The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes

Shengyi Liu1,*, Yumei Liu2,*, Xinhua Yang3,*, Chaobo Tong1,*, David Edwards4,*, Isobel A.P. Parkin5,*, Meixia Zhao16, Jianxin Ma6, Jingyin Yu1, Shunmou Huang1, Xiying Wang1,8, Junyi Wang5, Kun Lu9, Zhiyuan Fang2, Ian Bancroft10, Tae-Jin Yang11, Qiong Hu1, Xinfa Wang1, Zhen Yue3, Haojie Li12, Linfeng Yang3, Jian Wu2, Qing Zhou3, Wanxiang Wang2, Graham J. King12, J. Chris Pires14, Changxin Lu3, Zhangyan Wu3, Perumal Sampath11, Zhuo Wang3, Hui Guo7, Shengkai Pan3, Limei Yang2, Jiuming Min3, Dong Zhang7, Dianchuan Jin8, Wanshun Li3, Harry Belcrum15, Jinxing Tu16, Mei Guan17, Cunkou Qi18, Dezhi Du19, Jina Li9, Liangcai Jiang12, Jacqueline Batley20, Andrew G. Sharpe21, Beom-Seok Park22, Pradeep Ruperao4, Feng Cheng2, Nomar Espinosa Waminal11,23, Yin Huang3, Caihua Dong1, Li Wang3, Jingping Li7, Zhiyong Hu1, Mu Zhuang2, Yi Huang1, Junyuan Huang1, Jiaqin Shi1, Desheng Mei1, Jiaqin Li1, Tae-Ho Lee7, Jinpeng Wang8, Huizhe Jing1, Zaiyun Li16, Xun Li17, Jiefu Zhang10, Lu Xiao9, Yongming Zhou16, Zhengsong Liu17, Xueqin Liu24, Rui Qin24, Xu Tang7, Wenbin Liu3, Yupeng Wang7, Yangyong Zhang7, Jonghoon Lee11, Hyun Hee Kim23, France Denoeud25,26, Xun Xu3, Xinning Liang3, Wei Hua1, Xiaowu Wang2, Jun Wang3, Zhiyong Hu1,*, Yumei Liu2,*, Xinhua Yang3,*, Chaobo Tong1,*, David Edwards4,*, Isobel A.P. Parkin5,*, Meixia Zhao16, Jianxin Ma6, Jingyin Yu1, Shunmou Huang1, Xiying Wang1,8, Junyi Wang5, Kun Lu9, Zhiyuan Fang2, Ian Bancroft10, Tae-Jin Yang11, Qiong Hu1, Xinfa Wang1, Zhen Yue3, Haojie Li12, Linfeng Yang3, Jian Wu2, Qing Zhou3, Wanxiang Wang2, Graham J. King12, J. Chris Pires14, Changxin Lu3, Zhangyan Wu3, Perumal Sampath11, Zhuo Wang3, Hui Guo7, Shengkai Pan3, Limei Yang2, Jiuming Min3, Dong Zhang7, Dianchuan Jin8, Wanshun Li3, Harry Belcrum15, Jinxing Tu16, Mei Guan17, Cunkou Qi18, Dezhi Du19, Jina Li9, Liangcai Jiang12, Jacqueline Batley20, Andrew G. Sharpe21, Beom-Seok Park22, Pradeep Ruperao4, Feng Cheng2, Nomar Espinosa Waminal11,23, Yin Huang3, Caihua Dong1, Li Wang3, Jingping Li7, Zhiyong Hu1, Mu Zhuang2, Yi Huang1, Junyuan Huang1, Jiaqin Shi1, Desheng Mei1, Jiaqin Li1, Tae-Ho Lee7, Jinpeng Wang8, Huizhe Jing1, Zaiyun Li16, Xun Li17, Jiefu Zhang10, Lu Xiao9, Yongming Zhou16, Zhengsong Liu17, Xueqin Liu24, Rui Qin24, Xu Tang7, Wenbin Liu3, Yupeng Wang7, Yangyong Zhang7, Jonghoon Lee11, Hyun Hee Kim23, France Denoeud25,26, Xun Xu3, Xinning Liang3, Wei Hua1, Xiaowu Wang2, Jun Wang3,27,28,29, Boulos Chalhoub15 & Andrew H. Paterson7

Polyplodization has provided much genetic variation for plant adaptive evolution, but the mechanisms by which the molecular evolution of polyploid genomes establishes genetic architecture underlying species differentiation are unclear. *Brassica* is an ideal model to increase knowledge of polyploid evolution. Here we describe a draft genome sequence of *Brassica oleracea*, comparing it with that of its sister species *B. rapa* to reveal numerous chromosome rearrangements and asymmetrical gene loss in duplicated genomic blocks, asymmetrical amplification of transposable elements, differential gene co-retention for specific pathways and variation in gene expression, including alternative splicing, among a large number of paralogous and orthologous genes. Genes related to the production of anticancer phytochemicals and morphological variations illustrate consequences of genome duplication and gene divergence, imparting biochemical and morphological variation to *B. oleracea*. This study provides insights into Brassica genome evolution and will underpin research into the many important crops in this genus.

1 The Key Laboratory of Biology and Genetic Improvement of Oil Crops, The Ministry of Agriculture of PRC, Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan 430062, China. 2 The Key Laboratory of Biology and Genetic Improvement of Horticultural Crops, The Ministry of Agriculture, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 10081, China. 3 Beijing Genome Institute-Shenzhen, Shenzhen 518083, China. 4 Australian Centre for Plant Functional Genomics, School of Agriculture and Food Sciences, University of Queensland, Brisbane, Queensland 4072, Australia. 5 Agriculture and Agri-Food Canada, Saskatoon, Saskatchewan, Canada S7N 0X2. 6 Department of Agronomy, Purdue University, WSLR Building B018, West Lafayette, Indiana 47907, USA. 7 Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30605, USA. 8 Center for Genomics and Computational Biology, School of Life Sciences, and School of Sciences, Hebei United University, Tangshan 063000, China. 9 College of Agronomy and Biotechnology, Southwest University, Beibei District, Chongqing 400715, China. 10 Center for Novel Agricultural Products (CNAP), Department of Biology, University of York, Wentworth Way, Heslington, York YO10 5DD, UK. 11 Department of Plant Sciences, Plant Genomics and Breeding Institute and Research Institute for Agriculture and Life Sciences, College of Agriculture & Life Sciences, Seoul National University, Seoul 151-921, Republic of Korea. 12 Sichuan Academy of Agricultural Sciences, Chengdu 610046, China. 13 Southern Cross Plant Science, Southern Cross University, Lismore, New South Wales 2480, Australia. 14 Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65271-3700, USA. 15 Organization and Evolution of Plant Genomes, Unité de Recherche en Génomique Végétale, Unité Mixte de Recherche 165 (Institut National de Recherche Agronomique, Centre National de la Recherche Scientifique, Université Evry Val d’Essonne), Evry 91057, France. 16 National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China. 17 College of Agronomy, Hunan Agricultural University, Changsha 410128, China. 18 Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China. 19 Qinghai Academy of Agriculture and Forestry Sciences, National Key Laboratory Breeding Base for Innovation and Utilization of Plateau Crop Germplasm, Xining 810016, China. 20 Australian Research Council Centre of Excellence for Integrative Legume Research, University of Queensland, Brisbane, Queensland 4072, Australia. 21 National Research Council Canada, Saskatoon, Saskatchewan, Canada S7N 0W9. 22 The Agricultural Genome Center, National Academy of Agricultural Science, RDA, 126 Suin-Ro, Suwon 441-707, Republic of Korea. 23 Department of Life Science, Plant Biotechnology Institute, Shahyonak University, Seoul 139-742, Republic of Korea. 24 School of Life Sciences, South-Central University for Nationality, Wuhan 430074, China. 25 Commissariat à l’Energie Atomique (CEA), Genoscope, Institut de Génomique, BP5706 Evry 91057, France. 26 Centre National de Recherche Scientifique (CNRS), Université d’Evry, UMR 8030, CPS706, Evry 91057, France. 27 Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200, Copenhagen, Denmark. 28 King Abdulaziz University, Jeddah, 21589, Saudi Arabia. 29 Department of Medicine and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, 21 Sassoon Road, Hong Kong. * These are joint first authors. Correspondence and requests for materials should be addressed to S.L. (email: liusy@oikocrops.cn).
Brassica oleracea comprises many important vegetable crops including cauliflower, broccoli, cabbages, Brussels sprouts, kohlrabi and kales. The species demonstrates extreme morphological diversity and crop forms, with various members grown for their leaves, flowers and stems. About 76 million tons of Brassica vegetables were produced in 2010, with a value of 14.85 billion dollars (http://faostat.fao.org/). Most B. oleracea crops are high in protein and carotenoids, and contain diverse glucosinolates (GSLs) that function as unique phytochemicals for plant defence against fungal and bacterial pathogens and on consumption have been shown to have potent anticancer properties.

B. oleracea is a member of the family Brassicaceae (~338 genera and 3,709 species) and one of three diploid Brassica species in the classical triangle of U that also includes diploids B. rapa (AA) and B. nigra (BB) and allotetraploids B. juncea (AABB), B. napus (AACC) and B. carinata (BBCC). These allotetraploid species are important oilseed crops, accounting for 12% of world edible oil production (http://faostat.fao.org/). As the origin and relationship between these species is clear, the timing and nature of the evolutionary events associated with Brassica divergence and speciation can be revealed by interspecific genome comparison. Each of the Brassica genomes retains evidence of recursive whole-genome duplication (WGD) events and have undergone a Brassicaceae-lineage-specific whole-genome triplication (WGT) since their divergence from the Arabidopsis lineage. These events were followed by diploidization that involved substantial genome reshuffling and gene losses. Because of this, Brassica species are a model for the study of polyploid genome evolution and divergence, and serve as important resources for Brassica vegetable and oilseed crop breeding.

Results

B. oleracea genome assembly and annotation. Complementing the sequencing of the smaller B. rapa genome, a draft genome assembly of B. oleracea var. capitata line 02–12 was produced by interleaving Illumina, Roche 454 and Sanger sequence data. This assembly represents 85% of the estimated 630 Mb genome, and includes >98% of the gene space (Supplementary Methods, Supplementary Tables 1–3, 7 and 8 and Supplementary Fig. 3). The assembly was anchored to a new genetic map to produce nine pseudo-chromosomes that account for 72% of the assembly, and validated by comparison with a B. oleracea physical map, a high-density B. napus genetic map and complete BAC sequences (Supplementary Figs 4–9 and Supplementary Tables 4 and 5). For comparative analyses, identical genome annotation pipelines were used for annotation of protein-coding genes and transposable elements (TEs) for B. oleracea and B. rapa.

Table 1 | Summary of genome assembly and annotation of B. oleracea.

| B. oleracea genome assembly | N90 | N50 | Longest | Total size |
|----------------------------|-----|-----|---------|------------|
| Contig size (bp)           | 3,527 | 26,828 | 199,461 | 502,114,421 |
| Contig number              | 22,669 | 5,425 |         |            |
| Scaffold size (bp)         | 258,906 | 1,457,055 | 8,788,225 | 539,907,250 |
| Scaffold number            | 388 | 224 |         |            |

B. oleracea genome annotation

|          | B. oleracea | In the assembly | In WG short reads |
|----------|-------------|----------------|------------------|
| Size (bp) | Copy number | % assembly    |                  |
| Retrotransposon | 105,755,173 | 108,948 | 22.13 | 23.60 |
| DNA transposon   | 79,675,583 | 170,500 | 16.67 | 12.71 |
| Total            | 185,430,756 | 279,448 | 38.80 | 36.31 |
| Gene models      | 45,758 | 98%   | 91.6% | 99.0% |
| Non-coding RNA   | 1,762 bp | 1,037 bp | 4.6 | 30,932 |
| Protein-coding genes | 336 | 1,425 | 553 | 1,442 |
| Average transcript length | 1,19 | 75 | 166 | 110 |

1WG, whole genome. 20× coverage reads were randomly sampled from all the genomic short reads libraries.

1The copy number of TEs was from the RepeatMasker results.

1The untagged regions were used to detect the percentage of TEs in the assembly. TE sizes are from the untagged regions of B. oleracea 477,847,347 bp.

External estimates by public Brusica ESTs and RNA-seq data.
A total of 45,758 protein-coding genes were predicted, with a mean transcript length of 1,761 bp, a mean coding length of 1,037 bp, and a mean of 4.55 exons per gene (Table 1, Supplementary Methods, Supplementary Table 6 and Supplementary Fig. 10), similar to A. thaliana\textsuperscript{19} and B. rapa\textsuperscript{21}. Publicly available ESTs, together with RNA sequencing (RNA-seq) data generated in this study, support 94% of predicted gene models (Supplementary Tables 7 and 8), and 91.6% of predicted genes have a match in at least one public protein database (Supplementary Tables 9 and 10, and Supplementary Fig. 11). The 45,758 predicted genes, 13,032 produce alternative splicing (AS) variants with intron retention and exon skipping (Supplementary Table 11). Genome annotation also predicted 3,756 non-coding RNAs (miRNA, tRNA, rRNA and snRNA) (Supplementary Table 12).

A combination of structure-based analyses and homology-based comparisons resulted in the identification of 13,382 TEs with clearly identified terminal boundaries, including 5,107 retrotransposons and 8,275 DNA transposons (Supplementary Methods, Supplementary Fig. 12 and Supplementary Table 13). These elements together with numerous truncated elements or TE remnants make up 38.80% of the assembled portion of the B. oleracea genome, whereas TEs account for only 21.47% of the B. rapa genome, whereas TEs account for only 21.47% of the B. oleracea genome, whereas TEs account for only 21.47% of the B. rapa genome. Copia (11.64%) and gypsy (7.84%) retroelements are the major constituents of the repetitive fraction, and are unevenly distributed across each chromosome, with retrotransposons predominantly found in pericentromeric or heterochromatic regions (Supplementary Fig. 13) in B. oleracea.

Tentative physical positions of some of the centromeres were determined based on homologue and phylogenetic analysis of the centromere-specific 76 bp tandem repeats CentBo-1 and CentBo-2 and copia-type retrotransposon (CentCRBo) (Supplementary Table 14 and Supplementary Figs 14–17). The distribution of 45S and 5S rRNA sequences were also visualized by fluorescent in situ hybridization (Supplementary Figs 18 and 19), leading to a predicted karyotype ideogram for B. oleracea (Supplementary Fig. 20). An extra-centromeric locus with colocalized centromeric satellite repeat CentBo-1 and the centromeric retrotransposon CRBo-1 was observed on the long arm of chromosome 6 (Supplementary Figs 18–20). A comprehensive database for the genome information is accessible at http://www.ocri-genomics.org/bolbase/index.html.

Conserved syntenic blocks and genome rearrangement after WGT. The relatively complete triplicated regions in B. oleracea and B. rapa were constructed and they relate to the 24 ancestral crucifer blocks (A–X) of A. thaliana\textsuperscript{20}. Further the triplicated blocks resulting from WGT in the two Brassica species were partitioned into three subgenomes: LF (Least-fractionated), MF1 (Medium-fractionated) and MF2 (Most-fractionated)\textsuperscript{11} (Fig. 1a, Supplementary Methods, Supplementary Tables 15 and 16, and Supplementary Figs 21–26). These syntenic blocks occupy the majority of the genome assemblies of A. thaliana (19,628 genes, 72.24% of 27,169 genes), B. oleracea (26,485 genes, 57.88%) and B. rapa (26,698 genes, 64.84%), and provide a foundation for
comparative analyses of chromosomal rearrangement, gene loss and divergence of retained paralogues after WGT. Massive gene loss occurred in an asymmetrical and reciprocal fashion in the three subgenomes of each species and was largely completed before the \( B. \text{oleracea} - B. \text{rapa} \) divergence (Fig. 1c, Supplementary Tables 17–19 and Supplementary Figs 25–27). The timing of this evolutionary process was supported by the estimated timing of WGT \( \sim 15.9 \) million years ago (MYA), and species divergence \( \sim 4.6 \) MYA, based on synonymous substitution (Ks) rates of genes located in the blocks (Fig. 1b and Supplementary Table 20). Gene loss occurred mainly through small deletions that may be caused by illegitimate recombination \(^{21,22} \) (Supplementary Fig. 27), consistent with observations in other plant genomes.

Abundant genome rearrangement following WGT and subsequent \( \text{Brassica} \) species divergence resulted in complex mosaics of triplicated ancestral genomic blocks in the A and C genomes (Fig. 1a and Supplementary Fig. 28). At least 19 major, and numerous fine-scale, chromosome rearrangements occurred, which differentiate the two \( \text{Brassica} \) species (Supplementary Fig. 29). This is in agreement with previous comparative studies based on chromosome painting\(^ {12,23} \) and genetic mapping\(^ {24,25} \). The extensive chromosome reshuffling in \( \text{Brassica} \) is in contrast to that observed in other taxa, such as the highly syntenic tomato–potato and pear–apple genomes, each with longer divergence times and less genome rearrangement\(^ {26,27} \). This difference may be a consequence of mesopolyploidy in \( \text{Brassica} \).

Greater TEs accumulation in \( B. \text{oleracea} \) than \( B. \text{rapa} \). Both retro- (22.13%) and DNA (16.67%) TEs appear to be greater amplified in \( B. \text{oleracea} \) relative to \( B. \text{rapa} \) (9.43 and 12.04%) (Fig. 2a and Supplementary Table 13). We constructed 1,362 gap-free contig-contig syntenic regions by clustering orthologous \( B. \text{rapa}—B. \text{oleracea} \) genes using MCscan (Supplementary Figs 29 and 30). The \( B. \text{oleracea} \) TE length (34.03% of the 259.6M) is 3.4 times greater than that of the syntenic \( B. \text{rapa} \) regions (16.73% of the 155.0M) (Fig. 2c, Supplementary Tables 21 and 22, and Supplementary Fig. 31). Phylogenetic analysis revealed that \( B. \text{oleracea} \) has both more LTR retrotransposon (LTR-RT) families, and more members in most families than \( B. \text{rapa} \) (Fig. 2d and Supplementary Figs 12, 32 and 33). Furthermore, two new lineages of LTR-RTs, \( \text{Brassica Copia} \) Retrotransposon and \( \text{Brassica Gypsy} \) Retrotransposon, were defined in both \( \text{Brassica} \) species (Supplementary Fig. 33). Analysis of LTR insertion time revealed that \( \sim 98\% \) of \( B. \text{oleracea} \) intact LTR-RT’s amplified continuously over the \( \sim 4 \) million years (MY) since the \( B. \text{oleracea}—B. \text{rapa} \) split, whereas \( \sim 68\% \) of \( B. \text{rapa} \) intact LTR-RT’s amplified rapidly within the last 1 MY, predominantly in the recent 0.2 MY (Fig. 2b and Supplementary Fig. 34). Hence, LTR-RTs expanded more in the intergenic space of euchromatic regions in \( B. \text{oleracea} \) than \( B. \text{rapa} \). This agrees with previous observations based on comparison of BAC sequences between the A and C genomes\(^ {28} \). As a consequence of continuous TE amplification over the last 4 MY, the genome size of \( B. \text{oleracea} \) is \( \sim 30\% \) larger than that of \( B. \text{rapa} \).

**Figure 2 | TE comparison analyses in \( B. \text{oleracea} \) and \( B. \text{rapa} \).** (a) TE copy number and total length in each assembly and \( B. \text{oleracea}—B. \text{rapa} \) syntenic blocks. (b) The number of intact LTR (Copia-like and Gypsy-like) birthed at different times (million years ago, MYA) in the syntenic regions of \( B. \text{oleracea} \) and \( B. \text{rapa} \). (c) The comparison of TE distribution and composition in \( B. \text{oleracea}—B. \text{rapa} \) syntenic blocks along \( B. \text{oleracea} \) chromosomes. We divided \( B. \text{oleracea}—B. \text{rapa} \) syntenic region into non-overlapping sliding 200 kb windows to compare TE contents. For each window, the ratio \( \log_{10} \) (\( B. \text{oleracea}/B. \text{rapa} \)) was calculated for total syntenic block length (blue line), LTR length (purple line), gene length (yellow point), exons length (red point) and intron length (green point). If \( B. \text{oleracea} > B. \text{rapa} \) in absolute length of TE composition in a compared window, the dot or line is above the line \( y = 0 \). The corresponding \( B. \text{rapa} \) chromosome segments along \( B. \text{oleracea} \) CO8 were indicated by coloured bars. All other \( B. \text{oleracea} \) chromosomes are showed in Supplementary Fig. 31. (d) Phylogeny of the Copia-like elements as an example of LTR-RTs of the syntenic regions in \( B. \text{rapa} \) and \( B. \text{oleracea} \). The neighbor-joining (NJ) trees were generated based on the conserved RT domain nucleotide sequences using the Kimura two-parameter method\(^ {68} \) in MEGA4 (ref. 69).
Species-specific genes and tandemly duplicated genes. While the genomes of B. oleracea and B. rapa are highly similar in terms of total gene clusters/sequences and the gene number in each cluster, there are also a large number of species-specific genes in the two species. A total of 66.5% (34,237 genes) of B. oleracea genes and 74.9% (34,324) of B. rapa genes were clustered into OrthoMCL groups (Supplementary Table 23 and Supplementary Fig. 35). We identified 9,832 B. oleracea-specific and 5,735 B. rapa-specific genes, of which 77% were supported by gene expression and/or a clear Arabidopsis homologue (Supplementary Table 25). Of them, >90% of these specific genes were validated for their absence in the counterpart genomes by reciprocal mapping of raw clean reads (Supplementary Tables 25 and 26). Species-specific genes and tandemly duplicated genes

Biased loss and retention of genes after WGT/WGD. Following polyploidization, reversion of gene numbers towards diploid levels through gene loss has been widely observed in plants. However, in Brassica this only appears to be true for collinear genes in the conserved syntenic regions, with a loss of ~60%
of the predicted post-triplication gene set, nearly restoring the pre-triplication gene number. This is reflected in an overall retention rate of 1.2-fold of *A. thaliana* orthologous genes in corresponding syntenic regions (Fig. 1c and Supplementary Table 18). In contrast, in terms of genes that have no collinear gene in *A. thaliana* and either *Brassica* species (hereafter called non-collinear genes), gene retention rates is 2.5-fold the *A. thaliana* gene number in *B. oleracea* and 1.9-fold in *B. rapa*, both significantly higher than the expected rates (P value < 2.2e–16; Supplementary Table 34). For these retained genes, the numbers of the genes that are common in the two *Brassica* species are 11,746 in *B. oleracea* and 10,411 in *B. rapa*. Most of these genes are supported by expression and/or the presence of an *Arabidopsis* homologue (Supplementary Table 35). More than 61% of these genes have homologues present as collinear genes and 16% also are homologous (Supplementary Table 35). For these retained genes, the numbers of *B. oleracea* by expression and/or the presence of an the genes that are common in the two *Brassica* species are 11,746 in *B. oleracea* and 10,411 in *B. rapa*. Most of these genes are supported by expression and/or the presence of an *Arabidopsis* homologue (Supplementary Table 35). More than 61% of these genes have homologues present as collinear genes and 16% also are homologous to other non-collinear genes, indicating gene movement from triplicated syntenic regions and being similar to observations in *A. thaliana*, where half of the genes are nonsynteny within rosids31. This suggests that the breakdown of the triplicated syntenic relationship has not only prevented gene loss and a move towards pre-triplication gene numbers but has also maintained a higher gene density, and thus maintained WGT-derived genes for species evolution.

The presence of a large number of the retained paralogous genes in the syntenic regions led us to examine whether genes in some functional categories have preferentially been over-retained, as observed in other plants39. The results indicate that WGT-produced paralogous genes are over-retained in GO categories associated with regulation of metabolic and bio-synthetic processes, RNA metabolism and transcription factors (Supplementary Table 36 and Supplementary Figs 43–45), and the two *Brassica* species exhibit similar patterns of gene category retention. From a study of KEGG pathways, we also found that WGT-produced *Brassica* paralogous genes contribute 40–60% of total genes for 90% of KEGG pathways (Fig. 3c and Supplementary Fig. 43), and are functionally enriched in primary or core metabolic processes such as oxidative phosphorylation, carbon fixation, photosynthesis, circadian rhythm12 and lipid metabolism (Supplementary Tables 36 and 37 and Supplementary Figs 43–45). Notably, the pathways associated with energy metabolism have been enhanced in both *Brassica* species. For instance, in the oxidative phosphorylation pathway, there are 161 genes in *A. thaliana*, but 241 in *B. oleracea* and 208 in *B. rapa*. The majority (143/241 and 142/208) of these *Brassica* genes are multiple paralogues residing in the triplicated syntenic regions, and more than half of these paralogues have been retained as three copies, significantly higher than observed for other genes in the triplication regions (Fig. 3d and Supplementary Fig. 43).

Phylogenetic analyses show that WGT led to an expansion of genes involved in auxin functioning (AUX, IAA, GH3, PIN, SAUR, TAA, TIR, TPL and YUCCA), morphology specification (TCP), and flowering time control (FLC, CO, VRN1, LFY, API and GI) (Supplementary Table 38 and Supplementary Figs 46–61), and that most *Arabidopsis* genes in these families have two or three orthologs in *Brassica* species. These WGT-produced duplicated genes may provide important sources of evolutionary innovation33 and contribute to the extreme morphological diversity in *Brassica* species.

Divergence of duplicated genes in the *Brassica* genomes. The largest genetic foundation for plant genome evolution and new species formation is the differentiation of retained paralogous and orthologous genes. Around 38% (4,302/11,493) of all paralogous gene pairs in *B. oleracea* and ~36% (4,089/11,448) in *B. rapa* have different predicted exon numbers (Supplementary Data 1, Supplementary Tables 39 and 40 and Supplementary Fig. 62). There are 6,571 orthologous gene pairs with different exon numbers, accounting for 27.6% of total gene pairs (23,823). Some paralogous or orthologous pairs have high Ks values and low sequence similarity (Supplementary Fig. 63), indicating sequence differentiation. Of these paralogous genes, some offer appreciable opportunity for non-reciprocal DNA exchanges (gene conversion). About 8% of the 4,296 homologous quartets in *B. rapa* and *B. oleracea* have been affected by gene conversion (Fig. 4a, Supplementary Table 41 and Supplementary Fig. 64) and about one-sixth (53) of converted genes were inferred to have experienced independent conversion events in both *Brassica* species, a parallelism sometimes observed in other plants11,12,14. Around 40–44% of conversion events involved paralogues in the less-fractionated subgenomes LF in both species, substantially higher than the other two subgenomes (Supplementary Table 41). This finding suggests that gene conversion is related to homologous gene density, which determines the likelihood of illegitimate recombination. Analysis of RNA-seq data generated from callus, root, leaf, stem, flower and siliques of *B. oleracea* and *B. rapa* suggests that ~40% of WGT paralogous gene pairs are differentially expressed in these species (Fig. 4b and Supplementary Fig. 65), suggesting potential subfunctionalization of these genes. In both species, a general trend of expression differentiation was alpha-WGD paralogous genes (~46%) > WGT paralogous genes (~42%) > tandemly duplicated genes (~35%) (Fig. 4b, Supplementary Fig. 66 and Supplementary Tables 42 and 43). Different tissues harbour approximately the same number of differentially expressed duplicates, but this number was slightly higher in flower tissue. The expression level of genes in the LF subgenome was significantly higher than corresponding syntenic genes in the more fractionated subgenomes (MF1 and MF2) while no expression dominance relationship was observed between the subgenomes MF1 and MF2 (Fig. 4c, Supplementary Table 44 and Supplementary Fig. 67). Duplicated transcription factor gene pairs showed less differentiated expression (~38%) than the expected ratio at the genome-wide level (Fig. 4d and Supplementary Table 45), while paralogues with GO categories related to membrane, catalytic activity and defence response exhibited a higher ratio of differentiated expression (Fig. 4e and Supplementary Table 46). Of *B. oleracea*- *B. rapa* orthologous gene pairs (23,823 in total), ~42% were differentially expressed across all tissues (Supplementary Tables 42 and 43).

Furthermore, many paralogues generate different transcripts, resulting in expression differentiation. Analysis of AS variants of paralogous gene pairs that have identical numbers of exons demonstrated that these variants (either different variants or differential expression of the same variants) cause >20% and >44% of such paralogous genes to be differentially expressed in *B. oleracea* and *B. rapa*, respectively (Fig. 4f and Supplementary Table 47). For orthologous gene pairs of *B. oleracea* and *B. rapa*, 35.5% (8,467) of gene pairs showed differential expression due to AS variation. When only counting intron retention and exon skipping, 9.3% (2,215) of gene pairs differ. Divergence in AS variants of gene pairs presents an important layer of gene regulation, as reported28–30, and thus provides a genetic basis for species evolution and new species formation.

Unique GSLs metabolism pathways. GSLs and hydrolysis products have been of long-standing interest due to their role in plant defence and anticancer properties. Compared with *B. rapa* and *B. napus*, *B. oleracea* has the greatest GSL profile diversity, with wide qualitative and quantitative variation37,40. We identified 101
and 105 GSL biosynthesis genes in *B. rapa* and *B. oleracea*, respectively, and 22 GSL catabolism genes in each species (Fig. 5a, Supplementary Table 48 and Supplementary Data 2). In the GSL biosynthesis and catabolism pathways, tandem genes (41.4%, 40.7% and 33.9% in *A. thaliana*, *B. oleracea* and *B. rapa*, respectively) were present in a much higher proportion than the genome-wide average (Supplementary Table 32). The observed variation of GSL profiles is mainly attributed to the duplication of two genes, methylthioalkylmalate (MAM) synthase and 2-oxoglutarate-dependent dioxygenase (AOP).

In *Arabidopsis*, the MAM family contains three tandemly duplicated and functionally diverse members (MAM1, MAM2...
and MAM3), and functional analysis demonstrated that MAM2 (absent in ecotype Columbia) and MAM1 catalyses the condensation reaction of the first and the first two elongation cycles for the synthesis of dominant 3 and 4 carbon (C) side-chain aliphatic GSLs, respectively\(^{40,41}\), while MAM3 is assumed to contribute to the production of all GSL chain lengths\(^5\). In B. rapa and B. oleracea, MAM1/MAM2 genes experienced independent tandem duplication to produce 6 and 5 orthologs (glucoiberin)\(^{39}\) (Fig. 5a). Based on the results of expression and phylogenetic analyses, we found a pair of genes Bol017070 and Bra013007, which are the only orthologous genes showing high expression in B. oleracea but silenced in B. rapa (Fig. 5a).

This expression difference most likely leads to greater accumulation of the 3C GSL anticancer precursor sinigrin in B. oleracea. Meanwhile, the expression level of MAM3 in B. rapa is much higher than in B. oleracea, explaining the accumulation of 5C GSL glucobrassicacin in B. rapa. Other genes affecting specific anticancer GSL products are AOPs. Previously, research has reported four gene loci involved in the side-chain modifications of aliphatic GSLs in Arabidopsis. Two tandemly duplicated genes AOP2 and AOP3 catalyse the formation of alkenyl and hydroxyalkyl GSLs, respectively. When both AOPs are non-functional, the plant accumulates the precursor methysulfinyl alkyl GSL. We identified three AOP2 genes in B. oleracea (Fig. 5d), but two are non-functional due to the presence of premature stop codons. In contrast, all three AOP2 copies are functional in B. rapa\(^34\). No AOP3 homologue has been identified in B. rapa.

**Figure 5 | Whole-genome-wide comparison of genes involved in glucosinolate metabolism pathways in B. oleracea and its relatives.** (a) Aliphatic and indolic GSL biosynthesis and catabolism pathways in A. thaliana, B. oleracea and B. rapa. The copy number of GSL biosynthetic genes in A. thaliana, B. rapa and B. oleracea are listed in square brackets, respectively. Potential anticancer substances/precursors are highlighted in red bold, within the number in the green bracket representing the number of non-functional genes. (b,c) The neighbour-joining (NJ) trees of MAM and AOP genes were generated based on the aligned coding sequences and 100 bootstrap repeats. The silenced genes are indicated by red hollow circle, expressed functional genes are represented by red solid disc and green rectangle. In A. thaliana ecotype Columbia there are just MAM1 and MAM3. (d) Three B. oleracea AOP2 loci among which are one functional AOP2 and two mutated AOP2. 1MOI3M: 1-methoxyindol-3-ylmethyl GSL; 1H313M: 1-hydroxyindol-3-ylmethyl GSL; 3MSOP: 3-methysulfinylpropyl GSL; 3MTP: 3-methylthiopropyl GSL; 3PREY: 2-Propenyl GSL; 4BTEY: 3-butenyl GSL; 4MOI3M: 4-methoxyindol-3-ylmethyl GSL; 4OHB: 4-hydroxybutyl GSL; 4H313M: 4-hydroxyindol-3-ylmethyl GSL; 4MSOB: 4-methysulfinylbutyl GSL; 4MTB, 4-methylthiobutyl GSL; AITC: allyl isothiocyanate; I3C: indole-3-carbinol; I3M: indolyl-3-methyl GSL; DIM: 3,3'-diindolylmethane; MAM: methlythioalkylmalate; AOP: 2-oxoglutarate-dependent dioxygenase.
identified in *Brassica*. This analysis supports GSL content surveys and explains why glucoraphanin is abundant in *B. oleracea*, but not in *B. rapa*.

**Discussion**

The *Brassica* genomes experienced WGT followed by massive gene loss and frequent reshuffling of triplicated genomic blocks. Analysis of retained or lost genes following triplication identified over-retention of genes for metabolic pathways such as oxidative phosphorylation, carbon fixation, photosynthesis and circadian rhythm, which may contribute to polyploid vigour.

Fewer lost genes were observed in the less-fractionated sub-genome, possibly due to expression dominance as reported in maize.

Gene expression analysis revealed extensive divergence and AS variants between duplicate genes. This subfunctionalization or neofunctionalization of duplicated genes provides genetic novelty and a basis for species evolution and new species formation. For example, TF genes that are considered to be conserved still have more than 38% of paralogous pairs showing differential expression across tissues although this percentage is lower than the average from all duplicated genes. Gene expression variation may contribute to an increased complexity of regulatory networks after polyploidization.

The multi-layered asymmetrical evolution of the *Brassica* genomes revealed in this study suggests mechanisms of polyploid genome evolution underlying speciation. Asymmetrical gene loss between the *Brassica* subgenomes, the asymmetrical amplification of TEs and tandem duplications, preferential enrichment of genes for certain pathways or functional categories, and divergence in DNA sequence and expression, including alternative splicing among a large number of paralogous and orthologous genes, together give a rough outline for genome evolution after polyploidization. A molecular model of polyploid genome evolution through these asymmetrical mechanisms is summarized in Supplementary Fig. 2. The additional information of accessible large datasets and resource was provided in Supplementary Table 49.

In summary, the *B. oleracea* genomic sequence, its features in comparison with its relatives, and the genome evolution mechanisms revealed, provide a fundamental resource for the genetic improvement of important traits, including components of GSLs for anticancer pharmaceuticals. The genome sequence has also laid a foundation for investigation of the tremendous range of morphological variation in *B. oleracea* as well as supporting genome analysis of the important allotetraploid crop *B. napus* (canola or rapeseed).

**Methods**

Sample preparation and genome sequencing. A *B. oleracea* sp. *capitata* homozygous line 02–12 with elite agronomic characters and widely used as a parent in hybrid breeding was used for the reference genome sequencing (Supplementary Methods). The seedlings of plants were collected and genomic DNA was extracted from leaves with a standard CTAB extraction method. Illumina Genome Analyzer whole-genome shotgun sequencing combined with GS FLX Titanium sequencing technology was used to achieve a *B. oleracea* draft genome.

We constructed a total of 35 paired-end sequencing libraries with insertion sizes of 180–200 bp, 200–250 bp, 500–600 bp, 650–800 bp, 800 bp, 2 kb, 5 kb, 10 kb and 20 kb following a standard protocol provided by Illumina (Supplementary Methods). Sequencing was performed using Illumina Genome Analyser II according to the manufacturer's standard protocol.

Genome assembly and validation. We took a series of checking and filtering measures on reads following the Illumina-Pipeline, and low-quality reads, adaptor sequences and duplicates were removed (Supplementary Methods). The reads after the above filtering and correction steps were used to perform assembly including contig construction, scaffold construction and gap filling using SOAPdenovo2.0 (http://soap.genomics.org.cn/) (Supplementary Methods). Finally, we used 20-kb span-paired-end data generated from the 454 platform and 105-kb-span BAC-end data downloaded from NCBI to extend scaffold length (Supplementary Methods). The *B. oleracea* genome size was estimated using the distribution curve of 17-mer frequency (Supplementary Methods).

To anchor the assembled scaffolds onto pseudo-chromosomes, we developed a genetic map using a double haploid population with 165 lines derived from a F1 cross between two homozygous lines 02–12 (sequenced) and 0188 (re-sequenced). The genetic map contains 1,227 simple sequence repeat markers and single nucleotide polymorphism markers in nine linkage groups, which span a total of 1,180.2 cm with an average of 0.96 cm between the adjacent loci. To position these markers to the scaffolds, marker primers were compared with the scaffold sequences using e-PCR (parameters -n2 -g1 –d 400–800), with the best-scoring match chosen in case of multiple matches.

We validated the *B. oleracea* genome assembly by comparing it with the published physical map constructed using 73,728 BAC clones (http://lulu.pgm, uga.edu/kpc/WebAGCol/brassica/WebFPC/1) and a genetic map from *B. napus* (Supplementary Methods). Eleven Sanger-sequenced *B. oleracea* BAC clones were used to assess the assembled genome using MUMmer-3.22 (http://mummer.sourceforge.net/) (Supplementary Methods).

**Gene prediction and annotation.** Gene prediction was performed on the genome sequence after pre-masking for TEs (Supplementary Methods). Gene prediction was processed with the following steps: (i) De novo gene prediction used AUGUSTUS and GlimmerHMM with parameters trained from *A. thaliana* and *B. rapa* homologous prediction, we mapped the genome using blastn with from *A. thaliana*, O. sativa, C. papaya, V. vinifera and P. trichocarpa to the *B. oleracea* genome using blastn with an E-value cutoff of 10−5 and 3, and used GeneWise (Version 2.2.0) for gene annotation. (ii) For EST-aided annotation, the *Brassica* ESTs from NCBI were aligned to the *B. oleracea* genome using BLAT (identity ≥ 95, coverage ≥ 90%) and further assembled using PASA. Finally, all the predictions were combined using GLEAN to produce the consensus gene sets.

Functional annotation of *B. oleracea* genes was based on comparison with SwissProt, TrEMBL, Interproscan and KEGG proteins databases. The tRNA genes were identified by tRNAscan-SE using default parameters. Then rRNAs were identified with the genome using blastn. Other non-coding RNAs, including miRNA, snRNA, were identified using INFERNAL by comparison with the Rfam database.

**TE annotation.** LTR-RTs were initially identified using the LTR_STRUCT programme, and then manually annotated and checked based on structure characteristics and sequence homology. Identified intact elements were then used to identify other intact elements and solo LTRs. All the LTR-RTs with clear boundaries and insertion sites were classified into superfamilies (Copia-like, Gypsy-like, and Unclassified retroelements) and families relying on the internal protein sequence, 5′, 3′ LTRs, primer-binding site and polyuridine tracts. Non-LTR-RTs (Long interspersed nuclear element, LINE and Short interspersed nuclear element, SINE) and DNA transposons ( Tcl1-Mariner, hAT, Mutator, Pong, Pit-Harbinger, CACTA and miniature inverted repeat TE) were identified using conserved protein domains of reverse transposase or transposase as queries to search against the assembled genome using blastn. Further upstream and downstream sequences of the candidate matches were compared with each other to define their boundaries and structure. Helitron elements were identified by the HelSearch 1.0 programme and manually inspected. All the TE categories were identified according to the criteria described previously. Typical examples of each category were selected and mixed together as a database for RepeatMasker analysis. Around 20× coverage of shotgun reads randomly sampled from the two *Brassica* genomes were masked by the same TE data set to confirm the different accumulation of TEs between the two genomes.

**Syntenic block construction of *B. oleracea* and its relatives.** We used the same strategy as described in the *B. rapa* genome paper to construct syntenic blocks between species (Supplementary Methods). The all-against-all blastp comparison (E-value ≤ 1e−5) provided the gene pairs for syntenic clustering determined by MCScan (MATCH_SCORE: 50, MATCH_SIZE: 5, GAP_SCORE: –3, E_VALUE: 1e−05). As applied in *B. rapa* data set, we assigned and partitioned multiple *B. oleracea* or *B. rapa* orthologous segments that matched the same *A. thaliana* segment (A to X) numbering system in *A. thaliana* into three subgenomes: LF, MF1 and MF2.

**OrthoMCL clustering.** To identify and estimate the number of potential orthologous gene families between *B. oleracea*, *B. rapa*, *A. thaliana*, *C. papaya*, *P. trichocarpa*, *V. vinifera*, *S. bicolor* and *O. sativa*, and also between *B. oleracea* and *B. rapa*, we applied the OrthoMCL pipeline using standard settings (blastp E-value < 1e−4 and inflation factor = 1.5) to compute the all-against-all similarities.

**Phylogenetic analysis of gene families.** We performed comparative analysis of trait-related gene families. Genes from grape, papaya and Arabidopsis were downloaded from the GenoScope database (http://www.genoscope.cns.fr/externe/).
Differential expression of duplicated genes across tissues. RNA-seq reads were mapped to their respective locations on the reference genome using TopHat2. Uniquely aligned read counts were calculated for each gene for each tissue sample. We performed the exact conditional test of two Poisson rates on read counts of duplicated genes to test the differential expression of duplicated genes, according to the method applied in soybean65,66. For each duplicated gene pair (for example, genes A and B), read counts and gene length were denoted as Ea, Eb, La, and Lb for gene A and B, respectively. The read counts of the genes A and B were assumed to follow the Poisson distributions with rates $\lambda_A = Ra/La$ and $\lambda_B = Rb/Lb$. Under the null hypothesis of equal expression of the genes A and B, that is, $Ra = Rb$, the conditional distribution of Ea given $Ea + Eb = k$ follows a binomial distribution. The success probability $P = (La/a + Lb / b) / (La + Lb)$. The $P$ values were computed and further adjusted to maintain the false discovery rate at 0.05 across gene pairs using the Benjamini–Hochberg method67.

Statistical analysis. The average number of all retained orthologues in the three subgenomes was used to estimate the expected retained gene number, for the gene retention disparity statistics using the $w$ test ($\alpha = 0.05$ across gene pairs using the Benjamini–Hochberg method67).

References

1. US Department of Agriculture, Agricultural Research Service. USDA National Nutrient Database for Standard Reference, Release 26: Vegetables and Vegetable Products (2013).

2. Kopsell, D. A. & Kopsell, D. E. Accumulation and bioavailability of dietary carotenoids in vegetable crops. Trends Plant Sci. 11, 499–507 (2006).

3. Halkier, B. A. & Gershenzon, J. Biology and biochemistry of glucosinolates. Plant Cell 24, 433–458 (2012).

4. Wang, X. A. et al. Comparative genomics of Brassica rapa and Arabidopsis thaliana reveal gene loss, fragmentation, and dispersal after polyploidy. Plant Cell 18, 1348–1359 (2006).

5. 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 duplication. Genome Biol. 10, R111 (2009).

6. Wang, W. et al. Construction and analysis of a high-density genetic linkage map in cabbage (Brassica oleracea L. var. capitata). BMC Genomics 13, 523 (2012).

7. Wang, X. et al. A physical map of Brassica oleracea shows complexity of chromosomal changes following recursive paleopolyploidizations. BMC Genomics 12, 470 (2011).

8. Bancroft, J. et al. Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. Nat. Biotechnol. 29, 762–766 (2011).

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

10. Schranz, M. E., Lysak, M. A. & Mitchell-Olds, T. The ABCs of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci. 11, 535–542 (2006).

11. 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).

12. Lysak, M. A., Cheung, K., Kotschke, M. & Bures, P. Ancestral chromosomal blocks are triplicated in Brassicaceae species with varying chromosome number and genome size. Plant Physiol. 145, 402–410 (2007).

13. Panjabi, P. et al. Comparative mapping of Brassica juncea and Arabidopsis thaliana using Intron Polymorphism (IP) markers: homologous relationships, diversification and evolution of the A, B and C Brassica genomes. BMC Genomics 9, 113 (2008).

14. Parkin, I. A. et al. Segmental structure of the Brassica napus genome based on comparative analysis with Arabidopsis thaliana. Genetics 171, 765–781 (2005).

15. Wu, J. et al. The genome of the pear (Pyrus brentschneideri Rebh.). Genome Res. 22, 496–408 (2012).

16. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635–641 (2012).

17. Cheung, F. et al. Comparative analysis between homoeologous genome segments of Brassica napus and its progenitor species reveals extensive sequence-level divergence. Plant Cell 21, 1912–1928 (2009).

18. Freeling, M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu. Rev. Plant Biol. 60, 433–453 (2009).

19. Sankoff, D., Zheng, C. & Zhu, Q. The collapse of gene complement following whole genome duplication. BMC Genomics 11, 313 (2010).

20. Woodhouse, M. R., Tang, H. & Freeling, M. Different gene families in Arabidopsis thaliana transposed in different epochs and at different frequencies throughout the rosids. Plant Cell 23, 4241–4253 (2011).

21. Lou, P. et al. Preferential retention of circadian clock genes during diploidization following whole genome triplication in Brassica rapa. Plant Cell 24, 2415–2426 (2012).

22. Doyle, J. I. et al. Evolutionary genetics of gene merger and doubling in plants. Annu. Rev. Genet. 42, 443–461 (2008).

23. Wang, X. Tang, H. & 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).

24. Syed, N. H., Kalyna, M., Marquez, Y., Barta, A. & Brown, J. W. Alternative splicing in plants—coming of age. Trends Plant Sci. 17, 616–623 (2012).

25. Gabut, M. et al. Alternative splicing switches regulate embryonic stem cell pluripotency and reprogramming. Cell 147, 132–146 (2011).

26. Zhang, P. G., Huang, S. Z., Pin, A. L. & Adams, K. L. Extensive divergence in alternative splicing patterns after gene and genome duplication during the evolutionary history of Arabidopsis. Mol. Biol. Evol. 27, 1866–1897 (2010).

27. Filichkin, S. A. et al. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genescience 15, 910–9123 (2006).

28. Yang, B. & Quinones, C. F. Survey of glucosinolate variation in leaves of Brassica rapa crops. Genet. Res. Crop Evol. 57, 1079–1089 (2010).

29. Benderoth, M., Pfälz, M. & Kroymann, J. Methylthioalkylamylate synthetases: genetics, ecology and evolution. Phytochemistry. 80, 255–268 (2009).

30. Benderoth, M. et al. Positive selection driving diversification in plant secondary metabolism. Proc. Natl. Acad. Sci. USA 103, 9110–9123 (2006).

31. TeSor, S. de Kranz, J. W., Hause, B., Gershenzon, J. & Tokuhisa, G. J. MAM3 catalyses the formation of all aliphatic glucosinolate chain lengths in Arabidopsis. Plant Physiol. 144, 60–71 (2007).
Acknowledgements
This work was supported by the National Basic Research Program of China (2011CB109300, 2012CB113906, 2012CB723007 and 2006CB101600), the National Natural Science Foundation of China (3067134, 30671119 and 31301309), the National High Technology Research and Development Program (2013AA102602, 2012AA100105 and 2012AA100104), the China Agriculture Research System (CARS-13 and CARS-25-A), the Core Research Budget of the Non-profit Governmental Research Institution (1601172010005), the Special Fund for Agro-scientific Research in the Public Interest (20100316), China–Australia collaboration project (2010DFA31730), UK Biotechnology and Biological Sciences Research Council (BB/E017363/1), the Australian Research Council (LP0808295, LP0808346, DP0809553 and LP11010020), the Next-Generation BioGreen 21 Program (PJ008944 and PJ008202), and the US National Science Foundation (IOS 0863841, DBI 0849896, MCB 1021718).

Author contributions
I.B., L.C., D.E., Q.H., W.H., G.J.K., S.L., Y.H., J.M., A.H.P., J.C.P., L.A.P., J.W., JunW., XiaoWuW., XiYinW. and T.-J.Y. are principal investigators (alphabetical order). B.C., W.H., A.H.P., JunW. and XiaoWuW. are equally contributing senior authors. S.L., J.W., W.H., XX and Z.Y. planned and managed the project. S.L., C.T., A.H.P. and D.E., X.Y. and M.Z. wrote this manuscript and I.B., J. Ma., G.J.K., J.C.P., B.C., T.-J.Y., L.A.P., XiYinW., XiaoWuW., K.L., Y.B. and A.G.S. made revision or edits or comments. J.W. (leader), W.H. (co-leader), JunW., L.Y. and Z.Y. performed DNA sequencing. L.Y. (leader), W.H. (co-leader), S.H., J.W. and L.Y. conducted genomic sequence assembly. S.H. (leader), XiYinW. (co-leader), J.Min, L.B., W.H., J.B., D.E., P.R., S.L., I.S., Y.L. and W.W. conducted scaffold anchoring to linkage maps and assembly validation. X.Y. (leader), J.Y. (co-leader), S.L., Q.Z., S.H. and J. Min performed annotation. C.T. (leader), Wanshun L., W.H., Y.B., C.L., W.W., J. Wu, S.L., C.D. and M.Z. performed transcriptome sequencing. S.L. conceived analysis of comparison and evolution. S.L. (leader), C.T., X.Y., Zhan-gyanW., C.L., S.H., J. Ma., J.Y., M.Z., Zhou W., Q.Z., S.P., I.A.P., A.G.S., L.Y., I.B., G.J.K., J.C.P., XiaoWuW., B.C., F.C., YinH. and WenbinL. and X.Liang performed analysis of comparative genomics and evolution. J. Ma (leader), M.Z., Q.Z., C.T., S.L., B.C., S.H., H.B., C.L. and Jianal. conducted TE analysis. XiYinW. (leader), Y.Y., T.-J.Y., Zhan-gyanW., L.W., J. Li, T.-H.L., Jinheng W., H.L., X.T., M.G. and L.I. conducted gene family analysis. K.L. (leader), J.Y., S.L., C.T., H.L., G.H., S.P., D.Z., F.C., Q.H. and Xinfa W., C.Q., D.D., Z.H., J.H., D.M., L.I., Z. Li, J.Z., L.X., Y.Zhou., Z.L. and Y.Zhang conducted trait-related gene analysis. A.H.P. (leader), XiYinW., D.J., Y.W. and T.-H.L. conducted gene conversion analysis. T.-J. Y. (leader), M.Z., P.S., B.S.P., J.Ma., N.E.W., R.Q., X.L., J.Lee and H.H.K. conducted centromere analysis. C.T. (leader), S.L., X.Y., S.H., C.L., Zhangyan W., Q.Z., J.Y., J.T. and B.C. conducted tandemly duplicated gene analysis. ZhangyanW. and J.Y. performed data submission.

Additional information
Accession codes: Genome sequence data for B. oleracea have been deposited in the DDBJ/EMBL/GenBank nucleotide core database under the accession codes AOXX00000000. Transcriptome sequence data for B. rapa and B. oleracea have been deposited in the DDBJ/EMBL/GenBank Sequence Read Archive (SRA) under the accession codes GSE43245 and GSE42891 respectively.

Supplementary information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Liu, S. et al. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nat. Commun. 5:3930 (2014). doi: 10.1038/ncomms4930

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/

© 2014 Macmillan Publishers Limited. All rights reserved.