Deciphering the Diploid Ancestral Genome of the Mesohexaploid *Brassica rapa* 

Feng Cheng, a Terezie Mandáková, b Jian Wu, a Qi Xie, c Martin A. Lysak, b, 1 and Xiaowu Wang a, 1, 2

a Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China
b Plant Cytogenomics, Central European Institute of Technology (CEITEC) and Faculty of Science, Masaryk University, CZ-625 00 Brno, Czech Republic
c State Key Laboratory of Plant Genomics, National Center for Plant Gene Research (Beijing), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

The genus *Brassica* includes several important agricultural and horticultural crops. Their current genome structures were shaped by whole-genome triplication followed by extensive diploidization. The availability of several crucifer genome sequences, especially that of Chinese cabbage (*Brassica rapa*), enables study of the evolution of the mesohexaploid *Brassica* genomes from their diploid progenitors. We reconstructed three ancestral subgenomes of *B. rapa* (*n* = 10) by comparing its whole-genome sequence to ancestral and extant Brassicaceae genomes. All three *B. rapa* paleogenomes apparently consisted of seven chromosomes, similar to the ancestral translocation Proto-Calepineae Karyotype (tPCK; *n* = 7), which is the evolutionarily younger variant of the Proto-Calepineae Karyotype (*n* = 7). Based on comparative analysis of genome sequences or linkage maps of *Brassica oleracea*, *Brassica nigra*, radish (*Raphanus sativus*), and other closely related species, we propose a two-step merging of three tPCK-like genomes to form the hexaploid ancestor of the tribe Brassiceae with 42 chromosomes. Subsequent diversification of the Brassiceae was marked by extensive genome reshuffling and chromosome number reduction mediated by translocation events and followed by loss and/or inactivation of centromeres. Furthermore, via interspecies genome comparison, we refined intervals for seven of the genomic blocks of the Ancestral Crucifer Karyotype (*n* = 8), thus revising the key reference genome for evolutionary genomics of crucifers.

INTRODUCTION

The genus *Brassica* currently comprises 38 species and numerous varieties, many of which are important crops or weeds. The six species that make up the so-called “Triangle of U” are of particular economic importance because of their cultivation as vegetables, condiments, and sources of oilseed. These species are the diploids *Brassica rapa* (AA genome), *Brassica nigra* (BB), and *Brassica oleracea* (CC) as well as the allotetraploid species *Brassica juncea* (AABB), *Brassica napus* (AACC), and *Brassica carinata* (BBCC). *Brassica* is one of 47 genera in the monophyletic tribe Brassiceae. This tribe, which contains ~238 species primarily distributed in the Mediterranean region, southwestern Asia, and northern Africa (Prakash and Hinata, 1980; Al-Shehbaz, 2012), belongs to lineage II, one of three well-related tribes Eutremeae and Thlaspideae, remain unclear, and more than 20 loosely interrelated tribes have been described within an expanded lineage II (Figure 1B) (Franzke et al., 2011). Early chromosomal studies indicated that the diploid *Brassica* species might actually represent balanced secondary polyploids (Catcheside, 1934; Roebbelen, 1960). Later comparative restriction fragment length polymorphism mapping among these species and between *Brassica* species and *Arabidopsis thaliana* (Lagercrantz and Lydiate, 1996; Lagercrantz, 1998) suggested that diploid *Brassica* genomes descended from a hexaploid ancestor. All Brassiceae species analyzed to date contain either three or six copies of orthologous genomic regions of *A. thaliana* (Lysak et al., 2005, 2007; Parkin et al., 2005). These findings provide compelling evidence for an ancestral whole-genome triplication in the ancestry of the monophyletic tribe Brassiceae. More recently, analysis of the draft genome sequence of *B. rapa* (*n* = 10) revealed the three subgenomes of the hexaploid ancestor (Wang et al., 2011). A comparison of collinearity between the 10 *B. rapa* chromosomes and the five *A. thaliana* chromosomes using 24 ancestral genomic blocks (GBs) of the Ancestral Crucifer Karyotype (ACK; *n* = 8; Schranz et al., 2006) revealed three syntenic copies of each *A. thaliana* GB, with only one exception, in the *B. rapa* genome (Wang et al., 2011). Because the three GB copies differ regarding their rate of gene loss (fractionation), blocks were classified according to their gene density as belonging to the least fractionated (LF), the medium fractionated (MF1), and the most fractionated (MF2)
subgenome. Wang et al. (2011) suggested a two-step origin of *B. rapa* involving a tetraploidization followed by substantial genome fractionation (subgenomes MF1 and MF2) and subsequent hybridization with a third, less-fractionated genome. This two-step model gained support from a comparison of the frequency of deletion-sine sequences between the three *B. rapa* subgenomes (Tang et al., 2012).

Karyotype and genome evolution in the expanded lineage II was analyzed by Mandáková and Lysak (2008). Comparative chromosome painting using *A. thaliana* BAC contigs demonstrated that the six analyzed tribes (i.e., Calepineae, Coluteocarpeae, Conringieae, Eutremeae, Isatideae, and Sisymbrieae) descended from a common ancestor with the Proto-Calepineae Karyotype (PCK; \(n = 7\)). While Calepineae, Coluteocarpeae (synonym of Noccaeeae), and Conringieae taxa retain the PCK genome structure, the tribes Eutremeae, Isatideae, and Sisymbrieae display an additional whole-arm translocation (Figures 1B and 2A). This younger ancestral genome is henceforth referred to as the translocation PCK (tPCK). The tPCK genome structure was recently confirmed by whole-genome sequencing of *Schrenkiella parvula* (synonym of *Thellungiella parvula*, formerly placed in the Eutremeae) (Dassanayake et al., 2011) and *Thellungiella salsuginea* (Eutremeae) (Wu et al., 2012). The PCK shares five chromosomes with the ACK and structurally closely resembles the extant karyotypes of *Arabidopsis lyrata* and *Capsella rubella* (Lysak et al., 2006; Schranz et al., 2006; Hu et al., 2011). It has therefore been suggested that either the PCK or ACK is descended from a common ancestor or, more likely, that the seven PCK chromosomes are derived from the eight chromosomes of the ACK (Mandáková and Lysak, 2008) (Figure 1B).

For more than half a century, the chromosome number and genomic structure of the hexaploid *Brassica* ancestor and its three parental subgenomes have been a matter of debate. Several chromosome numbers, ranging from \(x = 3\) to \(7\), have been proposed for ancestral *Brassica* subgenomes (Roebbelen, 1960; Truco et al., 1996; reviewed in Prakash and Hinata, 1980). Based on a comparison of *B. rapa*, *B. oleracea*, and *B. nigra* linkage maps, Truco et al. (1996) suggested that all three subgenomes were derived from an ancestor with five, six, or seven chromosomes. Although those authors favored six ancestral chromosomes (W1 to W6), a seventh chromosome (W7) could not be ruled out. Comparison of the PCK with the structure of the *B. rapa* subgenome in *B. napus* (Parkin et al., 2005) revealed...
two shared rearrangements, thus suggesting that two or all three subgenomes of the hexaploid ancestor of the Brassiceae corresponded to either PCK or tPCK (Mandáková and Lysak, 2008).

The aim of this study was to elucidate the structure of the three ancestral subgenomes of the hexaploid ancestor of the Brassiceae. Toward that aim, structural genome analyses of *B. rapa* in comparison with ancestral and extant crucifer genomes were performed. Specifically, we tested whether the PCK and/or tPCK represent the karyotype of the diploid ancestors of the mesohexaploid *B. rapa*, the genus *Brassica*, and the tribe Brassiceae.

RESULTS

Deciphering the Ancestral Diploid Karyotypes of *B. rapa*

Based on the previously determined syntenic relationship of the three *B. rapa* subgenomes (LF, MF1, and MF2) to the *A. thaliana* genome (Wang et al., 2011; Cheng et al., 2012b), 71 of 72 expected GBs (3 × 24) were detected in the *B. rapa* genome (Figure 2B; see Supplemental Table 1 online). Of the 24 blocks, 23 were present in three copies and one (block G) in only two copies. Among the 71 GBs, 48 (67.6%) exhibited continuous sequence collinearity between *B. rapa* and *A. thaliana*, 21 (29.6%) were disrupted into at least two segments, and two (2.8%; one copy each of blocks E and T) contained sequence deletions (see Supplemental Tables 1 and 2 online). Analysis of GB integrity did not reveal significant differences in the degree of rearrangements among the three *B. rapa* subgenomes (see Supplemental Table 3 and Supplemental Data Set 1 online).

We compared the structure of the ACK, PCK, and tPCK (Figure 2A) with that of the three *B. rapa* subgenomes (Figure 2B). Using gene densities characteristic for each of the three *B. rapa* subgenomes and methods previously described (Wang et al., 2011; Cheng et al., 2012b), we then reconstructed three tentative *B. rapa* ancestral subgenomes by projecting the 71 GBs onto the ACK genome and the seven chromosomes of the PCK and tPCK. When a modern genome is compared with ancestral and extant crucifer genomes, the more recent the ancestor, the greater the expected degree of genomic continuity with the present-day genome. To examine genome continuity, we defined two adjacent GBs as a GB association and compared all GB associations between *B. rapa* and the ancestral genomes ACK, PCK, and tPCK. There are 16 GB associations in the ACK genome and 17 associations in the PCK and tPCK genomes (Table 1, Figure 2A). In *B. rapa*, 10, 15, and 16 GB associations were identified between ACK, PCK, and tPCK, respectively. Of the 10 ACK block associations, four were present in three copies in *B. rapa*, while six were present in two copies. Of the 15 GB associations in PCK, eight were present in three copies, six in two, and one as a single copy. Of the 16 tPCK block associations, eight were present in three copies, two in two copies, and as single copies (Table 1; see Supplemental Table 4 online).

Block associations V/K/L/Wa/O/X and O/P/W/R, specific for both PCK and tPCK, were identified in all three *B. rapa* subgenomes (Figure 2B, Table 1). Three genomic copies of the V/K/L/Wa/O/X association were found on chromosomes A02 (MF1), A06 (LF), and A09 (MF2), although further rearrangements occurred on A06 and A09. Copies of the O/P/W/R association were found but rearranged on chromosomes A02 (MF2) and A03 (MF1). A third O/P/W/R copy was split between chromosomes A09 (blocks O/P) and A10 (W/R). tPCK contained two specific GB associations, D/V and M/E, resulting from a whole-arm reciprocal translocation between chromosomes AK2 and AK5/6/8 within the PCK genome (Figure 2A; Mandáková and Lysak, 2008). In the MF2 subgenome, the D/V association was identified on chromosome A09. In the LF subgenome, blocks D and V were split between chromosomes A03 and A06 due to an intersubgenomic translocation between block D (LF) and a part of block U (MF1) (see Supplemental Figure 1A online). The M/E association is not present in *B. rapa*.

To facilitate the retrieval of gene sets of the three tPCK-like subgenomes in *B. rapa* (see Supplemental Figure 3 online) with mutual homoeology and synteny to *A. thaliana* orthologs, we developed a Web tool, Search Syntenic Genes Br-tPCK, embedded in the BRAD Brassica Database (BRAD; http://brassicadb.org/brad/searchsyntenytpck.php). There were a total of 12,914, 8905, and 7719 genes in the LF, MF1, and MF2 subgenomes, respectively, syntenic to *A. thaliana* (see Supplemental Table 5 online).

Paleocentromeres in the *B. rapa* Genome

Paleocentromere sequence analysis indicated that the 10 extant *B. rapa* centromeres were all inherited from the triplicated tPCK-like genomes. Using the *B. rapa* genome sequence, we analyzed the fate of paleocentromeres corresponding to 21 ancestral chromosomes of the three tPCK-like genomes. After aligning (pericentromere-specific repeats, including centromeric satellite repeats CentBr (Lim et al., 2005; Koo et al., 2011), PCBR, and TR238 (Lim et al., 2007), to the *B. rapa* genome sequence, signals were detected for 18 paleocentromere regions. The paleocentromeres of three ancestral chromosomes {AK3\(_{MF1}\), AK3\(_{MF2}\)}, and AK6\(_{MF2}\) could not be located (Figures 2B and 3; see Supplemental Data Set 2 online).

Among the 18 detected paleocentromeres, 10 were involved in the origin of the 10 *B. rapa* centromeres (Table 2, Figures 2B and 3).
Figure 2. Ancestral Brassicaceae Genomes and the Distribution of Ancestral GBs along 10 Chromosomes of *B. rapa*.

(A) The ancestral genomes ACK, PCK, and tPCK, each comprising 24 ancestral GBs. In the PCK and tPCK, chromosome labels reflect the presumed origin of the two karyotypes from the more ancestral ACK genome.

(B) Ten *B. rapa* chromosomes (A01 to A10) consisting of 24 triplicated GBs (71 blocks in total, as one copy of block G was lost) assigned to subgenomes LF (red), MF1 (green), and MF2 (blue). Two or more segments of a single block are labeled by lowercase letters (a, b, etc.). Downward-pointing arrows are adjacent to GBs that were inverted relative to other block(s), which originated from a single ACK chromosome. Detected and undetected ancestral centromeres in the *B. rapa* genome are shown as black and white small rectangles, respectively.
The centromeres on chromosomes A04, A05, A07, and A09 correspond to ancestral centromeres of AK7, AK1, AK3, and AK2/5/6/8, respectively, with two ancestrally associated GBs flanking the centromeres. The centromeres of A01, A03, and A10 were derived via a translocation or pericentric inversion event from paleocentromeres of AK2/5/6/8, AK4, and AK6/8, respectively, leaving only one of the two expected ancestral GBs associated with the centromere. The origin of centromeres on A06, A08, and A05 is directly linked to a reduction in chromosome number. For these centromeres, we inferred translocation events with breakpoints within the (peri)centromeric regions of two different ancestral genomes resulting in the loss of one of the two paleocentromeres. The available data do not allow us to determine which three of the six participating paleocentromeres remained active in B. rapa (Table 2, Figure 3).

Of the eight remaining ancestral centromeres, six were disrupted by translocation events, with sequence traces detectable at four paleocentromere regions (Figure 3). Chromosomes A03 and A04 contained two entirely conserved ancestral chromosomes with inactivated/deleted centromeres. The AK3 centromere left no sequence remnants between blocks G_{MF2} and H_{MF2} within A03, whereas centromere sequence traces (PCRBr and TR238 in the boundary of J_{NF2}) corresponding to AK4 (see Supplemental Data Set 2 online) were detected on A04 (Figures 2B and 3).

Comparison of B. rapa Ancestral Genomes with Other Brassicaceae Genomes

To obtain further insights into the origin and phylogenetic context of the B. rapa ancestral genomes, we investigated whether the tPCK-specific GB associations D/V/K/L/Wa/Q/X, O/P/W/R, M/E, and D/V are present in genomes of other Brassica and Brassicaceae species, as well as other lineage II species (i.e., B. oleracea, B. nigra, radish (Raphanus sativus), Sinapis alba, Caulanthus amplexicaulis, S. parvula, and T. salsuginea; Figure 1B). Three copies of the first two tPCK-specific block associations were identified in B. oleracea assembled scaffolds (see Supplemental Table 6 online). In the B. nigra genome analyzed within the allotetraploid B. juncea, two copies of V/K/L/Wa/Q/X were located on chromosomes B1 and B6, whereas the third copy was split between B4 and B8. Similarly, two copies of O/P/W/R were located on B2 and B3, while the third copy was rearranged on chromosome B8 (Panjabi et al., 2008) (see Supplemental Figure 2 online). Furthermore, the D/V association was found on one scaffold of B. oleracea (see Supplemental Table 6 online) as well as on chromosomes B1 and B6 in B. nigra (Panjabi et al., 2008) (see Supplemental Figure 2 online). Similar to B. rapa, the M/E association was not found in B. oleracea; however, one copy of M/E was rearranged with O/P/W/R on B2 in B. nigra (Panjabi et al., 2008) (see Supplemental Figure 2 online). These analyses suggest that the three Brassica genomes (A, B, and C) descended from a common hexaploid ancestor that originated from the merger of three tPCK-like ancestral genomes.

For radish, we used an EST linkage map (Shirasawa et al., 2011) to align EST sequences to A. thaliana genes. These results were then used to determine the position of 24 GBs on the nine radish chromosomes (see Supplemental Table 7 online). Most of the 24 GBs (75%) were present in three copies in the radish genome. The two tPCK-specific block associations [V/K/L/Wa/Q/X and O/P/W/R] were detected in triplicate on chromosomes LG2, LG4, and LG6, and LG5, LG6, and LG8, respectively. In addition, two copies of V/K/L/Wa/Q/X blocks were associated

Table 1. Comparison of GB Associations Shared between B. rapa and Three Ancestral Crucifer Genomes (ACK, PCK, and tPCK)

<table>
<thead>
<tr>
<th>Chr.</th>
<th>GB Assoc.</th>
<th>No.</th>
<th>B. rapa Chr.</th>
<th>Chr.</th>
<th>GB Assoc.</th>
<th>No.</th>
<th>B. rapa Chr.</th>
<th>Chr.</th>
<th>GB Assoc.</th>
<th>No.</th>
<th>B. rapa Chr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AK1</td>
<td>A/B</td>
<td>2</td>
<td>A06, A08</td>
<td>A06</td>
<td>A/B</td>
<td>2</td>
<td>A06, A08</td>
<td>A06</td>
<td>A/B</td>
<td>2</td>
<td>A06, A08</td>
</tr>
<tr>
<td>B/C</td>
<td>2</td>
<td>A05, A08</td>
<td>B/C</td>
<td>2</td>
<td>A05, A08</td>
<td>2</td>
<td>A05, A08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK2</td>
<td>D/E</td>
<td>0</td>
<td>–</td>
<td>D/E</td>
<td>0</td>
<td>–</td>
<td>AK2/5</td>
<td>N/M</td>
<td>3</td>
<td>A01, A03, A09</td>
<td></td>
</tr>
<tr>
<td>AK3</td>
<td>F/G</td>
<td>2</td>
<td>A03, A07</td>
<td>A03</td>
<td>F/G</td>
<td>2</td>
<td>A03, A07</td>
<td>A03</td>
<td>F/G</td>
<td>2</td>
<td>A03, A07</td>
</tr>
<tr>
<td>G/H</td>
<td>2</td>
<td>A03, A07</td>
<td>G/H</td>
<td>2</td>
<td>A03, A07</td>
<td>2</td>
<td>A03, A07</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK4</td>
<td>I/J</td>
<td>2</td>
<td>A03, A04</td>
<td>A03</td>
<td>I/J</td>
<td>2</td>
<td>A03, A04</td>
<td>A03</td>
<td>I/J</td>
<td>2</td>
<td>A03, A04</td>
</tr>
<tr>
<td>AK5</td>
<td>K/L</td>
<td>3</td>
<td>A02, A06, A09</td>
<td>A02</td>
<td>S/T</td>
<td>2</td>
<td>A04, A08</td>
<td>A04</td>
<td>S/T</td>
<td>2</td>
<td>A04, A08</td>
</tr>
<tr>
<td>AK6</td>
<td>O/P</td>
<td>3</td>
<td>A02, A03, A09</td>
<td>A02</td>
<td>P/W</td>
<td>1</td>
<td>A03</td>
<td>A03</td>
<td>P/W</td>
<td>1</td>
<td>A03</td>
</tr>
<tr>
<td>M/N</td>
<td>3</td>
<td>A01, A03, A09</td>
<td>M/N</td>
<td>3</td>
<td>A01, A03, A09</td>
<td>3</td>
<td>A01, A03, A09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK7</td>
<td>S/T</td>
<td>2</td>
<td>A04, A08</td>
<td>M/V</td>
<td>0</td>
<td>–</td>
<td>AK2/5/6/8</td>
<td>D/V</td>
<td>1</td>
<td>A09</td>
<td></td>
</tr>
<tr>
<td>AK8</td>
<td>V/W</td>
<td>0</td>
<td>–</td>
<td>V/W</td>
<td>3</td>
<td>A02, A06, A09</td>
<td>V/W</td>
<td>3</td>
<td>A02, A06, A09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W/X</td>
<td>0</td>
<td>–</td>
<td>Q/X</td>
<td>Q/X</td>
<td>3</td>
<td>A02, A03, A09</td>
<td>Q/X</td>
<td>3</td>
<td>A02, A03, A09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Ancestral chromosome (Chr.).
GB association.
Copy number of GB association in the B. rapa genome.
–, nonexistent GB association.
with D on chromosomes LG2 and LG6, and one copy of O/P/W/R was rearranged with M/E on LG5 (see Supplemental Table 7 online). This suggests that at least two tPCK-like genomes were involved in the origin of the radish genome.

In the diploid genome of *S. parvula*, tPCK-specific GB associations V/K/L/Wa/Q/X, O/P/W/R, and M/E were located on chromosomes SpChr2, SpChr6, and SpChr5, respectively (Dassanayake et al., 2011). tPCK-specific block associations were found to be conserved also in the *T. salsuginea* genome (Wu et al., 2012). Comparative genome mapping in white mustard (*S. alba*; Nelson et al., 2011) and *C. amplexicaulis* (Burrell et al., 2011) also revealed tPCK-specific arrangements of GBs. However, because of extensive genome rearrangements and the limited resolution of the linkage maps, no clear conclusions could be drawn for the two species. In *S. alba*, V/K/L/Wa/Q/X on chromosomes S10 and S11 and O/P/W/R on S09 were rearranged and not readily identifiable. The M/E association was identified on S02 (Nelson et al., 2011), but the D/V association was not found. In *Caulanthus*, the block association D/V/K/L/(Wa)/Q/X on LG3 was fragmented to such extent that it could not be clearly identified (i.e., blocks V, L, and Q were disrupted, while K remained undetected). The O/P/W/R association was observed on LG2 (with blocks P and R fragmented), whereas the M/E association was not present (Burrell et al., 2011).

### Table 2. The Inheritance of 10 *B. rapa* Centromeres from the Triplicated Ancestral tPCK-Like Genome (*n* = 21)

<table>
<thead>
<tr>
<th><em>B. rapa</em> Chromosome</th>
<th>Ancestral Chromosome</th>
<th><em>B. rapa</em> Subgenome</th>
</tr>
</thead>
<tbody>
<tr>
<td>A01</td>
<td>AK2/5/6/8</td>
<td>MF1</td>
</tr>
<tr>
<td>A02</td>
<td>AK6/8 or AK2/5</td>
<td>MF2 or MF1</td>
</tr>
<tr>
<td>A03</td>
<td>AK2/5/6/8</td>
<td>LF</td>
</tr>
<tr>
<td>A04</td>
<td>AK7</td>
<td>LF</td>
</tr>
<tr>
<td>A05</td>
<td>AK1</td>
<td>MF2</td>
</tr>
<tr>
<td>A06</td>
<td>AK2/5 or AK2/5</td>
<td>LF or MF2</td>
</tr>
<tr>
<td>A07</td>
<td>AK3</td>
<td>LF</td>
</tr>
<tr>
<td>A08</td>
<td>AK1 or AK7</td>
<td>MF1 or MF2</td>
</tr>
<tr>
<td>A09</td>
<td>AK2/5/6/8</td>
<td>MF2</td>
</tr>
<tr>
<td>A10</td>
<td>AK6/8</td>
<td>LF</td>
</tr>
</tbody>
</table>
Analysis of Synonymous Substitution Rates

Using the syntenic orthologs identified for B. rapa, A. thaliana, A. lyrata, S. parvula, and T. salsuginea, we calculated \( K_s \) values (the number of substitutions per synonymous site) for B. rapa relative to the other crucifer species. \( K_s \) for B. rapa relative to S. parvula was smaller (−0.29 to 0.31, −10 million years of divergence; Koch et al., 2000) than that relative to T. salsuginea (−0.34 to 0.36, −11.7 million years), A. lyrata (−0.41 to 0.45, −14.3 million years), and A. thaliana (−0.42 to 0.45, −14.5 million years) (Figure 4A). Based on these calculations, B. rapa is thus phylogenetically closer to S. parvula than to T. salsuginea or to the genus Arabidopsis. \( K_s \) values for each B. rapa subgenome relative to S. parvula ranged from 0.29 to 0.31 (Figure 4B), while those for the subgenomes relative to one another ranged from 0.29 to 0.33 (i.e., −10.3 million years) (Figure 4C).

To investigate the evolutionary relationship of A. thaliana, A. lyrata, S. parvula, T. salsuginea, and the three B. rapa subgenomes, we performed a phylogenetic analysis using shared synonymous single nucleotide polymorphisms for these five genomes and the two outgroup species papaya (Carica papaya) and grape (Vitis vinifera). From the 591 syntenic orthologs shared by the three B. rapa subgenomes and the six other species, 63,239 \( K_s \) loci (i.e., synonymous single nucleotide polymorphisms; see Methods) were identified and concatenated into a data set that was subjected to phylogenetic analysis. Based on the resulting tree, Brassicaceae and Arabidopsis ancestors diverged prior to the Brassicaceae-Thellungiella split. The Brassicaceae-specific whole-genome triplication occurred near the time of the split between the progenitors of the Brassicaceae and Schrenkiella (Figure 4D).

Reevaluation of Block Association V/K/L/Wa/Q/X and Interval Refinement of Seven GBs

The results of the multispecies comparison prompted a reevaluation of the V/K/L/Q/X block association considered specific for the \( x = 7 \) tribes of expanded lineage II (Mandáková and Lysak, 2008). This association was identified on chromosome Sp2 in S. parvula, albeit with a fragment of block V between L and Q (V/K/L/V/Q/X) (Dassanayake et al., 2011). The same rearrangement was found on B. rapa chromosomes A02, A06, and A09 (Figure 5A), B. oleracea (see Supplemental Table 6 online), and radish (Li et al., 2011; Shirasawa et al., 2011). We analyzed this chromosome region by performing an interspecies comparison between B. rapa, B. oleracea, S. parvula, A. thaliana, and A. lyrata, with the A. lyrata genome (Hu et al., 2011) serving as a proxy for the ACK genome (\( n = 8 \)). The presumed V’ region turned out to be part of block W, located on the bottom arm of the AK8 (A18) chromosome and adjacent to the centromere (Figure 5A). The origin of the V/K/L/Wa/Q/X rearrangement, which is specific to the ancestral PCK genome, is consistent with the previously proposed scenario (Mandáková and Lysak, 2008), except that one of the breakpoints is actually positioned within block W (Figure 5B; see Supplemental Figure 4A online).

An alternative scenario involves a paracentric inversion of blocks V/K/L followed by inactivation/removal of the AK8 paleocentromere (see Supplemental Figure 4B online). These results, together with multispecies comparisons of syntenic chromosomes, allowed us to redefine the boundaries of blocks V and W (Table 3, Figure 5A).

By comparing the boundaries of blocks H, I, Q, R, V, W, and X reported by Schranz et al. (2006) with corresponding regions and/or breakpoints within genomes of B. rapa, A. lyrata, S. parvula, and A. thaliana (see Supplemental Data Set 3 online), we revised intervals of the seven blocks of ACK (Table 3, Figure 5B).

DISCUSSION

We compared the whole-genome sequence of B. rapa (Wang et al., 2011) with genome sequences or genetic maps of other crucifer species, including A. thaliana, A. lyrata, B. oleracea, B. nigra, C. amplexicaulis, radish, S. alba, S. parvula, and T. salsuginea, as well as with previously proposed Brassicaceae ancestral genomes (Schranz et al., 2006; Mandáková and Lysak, 2008; Panjabi et al., 2008; Burrell et al., 2011; Dassanayake et al., 2011; Hu et al., 2011; Li et al., 2011; Nelson et al., 2011; Shirasawa et al., 2011; Wu et al., 2012). We provided conclusive evidence that the B. rapa genome arose via a whole-genome triplication event involving three \( n = 7 \) genomes structurally resembling the ancestral tPCK.

tPCK-Like Ancestral Genomes and the Origin of the Mesohexaploid Brassica Genome

Since diploid Brassica species were proposed to represent balanced secondary polyplody (Catcheside, 1934; Roebben, 1960), numerous lines of evidence have supported the occurrence of an ancestral whole-genome triplication prior to species radiation of the tribe Brassicaceae (Lagercrantz, 1998; Lysak et al., 2005; Parkin et al., 2005; Wang et al., 2011). In this study, we inferred the tPCK-like genome with seven chromosomes to be the basic diploid genome of the hexaploid ancestor of Brassicaceae. The hexaploid genome likely evolved through hybridization between a tetraploid and diploid progenitor genome. Evidence for this is based on genetic mapping (Lagercrantz, 1998; Babula et al., 2003), comparative cytogenetic analyses (Lysak et al., 2005; Ziołkowski et al., 2006), and, most convincingly, whole-genome sequence analysis. Wang et al. (2011) found that the three subgenomes in B. rapa have experienced different levels of gene loss. The more extensive fractionation of two subgenomes (MF1 and MF2) compared with the third subgenome (LF) suggest a two-step origin for the hexaploid genome. Several recent investigations (Cheng et al., 2012b; Tang and Lyons, 2012; Tang et al., 2012) convincingly support the origin of the Brassicaceae ancestral genome according to the equation \( 4x \times 2x \rightarrow 3x \rightarrow 6x \). Here, we provide evidence that all three participating ancestral genomes had seven chromosomes and that the hexaploid ancestor probably possessed 42 chromosomes (\( 2n = 6x = 42 \)).

tPCK in a Phylogenetic Context

Because species of the genus Brassica and closely related genera, such as Raphanus and Sinapis, can be intercrossed, and generic circumscriptions based on morphological characters are unreliable, congeneric species have frequently been
Figure 4. Pairwise Comparison of $K_s$ Values and a Phylogenetic Tree Based on $K_s$ Loci.

(A) The distribution of $K_s$ values between B. rapa and crucifer species T. salsuginea, S. parvula, A. thaliana, and A. lyrata.

(B) The distribution of $K_s$ values between each of the three B. rapa subgenomes and S. parvula.

(C) The distribution of $K_s$ values between pairs of the three B. rapa subgenomes.

(D) Phylogenetic tree for the four crucifer species and the three B. rapa subgenomes, with non-Brassicaceae species papaya and grape used as outgroups. The tree was constructed based on $K_s$ loci for all syntenic orthologs shared among the seven species. Numbers refer to bootstrap values.
assigned to two different sister phylogenetic clades within Brassiceae: Rapa/Oleracea and Nigra (Warwick and Black, 1991). To better understand the structure of Brassica progenitor genomes, we compared tPCK-specific block associations with chromosome structures of Rapa/Oleracea (B. oleracea, B. rapa, and radish) and Nigra (B. nigra) species. We found that all three Rapa/Oleracea species (B. oleracea, B. rapa, and radish) retained three copies of the V/K/L/Wa/Q/X and O/P/W/R associations. D/V or M/E associations were also found in the three species. In addition, chromosome synteny among B. nigra, B. oleracea, and B. rapa and the existence of the tPCK-specific associations D/V and M/E in B. nigra and D/V in B. oleracea and radish indicate that B. nigra, B. oleracea, and radish had the same origin as B. rapa. Furthermore, some tPCK-specific associations observed within the comparative linkage map of S. alba (Nelson et al., 2011), a species from the Nigra clade,
suggest that the genome of this species is also derived from a triplicated TPCK-like ancestral genome. The tribes Brassiceae, Isatideae, Thelypodieae, and Sisymbrieae were designated as lineage II of the Brassicaceae (Al-Shehbaz et al., 2006, Beilstein et al., 2006, 2008). However, the relationship of these tribes to the two other Brassicaceae lineages and to the remaining 45 tribes remains uncertain. Franzke et al. (2011) grouped lineage II with 18 other tribes to form an expanded lineage II. They demonstrated that tribes Calepinae, Coluteocarpeae, Conringieae, Eutremeae, Isatideae, Schrenkiella, Ochthodium, and Eutremeae, and probably also Thelypodieae (Caulanthus), descended from the same diploid TPCK-like ancestral genome. While the triplicated TPCK genome has undergone extensive repatterning in Brassicaceae, the ancestral structure of the TPCK genome has been conserved in diploid x = 7 taxa, such as Schrenkiella or Thellungiella, where no diploidization of a polyploid genome could occur.

### Descending Dysploidy in the Hexaploid Ancestor of Brassicaceae

The chromosome number of 2n = 42 of the hexaploid ancestor has been reduced by a factor of 2 during the evolution of Brassicaceae (Warwick and Al-Shehbaz, 2006). Such reductions are associated with the origin of composite chromosomes and loss of chromosome segments, such as centromeres, telomeres, and nucleolar organizing regions. In B. rapa, descending dysploidy led to the origin of 10 fusion chromosomes and the elimination of 11 paleocentromeres. Some paleocentromeres were eliminated through symmetric translocations resulting in products of unequal size: a large fusion chromosome and a minichromosome, which contains mainly a centromere and telomeres and usually gets lost during meiosis (Lysak et al., 2006; Schranz et al., 2006; Schubert and Lysak, 2011). Other ancestral chromosomes were combined by asymmetric end-to-end translocations (e.g., between AK6/8 and AK3 and in A03; Figure 2B). End-to-end translocation events must be accompanied by centromere loss or inactivation on one of the participating chromosomes to prevent unstable dicentric chromosomes. Centromeres were apparently lost from the collinear GBs F/G/H (AK3) of A03 and I/J (AK4) of A03 and A04 (Figure 2B). Nested chromosome fusions (i.e., insertion of one chromosome into the centromere of a recipient chromosome), which occurred frequently in the evolution of grass species (Luo et al., 2009; Salse et al., 2009; Murat et al., 2010), have not been detected in the karyotype of B. rapa. In Brassicaceae, single cases of such events apparently occurred in Pachycladon species (Mandáková et al., 2010) and in Homuncia alpina (Lysak et al., 2006).

We also analyzed whether recombination between homoeologous chromosome segments of the three B. rapa subgenomes exceeded recombination between nonhomoeologous segments. If recombination between homoeologous segments was prevalent, we would expect to find the same GBs located adjacent to one another (e.g., association A MF1/A MF2). In B. rapa, only four such associations were detected (S LF/S MF2 on A04, M LF/M MF3 on A06, E MF2/E LF on A07, and D MF1/D MF2 on A09). These data suggest that interchromosomal rearrangements were not mediated preferentially by recombination between homoeologous GBs. Nevertheless, the true picture may have been blurred by numerous rearrangements (typically inversions) that may have occurred later and reshuffled originally adjacent homoeologous GBs.

### Whole-Genome Sequence Assembly Improved the Resolution of Comparative Paleogenomics in Brassicaceae

The B. rapa genome sequence (Wang et al., 2011) provided much more detailed information on genome structure than previously...
constructed linkage maps. Although most GBs and parts thereof (68%) were detected by genetic mapping (Parkin et al., 2005; Panjabi et al., 2008; Kim et al., 2009; Trick et al., 2009), some blocks and block associations were missed or reported as being present in less than three copies. A comparison of the B. rapa genome sequence with previously reported linkage maps shows that block D on chromosome A01; blocks W, Q, X, M, and D on A03; S on A05; M, H, and U on A06; F, X, and S on A07; K, L, and M on A09; and C on A10 were not detected by genetic mapping (Parkin et al., 2005; Panjabi et al., 2008; Kim et al., 2009; Trick et al., 2009). Although the assembly of B. rapa chromosome A03 by Mun et al. (2010) identified 14 GBs or block segments, three additional blocks on the same chromosome were detected within the B. rapa reference genome (Wang et al., 2011). Similarly, the 2.4-Mb fragment of block W within the V/K/L/Wa/Q/X association was not detected by genetic mapping. Moreover, two adjacent copies of the same GB belonging to different subgenomes have often been regarded as a single block in previous genetic maps, including blocks S_{UL}^{MF2}/S_{MF2}^{MF2}, E_{UL}/E_{MF2}, and D_{MF2}/D_{MF2} on chromosomes A04, A07, and A09, respectively (Panjabi et al., 2008; Figure 2B). Furthermore, we found that earlier reports on GBs represented by more than three copies were mostly due to misinterpretation of block segments as entire blocks. For instance, Panjabi et al. (2008) observed four copies of block A located on chromosomes A06, A08, A09, and A10. However, analysis of B. rapa genome sequences revealed that the two copies of block A on chromosomes A06 and A10 were in fact two segments of this block (Figure 2B). These examples demonstrate that whole-genome sequence assemblies largely eliminate the risk of misinterpreting low-resolution genetic data.

**Refining Intervals of Ancestral GBs**

The original concept of 24 GBs (Schranz et al., 2006) resulted from the synthesis of comparative genetic and cytogenetic data for A. thaliana, A. lyrata, C. rubella, the A and C genomes of B. rapus, and other lineage I species (Bovin et al., 2004; Kuittinen et al., 2004; Parkin et al., 2005; Lysak et al., 2006). Because A. thaliana was the only crucifer species with a sequenced genome at that time, chromosome collinearity between all Brassicaceae species was inferred by mapping homoeologous genetic markers from A. thaliana and by chromosomal location of A. thaliana BAC clones and contigs. This approach was highly influenced by the density of genetic markers and the selection of chromosome-specific BAC contigs. Without sufficiently accurate information on cross-species collinearity, only approximate or minimal intervals could be defined for some GBs (Schranz et al., 2006) and GB associations. This is exemplified by the PCK/IPCK-specific chromosome rearrangement initially described as V/K/L/Q/X (Mandáková and Lysak, 2008) and V/K/L/V/Q/X (Dassanayake et al., 2011) and eventually refined as V/K/L/Wa/Q/X in this study. The extensive repatterning of ancestral GBs within the B. rapa genome, along with newly available sequence information for Brassicaceae genomes, allowed us to refine the intervals of seven GBs. This was particularly feasible when individual GBs were found to be relocated and formed a nonancestral GB association with breakpoints confirmed by syntenic positions in two or three B. rapa subgenomes. The interval refinement of the ancestral GB structure will improve the resolution of interspecies comparisons of genome collinearity across Brassicaceae.

**METHODS**

**Syntenic Ortholog Determination**

Syntenic orthologs between Brassica rapa, Arabidopsis thaliana, Arabidopsis lyrata, Schrenkiiella parvula, Thellungiella salsuginea, papaya (Carica papaya), and grape (Vitis vinifera) were identified using the tool SynOrths (Cheng et al., 2012a), with B. rapa as the query genome and each of the other species as the subject genome. SynOrths was run with default parameters (m = 20, n = 100, and r = 0.2). Under these settings, SynOrths first identifies homoeologous gene pairs between the two genomes using Basic Local Alignment Search Tool for Protein, defining gene pairs as those with an E-value more significant than $1 \times 10^{-20}$ or representing the best hits through the entire genome. The 20 closest genes (m = 20) flanking either side of the gene in the query B. rapa genome are then compared with the 100 closest genes (n = 100) flanking either side of the gene in the subject genome. If at least 20% (r = 0.2) of the best hits for the 40 genes (20 \times 2 for both sides) in the query B. rapa genome are found within the 200 genes (100 \times 2 for both sides) in the subject genome, the original pair of homoeologous genes are designated as a syntenic ortholog candidate. The increase in the number of flanking genes searched in the subject genomes was necessary to compensate for the fractionation that has reduced the gene content of each B. rapa subgenome.

**Comparison of Arrangements of GB Associations**

We searched the distribution of predefined GBs in the three genomes of the ACK, PCK, and IPCK for adjacent pairs of GBs (GB associations). These GB associations were compared with the extant arrangement of GBs within B. rapa. GB associations in ACK, PCK, and IPCK that were also immediately adjacent to each other in the B. rapa genome were counted and recorded.

**Reconstruction of Subgenomes Based on the IPCK**

As previously reported (Wang et al., 2011; Cheng et al., 2012b), the B. rapa genome is composed of three subgenomes each of the 24 GBs. We annotated these blocks along the seven chromosomes of the IPCK and reconstructed three IPCK-like subgenomes. At first we identified all continuous chromosome segments corresponding to GBs and belonging to the same chromosome of the IPCK. Then, chromosome segments were concatenated according to the following rules: (1) the reconstructed chromosome could not contain redundant copies of the same syntenic segment, and (2) the concatenated chromosome segments had to have a comparable gene density corresponding to the three B. rapa subgenomes (LF, MF1, and MF2).

**Ancient Centromere Detection**

Centromere-specific sequences, including CentBr, PCRBBr, and TR238, were obtained from previous studies (Koo et al., 2004, 2011; Lim et al., 2005, 2007). These sequences were aligned to the B. rapa genome using Nucmer (Kurtz et al., 2004) with parameters “-maxmatch -g 500 -c 16 -16” (i.e., find all matches with minimal exact match size of 16 bp regardless of uniqueness, merge exact matches distributed <500 bp apart into clusters, and report all clusters >16 bp). The nearest genes to the best Nucmer hits were determined and used as labels of genomic coordinates for the detected centromere traces (see Supplemental Data Set 2 online).
Identification of tPCK-Specific Block Associations in *B. oleracea*

Because a pseudomolecule-level assembly for the *B. oleracea* genome is not yet available, we used a different method to determine local syntenic relationships between *B. rapa* and *B. oleracea*. Thirty *B. rapa* genes located on either side of the boundaries between GBs within the GB associations O/P/W/R and D/V/K/L/Q/X were selected. These genes were compared using BLAST against *B. oleracea* genes via the Web interface provided by BRAD (http://brassicadb.org/brad/blastPage.php) (Cheng et al., 2011). Two blocks were considered as syntenic in *B. oleracea* when >60% of the *B. rapa* block boundary genes from each GB had BLAST hits (identity and coverage) on the same *B. oleracea* scaffold. For example, if (1) 60% of 30 *B. rapa* genes at the boundary of block O (near P) were homoeologous to genes on Scaffold000203 of *B. oleracea*, (2) 30 *B. rapa* genes at the block P boundary (near O) were also homoeologous to *B. oleracea* genes in Scaffold000203, and (3) these genes were adjacent to each other on Scaffold000203, then the O/P association was considered to be present also in the *B. oleracea* genome.

**Phylogenetic Analysis**

For phylogenetic analysis, the same MUSCLE alignments as for *K* s analysis were used. The genotypes of synonymous loci in 591 syntenic orthologs among *B. rapa*, *S. parvula*, and *T. salsuginea* were extracted from multiple alignments and concatenated into a single sequence for each species. Phylogenies were constructed, based on the character states of these 63,239 synonymous loci using the neighbor-joining method, and plotted under a linearized tree model as implemented in MEGA (Tamura et al., 2011).

**Accession Numbers**

Gene and protein sequences for *B. rapa* were obtained from BRAD (V1.2; http://brassicadb.org) (Cheng et al., 2011). Gene and genome data sets for *A. thaliana*, *A. lyrata*, *S. parvula*, and *T. salsuginea* were downloaded from The Arabidopsis Information Resource (TAIR9; http://www.Arabidopsis.org/index.jsp). Data sets for *A. lyrata* were downloaded from the Joint Genome Initiative database (Gene model 6; http://genome.jgi-psf.org/Araly1/Araly1.home.html) (Hu et al., 2011). *S. parvula* data sets (TpV7) were obtained from Dassanayake et al. (2011). *T. salsuginea* data were downloaded from the National Center for Biotechnology Information (accession number AH1U00000000) (Wu et al., 2012).

**Supplemental Data**

The following materials are available in the online version of this article.

**Supplemental Figure 1.** Purported Translocation Events Reshuffling tPCK-Specific Associations D/V and M/E in the *B. rapa* Genome.

**Supplemental Figure 2.** Chromosome Collinearity Comparison among *A. lyrata*, *B. rapa*, and *C. Genomes*, Revised from Panjabi et al. (2008).

**Supplemental Figure 3.** Reconstruction of Three tPCK-Like Ancestral Subgenomes in *B. rapa*.

**Supplemental Figure 4.** Two Alternative Origins of Translocation Chromosomes AK5/6/8 (PCK) and AK2/5/6/8 (tPCK).

**Supplemental Table 1.** Summary of Genomic Blocks Identified in *B. rapa*.

**Supplemental Table 2.** Genomic Blocks E, G, and T Showing Sequence Deletion in *B. rapa* (Br) as Compared with Arabidopsis thaliana (At).

**Supplemental Table 3.** Number of Genomic Blocks with Different Fragmental Status in *B. rapa* Subgenomes.

**Supplemental Table 4.** Frequency of tPCK-Specific Block Associations in Each of the Three *B. rapa* Subgenomes.

**Supplemental Table 5.** Number of Syntenic Genes Shared between Each *B. rapa* Subgenome and the Arabidopsis thaliana Genome (At).

**Supplemental Table 6.** The Frequency of tPCK-Specific Block Associations in *B. rapa* and *B. oleracea*.

**Supplemental Table 7.** List of Genomic Blocks Comprising the Nine Linkage Groups of Raphanus sativus (Shirasawa et al., 2011).

**Supplemental Data Set 1.** Syntenic Fragment Comparison between *B. rapa* and the Ancestral Crucifer Karyotype.

**Supplemental Data Set 2.** Centromere-Specific Sequences Detected in the *B. rapa* Genome.

**Supplemental Data Set 3.** Redefined Intervals of Seven Ancestral Genomic Blocks.

**ACKNOWLEDGMENTS**

We thank Michael Freeling, Isobel A. Parkin, James Schnable, and Ingo Schubert for their helpful suggestions concerning this work. This study was funded by the National Program on Key Basic Research Projects (The 973 Program: 2012CB113900 and 2013CB127000) and the National High Technology R&D Program of China (2012AA100101). Research was carried out in the Key Laboratory of Biology and Genetic Improvement of Horticultural Crops, Ministry of Agriculture, P.R. China. T.M. and M.A.L. were supported by a research grant from the Czech Science Foundation (Excellence Cluster P50112/G090) and by the European Regional Development Fund (CZ.1.05/1.1.00/02.0068).

**AUTHOR CONTRIBUTIONS**

X.W., F.C., and M.A.L. designed the research. F.C., T.M., and J.W. performed the research and analyzed the data. F.C., M.A.L., and Q.X. contributed new computational tools and data. F.C., M.A.L., and X.W. wrote the article.

Received February 6, 2013; revised March 23, 2013; accepted April 17, 2013; published May 7, 2013.

**REFERENCES**


