nature genetics

The genome of the mesopolyploid crop species *Brassica rapa*

The *Brassica rapa* Genome Sequencing Project Consortium

We report the annotation and analysis of the draft genome sequence of *Brassica rapa* accession Chiifu-401-42, a Chinese cabbage. We modeled 41,174 protein coding genes in the *B. rapa* genome, which has undergone genome triplication. We used *Arabidopsis thaliana* as an outgroup for investigating the consequences of genome triplication, such as structural and functional evolution. The extent of gene loss (fractionation) among triplicated genome segments varies, with one of the three copies consistently retaining a disproportionately large fraction of the genes expected to have been present in its ancestor. Variation in the number of members of gene families present in the genome may contribute to the remarkable morphological plasticity of *Brassica* species. The *B. rapa* genome sequence provides an important resource for studying the evolution of polyploid genomes and underpins the genetic improvement of *Brassica* oil and vegetable crops.

Model species have provided valuable insights into angiosperm (flowering plant) genome structure, function and evolution. 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]. Compared with an ancestor at just a few million years ago, *A. thaliana* has undergone a ~30% reduction in genome size[2] and 9–10 chromosomal rearrangements[3,4] that differentiate it from its sister species *Arabidopsis lyrata*. Whole-genome duplication has been observed in all plant genomes sequenced to date. *A. thaliana* has undergone three paleo-polyploidy events[5]: a paleohexaploidy (γ) event shared with most dicots (asterids and rosids) and two paleotetraploidy events (β then α) shared with other members of the order Brassicales. *B. rapa* shares this complex history but with the addition of a whole-genome triplication (WGT) thought to have occurred between 13 and 17 million years ago (MYA)[6,7], making 'mesohexaploidy' a characteristic of the Brassiceae tribe of the Brassicaceae[8].

Brassica crops are used for human nutrition and provide opportunities for the study of genome evolution. These crops include important vegetables (*B. rapa* (Chinese cabbage, pak choi and turnip) and *Brassica oleracea* (broccoli, cabbage and cauliflower)) as well as oilseed crops (*Brassica napus*, *B. rapa*, *Brassica juncea* and *Brassica carinata*), which provide collectively 12% of the world's edible vegetable oil production[9]. The six widely cultivated *Brassica* species are also a classical example of the importance of polyploidy in botanical evolution, described by 'U's triangle'[10], with the three diploid species *B. rapa* (A genome), *Brassica 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*[11] and *B. rapa*[12] since its divergence from the *A. thaliana* lineage at least 13–17 MYA[6,7,13].

Using 72× coverage of paired short read sequences generated by Illumina GA II technology and stringent assembly parameters, we assembled the genome of the *B. rapa* ssp. *pekinensis* line Chiifu-401-42 and analyzed the assembly (Online Methods and **Supplementary Note**). The final assembly statistics are summarized in **Table 1**. The assembled sequence of 283.8 Mb was estimated to cover >98% of the gene space (**Supplementary Table 1**) and is greater than the previous estimated size of the euchromatic space, 220 Mb[14]. The assembly showed excellent agreement with the previously reported chromosome A03 (ref. 15) and with 647 bacterial artificial chromosomes (BACs)[14] (Online Methods) sequenced by Sanger technology. Integration with 199,452 BAC-end sequences produced 159 super scaffolds representing 90% of the assembled sequences, with an N50 scaffold (N50 scaffold is a weighted median statistic indicating that 50% of the entire assembly is contained in scaffolds equal to or larger than this value) size of 1.97 Mb. Genetic mapping of 1,427 markers in *B. rapa* allowed us to produce ten pseudo chromosomes that included 90% of the assembly (**Supplementary Table 2**).

We found the difference in the physical sizes of the *A. thaliana* and *B. rapa* genomes to be largely because of transposable elements (**Supplementary Table 3**). Although widely dispersed throughout the genome, as shown in **Figure 1**, the transposon-related sequences were most abundant in the vicinity of the centromeres. We estimated that transposon-related sequences occupy 39.5% of the genome, with the proportions of retrotransposons (with long terminal repeats), DNA transposons and long interspersed elements being 27.1%, 3.2% and 2.8%, respectively (**Supplementary Tables 4** and **5**).

We modeled and analyzed protein coding genes (described in the Online Methods and the **Supplementary Note**). We identified 41,174 protein coding genes, distributed as shown in **Figure 1**. The gene models have an average transcript length of 2,015 bp, a coding length of 1,172 bp and a mean of 5.03 exons per gene, both similar to that observed in *A. thaliana*[16]. A total of 95.8% of gene models have a match in at least one of the public protein databases and 99.3% are represented among the public EST collections or *de novo* Illumina mRNA-Seq data. Among the total 16,917 *B. rapa* gene families, only 1,003 (5.9%) appear to be lineage specific, with 15,725 (93.0%) shared with *A. thaliana*[16] and 9,909 (58.6%) also shared by *Carica papaya*[17] and *Vitis vinifera*[18] (**Fig. 2**).

A full list of members appears at the end of the paper.

Received 7 March 2011; accepted 3 August 2011; published online 28 August 2011; doi:10.1038/ng.919

Table 1 Summary of the final assembly statistics

	Contig size	Contig number	Scaffold size	Scaffold number
N90	5,593	10,564	357,979	159
N80	10,984	7,292	773,703	104
N70	15,947	5,308	1,257,653	77
N60	21,229	3,874	1,452,355	56
N50	27,294	2,778	1,971,137	39
Total size	264,110,991		283,823,632	
Total number (>100 bp)		60,521		40,549
Total number (>2 kb)		14,207		794

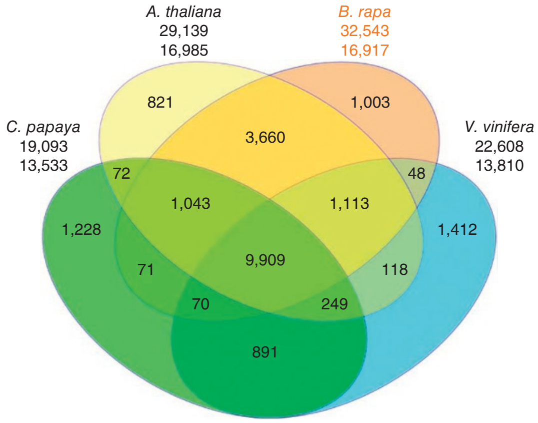


Figure 2 Venn diagram showing unique and shared gene families between and among four sequenced dicotyledonous species (*B. rapa*, *A. thaliana*, *C. papaya* and *V. vinifera*).

We analyzed the organization and evolution of the genome (as described in the Online Methods and the **Supplementary Note**). *B. rapa*'s close relationship to *A. thaliana* allows *Arabidopsis* to be used as an 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 were contained within collinear blocks. We confirmed the almost complete triplication of the *B. rapa* genome relative to *A. thaliana* (**Fig. 3**) and (by inference) to the postulated Brassicaceae ancestral genome ($n = 8$). The gene paralogues anchored in the triplicated segments (**Supplementary Fig. 1**) and their orthologs (**Supplementary Table 6**) dated the mesohexaploidy event to between 5 and 9 MYA (**Supplementary Fig. 2**), which is more recent than has been reported previously[13].

The *Brassica* mesohexaploidy offers an opportunity to study gene retention in triplicated genomes. Assuming an initial count of protein coding genes similar to that of *A. thaliana* (around 30,000), the newly formed hexaploid would have about 90,000 genes, of which we can now identify only 41,174. This is typical of the substantial gene loss that occurs following polyploid formation in eukaryotes[19–21]. We identified each of the orthologous blocks in the *B. rapa* genome corresponding to ancestral blocks using collinearity between orthologs on the genomes of *B. rapa* and *A. thaliana* and found significant disparity in gene loss across the triplicated blocks (**Supplementary Fig. 3**). Of the 21 regions of conserved synteny, 20 showed significant deviations from equivalent gene frequencies ($P < 0.05$) (**Supplementary Fig. 4**). To illustrate this variation, we concatenated the least fractionated blocks (LF), the medium fractionated blocks (MF1) and the most fractionated blocks (MF2) and calculated the proportions of genes retained in each of these sub-genomes relative to *A. thaliana*. The LF sub-genome retains 70% of the genes found in *A. thaliana*, whereas the MF1 and MF2 sub-genomes retain substantially lower proportions of retained genes (46% and 36%, respectively; **Fig. 4**). Based on the analysis of synonymous base substitution rates (K_s values), the pairwise divergences between the three sub-genomes are indistinguishable from each other (**Supplementary Table 7**). Our observation of differentially fractionated sub-genomes is consistent with the hypothesis that the sub-genomes MF1 and MF2 underwent substantial fractionation in a tetraploid nucleus before fractionation commenced in the LF genome in a more recently formed hexaploid. However, biased fractionation following tetraploidy (albeit less extreme than we observed) has been reported in *A. thaliana*[22] and maize[23], where it was hypothesized to be the result of differential epigenetic marking of the parent genomes (resulting in differential gene silencing and consequential fraction), representing an alternative hypothesis.

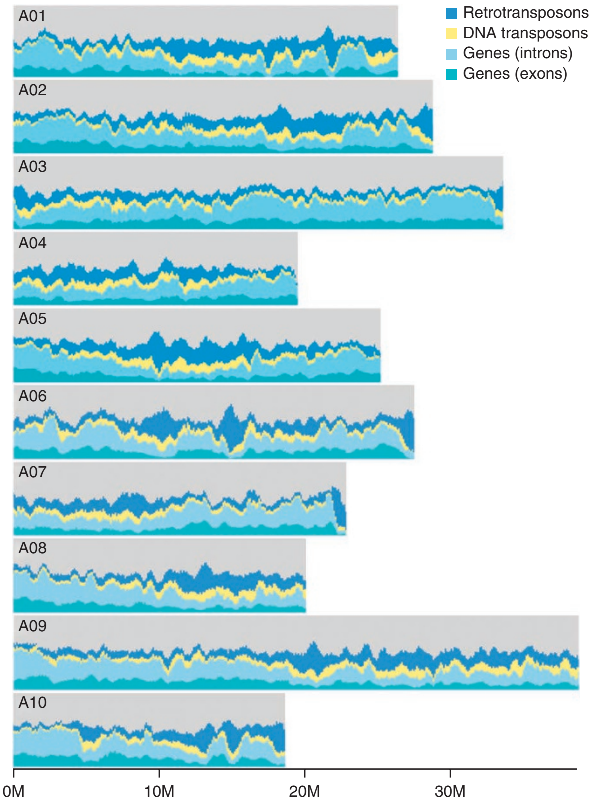


Figure 1 Chromosomal distribution of the main *B. rapa* genome features. Area charts quantify retrotransposons, genes (exons and introns) and DNA transposons. The *x* axis denotes the physical position along the *B. rapa* chromosomes in units of million (M) bases.

The retention of extensive collinear genome blocks provides a potential opportunity for ectopic DNA recombination. By finding and comparing homologous gene quartets, including two α or β duplicates in *Brassica* and their respective orthologs in *Arabidopsis*, we noted that, respectively, 25% and 30% of *Brassica* and *Arabidopsis* duplicates are more similar to their intragenomic paralog than to their intergenomic ortholog, suggesting appreciable gene conversion since the divergence of these lineages (**Supplementary Note**). The sizes of the affected regions vary from 10 bp to >2 kb, with a majority of these apparent conversion events occurring in parallel in both species. Genes proximal to telomeres tend to have lower nucleotide substitution rates than distal genes ($P = 0.0004$), which is likely to be a result of higher conversion rates in the former and is consistent with prior findings in grasses[24,25].

The gene dosage hypothesis[26] predicts that gene functional categories encoding products that interact with one another or in networks

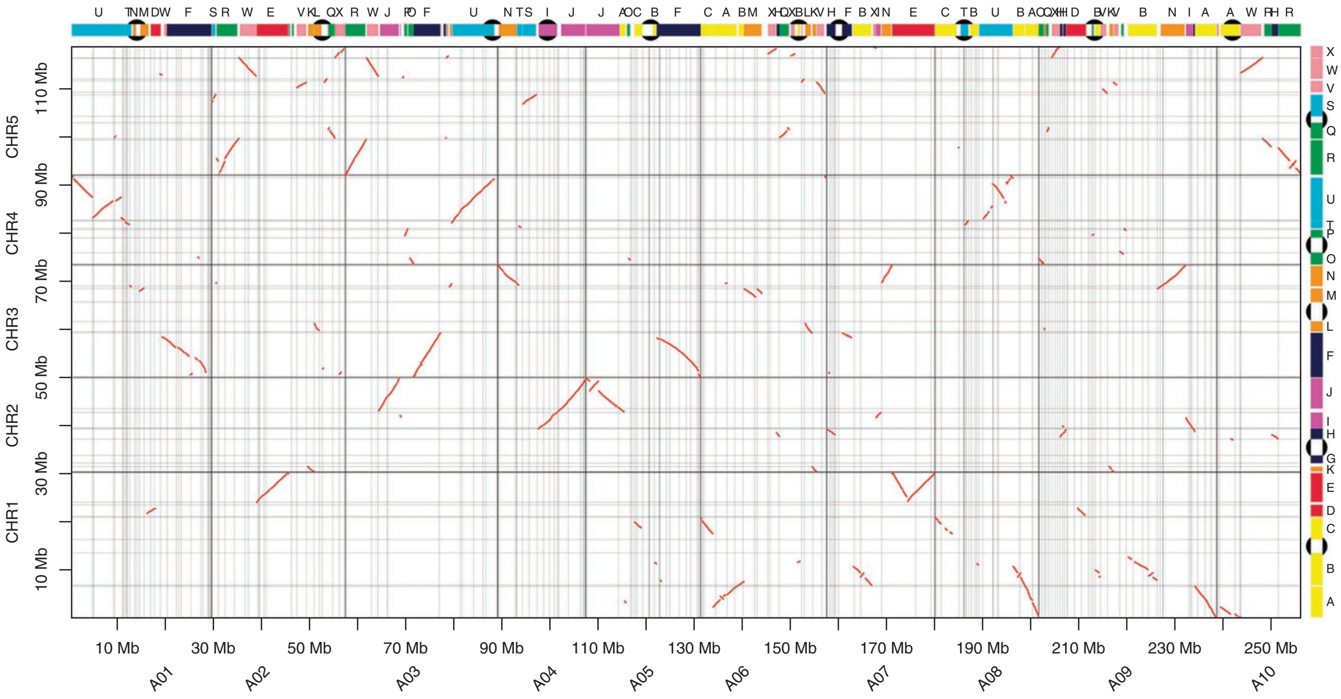


Figure 3 Segmental collinearity of the genomes of *B. rapa* and *A. thaliana*. Conserved collinear blocks of gene models are shown between the ten chromosomes of the *B. rapa* genome (horizontal axis) and the five chromosomes of the *A. thaliana* genome (vertical axis). These blocks are labeled A to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

should be over retained and genes with products that do not interact with other gene products should be under retained. In accordance with this hypothesis, we found *B. rapa* transcription factors with a detectable ortholog in *A. thaliana* to be significantly over retained (**Supplementary Table 8** and **Supplementary Note**). We obtained similarly consistent results for genes encoding known protein subunits of cytoplasmic ribosomes and for genes known to be involved with the proteosome. We found under retention of genes encoding products with few interactions, specifically those associated with DNA repair, nuclease activity, binding and the chloroplast (**Supplementary Table 9**). The Gene Ontology 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 Gene Ontology 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 (**Fig. 5**). Genes responding to plant hormones (jasmonic acid, auxin, salicylic acid, ethylene, brassinosteroid, cytokinin and abscisic acid) were also over retained.

Under selection, *Brassica* species have a remarkable propensity for the development of morphological variants[28]; we analyzed factors potentially involved in this development (**Supplementary Note**). One factor may be a general acceleration of nucleotide substitution rates. For 2,275 orthologous groups of genes in *B. rapa*, *A. thaliana*, papaya and grape (**Supplementary Table 10**), the nucleotide substitution rates in *B. rapa* were greater than in the other plants, with average K_s (K_s is the ratio of the number of synonymous substitutions per synonymous site) and K_a (K_a is the ratio of the number of non-synonymous substitutions per non-synonymous site) values 69% and 24%, respectively, greater than papaya and 1% and 7%, respectively, greater than *A. thaliana* (**Supplementary Table 11**). The much slower evolutionary rate in papaya may be explained by its longer generation time as a perennial. Another factor may be expansion of auxin-related gene families, as auxin controls many plant growth and morphological developmental processes[29–31]. We identified 347 *B. rapa* genes related to auxin synthesis, transportation, signal transduction and inactivation, in contrast to 187 such genes present in *A. thaliana* (**Supplementary Tables 12** and **13** and **Supplementary Figs. 5–14**). The TCP gene family is important in the evolution and specification of plant morphology[32]. This family has been amplified in *B. rapa*, which contains 39 TCP genes, which is more than *A. thaliana* (24), grape (19) or papaya (21) (**Supplementary Fig. 15**). The regulation of flowering is key to many *Brassica* morphotypes. Mesohexaploidy has had contrasting effects on the genes involved. *FLC* (*FLOWERING LOCUS C*)[33] has three orthologs in *B. rapa* as a consequence of the WGT (**Supplementary Fig. 16**). Likewise, five of six *B. rapa VRN1* (*VERNALIZATION1*) genes[34]

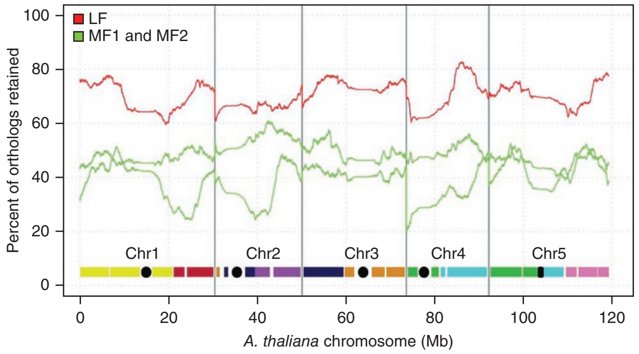
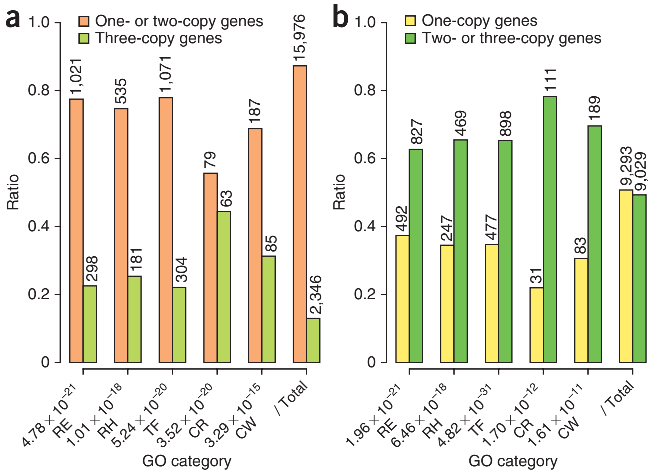


Figure 4 The density of orthologous genes in three subgenomes (LF, MF1 and MF2) of *B. rapa* compared to *A. thaliana*. The *x* axis denotes the physical position of each *A. thaliana* gene locus. The *y* axis denotes the percentage of retained orthologous genes in *B. rapa* subgenomes around each *A. thaliana* gene, where 500 genes flanking each side of a certain gene locus were analyzed, giving a total window size of 1,001 genes.

Figure 5 The over retention genes in *B. rapa* showing strong bias. The *x* axis denotes the gene category. The *y* axis denotes the ratio of different copies in each category. The number of *B. rapa* orthologs of each class is indicated above each bar. RE, response to environment; RH, response to hormone; TF, transcription factor; CR, cytosolic ribosome; CW, cell wall. (**a**) The orange bar is the ratio of one- or two-copy orthologs, and the light green bar is the ratio of three copies. (**b**) The yellow bar is the ratio of one-copy orthologs, and the dark green bar is the ratio of two- or three-copy orthologs. The last category is the total sets of all orthologs listed as a control. The *P* value of each category is indicated under the bars. GO, Gene Ontology.



produced by the WGT have been preserved (**Supplementary Fig. 17**). However, *GI* (*GIGANTEA*) genes[35] have been limited to only one copy (**Supplementary Fig. 18**), as have the *SVP* (*SHORT VEGETATIVE PHASE*) genes[36] (**Supplementary Fig. 19**) and each of the three *COL* (*CONSTANS-LIKE*) genes[37] (**Supplementary Fig. 20**).

The comparison of the genomes of *B. rapa* and *A. thaliana*, as for previous comparisons of the cereals sorghum and rice[38], sheds new light on the evolution of genome evolution in plants important for human nutrition. Our growing understanding of the processes shaping the triplicated genome of the mesopolyploid *B. rapa* is of relevance not only for closely related crops species, such as *B. oleracea* and *B. napus*, but also for other important crops with triplicated genomes, such as bread wheat.

URLs. *Brassica* info, http://www.brassica.info/; GenoScope database, http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/; Hawaii Papaya Genome Project, http://asgpb.mhpcc.hawaii.edu/papaya/; *Arabidopsis* Information Resource, http://www.arabidopsis.org/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. 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. Full annotation is available at http://brassicadb.org/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

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; funds from Shenzhen Municipal Government of China; the Danish Natural Science Research Council; National Academy of Agricultural Science and the Next-Generation Biogreen21 Program, Rural Development Administration, Korea; the Technology Development Program for Agriculture and Forestry, Ministry for Food, Agriculture, Forestry and Fisheries, Korea; United Kingdom's Biotechnology and Biological Sciences Research Council; Institute National de la Recherche Agronomique, France; Japanese Kazusa DNA Research Institute Foundation; National Science Foundation, USA; Bielefeld University, Germany; the Australian Research Council; the Australian Grains Research and Development Corporation; Agriculture and Agri-Food Canada; and the National Research Council of Canada's Plant Biotechnology Institute. See the **Supplementary Note** for a full list of support and acknowledgments.

AUTHOR CONTRIBUTIONS

Principal investigators: Xiaowu Wang, J. Wu, S.L., Y.B., J.-H.M. and I.B. **DNA and transcriptome sequencing:** Bo Wang (group leader), Xiaowu Wang (group leader), B.C. (group leader), Jun Wang (BGI), K.W., J. Wu, S.L., W.H., B.-S.P., I.B., D.E., I.A.P.P., J.-H.M., H.A., Bernd Weisshaar, Shusei Sato, H.H., S.T., A.G.S., Y. Lim, G.B., J.B., C.L., C.G., J.P., S.-J.K., J.A.K., M.T., F.F., E.S., M.G.L., C.K., K.H., Y.N., P.J.B. and C.D. **Sequence assembly:** Junyi Wang (group leader), Jun Wang (BGI), D.M., Y. Li, X.X., Bo Liu, Silong Sun, Z.Z., Z.L., Binghang Liu, Q.C., Shu Zhang, Y.B., Zhiwen Wang, X.Z., C.S., J.Y. and J.J. **Anchoring to linkage maps:** J. Wu (group leader), W.H. (group leader), G.J.K., Y. Lim, B.-S.P., I.B., J.B., D.E., Yan Wang, Bo Liu, Silong Sun, Jun Wang (Rothamsted), I.A.P.P., J. Meng, Hui Wang, J.D., Y. Liao, Y.B., Haiping Wang, M.J., J.-S.K., S.-R.C., N.R. and A.H. **Annotation:** Y.B. (group leader), S.L. (group leader), R.L., W.F., Q.H., F.C., Bo Liu, D.E., J. Min, Jianwen Li, C.P., H.Z., Shunmou Huang, B.C., J.J., H.B., G.L., N.D. and M.T. **Stabilizing the genome of a polyploidy dicotyledonous species:** F.C. (group leader), Sanwen Huang (group leader), Y.B., Xiaowu Wang, B. Li, S.C., Y.Y., J.X. and C.T. **Comparative genomics:** Xiaowu Wang (group leader), J.C.P. (group leader), Xiyin Wang (group leader), I.B., F.C., H.T., G.C., H.G., T.-H.L., Jinpeng Wang and Zhenyi Wang. **Retention of genes duplicated by polyploidy:** M.F. (group leader), A.H.P. (group leader), F.C., H.T., Bo Liu, Silong Sun, L.F., Z.X., M.Z., Jingping Li, H.J. and X.T. **Characteristics of a crop genome:** J. Wu (group leader), X.L. (group leader), R.S., Hanzhong Wang, Y.D., Xiaowu Wang, Hui Wang, J.D., D.S., Y.Q., Shujiang Zhang, F.L., L.W. and Yupeng Wang.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954 (2008).
2. Johnston, J.S. *et al.* Evolution of genome size in Brassicaceae. *Ann. Bot.* **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*. *Am. J. Bot.* **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 Res.* **15**, 505–515 (2005).
5. Bowers, J.E., Chapman, B.A., Rong, J. & Paterson, A.H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438 (2003).
6. 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. *J. Mol. Evol.* **48**, 597–604 (1999).
7. Town, C.D. *et al.* Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell* **18**, 1348–1359 (2006).
8. Lysak, M.A., Koch, M.A., Pecinka, A. & Schubert, I. Chromosome triplication found across the tribe Brassiceae. *Genome Res.* **15**, 516–525 (2005).
9. 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, Germany, 1993).
10. Nagaharu, U. Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Jap. J. Bot.* **7**, 389–452 (1935).
11. 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**, 233–243 (2000).
12. 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**, 579–588 (2005).
13. Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **107**, 18724–18728 (2010).

14. 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 Biol.* **10**, R111 (2009).
15. Mun, J.H. *et al.* Sequence and structure of *Brassica rapa* chromosome A3. *Genome Biol.* **11**, R94 (2010).
16. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* **408**, 796–815 (2000).
17. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991–996 (2008).
18. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
19. Sankoff, D., Zheng, C. & Zhu, Q. The collapse of gene complement following whole genome duplication. *BMC Genomics* **11**, 313 (2010).
20. Messing, J. *et al.* Sequence composition and genome organization of maize. *Proc. Natl. Acad. Sci. USA* **101**, 14349–14354 (2004).
21. Schnable, J.C., Springer, N.M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. USA* **108**, 4069–4074 (2011).
22. 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 Res.* **16**, 934–946 (2006).
23. Woodhouse, M.R. *et al.* Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. *PLoS Biol.* **8**, e1000409 (2010).
24. 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**, 1026–1032 (2009).
25. 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* **23**, 27–37 (2011).
26. Birchler, J.A. & Veitia, R.A. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell* **19**, 395–402 (2007).
27. Ha, M., Kim, E.D. & Chen, Z.J. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc. Natl. Acad. Sci. USA* **106**, 2295–2300 (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 Biol.* **2**, R1011 (2001).
29. Teale, W.D., Paponov, I.A. & Palme, K. Auxin in action: signalling, transport and the control of plant growth and development. *Nat. Rev. Mol. Cell Biol.* **7**, 847–859 (2006).
30. Theologis, A. *et al.* Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana. Nature* **408**, 816–820 (2000).
31. Vanneste, S. & Friml, J. Auxin: a trigger for change in plant development. *Cell* **136**, 1005–1016 (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. *Mol. Biol. Evol.* **20**, 1997–2009 (2003).
33. 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**, 949–956 (1999).
34. 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**, 243–246 (2002).
35. Günl, 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).
36. Li, D. *et al.* A repressor complex governs the integration of flowering signals in *Arabidopsis. Dev. Cell* **15**, 110–120 (2008).
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. Paterson, A.H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).

Xiaowu Wang[1], Hanzhong Wang[2], Jun Wang[3,4], Rifei Sun[1], Jian Wu[1], Shengyi Liu[2], Yinqi Bai[3], Jeong-Hwan Mun[5], Ian Bancroft[6], Feng Cheng[1], Sanwen Huang[1], Xixiang Li[1], Wei Hua[2], Junyi Wang[3], Xiyin Wang[7–9], Michael Freeling[10], J Chris Pires[11], Andrew H Paterson[9], Boulos Chalhoub[12], Bo Wang[3], Alice Hayward[13,14], Andrew G Sharpe[15], Beom-Seok Park[5], Bernd Weisshaar[16], Binghang Liu[3], Bo Li[3], Bo Liu[1], Chaobo Tong[2], Chi Song[3], Christopher Duran[13,17], Chunfang Peng[3], Chunyu Geng[3], Chushin Koh[15], Chuyu Lin[3], David Edwards[13,17], Desheng Mu[3], Di Shen[1], Eleni Soumpourou[6], Fei Li[1], Fiona Fraser[6], Gavin Conant[18], Gilles Lassalle[19], Graham J King[20], Guusje Bonnema[21], Haibao Tang[10], Haiping Wang[1], Harry Belcram[12], Heling Zhou[3], Hideki Hirakawa[22], Hiroshi Abe[23], Hui Guo[9], Hui Wang[1], Huizhe Jin[9], Isobel A P Parkin[24], Jacqueline Batley[13,14], Jeong-Sun Kim[5], Jérémy Just[12], Jianwen Li[3], Jiaohui Xu[3], Jie Deng[1], Jin A Kim[5], Jingping Li[9], Jingyin Yu[2], Jinling Meng[25], Jinpeng Wang[7,8], Jiumeng Min[3], Julie Poulain[26], Jun Wang[20], Katsunori Hatakeyama[27], Kui Wu[3], Li Wang[7,8], Lu Fang[1], Martin Trick[6], Matthew G Links[24], Meixia Zhao[2], Mina Jin[5], Nirala Ramchiary[28], Nizar Drou[6], Paul J Berkman[13,17], Qingle Cai[3], Quanfei Huang[3], Ruiqiang Li[3], Satoshi Tabata[22], Shifeng Cheng[3], Shu Zhang[3], Shujiang Zhang[1], Shunmou Huang[2], Shusei Sato[22], Silong Sun[1], Soo-Jin Kwon[5], Su-Ryun Choi[28], Tae-Ho Lee[9], Wei Fan[3], Xiang Zhao[3], Xu Tan[9], Xun Xu[3], Yan Wang[1], Yang Qiu[1], Ye Yin[3], Yingrui Li[3], Yongchen Du[1], Yongcui Liao[1], Yongpyo Lim[28], Yoshihiro Narusaka[29], Yupeng Wang[8], Zhenyi Wang[7,8], Zhenyu Li[3], Zhiwen Wang[3], Zhiyong Xiong[11] & Zhonghua Zhang[1]

[1]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences (IVF, CAAS), Beijing, China. [2]Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan, Hubei, China. [3]BGI-Shenzhen, Shenzhen, China. [4]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [5]Department of Agricultural Biotechnology, National Academy of Agricultural Science, Rural Development Administration, Suwon, Korea. [6]John Innes Centre, Norwich Research Park, Colney, Norwich, UK. [7]Center for Genomics and Computational Biology, School of Life Sciences, Hebei United University, Tangshan, Hebei, China. [8]School of Sciences, Hebei United University, Tangshan, Hebei, China. [9]Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia, USA. [10]Department of Plant and Microbial Biology, University of California, Berkeley, California, USA. [11]Division of Biological Sciences, Bond Life Sciences Center, University of Missouri, Columbia, Missouri, USA. [12]Organization and Evolution of Plant Genomes, Unité de Recherche en Génomique Végétale, Unité Mixte de Recherché 1165, (Inland Northwest Research Alliance-Centre National de la Recherche Scientifique, Université Evry Val d'Essonne), Evry, France. [13]University of Queensland, School of Agriculture and Food Sciences, Brisbane, Queensland, Australia. [14]Australian Research Council Centre of Excellence for Integrative Legume Research, Brisbane, Queensland, Australia. [15]National Research Council-Plant Biotechnology Institute, Saskatoon, Saskatchewan, Canada. [16]Center for Biotechnology, Bielefeld University, Bielefeld, Germany. [17]Australian Centre for Plant Functional Genomics, Brisbane, Queensland, Australia. [18]Division of Animal Sciences, University of Missouri, Columbia, Missouri, USA. [19]Inland Northwest Research Alliance-Agrocampus Rennes–University of Rennes 1, Unité Mixte de Recherché 118 Amélioration des Plantes et Biotechnologies Végétales, Le Rheu Cedex, France. [20]Centre for Crop Genetic Improvement, Rothamsted Research, West Common, Harpenden, UK. [21]Droevendaalsesteeg 1, Wageningen University, Wageningen, The Netherlands. [22]Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba, Japan. [23]Experimental Plant Division, RIKEN BioResource Center, Tsukuba, Japan. [24]Agriculture and Agri-Food Canada, Saskatoon, Saskatchewan, Canada. [25]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China. [26]Genoscope, Institut de Génomique du Commissariat à l'Energie Atomique, 2 rue Gaston Crémieux, Evry, France. [27]National Institute of Vegetable and Tea Science, Tsu, Japan. [28]Molecular Genetics and Genomics Lab, Department of Horticulture, Chungnam National University, Daejeon, Republic of Korea. [29]Research Institute for Biological Sciences, Okayama, Japan. Correspondence should be addressed to Xiaowu Wang (wangxw@mail.caas.net.cn), Jun Wang (wangj@genomics.org.cn), Hanzhong Wang (wanghz@oilcrops.cn) or R.S. (rifei.sun@caas.net.cn).

ONLINE METHODS

Genome sequencing and assembly. Approximately 72-fold shotgun coverage was generated using Illumina GA II sequencing from short (~200 bp), medium (~500 bp) and long (~2 kb, 5 kb and 10 kb) insert libraries (**Supplementary Note** and **Supplementary Table 14**). The raw Illumina reads were filtered for duplicates, adaptor contamination and low quality before assembly into preliminary scaffolds using SOAPdenovo[39] run with default parameters. We first assembled the reads from the short insert size (≤500 bp) libraries into contigs using Kmer (de bruijn graph kmer) overlap information and ensured the resulting contigs were unique by determining an unambiguous path in the de bruijn graph. This resulted in contigs with an N50 length of 1.1 kb, achieving a total length of 222 Mb; the long insert size mate-paired libraries (≥2 kb) were not used initially because the chimaeric reads common to such libraries can generate incorrect sequence overlaps. After obtaining the unique contigs, we mapped all available paired-end reads to these contigs to connect adjacent contigs. In order to avoid interleaving and to reduce the impact of the insert-size deviation of any sequencing library, we used a hierarchical assembly method, constructing the scaffolds step by step by adding data from each library separately ranked according to insert size from smallest to largest. This obtained scaffolds with an N50 length of 347 kb and a total genome length of 288 Mb. Most of the remaining gaps between contigs probably occur in repetitive regions, so we identified the paired-end reads with only one end mapped to a unique contig and performed local assembly with the unmapped end to fill small gaps within the scaffolds. The resulting assembly had a final contig N50 length of 27 kb (**Supplementary Table 15**). In total, 32-Mb gaps were closed. A total of 199,452 BAC-end Sanger sequences retrieved from http://www.brassica-rapa.org/BRGP/bacEndList.jsp were used to construct the super scaffolds. The gaps within the scaffolds were filled in as previously described[40]. The expected genome size of *B. rapa* was estimated from the distribution of 17-mer depth as assessed from the filtered sequence data using methods previously described[40]. The peak depth of 17-mers was at 15-folds and a total 7,287,899,150 17-mers were obtained. We obtained an estimated genome size of 485 Mb by dividing the total number of 17-mers by the peak depth.

Validation of assembly. NUCmer[41] was used to compare the sequence of chromosome A03 assembled here by whole-genome shotgun sequencing (WGS A03) to the same chromosome assembled by BAC Sanger sequencing (BAC A03) previously reported[15] (**Supplementary Note** and **Supplementary Fig. 21**). The total sizes of WGS A03 and BAC A03 are approximately 31.72 Mb and 32.70 Mb, respectively, with slightly more repeat sequences assembled using the BAC approach (9.82 Mb in BAC A03 and 5.68 Mb in WGS A03) (**Supplementary Table 16**). There were more gaps observed in BAC A03 (1,035/1,358,889 bp, number of gaps/total size of gaps) than in WGS A03 (858/844,319 bp) (**Supplementary Table 17**). We identified 44 obvious inversions (>1 kb) between the two assemblies. Evidence provided by studying the mapped paired ends, the depth of the mapped reads and gaps at the boundaries for 38 inversions supported the WGS assembly (**Supplementary Fig. 22a**,**b**), and 6 inversions remained ambiguous (**Supplementary Fig. 23c**). To evaluate the accuracy of the assembly on a local scale, the sequence of 647 complete BAC clones (phase 2 and phase 3) that had been deposited in NCBI and had been genetically mapped (see URLs) were compared with their equivalent WGS sequence (**Supplementary Table 18**).

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 (**Supplementary Table 19**). 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 from Parkin *et al.*[42] (equivalent to N1-N10). Where genetic information was not available from *Brassica* maps, scaffold order and/or orientation was inferred based on evidence of conserved collinearity with the *A. thaliana* gene order.

Protein coding gene annotation. In addition to available *Brassica* EST data (downloaded from dbEST at NCBI 10 July 2010), we generated a total of 27.1 million Illumina RNA-Seq paired-end reads, 19.9 million of which were from Chiifu-401-42 and 7.2 million of which were from a Caixin accession, L58, to verify the predicted gene models (**Supplementary Fig. 24**). For Chiifu-401-42, equally mixed total RNA isolated from eight different tissues and growth conditions was used: leaves, roots and floral stems from plants grown in pots; 2-week-old etiolated seedlings; shoots from plants grown hydroponically under normal conditions; and leaves from plants treated with 0.5% NaCl at 4 °C and 37 °C for 24 h. For L58, equally mixed total RNA was isolated from similar tissues with the addition of germinating seeds, callus and pods.

The genome assembly was premasked for class I and class II transposable elements, and Genscan and Augustus were used to carry out *de novo* predictions with gene model parameters trained from *A. thaliana*. Genes with less than 150 bp of coding sequence were filtered out. For homology-based gene prediction, we aligned *A. thaliana*, *C. papaya*, *Populus trichocarpa*, *V. vinifera* and *Oryza sativa* protein sequences to the *B. rapa* genome using TBLASTN (at an *E* value of 1×10^{-5}) for fast alignment and Genewise[43] for precise alignment. The Unigene sequences of *B. rapa* and the *Brassica* ESTs downloaded from NCBI were aligned to the *B. rapa* genome using BLAT and assembled by PASA[44] based on genomic location. As the fragmental exons in ESTs data might lead to pseudo alignments, we filtered out the results with intron(s) more than 10,000 bp. GLEAN[45] was used to combine *de novo* gene sets and homology-based gene sets and incorporated the expressed sequence data described above as supporting evidence (**Supplementary Tables 20** and **21** and **Supplementary Figs. 24** and **25**). In addition, those predicted *B. rapa* proteins that aligned to the Repbase transposable element protein database (*E* value 1×10^{-5} at ≥50%) were filtered out.

The *B. rapa* predicted proteins were annotated based on alignment to the Swiss-Prot and TrEMBL databases with BLASTP at *E* value 1×10^{-5}. InterPro was used to annotate motifs and domains by comparison with publicly available databases including Pfam, PRINTS, PROSITE, ProDom and SMART. The Gene Ontology information for each gene code was extracted from the InterPro results.

To identify and estimate the number of potential orthologous gene families between *B. rapa, V. vinifera*, *A. thaliana* and *C. papaya*, we applied the OrthoMCL pipeline[46] using standard settings (BLASTP *E* value $< 1 \times 10^{-5}$) to compute the all-against-all similarities.

Inter- and intra-genomic alignments. The synteny within and between species was constructed by McScan (MATCH_SCORE: 50, MATCH_SIZE: 5, GAP_SCORE: –3, E_VALUE: 1E–05). An all-against-all BLASTP comparison provided the pairwise gene information and *P* value for a primary clustering. Then, paired segments were extended by identifying clustered genes using dynamic programming. This method was used to build the genome synteny blocks of *B. rapa* versus *B. rapa*, *A. thaliana*, *C. papaya*, *P. trichocarpa*, *V. vinifera* and *O. sativa*, and *A. thaliana* versus *A. thaliana*.

Phylogenetic analyses of biologically important gene families. Gene sequences from grape, papaya and *Arabidopsis* were downloaded from the GenoScope database, the Hawaii Papaya Genome Project and the *Arabidopsis* Information Resource, respectively (see URLs). Previously reported *Arabidopsis* and *Brassica* gene sequences were downloaded from GenBank. The protein sequences of the genes were used to determine homologs in grape, papaya, *A. thaliana* and *B. rapa* by performing BLASTP searches at *E* value 1×10^{-10}. Alignment of each family was performed by running MEGA with default parameters[47] and subjected to careful manual checking to remove highly divergent sequences from further analysis. The retrieved protein sequences were used to reconstruct phylogenies using the neighbor-joining approach implemented in MEGA[47]. A bootstrapping test was performed using 100 repetitive samplings for each gene family.

39. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
40. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
41. Delcher, A.L., Phillippy, A., Carlton, J. & Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483 (2002).
42. Parkin, I.A. *et al.* Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. *Genetics* **171**, 765–781 (2005).
43. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
44. Haas, B.J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
45. Elsik, C.G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
46. Li, L., Stoeckert, C.J. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
47. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).

doi:10.1038/ng.919