

# Subgenome-aware analyses suggest a reticulate allopolyploidization origin in three *Papaver* genomes

Received: 18 July 2022

Accepted: 5 April 2023

Published online: 19 April 2023

Check for updates

Ren-Gang Zhang<sup>1,2,12</sup>, Chaoxia Lu<sup>3,12</sup>, Guang-Yuan Li<sup>4,12</sup>, Jie Lv<sup>5</sup>, Longxin Wang<sup>6</sup>, Zhao-Xuan Wang<sup>7</sup>, Zhe Chen<sup>8</sup>, Dan Liu<sup>9</sup>, Ye Zhao<sup>10</sup>, Tian-Le Shi<sup>10</sup>, Wei Zhang<sup>4</sup>, Zhao-Hui Tang<sup>3</sup>, Jian-Feng Mao<sup>10</sup>, Yong-Peng Ma<sup>1</sup>✉, Kai-Hua Jia<sup>3</sup>✉ & Wei Zhao<sup>11</sup>✉

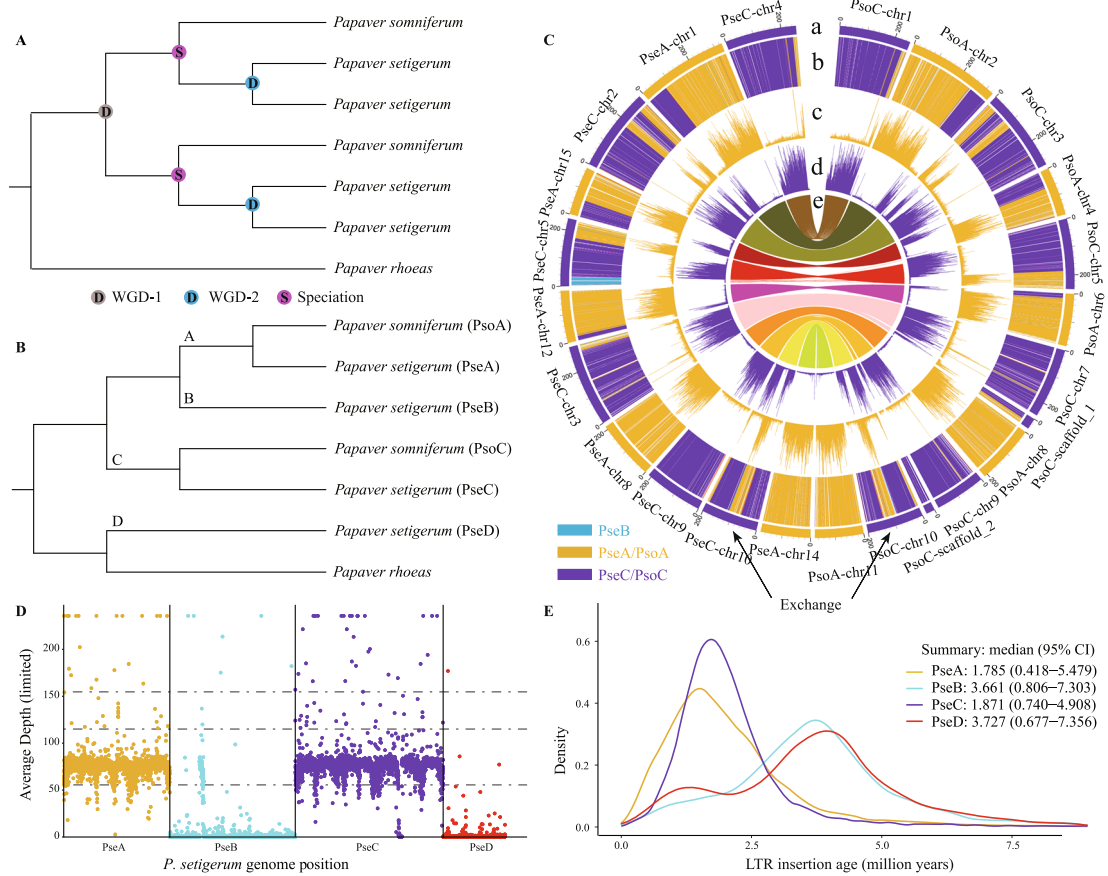
ARISING FROM X. Yang et al. *Nature Communications* <https://doi.org/10.1038/s41467-021-26330-8> (2021)

Hybridization and polyploidization are important driving forces in angiosperm evolution, resulting in novel phenotypes capable of prompting ecological diversification and invasion of new niches<sup>1</sup>. The genus *Papaver* (Papaveraceae) contains many taxa used in the pharmaceutical and culinary industries and as ornamental plants<sup>2</sup>. Yang et al. assembled de novo two chromosome-level genomes of *P. rhoeas* (common poppy,  $2n = 14$ ) and *P. setigerum* (Troy poppy,  $2n = 44$ ), and improved the *P. somniferum* genome assembly (opium poppy,  $2n = 22$ )<sup>3</sup>. These high-quality, chromosome-scale genome assemblies represent a valuable resource for studying the early evolutionary history of eudicots and the evolution of morphinan biosynthesis. Based on synteny and phylogenomic analyses, the authors identified two rounds of whole-genome duplication (WGD), one in the ancestor to *P. setigerum* and *P. somniferum* (WGD-1) at -7.2 million years ago (MYA), and one lineage-specific WGD-2 in *P. setigerum* at -4.0 MYA. In the absence of effective subgenome-phasing techniques, they proposed complex models to explain the extensive genome reorganization and gene family evolution built upon the duplication of the genome itself (their Figs. 2–4 and Supplementary Figs 27–34). Leveraging the recent developments in subgenome-phasing method published by Jia et al.<sup>4</sup>, we propose an alternative model, i.e., reticulate allopolyploidization, to account for the evolution and the genomic diversity of these three

*Papaver* species. Our hypothesis is supported by the following lines of evidence:

1. We extracted 4,791 anchor genes from the inter-genomic syntenic blocks at a ratio of 1:2:4 in *P. rhoeas*, *P. somniferum* and *P. setigerum* using OrthoFinder v2.3.1<sup>5</sup> and MCScanX<sup>6</sup> (Supplementary Figs. 1A, 2–3). According to the WGD model proposed by Yang et al.<sup>3</sup>, *P. setigerum* should have two sister-pairs of homoeologous subgenomes appearing as sisters to the subgenomes of *P. somniferum* (Fig. 1A). We inferred the maximum likelihood (ML) trees for each gene and the concatenated sequences of all genes in the same homoeologous chromosome sets (macro-synteny) using IQ-TREE v1.6.12<sup>7</sup>, with *P. rhoeas* as the outgroup. The top six gene tree topologies, supported by 4,231 (88%) of the 4,791 gene trees (Supplementary Fig. 4), show that orthologous gene pairs from *P. somniferum* and *P. setigerum* group together, and are sister to the homoeologous genes from *P. setigerum*. None of the topologies comprising at least 50 gene trees (Supplementary Fig. 4) agree with the WGD model shown in Fig. 1A, and most gene trees (43% of the 4791) support the hypothesis that *P. somniferum* and *P. setigerum* were derived from a reticulate origin (Supplementary Fig. 4; Fig. 1B). In addition, we obtained 15 groups of concatenated gene trees (macro-synteny

<sup>1</sup>Yunnan Key Laboratory for Integrative Conservation of Plant Species with Extremely Small Populations / Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201 Yunnan, China. <sup>2</sup>University of the Chinese Academy of Sciences, 100049 Beijing, China. <sup>3</sup>Key Laboratory of Crop Genetic Improvement & Ecology and Physiology, Institute of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences, Jinan 250100 Shandong, China. <sup>4</sup>Department of Bioinformatics, Ori (Shandong) Gene Science and Technology Co., Ltd., Weifang 261322 Shandong, China. <sup>5</sup>College of Life Science and Technology, Beijing University of Chemical Technology, 100029 Beijing, China. <sup>6</sup>School of Biological Science and Technology, University of Jinan, Jinan 250022 Shandong, China. <sup>7</sup>Shijiazhuang People's Medical College, Shijiazhuang 050091 Hebei, China. <sup>8</sup>InvoGenomics Biotechnology Co., Ltd., Jinan 250109 Shandong, China. <sup>9</sup>Shandong Provincial Center of Forest and Grass Germplasm Resources, Jinan 250102 Shandong, China. <sup>10</sup>National Engineering Research Center of Tree Breeding and Ecological Restoration, State Key Laboratory of Tree Genetics and Breeding, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, College of Biological Sciences and Technology, Beijing Forestry University, 100083 Beijing, China. <sup>11</sup>Department of Ecology and Environmental Science, Umeå University, SE-901 87 Umeå, Sweden. <sup>12</sup>These authors contributed equally: Ren-Gang Zhang, Chaoxia Lu, Guang-Yuan Li. ✉e-mail: [mayongpeng@mail.kib.ac.cn](mailto:mayongpeng@mail.kib.ac.cn); [kaihuajia\\_saas@163.com](mailto:kaihuajia_saas@163.com); [zhao.wei@umu.se](mailto:zhao.wei@umu.se)



**Fig. 1 | The origin and evolution of the subgenomes in the three studied *Papaver* species.** **A** Phylogenomic relationships among the subgenomes assuming the whole-genome duplication (WGD) model of Yang et al.<sup>3</sup>. **B** Tree topology recovered by gene trees, macro-synteny trees, and species/subgenome trees (see Supplementary Figs. 4, 5, 8 for details). The four subgenomes of *P. setigerum* are designated PseA, PseB, PseC and PseD; the two subgenomes of *P. somniferum* are designated PsoA and PsoC, and their ancestors are designated A–D. **C** Circos plot of subgenome partitions of *P. somniferum* and *P. setigerum* genomes (more details in Supplementary Figs. 6, 7) indicates that PseA and PsoA, and PseC and PsoC share almost all subgenomic exchanges except a segment in PseC-chr5 that shows exchange with PseB. (a) Subgenome assignment of chromosomes based on the *k*-means algorithm. (b) Significant enrichment of subgenome-specific *k*-mers

(subgenome partitions). Partitions with the same color as that of a subgenome indicates significant enrichment of *k*-mers specific to that subgenome. The white areas are not significantly enriched. (c–d) Count (absolute) of each subgenome-specific *k*-mer set. (e) Homoeologous blocks between the two species. All statistics (b–d) were computed in sliding windows of 1 Mb. Exchanges between subgenomes, such as that in the middle regions of PseC-chr10 and PsoC-chr10, are inferred from inconsistencies between subgenome assignments calculated using chromosomes (ring a) and windows (rings b–d). **D** The mapping depth of Illumina sequencing reads from *P. somniferum* to *P. setigerum* subgenomes. **E** Insertion times of subgenome-specific LTR-RTs. The 95% confidence intervals (CI) of the insertion times are used to infer the time boundary of divergence to hybridization period.

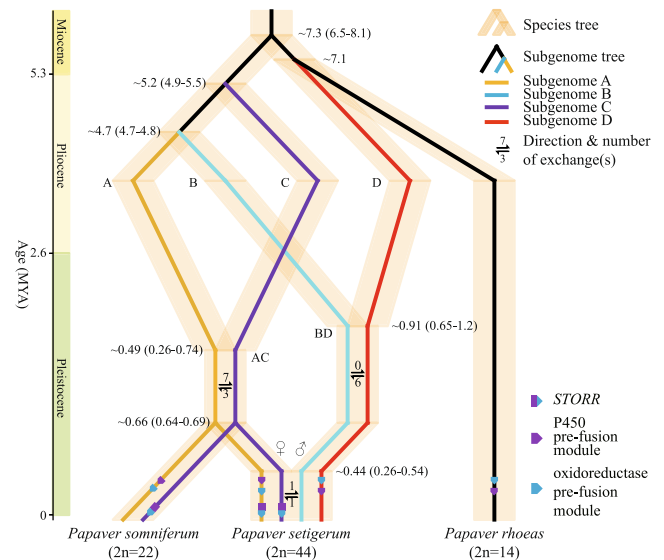
- trees) with at least 100 syntenic genes, and the topologies of these macro-synteny trees are identical to the most gene trees (Supplementary Fig. 5), which further supports the model presented in Fig. 1B rather than that in Fig. 1A.
- We phased the subgenomes of *P. somniferum* and *P. setigerum* using SubPhaser v1.2<sup>4</sup> (Supplementary Fig. 1B, C, Supplementary Figs. 6, 7), and extracted orthogroups across 23 species/subgenomes, including two subgenomes of *P. somniferum*, four subgenomes of *P. setigerum*, *P. rhoeas*, and representative lineages of other angiosperms, using OrthoFinder (Supplementary Fig. 8). We then inferred species/subgenome trees using the ML and coalescence-based methods (Supplementary Fig. 8). The topologies of these subgenome trees were consistent with those of the most gene trees (Supplementary Figs. 4, 5), which support the model presented in Fig. 1B. We named the four subgenomes of *P. setigerum* as PseA, PseB, PseC and PseD, and the two subgenomes of *P. somniferum* as PsoA and PsoC according to their phylogenetic relationships (Fig. 1B). Our data suggest that PseA and PsoA, and PseC and PsoC, are derived from separate common ancestors (designated A and C) (Fig. 1B).

- We identified exchanges between homoeologous subgenomes in *P. somniferum* and *P. setigerum* using SubPhaser (Supplementary Figs. 6, 7; Supplementary Tables 1, 2). We found that the pattern of exchanges on each chromosome between PsoA and PsoC is almost identical to that between PseA and PseC (except for a single exchange between PseB and PseC; Fig. 1C, Supplementary Figs. 6, 7). We then mapped the Illumina sequencing reads from *P. somniferum* to the *P. setigerum* subgenomes (Supplementary Fig. 1B) using sppIDer<sup>8</sup>. The coverage depth plot showed that almost all the *P. somniferum* reads mapped to PseA and PseC, and very few reads mapped to PseB (i.e. the region exchanged between PseB and PseC) (Fig. 1D). Syntenic dot plots between the subgenomes showed that PsoA and PsoC had greater similarity (lower *K*s) with PseA and PseC, respectively, but higher *K*s with PseB and PseD (Supplementary Fig. 2). These results strongly suggest that *P. somniferum* and the two subgenomes PseA and

PseC of *P. setigerum* were derived from a common allotetraploid ancestor (designated AC). This suggestion agrees with previous cytological evidence that hybrids between *P. somniferum* ( $2n = 22$ ) and *P. setigerum* ( $2n = 44$ ) had around 11 bivalents (mean 10.711 + 11.61) at metaphase I<sup>9</sup>.

4. There are two possible processes that could lead to the genomic pattern observed in *P. setigerum*: (i) AC hybridized with the ancestors of PseB and PseD separately in a stepwise process; or (ii) the ancestors of PseB and PseD hybridized, forming an allotetraploid (designated BD), then BD hybridized with AC forming the allooctoploid progenitor of *P. setigerum*. To test these two scenarios, we first removed all the potential exchanges between subgenomes of *P. setigerum*, and identified the subgenome-specific long terminal repeat retrotransposons (LTR-RTs) using SubPhaser (Supplementary Fig. 1D). Then we estimated the insertion times of subgenome-specific LTR-RTs in *P. setigerum* to represent the time boundaries from subgenomes differentiation to allohybridization. The estimated PseA- and PseC-specific LTR-RTs insertion times were similar, ranging from -5 to -0.5 MYA (95% confidence interval; Fig. 1E). Similarly, the PseB- and PseD-specific LTR-RTs insertion times were also similar (ranging from -7.3 to -0.7 MYA) but distinct from those of PseA and PseC (Fig. 1E), suggesting that PseB and PseD were more likely to have been introduced into the *P. setigerum* genome at the same time. Thus, we favored the second scenario, i.e. that the ancestors of PseB and PseD formed an allotetraploid BD, then BD hybridized with AC forming the allooctoploid progenitor of *P. setigerum*.
5. To test whether other potential progenitors were involved in the evolution of these three species, we downloaded all the available sequencing data of *Papaver* species from public databases (see Data Availability for details), and assembled the genes of each species using the HybPiper pipeline<sup>10</sup>. We then extracted 1,474 single-copy genes, and inferred a species tree using ASTRAL-MP v5.14.5<sup>11</sup>. The results suggested that subgenome PseD, *P. rhoeas* and *P. dubium* originated from a common ancestor (Supplementary Fig. 9). Similar to *P. rhoeas*<sup>3</sup>, *P. dubium* showed no evidence of recent WGD (Supplementary Fig. 10), suggesting it could not be a direct tetraploid progenitor (BD) of *P. setigerum*. We did not find closely related species for the subgenomes A, B and C, suggesting either the extinction of related ancestors or a sampling gap in taxon coverage. The tree inferred from the whole chloroplast genomes further suggested that *P. somniferum* was the most likely direct maternal progenitor of *P. setigerum* (Supplementary Fig. 11). Patterns of genome organization in *P. setigerum* and *P. somniferum* suggest that post-polyploidization diploidization is probably still ongoing within the two species as there was no largely biased gene fractionation observed in the subgenomes (Supplementary Fig. 12).

In summary, our comprehensive set of analyses confirmed the two rounds of WGDs previously documented<sup>3</sup>, but we uncovered a reticulate allopolyploidization scenario of evolution in the three studied *Papaver* species (Fig. 2A), involving four ancient diploid genomes (i.e. A, B, C, D) and two tetraploid genomes (i.e. AC and BD). Their most recent common ancestor (MRCA) first diverged into A, B, C, D and *P. rhoeas* at -4.7–7.3 MYA. B and D then hybridized, resulting in the allotetraploid BD at -0.91 MYA. The hybridization between A and C occurred -0.74–0.26 MYA, resulting in the allotetraploid AC, which led to the formation of *P. somniferum* -0.66 MYA. AC and BD hybridized, resulting in *P. setigerum* at -0.44 MYA. Genetic exchange between PseB and PseC occurred later. On-going post-polyploidization diploidization resulted in the genome structure we observe in present-day species. However, accurately reconstruction of the genome rearrangements during the allopolyploidization and re-diploidization history remains a challenge with our current methodologies and would



**Fig. 2 | The reticulate allopolyploidization model in the three *Papaver* species and the subgenomic locations of *STORR* and its pre-fusion loci.** The times of species/subgenomes divergence and hybridization shown on the topology were based on median *Ks* values and the subgenome-specific LTR-RT insertion times, respectively, see Supplementary Tables 3, 4 for details. The currently retained copies of *STORR* and its pre-fusion modules are sorted to subgenomes (see more details in Supplementary Figs. 13–15). Purple and blue denote the cytochrome P450 and oxidoreductase modules, respectively, to correspond with the scheme of Fig. 3a in Yang et al.<sup>3</sup>.

require further investigation. Our revision of the speciation and genome evolution model from Yang et al.<sup>3</sup> has implications for understanding not only the role of reticulation in *Papaver* diversification, but also the evolution of the morphinan and noscapine biosynthesis pathways. Under our genome evolution model, the *STORR* gene fusion event is therefore most likely to have occurred in the ancestor of A and C, or even earlier in the MRCA of this species complex, and was brought to the genomes of *P. somniferum* and *P. setigerum* via hybridization (Fig. 2, Supplementary Figs. 13–15; detailed explanations in Supplementary Note 1), rather than through a post-WGD-1 fusion-translocation event and then duplication following WGD-2, as proposed by Yang et al.<sup>3</sup>. Our model is consistent with a recent study which shows that the *STORR* gene fusion event occurred only once, taking place between 16.8–24.1 MYA, prior to the speciation of this species complex<sup>12</sup>.

## Methods

### Reconstruction of gene and macro-synteny trees for the three studied *Papaver* species

Syntenic blocks within the three *Papaver* species were identified with OrthoFinder v2.3.1<sup>5</sup>. Orthologous and paralogous relationships, as well as orthogroups, were inferred using the parameters “-M msa -T fastree” based on proteome sequences from multiple species. The resulting gene pairs were used to call collinear/syntenic blocks using MCScanX (parameters: -a -b 0 -c 0)<sup>6</sup>. For syntenic homologous gene pairs, *Ks* was calculated using the ParaAT pipeline<sup>13</sup> (Supplementary Figs. 2–3). Briefly, the protein sequences of each gene pair were first aligned in MUSCLE v3.8.425<sup>14</sup>, and the alignment was then converted to a codon alignment using PAL2NAL v14<sup>15</sup>. The *Ks* was finally calculated using KaKs\_Calculator v2.0<sup>16</sup> with the YN model<sup>17</sup>.

We then extracted 4,791 anchor genes from the inter-genomic syntenic blocks in *P. rhoeas*, *P. somniferum* and *P. setigerum* with a ratio of exactly 1:2:4. To reconstruct the gene trees, the homoeologous gene sequences were aligned with MAFFT v7.481<sup>18</sup> and trimmed with trimAl



v1.2<sup>19</sup> using a heuristic selection optimized for maximum likelihood (ML) phylogenetic tree reconstruction. Then, the ML tree (Supplementary Fig. 4) was inferred using IQ-TREE v1.6.12<sup>7</sup> with 1000 bootstraps<sup>20</sup>. The 1:2:4 genes located on the same chromosome set were considered as macro-synteny, and two methods were used to infer the macro-synteny trees (Supplementary Fig. 5): the ML and the coalescence-based method. For the ML method, the gene alignments generated earlier were concatenated, and a tree was reconstructed using IQ-TREE<sup>7</sup> with 1000 bootstraps<sup>20</sup>. For the coalescence-based method, the gene trees were input into ASTRAL (MP-5.14.5)<sup>11</sup> to infer the tree based on coalescence.

### Phasing the subgenomes of *P. somniferum* and *P. setigerum*, and inference of species/subgenome trees

We used SubPhaser (parameters: -q 150 -exclude\_exchanges)<sup>4</sup> to phase and partition the subgenomes of *P. somniferum* and *P. setigerum* (Supplementary Figs. 6–7). In brief, chromosomes of a neoallopolyploid were assigned to subgenomes based on differential repetitive *k*-mers that were assumed to have expanded during the period of independent evolution after divergence from the nearest common ancestor and before the hybridization of the parental progenitors (the so-called divergence–hybridization period). A subgenome is considered to be well-phased when it displays distinct patterns of both differential *k*-mers and homoeologous chromosomes, confirming the presence of subgenome-specific features, as expected.

We considered each subgenome as an independent pseudo-species for the subsequent phylogenomic analyses. We additionally collected genomic data from 15 other taxa in the Ranunculales and other angiosperm lineages, as well as RNA-Seq data from *P. bracteatum*, from published papers and public databases (see Data Availability for details). The transcriptome of *P. bracteatum* was first de novo assembled using Trinity v2.6.6<sup>21</sup> and the coding region of each transcript was annotated using TransDecoder v5.2.0 (<https://github.com/TransDecoder/TransDecoder/>). Only transcripts with the longest coding region for each gene were retained, and only genes with complete coding regions were used for downstream analyses. We inferred orthogroups from these data using OrthoFinder v2.3.1<sup>5</sup> as described above. Finally, we inferred the species/subgenome trees (Supplementary Fig. 8) using both ML and coalescence-based methods, as described earlier.

### Identification and validation of exchanges between subgenomes

The identification of exchanges between subgenomes (Supplementary Figs. 6–7, Fig. 1C) was carried out using SubPhaser<sup>4</sup> in a semi-automated process. SubPhaser assigned each 1 Mb genomic window to subgenomes and flagged windows with enrichments that did not match the subgenome assignments of their chromosome as potential exchanges. These were further checked manually to determine whether they were bona fide exchanges or not. For example, in the middle of C-Pse-chr10 (Supplementary Fig. 6D), subgenome A-specific *k*-mers showed continuous significant enrichments (2nd circle from outer to inner circles), and the abundance of subgenome A-specific *k*-mers was comparable to those on subgenome A chromosomes (4th circle) which contrasted with the other subgenomes (5–7th circles). Based on these observations, we confidently concluded that an exchange had occurred.

After manually checking, we excluded short exchanges with lengths of less than 5 Mb (Supplementary Tables 1–2). As unbalanced exchanges were expected to have syntenic blocks within subgenomes, we validated them through syntenic analyses (Supplementary Figs. 2–3). For example, we observed an exchange where the segment at the 5' end of PseC-15 had been exchanged to the 3' end of PseB-17, resulting in a large syntenic block between PseB-17-3' and naive PsoB-5-5' (Supplementary Fig. 2). Due to this imbalance, subgenome PseB now has two copies of this homoeologous segment, leading to a large

syntenic block between PseB-17-3' and PseB-13-3' where the PseB-naive homoeologous segment is located (Supplementary Fig. 3).

### Comparison of genomic composition of *P. somniferum* and *P. setigerum*

We used the sppIDer<sup>8</sup> pipeline to confirm the genomic composition of *P. somniferum* and *P. setigerum* (Fig. 1D). This involved mapping short-read sequencing data from *P. somniferum* to the genome of *P. setigerum* to assess the genomic contribution and relative ploidy of each of the subgenomes.

### Identification of potential progenitor species with other *Papaver* species

To identify potential progenitor species, all available sequencing data (i.e. the genome skimming data) from *Papaver* species were downloaded from public databases (see Data Availability for details). The genome skimming data were assembled using the HybPiper pipeline<sup>10</sup>, where the short sequencing reads were mapped to each homologous gene group using BWA-MEM v0.7.17<sup>22</sup>, and assembled with SPAdes v3.13.1<sup>23</sup>. The coding regions were then annotated with exonerate v2.2.0<sup>24</sup>. A total of 1474 single-copy genes were extracted, and a species tree (Supplementary Fig. 9) was inferred using methods described above.

*P. dubium* has the potential to be the tetraploid progenitor (BD) of *P. setigerum*. To explore this possibility further, we downloaded the transcript sequences of *P. dubium* subsp. *lecoqii* from a recent study<sup>12</sup> and annotated the coding regions using TransDecoder v5.2.0. Next, we inferred orthogroups with OrthoFinder v2.3.1<sup>5</sup> and calculated *Ks* for both orthologous and paralogous gene pairs using the ParaAT pipeline<sup>13</sup>. Using the *Ks*-based method, we inferred potential recent WGD events in *P. dubium* subsp. *lecoqii* (Supplementary Fig. 10).

### Identification of potential maternal parent using the chloroplast tree

To determine the potential maternal parent of *P. somniferum* and *P. setigerum*, we assembled chloroplast genomes from short-read sequencing data of *Papaver* species using GetOrganelle v1.6.2e (parameters: -w 115 -R 13)<sup>25</sup>. The assembled genomes were then annotated using the OGAP pipeline (<https://github.com/zhangrengang/ogap>). Based on whole plastome sequences of *Papaver* and related taxa, a phylogenomic tree (Supplementary Fig. 11) was inferred using IQ-TREE<sup>7</sup> with 1000 bootstraps<sup>20</sup>.

### Estimation of divergence and hybridization times

The timing of species/subgenome divergence and hybridization (Supplementary Tables 3–4, Supplementary Figs. 6–7, Fig. 1E) were estimated with two methods: the LTR-based method and the *Ks*-based method. Subgenome-specific long terminal repeat retrotransposons (LTR-RTs) are expected to undergo a burst of activity during the divergence–hybridization period. We employed SubPhaser, which uses subgenome-specific LTR-RTs to estimate the upper and lower boundaries of the divergence–hybridization period by applying a symmetric 95% percentile-based confidence interval to the subgenome-specific LTR insertion ages. The analysis excluded any potential exchanged LTR-RTs. Due to the large uncertainty in time estimation using LTR-RTs (particularly for the divergence time)<sup>4</sup>, a traditional *Ks*-based method<sup>3</sup> was also used to estimate the divergence time independently, based on a divergence time of *P. somniferum*–*P. rhoeas* (7.7 MYA<sup>3</sup>). The estimated times were calculated using the formula 1:

$$T = Ks/Ks(P.somniferum-P.rhoeas)*7.7 MYA \quad (1)$$

assuming an equal substitution rate per year.

## Assignment of subgenome and building gene phylogenies for *STORR*-related loci

Subgenomes for *STORR*-related loci (Supplementary Figs. 13–15, Fig. 2) were determined by their locations on subgenome segments (Supplementary Tables 1–2) using bedtools v2.27.1<sup>26</sup>. Gene trees (Supplementary Figs. 13–14) were reconstructed using IQ-TREE<sup>7</sup> with 1000 bootstraps<sup>20</sup>.

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The chloroplast genome sequences assembled in this study have been deposited in the GenBank database under the accession codes [OM174280–OM174296](https://doi.org/10.1038/s41467-023-37939-2). Assemblies of transcriptome and genome skimming data of *Papaver* generated in this study are available at figshare [<https://doi.org/10.6084/m9.figshare.20323995.v1>]. Genome assemblies of *P. rhoeas*, *P. somniferum* and *P. setigerum* were downloaded from the National Genomics Data Center (NGDC) Genome Warehouse (GWH) database under the BioProject accession [PRJCA004217](https://doi.org/10.1038/s41467-023-37939-2). Gene annotations of *P. rhoeas*, *P. somniferum* and *P. setigerum* were downloaded from GitHub [<https://github.com/xjtumomics/Papaver-Genomics/>]. Raw genome sequencing reads of *P. rhoeas*, *P. somniferum* and *P. setigerum* were downloaded from the NCBI Sequence Read Archive (SRA) database under the BioProject accession [PRJNA720042](https://doi.org/10.1038/s41467-023-37939-2). Gene annotations of *Macleaya cordata*, *Kingdonia uniflora*, *Tetracentron sinense*, *Coptis chinensis*, and *Prunus persica* were downloaded from the NCBI GenBank/RefSeq databases under the accessions [GCA\\_002174775.1](https://doi.org/10.1038/s41467-023-37939-2), [GCA\\_014058105.1](https://doi.org/10.1038/s41467-023-37939-2), [GCA\\_015143295.1](https://doi.org/10.1038/s41467-023-37939-2), [GCA\\_015680905.1](https://doi.org/10.1038/s41467-023-37939-2) and [GCF\\_000346465.2](https://doi.org/10.1038/s41467-023-37939-2), respectively. Gene annotations of *Vitis vinifera* v2.1 [[https://phytozome-next.jgi.doe.gov/info/Vvinifera\\_v2\\_1](https://phytozome-next.jgi.doe.gov/info/Vvinifera_v2_1)] and *Aquilegia coerulea* v3.1 [[https://phytozome-next.jgi.doe.gov/info/Acoerulea\\_v3\\_1](https://phytozome-next.jgi.doe.gov/info/Acoerulea_v3_1)] were taken from the Phytozome database. Gene annotations of *Macadamia integrifolia* were downloaded from the NGDC GWH database under the accession [GWHBAUK00000000.1](https://doi.org/10.1038/s41467-023-37939-2) [<https://ngdc.cnpc.ac.cn/gwh/Assembly/23196/show>]. Gene annotations of *Amborella trichopoda* v6.1 were downloaded from the CoGe database [<https://genomevolution.org/coge/GenomelInfo.pl?gid=50948>]. Gene annotations of *Trochodendron aralioides* were taken from the GigaDB database [<http://gigadb.org/dataset/100657>]. Gene annotation of *Coffea canephora* were downloaded from the Coffee Genome Hub [<https://coffee-genome-hub.southgreen.fr/node/1/2>]. Gene annotations of *Nelumbo nucifera* China Antique v2.0 were downloaded from the *Nelumbo* Genome Database [<http://nelumbo.biocloud.net/page/download/download>]. Gene annotations of *Eschscholzia californica* v1.0 were from the *Eschscholzia* Genome DataBase [[https://drive.google.com/drive/folders/1MIUdVBRBvalizy75JV19uh9afd\\_SYXLo](https://drive.google.com/drive/folders/1MIUdVBRBvalizy75JV19uh9afd_SYXLo)]. Gene annotations of *Aquilegia oxyssepala* [<https://doi.org/10.1038/s41438-020-0328-y>] and *Akebia trifoliata* [<https://doi.org/10.1038/s41438-020-00458-y>] were obtained from the corresponding authors. Raw genome skimming sequencing data of 16 *Papaver* species were downloaded from the NCBI SRA database under the BioProject accession [PRJEB43865](https://doi.org/10.1038/s41467-023-37939-2). Raw transcriptome sequencing data of *P. bracteatum* were downloaded from the NCBI SRA database under the BioProject accession [PRJEB21674](https://doi.org/10.1038/s41467-023-37939-2). A transcriptome shotgun assembly of *P. dubium* subsp. *lecoqii* was downloaded from the NCBI GenBank database under the accession [GJOS00000000.1](https://doi.org/10.1038/s41467-023-37939-2). Chloroplast genome sequences of *Papaver* and related taxa were downloaded from the NCBI GenBank/RefSeq databases with accessions [MK820043.1](https://doi.org/10.1038/s41467-023-37939-2), [NC\\_029434.1](https://doi.org/10.1038/s41467-023-37939-2), [NC\\_037831.1](https://doi.org/10.1038/s41467-023-37939-2), [NC\\_037832.1](https://doi.org/10.1038/s41467-023-37939-2), [MW411801.1](https://doi.org/10.1038/s41467-023-37939-2), [OK349678.1](https://doi.org/10.1038/s41467-023-37939-2), [MK533647.1](https://doi.org/10.1038/s41467-023-37939-2), [NC\\_050878.1](https://doi.org/10.1038/s41467-023-37939-2), [NC\\_056996.1](https://doi.org/10.1038/s41467-023-37939-2), [NC\\_050877.1](https://doi.org/10.1038/s41467-023-37939-2), [NC\\_056967.1](https://doi.org/10.1038/s41467-023-37939-2), [NC\\_039625.1](https://doi.org/10.1038/s41467-023-37939-2), [MK281585.1](https://doi.org/10.1038/s41467-023-37939-2), [NC\\_039623.1](https://doi.org/10.1038/s41467-023-37939-2), and [NC\\_029427.1](https://doi.org/10.1038/s41467-023-37939-2). Source data are provided with this paper.

## Code availability

The codes used for phasing subgenomes can be found at GitHub [<https://github.com/zhangrengang/SubPhaser>]<sup>27</sup>.

## References

1. Otto, S. P. & Whitton, J. Polyploid incidence and evolution. *Annu. Rev. Genet.* **34**, 401–437 (2000).
2. Butnariu, M. et al. *Papaver* plants: current insights on phytochemical and nutritional composition along with biotechnological applications. *Oxid. Med. Cell. Longev.* **2022**, 2041769 (2022).
3. Yang, X. et al. Three chromosome-scale *Papaver* genomes reveal punctuated patchwork evolution of the morphinan and noscapine biosynthesis pathway. *Nat. Commun.* **12**, 6030 (2021).
4. Jia, K. H. et al. SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. *New Phytol.* **235**, 801–809 (2022).
5. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
6. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
7. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
8. Langdon, Q. K., Peris, D., Kyle, B. & Hittinger, C. T. sppDer: a species identification tool to investigate hybrid genomes with high-throughput sequencing. *Mol. Biol. Evol.* **35**, 2835–2849 (2018).
9. Malik, C. P., Mary, T. N. & Grover, I. S. Cytogenetic studies in *Papaver* V. Cytogenetic studies on *P. somniferum* × *P. setigerum* hybrids and amphiploids. *Cytologia* **44**, 59–69 (1979).
10. Johnson, M. G. et al. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* **4**, 1600016 (2016).
11. Yin, J., Zhang, C. & Mirarab, S. ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics* **35**, 3961–3969 (2019).
12. Catania, T. et al. A functionally conserved *STORR* gene fusion in *Papaver* species that diverged 16.8 million years ago. *Nat. Commun.* **13**, 3150 (2022).
13. Zhang, Z. et al. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* **419**, 779–781 (2012).
14. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
15. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
16. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs\_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genom. Proteom. Bioinf.* **8**, 77–80 (2010).
17. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
18. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
19. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
20. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

21. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
22. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303**, 3997 (2013).
23. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
24. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* **6**, 31 (2005).
25. Jin, J. J. et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 241 (2020).
26. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
27. Zhang R. G. Subgenome-aware analyses suggest a reticulate allopolyploidization origin in three *Papaver* genomes. *Zenodo*, <https://doi.org/10.5281/zenodo.7790632> (2023).

## Acknowledgements

We thank Prof. Xiao-Ru Wang for helpful discussions and revision of the manuscript. YPM is supported by the National Key Research and Development Program (2022YFF1301702), the Natural Science Foundation and Ten Thousand Talent Program of Yunnan Province (202001AS070019, YNWR-QNBJ-2018-174) and the “Light of West China” Program. KHJ is supported by the Agricultural Science and Technology Innovation Project of SAAS (CXGC2023F13). WZ is supported by the Swedish Research Council (VR, 2017-04686).

## Author contributions

R.G.Z., K.H.J., Y.P.M. and W.Z. conceived and designed the study; R.G.Z., C.L., G.Y.L., J.L., L.W., Z.X.W., Z.C., D.L., Y.Z., T.L.S., W.Z., Z.H.T., and J.F.M. collected and analyzed the data; R.G.Z., K.H.J., W.Z. T.L.S. and Z.X.W. prepared figures and tables; W.Z., K.H.J. and R.G.Z. wrote the manuscript; R.G.Z., K.H.J., Y.P.M. and W.Z. revised the manuscript; all authors approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-023-37939-2>.

**Correspondence** and requests for materials should be addressed to Yong-Peng Ma, Kai-Hua Jia or Wei Zhao.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023