Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome

Boulos Chalhoub,1,† France Denoëud,2,3,4,5 Shengyi Liu,6,7 Isobel A. P. Parkin,6,† Haibo Tang,1,† Xiyan Wang,6,7,8 Julien Chiquet,1 Harry Beleram,1 Chaobo Tong,9 Birgit Samans,5 Margot Corrêa,1 Corinne Da Silva,2 Jérémy Just,1 Cyril Falentin,13 Chu Shin Koh,14 Isabelle Le Clainche,1 Maria Bernard,2 Pascal Bento,1 Benjamin Noel,2 Karine Labadie,2 Adriana Alberti,2 Mathieu Arnaud,3 Hui Guo,9 Christian Davidau,14 Salvador Alaniz,17,26 Jérémie Jourdan,11,26 Michela Deniot,13 Mei Guan,25 Zhongsong Liu,25 Fengming Sun,22 Yong Pyo Lim,26 Philipp E. Bayer,14 Agnieszka A. Golicz,17 Sahana Manoli,17 Tae-Ho Lee,9 Patrick P. Edger,20 Houda Chelaifa,1 David Tack,21 Gilles Lassalle,13 Imen Mestiri,1 Nicholas Schnel,13 Marie-Christine Le Paslier,13 Guangyi Fan,23 Victor Renault,23

**PLANT GENETICS**

**Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome**

The *B. napus* genome contains 101,040 gene models estimated from 353 Gb of RNA sequencing (RNA-seq) data (tables S15 and S16) in combination with ab initio gene prediction, protein and EST alignments, and transposon masking (7). Of these, 91,167 were confirmed by matches with *B. rapa* and/or *B. oleracea* predicted proteomes. Genes are abundant in distal euchromatin but sparse near centromeric heterochromatin (Fig. 2). RNA-seq data revealed alternative splicing in 48% of genes, with frequent intron retention (82%) and rare exon skipping (3%) (tables S17 and S18 and fig. S11).

The *B. napus* An and Cn subgenomes are largely colinear to the corresponding diploid An and Cn subgenomes (tables S2 and S3). Unique mapping of ~500,000 sequences from *B. rapa* (‘Chifu’) or *B. oleracea* (‘TO1000’) assigned most of the 207,020 *B. napus* scaffolds to either the An (8294) or the Cn (9884) subgenomes (tables S4 and S5 and fig. S3). The assembly covers ~79% of the 130-Mb genome and includes 85.6% of *Brassica* expressed sequence tags (ESTs) (7). A single-nucleotide polymorphism (SNP) map (tables S6 to S9 and fig. S9) was constructed from 207,020 *B. napus* scaffolds (table S14a and figs. S9 and S10). The *B. napus* subgenomes (325.8 Mb) is larger than the An subgenome (314.2 Mb), consistent with the relative sizes of the assembled An genome of *B. oleracea* (540 Mb, 85% of the ~630-Mb genome) and the *B. rapa* genome (312 Mb, 59% of the ~530-Mb genome) (9–11). The *B. napus* assembly contains 34.8% transposable elements (TEs), less than the 40% estimated from the progenitor genomes (9–11). A small TE fraction has proliferated since *B. rapa* separated from its progenitors (7), at lower rates in the *B. napus* subgenomes than the corresponding progenitor genomes (table S14 and figs. S9 and S10).

Oilseed rape (*Brassica napus* L.) was formed ~7500 years ago by hybridization between *B. rapa* and *B. oleracea*, followed by chromosome doubling, a process known as allopolyploidy. Together with more ancient polyploidizations, this conferred an aggregate 72× genome size. Homozygous chromosome mergers (genome AnAnCnCn) were formed ~7500 years ago by hybridization between the ancestors of *B. napus* and *B. rapa* (genome AnAnCn). These predate the earliest cultigens (10). The homozygous genome mergers (Brassica napus L.) was formed ~7500 years ago by hybridization between *B. rapa* and *B. oleracea*, followed by chromosome doubling, a process known as allopolyploidy. Incipient gene loss and expression divergence have begun. Selection in *B. napus* oilseed types has accelerated the loss of glucosinolate genes, while preserving expansion of oil biosynthesis genes. These processes provide insights into allopolyploid evolution and its relationship with crop domestication and improvement.

The *Brassicaceae* are a large eudicot family (1) and include the model plant *Arabidopsis thaliana*. *Brassicas* have a propensity for genome duplications (Fig. 1) and genome mergers (2). They are major contributors to the human diet and were among the earliest cultigens (3).

*B. napus* (genome AnAnCnCn) was formed by recent allopolyploidy between ancestors of *B. oleracea* (Mediterranean cabbage, genome CnCn) and *B. rapa* (Asian cabbage or turnip, genome AnAn) and is polyphyletic (2, 4), with spontaneous formation regarded by Darwin as an example of unconscious selection (5). The assembled Cn subgenome (525.8 Mb) is larger than the An subgenome (314.2 Mb), consistent with the relative sizes of the assembled An genome of *B. oleracea* (540 Mb, 85% of the ~630-Mb genome) and the *B. rapa* genome (312 Mb, 59% of the ~530-Mb genome) (9–11). The *B. napus* assembly contains 34.8% transposable elements (TEs), less than the 40% estimated from the progenitor genomes (9–11). A small TE fraction has proliferated since *B. rapa* separated from its progenitors (7), at lower rates in the *B. napus* subgenomes than the corresponding progenitor genomes (table S14 and figs. S9 and S10).
genomes, with asymmetric gene distribution (42,320 and 48,847, respectively) and 93% of the diploid gene space in orthologous blocks (fig. S12) (7). We identified 34,255 and 38,661 orthologous gene pairs between the An and Cn subgenomes and their respective progenitor genomes (fig. S13). Comparison of An-An and Cn-Cn orthologous gene pairs suggested a divergence 7500 to 12,500 years ago (fig. S14), indicating formation of B. napus after this date. Synteny with Arabidopsis (table S19) confirmed the triplicated mesoploid structure (9–11) of the An and Cn subgenomes, with the recent allopolyploidy conferring on B. napus an aggregate 72× genome multiplication since the origin of angiosperms (Fig. 1) (7).

**Fig. 1. Recurrent genome duplications in B. napus.**

Genomic alignments between the basal angiosperm Amborella trichopoda (24), the basal eudicot Vitis vinifera (25), and the model crucifer A. thaliana, as well as B. rapa (9), B. oleracea (10, 11), and B. napus, are shown. A typical ancestral region in Amborella is expected to match up to 72 regions in B. napus (69 were detected for this specific region). Gray wedges in the background highlight conserved synteny blocks with more than 10 gene pairs.

**Fig. 2. The genome of the B. napus oilseed cultivar ‘Darmor-bzh’**. The genome comprises 9 chromosomes belonging to the Cn subgenome and 10 to the An subgenome, scaled on the basis of their assembled lengths. Tracks displayed are (A) gene density (nonoverlapping, window size = 100 kb for all tracks). Positions showing loss of one or more consecutive genes are displayed (triangles) along with homeologous exchanges, detected as missing genomic segments that have been replaced by duplicates of corresponding homeologous segments (red rectangles). (B and C) Transcription states estimated by RNA-seq in leaves (B) and roots (C) (in nonoverlapping 100-kb windows). (D) DNA transposon density. (E) Retrotransposon density. (F) CpG methylation in leaves (green) and roots (brown); both curves are overlapping. (G) Centromeric repeats (densities exaggerated for visual clarity). Homeologous relationships between An and Cn chromosomes are displayed with connecting lines colored according to the Cn chromosomes.
Most orthologous gene pairs in *B. rapa* and *B. oleracea* remain as homeologous pairs in *B. napus* (tables S19 to S25 and figs. S12 to S17). DNA sequence analysis (7) confirmed the loss of 112 An and 91 Cn genes in *B. napus* ‘Darmor-bzh’ (tables S21 to S26), ~2.6 times higher than the 41 and 37 genes lost in *B. rapa* ‘Chiifu’ and *B. oleracea* ‘TO1000’ respectively (tables S26 and S27; \( \chi^2 \) test \( P = 5.3 \times 10^{-14} \)). Further analyses of a *Brassica* diversity set showed that ~47% of Darmor-bzh An and 31% of Cn deleted genes were also deleted in at least one additional progenitor genotype (tables S28 and S29), indicating that their deletion probably predated allopolyploidization of *B. napus* (7). A high proportion (27% to 54%) of the remaining Darmor-bzh deleted genes were also deleted from diverse *B. napus* genotypes (tables S28 and S29).

Homeologous exchanges (HEs), including crossovers and noncrossovers, are frequent between *B. napus* subgenomes and range in size from large segments to single SNPs (7) (Fig. 3, figs. S17 to S24, and tables S30 to S39). At the chromosome segment level, HEs are characterized by replacement of a chromosomal region with a duplicated copy from the corresponding homeologous subgenome (7). We identified 17 HEs, 14 Cn to An and 3 An to Cn (Fig. 3, fig. S19, and tables S30 and S31). Sequences from seven diverse *B. napus* genotypes revealed both shared and specific segmental HEs. These are of varying sizes and are most frequent between chromosomes An1-Cn1, An2-Cn2, and An9-Cn9 (table S32, Fig. 3, and fig. S19). Larger HEs found in the synthetic *B. napus* H165 affect, for example, most of chromosomes An1-Cn1 and An2-Cn2 (Fig. 3 and fig. S19). Functional annotation of genes within HEs suggests some have experienced selection, contributing to the

Fig. 3. HEs between *B. napus* chromosomes An2 and Cn2. (A) Coverage depth obtained along the An2 chromosome after mapping Illumina sequence reads from seven natural and one resynthesized *B. napus* genotypes (named on the right) to the reference genome of *B. napus* ‘Darmor-bzh’. (B) Coverage depth obtained for An2 and Cn2 chromosomes, respectively, after mapping >21 genome-equivalents of Illumina sequence reads from *B. napus* Darmor-bzh on the *B. rapa* and *B. oleracea* genome assemblies concatenated together. (D) Similar to (A), where the Cn2 chromosome of Darmor-bzh is displayed. Segmental HEs are revealed based on sequence read coverage analysis, where a duplication (red) is revealed by significantly greater coverage for a given segment than the rest of the genome (black) and a deletion (blue) by little or no coverage for the corresponding homeologous segment. Sizes of chromosomes are indicated in Mb. An-to-Cn converted genes (at 60% or more conversion sites) are plotted as blue dots on An2 (B) and red dots on Cn2 (C). Cn-to-An converted genes are plotted as blue dots on Cn2 (C) and red dots on An2 (B). Open circles denote entirely converted genes using the same color code. Light gray lines connecting (A), (B), (C), and (D) indicate orthology relationships, and dark gray lines highlight segmental HEs in Darmor-bzh (names and descriptions detailed in table S31). Further HEs occurring between other homeologous chromosomes are shown in fig. S19. Black arrows in (A) indicate HEs involving GSL and FLC genes.
The expansion of GSL biosynthesis in \( B. \) napus has undergone intensive breeding to optimize seed oil content and lipid composition, decrease nutritionally undesirable erucic acid and glucosinolates (GSLs), optimize flowering behavior, and improve pathogen resistance. The expansion of \( B. \) napus lipid biosynthesis genes exceeds that known in other oilseed plants, with 1097 and 1132 genes annotated in the \( A_n \) and \( C_n \) subgenomes, respectively (7) (tables S46 to S48). Most lipid biosynthesis genes identified in the progenitor genomes are conserved in \( B. \) napus. For 18 acyl lipid orthologs, 3 and 2 genes appeared to be deleted from \( A_n \) and \( C_n \) subgenomes, respectively. Another 15 have been converted by HEs, nine from \( A_n \) to \( C_n \), and four from \( C_n \) to \( A_n \) (tables S47 and S48) (7).

Genetic variation for reduced seed GSLs also appears to be under breeding-directed selection. GSLs are sulfur-rich secondary metabolites important for plant defense and human health (19). However, high levels in seeds can lead to breakdown products in animal feeds (20). All 22 GSL catabolism genes identified in \( B. \) rapa and \( B. \) oleracea (10) are conserved in \( B. \) napus (7), and orthologs of only three \( C_n \) and one \( A_n \) GSL biosynthesis genes are missing (table S49). One deleted homeologous pair, corresponding to orthologs of \( B. \) oleracea \( B22g161590 \) and \( B. \) rapa \( Brat02s931 \), contains two with quantitative trait loci (QTLs) for total aliphatic GSL content (2P) and corresponds to a HE in which a segment of \( A_2 \), with a missing \( C_n \) gene, has replaced the \( C_2 \) homeolog (Fig. 3). Two additional QTLs for aliphatic GSL content (2P) colocalize with a deletion of one \( B. \) rapa \( Brat03s0299 \) ortholog on \( A_9 \) and its nondeleted homeolog on \( C_9 \) (\( BnaC09g08300D \), fig. S17).

We identified 425 nucleotide binding site leucine-rich repeat (NBS-LRR) sequences encoding resistance gene homologs (245 on \( C_n \) and 180 on \( A_n \)). Of these, 75% (153 \( A_n \) and 224 \( C_n \)) are syntenic to \( A_n \) and \( C_n \) progenitors (7) (table S50 and figs. S33 and S44). We confirmed the absence of five NBS-LRR genes from the \( A_n \) subgenome, three from the \( C_n \) subgenome, and three from \( B. \) rapa \( A_n \), with none absent from \( B. \) oleracea \( C_n \). This variation may reflect differential selection for resistance to diseases.

\( B. \) napus morphotypes show broad adaptation to climatic and latitudinal conditions, and greater transposon density in the \( C_n \) subgenome (Fig. 2F). Of the ~3100 gene pairs with patterns of both genome dominance and tissue effects and tissue-by-subgenome interaction for resistance to diseases.

Human cultivation and breeding of \( B. \) napus morphotypes may have selected favorable HEs, causing subgenome restructuring of regions containing genes controlling valuable agronomic traits such as those shown here for oil biosynthesis, seed GSL content, disease resistance, and flowering. Because \( B. \) napus is a young allotetraploid beginning gene loss and genome reorganization, further partitioning of expression may become a key determinant for the long-term preservation of its duplicated genes (23).

The integrative genomic resources that we report provide unique perspectives on the early evolution of a domesticated polyploid and will facilitate the manipulation of useful variation, contributing to sustainable increases in oilseed crop production to meet growing demands for both edible and biofuel oils.
Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome


Science 345 (6199), 950-953.
DOI: 10.1126/science.1253435

The genomic origins of rape oilseed

Many domesticated plants arose through the meeting of multiple genomes through hybridization and genome doubling, known as polyploidy. Chalhoub et al. sequenced the polyploid genome of Brassica napus, which originated from a recent combination of two distinct genomes approximately 7500 years ago and gave rise to the crops of rape oilseed (canola), kale, and rutabaga. B. napus has undergone multiple events affecting differently sized genetic regions where a gene from one progenitor species has been converted to the copy from a second progenitor species. Some of these gene conversion events appear to have been selected by humans as part of the process of domestication and crop improvement.

Science, this issue p. 950