of Agricultural Science Rural Development Administration, and the Technology Development Program for Agriculture and Forestry, Ministry for Food, Agriculture, Forestry and Fisheries; United Kingdom's Biotechnology and Biological Sciences Research Council; Institute National de la Recherche Agronomique, France; Japanese Kazusa DNA Research Institute Foundation; National Sanitation Foundation, USA; Bielefeld University, German; Australian Research Council, the Grains Research and Development Corporation. A full list of support and acknowledgements is in the Supplementary Information.

References

1. Tang, H. et al. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* **18**, 1944-54 (2008).
2. Johnston, J.S. et al. Evolution of genome size in Brassicaceae. *American Journal of Botany* **95**, 229-235 (2005).
3. Koch, M.A. & Kiefer, M. Genome evolution among cruciferous plants: A lecture from the comparison of the genetic maps of three diploid species - Capsella rubella, Arabidopsis lyrata subsp Petraea, and A. thaliana. *American Journal of Botany* **92**, 761-767 (2005).
4. Yogeeswaran, K. et al. Comparative genome analyses of Arabidopsis spp.: Inferring chromosomal rearrangement events in the evolutionary history of A-thaliana. *Genome Research* **15**, 505-515 (2005).
5. Labana, K.S. & Gupta, M.L. Importance and origin. in *Breeding Oilseed Brassicas* (eds. Labana, K.S., Banga, S.S. & Banga, S.K.) 1-20 (Springer-Verlag, Berlin, 1993).
6. U, N. Genome analysis in Brassica with special reference to the experimental formation of *B. napus* and peculiar mode of fertilication. *Jap J Bot* **7**, 389-452 (1935).
7. O'Neill, C.M. & Bancroft, I. Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. *Plant J* **23**, 233-43 (2000).
8. Park, J.Y. et al. Physical mapping and microsynteny of *Brassica rapa* ssp. *pekinensis* genome corresponding to a 222 kbp gene-rich region of Arabidopsis chromosome 4 and partially duplicated on chromosome 5. *Mol Genet Genomics* **274**, 579-88 (2005).
9. Yang, Y.W., Lai, K.N., Tai, P.Y. & Li, W.H. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. *Journal of molecular evolution* **48**, 597-604 (1999).
10. Town, C.D. et al. Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell* **18**, 1348-59 (2006).
11. Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 18724-8 (2010).
12. Mun, J.H. et al. Genome-wide comparative analysis of the *Brassica rapa* gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. *Genome Biology* **10**, R111 (2009).
13. Mun, J.H. et al. Sequence and structure of *Brassica rapa* chromosome A3. *Genome Biol* **11**, R94.
14. Arabidopsis, Genome & Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).

15. Ming, R. et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991-6 (2008).
16. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-7 (2007).
17. Bowers, J.E., Chapman, B.A., Rong, J. & Paterson, A.H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433-8 (2003).
18. Lysak, M.A., Koch, M.A., Pecinka, A. & Schubert, I. Chromosome triplication found across the tribe Brassiceae. *Genome research* **15**, 516-25 (2005).
19. Sankoff, D., Zheng, C. & Zhu, Q. The collapse of gene complement following whole genome duplication. *BMC genomics* **11**, 313 (2010).
20. Thomas, B.C., Pedersen, B. & Freeling, M. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome research* **16**, 934-46 (2006).
21. Woodhouse, M.R. et al. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. *PLoS biology* **8**, e1000409 (2010).
22. Wang, X., Tang, H., Bowers, J.E. & Paterson, A.H. Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. *Genome Res* **19**, 1026-32 (2009).
23. Wang, X.Y., Tang, H.B. & Paterson, A.H. Seventy Million Years of Concerted Evolution of a Homoeologous Chromosome Pair, in Parallel, in Major Poaceae Lineages. *Plant Cell* (2011).
24. Birchler, J.A. & Veitia, R.A. The gene balance hypothesis: from classical genetics to modern genomics. *The Plant cell* **19**, 395-402 (2007).
25. Freeling, M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual review of plant biology* **60**, 433-53 (2009).
26. Edger, P.P. & Pires, J.C. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **17**, 699-717 (2009).
27. Ha, M., Kim, E.D. & Chen, Z.J. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 2295-300 (2009).
28. Paterson, A.H., Lan, T.H., Amasino, R., Osborn, T.C. & Quiros, C. Brassica genomics: a complement to, and early beneficiary of, the Arabidopsis sequence. *Genome biology* **2**, REVIEWS1011 (2001).
29. Teale, W.D., Paponov, I.A. & Palme, K. Auxin in action: signalling, transport and the control of plant growth and development. *Nature reviews. Molecular cell biology* **7**, 847-59 (2006).
30. Theologis, A. et al. Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature* **408**, 816-20 (2000).
31. Vanneste, S. & Friml, J. Auxin: a trigger for change in plant development. *Cell* **136**, 1005-16 (2009).

32. Reeves, P.A. & Olmstead, R.G. Evolution of the TCP Gene Family in Asteridae: Cladistic and Network Approaches to Understanding Regulatory Gene Family Diversification and Its Impact on Morphological Evolution. *Molecular Biology and Evolution* **20**, 1997-2009 (2003).
33. Martín-Trillo, M. & Cubas, P. TCP genes: a family snapshot ten years later. *Trends in Plant Science* **15**, 31-39 (2009).
34. Navaud, O., Dabos, P., Carnus, E., Tremousaygue, D. & Hervé, C. TCP Transcription Factors Predate the Emergence of Land Plants. *Journal of Molecular Evolution* **65**, 23-33 (2007).
35. Michaels, S.D. & Amasino, R.M. FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell* **11**, 949-56 (1999).
36. Levy, Y.Y., Mesnage, S., Mylne, J.S., Gendall, A.R. & Dean, C. Multiple roles of Arabidopsis VRN1 in vernalization and flowering time control. *Science* **297**, 243-6 (2002).
37. Ledger, S., Strayer, C., Ashton, F., Kay, S.A. & Putterill, J. Analysis of the function of two circadian-regulated CONSTANS-LIKE genes. *Plant J* **26**, 15-22 (2001).
38. Gunl, M., Liew, E.F., David, K. & Putterill, J. Analysis of a post-translational steroid induction system for GIGANTEA in Arabidopsis. *BMC Plant Biol* **9**, 141 (2009).
39. Li, D. et al. A repressor complex governs the integration of flowering signals in Arabidopsis. *Dev Cell* **15**, 110-20 (2008).
40. Yu, H., Ito, T., Wellmer, F. & Meyerowitz, E.M. Repression of AGAMOUS-LIKE 24 is a crucial step in promoting flower development. *Nat Genet* **36**, 157-61 (2004).
41. Paterson, A.H. et al. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-6 (2009).
42. Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20**, 265-72 (2010).
43. Parkin, I.A. et al. Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. *Genetics* **171**, 765-81 (2005).

Author Information

This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AENI00000000. The version described in this paper is the first version, AENI01000000. Correspondence and requests for materials should be addressed to X.W. (wangxw@mail.caas.net.cn), J.W. (wangj@genomics.org.cn), H.W. (wanghz@oilcrops.cn) and R.S. (rifei.sun@caas.net.cn).

Author Contributions See list of consortium authors below. "§" indicates group leader

The *Brassica rapa* Genome Sequencing Project Consortium:

*: **The *Brassica rapa* Genome Sequencing Project Consortium** ("§" indicates group leader)

Correspondence and requests for materials

Xiaowu Wang [1] (wangxw@mail.caas.net.cn), Hanzhong Wang [2] (wanghz@oilcrops.cn), Jun Wang [3,4] (wangj@genomics.org.cn) and Rifei Sun [1] (rifei.sun@caas.net.cn)

Principal investigators

Xiaowu Wang [1], Jian Wu [1], Shengyi Liu [2], Yinqi Bai [3], Jeong-Hwan Mun [5] and Ian Bancroft [6]

DNA and transcriptome sequencing

Bo Wang [3, §], Xiaowu Wang [1, §], Boulos Chalhoub [7, §], Jun Wang [3], Kui Wu [3], Jian Wu [1], Shengyi Liu [2], Wei Hua [2], Beom-Seok Park [5], Ian Bancroft [6], David Edwards [8], Isobel A.P. Parkin [9], Jeong-Hwan Mun [5], Hiroshi Abe [10], Bernd Weisshaar [11], Shusei Sato [12], Hideki Hirakawa [12], Satoshi Tabata [12], Andrew G. Sharpe [13], Yongpyo Lim [14], Guusje

Bonnema [15], Jacqueline Batley [16], Chuyu Lin [3], Chunyu Geng [3], Julie Poulain [17], Soo-Jin Kwon [5], Jin A Kim [5], Martin Trick [6], Fiona Fraser [6], Eleni Soumpourou [6], Matthew G. Links [9], Chushin Koh [7], Katsunori Hatakeyama [18], Yoshihiro Narusaka [19], Paul Berkman [8], Christopher Duran [8]

Sequence assembly

Junyi wang [3, §], Jun Wang [3,4], Desheng Mu [3], Yingrui Li [3], Xun Xu [3], Bo Liu [1], Silong Sun [1], Zhonghua Zhang [1], Zhenyu Li [3], Binghang Liu [3], Qingle Cai [3], Shu Zhang [3], Yinqi Bai [3], Zhiwen Wang [3], Xiang Zhao [3], Song Chi [3], Jingyin Yu [2], Jérémy Just [7]

Anchoring to linkage maps

Jian Wu [1, §], Wei Hua [2, §], Graham J King [20], Yong-Pyo Lim [14], Beom-Seok Park [5], Ian Bancroft [6], Jacqueline Batley [16], David Edwards [8], Yan Wang [1], Bo Liu [1], Silong Sun [1], Jun Wang [20], Isobel Parkin [9], Jinglin Meng [21], Hui Wang [1], Jie Deng [1], Yongcui Liao [1], Yinqi Bai [3], Haiping Wang [1], Mina Jin [5], Jeong-Sun Kim [5], Su-Ryun Choi [14], Nirala Ramchiary [14], Alice Hayward [16]

Annotation

Yinqi Bai [3, §], Shengyi Liu [4, §], Ruiqian Li [3], Fan Wei [3], Quanfei Huang [3], Feng Cheng [1], Bo liu [1], David Edwards [8], Jiumeng Min [3], Jianwen Li [3], Chunfang Peng [3], Heling Zhou [3], Shunmou Huang [2], Boulos Chalhoub [7], Jérémy Just [7], Harry Belcram [7], Gilles Lassalle [22], Nizar Drou [6], Martin Trick [6]

Stabilizing the genome of a polyploidy dicotyledonous species

Feng Cheng [1, §], Sanwen Huang [1, §], Yinqi Bai [3], Xiaowu Wang [1], Bo Li [1], Shifeng Cheng [3], Ye Yin [3], Jiaohui Xu [3], Chaobo Tong [2]

Comparative genomics

Xiaowu Wang [1, §], J. Chris Pires [23, §], Xiyin Wang [24, 25, §], Ian Bancroft [6], Feng Cheng [1], Haibao Tang [26], Gavin Conant [27], Tae-Ho Lee [25], Jinpeng Wang [24], Zhenyi Wang [24], Hui Guo [25]

Retention of genes duplicated by polyploidy

Michael Freeling [26, §], Andrew H. Paterson [25, §], Feng Cheng [1], Haibao Tang [26], Bo Liu [1], Silong Sun [1], Lu Fang [1], Zhiyong Xiong [23], Meixia Zhao [4], Jingping Li [25], Huizhe Jin [25], Xu Tan [25]

Characteristics of a crop genome

Jian Wu [1, §], Xixiang Li [1, §], Rifei Sun [1], Hanzhong Wang [2], Yongchen Du [1], Xiaowu Wang [1], Hui Wang [1], Jie Deng [1], Di Shen [1], Yang Qiu [1], Shujiang Zhang [1], Fei Li [1], Li Wang [24], Yupeng Wang [25]

Affiliations:

1. Key Laboratory of Horticultural Crop Genetic Improvement, MOA; Sino-Dutch Joint Lab of Horticultural Genomics Technology; Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences (IVF, CAAS) , Beijing, 100081, China
2. The Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan, Hubei, 430062, China
3. BGI-Shenzhen, Shenzhen, 518083, China
4. Department of Biology, University of Copenhagen, Copenhagen, Denmark
5. Department of Agricultural Biotechnology, National Academy of Agricultural Science, Rural Development Administration, Suwon 441-707, Korea
6. John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK
7. Organization and evolution of plant genomes, URGV, UMR1165, (INRA-CNRS, UEVE), 2 rue Gaston Crémieux, 91057 Evry, France
8. University of Queensland, School of Land, Crop and Food Sciences; and Australian Centre for Plant Functional Genomics, Brisbane, QLD 4072 Australia
9. Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK, S7N 0X2, Canada
10. Experimental Plant Division, RIKEN BioResource Center, Tsukuba 305-0074, Japan
11. Center for Biotechnology, Bielefeld University, Universitaetsstrasse 25, 33615 Bielefeld, Germany
12. Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818 Japan

13. NRC Plant Biotechnology Institute (NRC-PBI), 110 Gymnasium Place, Saskatoon, SK S7N 0W9
14. Molecular Genetics and Genomics Lab, Department of Horticulture, Chungnam National University, Daejeon 305 764, Republic of Korea
15. Droevendaalsesteeg 1, Wageningen University, 6708 PB Wageningen, The Netherlands
16. University of Queensland, School of Land, Crop and Food Sciences; and ARC Centre of Excellence for Integrative Legume Research, Brisbane, QLD 4072 Australia
17. Genoscope, Institut de Génomique du CEA, 2 rue Gaston Crémieux, 91057 Evry, France
18. National Institute of Vegetable and Tea Science, Tsu 514-2392, Japan
19. Research Institute for Biological Sciences, Okayama 716-1241, Japan
20. Centre for Crop Genetic Improvement, Rothamsted Research, West Common, Harpenden, AL5 2QJ, U.K.
21. National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, 430070, China
22. INRA - Agrocampus Rennes - Univ. Rennes 1, UMR118 Amélioration des Plantes et Biotechnologies Végétales, BP 35327, 35653 Le Rheu Cedex, France
23. Division of Biological Sciences, Bond Life Sciences Center, 1201 Rollins Street, University of Missouri, Columbia, MO 65211-7310 USA
24. Center for Genomics and Computational Biology, School of Life Sciences, and School of Sciences, Hebei United University, Tangshan, Hebei, China, 063009
25. Plant Genome Mapping Laboratory, Univ Georgia, Athens GA 30602 USA
26. Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720 USA
27. Division of Animal Sciences, 920 E. Campus Drive, University of Missouri, Columbia, MO 65211-5300 USA

Figures

Figure 1 | Chromosomal distribution of the main *Brassica rapa* genome features. Area charts quantify retrotransposons (RTs), genes (exons and introns), DNA transposons (DNA-TEs). Heat-map tracks detail the distribution of selected elements. LTR-RTs, long terminal repeat retrotransposons (gypsy and copia); DNA-TEs, DNA transposons (CACTA and MITE); Genes.

Figure 2 | Venn diagram showing unique and shared gene families between and among three sequenced dicotyledonous species (*Brassica rapa, Arabidopsis thaliana, Carica papaya, Vitis vinifera*).

Figure 3 | Segmental collinearity of the genomes of *Brassica rapa* and *Arabidopsis thaliana*. Conserved collinear blocks of gene models are shown between the ten chromosomes of the *B. rapa* genome (horizontal axis) and the five chromosomes of the *A. thaliana* genome (vertical axis). These blocks are labeled A to X and color-coded by inferred ancestral chromosome following established convention.

Figure 4 | The density of orthologous genes in three subgenomes (LF, MF1 and MF2) of *B. rapa* compared to *A. thaliana*. Axis x denotes the physical position of each *A. thaliana* gene locus. Axis y denotes the percentage of retained orthologous genes in *B. rapa* subgenomes around each *A. thaliana* gene, where 500 genes flanking each side of a certain gene locus were analyzed giving a total window size of 1001 genes.

Figure 5 | The over retention genes in *B. rapa* showing strong bias. The x axis is the gene category, and y axis is the ratio of different copies in each category. The digit above each bar is the number of orthologs of B. rapa in *A. thaliana* of each class. RE: response to environment, RH: response to hormone, TF: Transcription factor, CR: cytosolic ribosome, CW: cell wall. A, orange bar: ratio of 1 and 2-copy orthologs, light-green bar: ratio of 3 copies. B, yellow bar: ratio of 1-copy orthologs, blue bar: ratio of 2 or 3-copy orthologs. The last category is the total sets of all orthologs listed as a control. The P-value of each category was indicated under the bars.

Tables

[Insert Tables here]

B. rapa genome