Three chromosome-scale *Papaver* genomes reveal punctuated patchwork evolution of the morphinan and noscapine biosynthesis pathway

Xiaofei Yang1,2,3,9, Shenghan Gao2,4,9, Li Guo2,4,5,9, Bo Wang4, Yanyan Jia4, Jian Zhou6, Yizhuo Che2,4, Peng Jia2,4, Jiadong Lin2,4,7, Tun Xu2,4, Jianyong Sun8 & Kai Ye2,3,4,5,7✉

For millions of years, plants evolve plenty of structurally diverse secondary metabolites (SM) to support their sessile lifestyles through continuous biochemical pathway innovation. While new genes commonly drive the evolution of plant SM pathway, how a full biosynthetic pathway evolves remains poorly understood. The evolution of pathway involves recruiting new genes along the reaction cascade forwardly, backwardly, or in a patchwork manner. With three chromosome-scale *Papaver* genome assemblies, we here reveal whole-genome duplications (WGDs) apparently accelerate chromosomal rearrangements with a nonrandom distribution towards SM optimization. A burst of structural variants involving fusions, translocations and duplications within 7.7 million years have assembled nine genes into the benzylisoquinoline alkaloids gene cluster, following a punctuated patchwork model. Biosynthetic gene copies and their total expression matter to morphinan production. Our results demonstrate how new genes have been recruited from a WGD-induced repertoire of unregulated enzymes with promiscuous reactivities to innovate efficient metabolic pathways with spatiotemporal constraint.
Metabolic gene cluster (MGC) is a special genetic architecture well known in microorganisms involved in both primary and secondary metabolisms. While MGCs are relatively uncommon in plants, recent studies have discovered over 30 plant secondary metabolites whose biosynthetic pathways are encoded by MGCs, such as thalianol, noscapine, atotmatine, mamilactones, DIMBOA, 4,6-benzoxazin-3-one and taxadine etc. Furthermore, genomic mining using algorithms such as plantiSMASH and PhytoClust have identified many putative plant MGCs. As more plant genomes being sequenced and analyzed, the list of putative plant MGCs is expected to grow, although how many of these MGCs encode functional metabolic pathways awaits investigation. Functionally related genes in plant genomes are typically unlinked. However, MGCs are exceptional raising questions why they are necessary and widely distributed across multiple kingdoms. One hypothesis for cluster formation is co-inheritance arguing that colocalization of genes prevents losing key pathway genes due to recombination events. Another hypothesis is co-expression suggesting that co-located genes share common promoters and cis-regulatory elements, ensuring a spatially and temporally coordinated expression of biosynthetic enzymes and thus maximizing the yield of end products.

Despite discovering plant MGCs, our knowledge is limited regarding the mechanisms and history of their formation, maintenance, and diversification. Formation of microbial MGCs is often through relocation or duplication of native genes, or horizontal gene transfers. By contrast, plants are limited to vertical gene transmission, and thus must have their own as-yet-unknown mechanisms of gene cluster formation. Understanding the evolutionary history of plant MGCs can explain why many secondary metabolites are exclusive to plants of specific lineage or species. For example, opium poppy remains the sole natural source for opiates used to extract morphine for painkilling drugs, while the anti-cancer drug paclitaxel (trade name Taxol) mainly exists in the bark of Taxus trees. Additionally, some secondary metabolites have apparently undergone convergent evolution in distantly-related plant lineages. Driven by genetic variations, evolution occurs upon selection on traits that benefit organisms under special environment. Therefore, genetic variations and natural selection expectedly play critical roles in the independent evolution of plant MGCs, giving plants a plethora of structurally diverse secondary metabolites of specific bioactivities to support their sessile lifestyle.

Although ample instances of new genes driving the evolution of biosynthetic pathway have been well documented, how a full cascaded metabolic pathway evolved among different lineages is largely unknown. Recent studies suggest that plant gene clusters may have evolved through the recruitment of genes from elsewhere in the genome via duplications and neo-functionalization, genomic rearrangement, transposons, and other unknown mechanisms. Three main evolution models of pathway have been proposed during the last decades, including the forward, backward, and patchwork model. The forward model depicts the evolution of the pathway by recruiting enzymes catalyzing forwardly from earlier steps to later steps, while the backward model describes the pathway evolved by acquiring enzymes in a backward fashion, e.g., from later steps to earlier steps. The patchwork model suggests metabolic pathways are assembled by duplicating genes encoding enzymes reacting with diverse substrates. An evolutionary model for MGC-encoded plant metabolic pathway is overall lacking due to quite limited comparative analysis of multiple chromosome-scale plant genomes that are notoriously difficult to assemble. Without taking genomic synteny and large-scale rearrangement events into account, evolutionary analysis of plant gene cluster is likely prone to misleading conclusions on the evolution of pathway.

Opium poppy (Papaver somniferum L.) is a medicinal plant producing various benzylisoquinoline alkaloids (BIAs) such as morphine and noscapine. BIA biosynthetic pathway is one of the most completely elucidated in plants, with key enzymes identified and functionally validated. We and others previously reported that the biosynthetic pathways of morphinan and noscapine were primarily controlled by a BIA gene cluster of 584 kb in P. somniferum genome, representing one of the largest gene clusters encoding secondary metabolic pathways in eukaryotes. The BIA gene cluster includes a 10-gene cluster for noscapine biosynthesis and five genes (STORR, SALSYN, SALAT, SALR, and THS) for morphinan biosynthesis, and also several genes of unknown functions. Genes encoding the biosynthesis of precursor (S)-reticuline (TYDC, 4OMT, 6OMT, BBE etc.) as well as converting thebaine to morphine (CODM, COR, and T6ODM) are unclustered and dispersed in the genome. The morphinan biosynthetic pathway encoded by a single MGC yet producing two distinct subclasses of BIAs, morphinan and noscapine, presents an ideal model to study the complex evolutionary history of plant MGCs in early-diverging eudicots, given that the BIAs have experienced both natural selection and artificial selection by a human being.

In this work, we de novo assemble two chromosome-level genomes of P. rhoes (common poppy) and P. setigerum (Troy poppy) and improve the previous draft P. somniferum HN1 genome to gain insights into the evolution of BIA gene cluster among Papaver species. The three Papaver species are diverse in terms of morphinan and noscapine production level, and WGD events, therefore providing fresh insights into the evolution of the BIA biosynthetic pathway. Through comparative genomic analysis, we infer the evolutionary history of the morphinan biosynthetic pathway and BIA gene cluster which is most likely explained by the punctuated patchwork model.

Results

Quantification of noscapine and morphinans in three Papaver species. We evaluated BIA productions of three Papaver species by quantifying the noscapine and morphinans (thebaine, codeine, and morphine) in capsules using the HPLC-MS method. We accumulated the highest amount of morphinan and noscapine in P. somniferum (Supplementary Fig. 1) produced a trace amount of noscapine and morphinans in three species by quantifying the noscapine and morphinans (thebaine, codeine, and morphine) in capsules using the HPLC-MS method.

Genome assembly and annotation. To assemble P. setigerum and P. rhoes genomes, we used a combination of sequencing technologies including Oxford Nanopore (ONT) long reads, ONT ultra-long reads, Illumina paired-end reads, and high-throughput chromosome conformation capture (Hi-C) sequencing reads. The genome assemblies of P. setigerum and P. rhoes were highly contiguous with 97.6% and 87.9% of genome contigs anchored to chromosomes by using Hi-C scaffolding, respectively. The genome assembly size of 4.6 and 2.5 Gb, scaffold N50 values of 211.2 and 329.4 Mb, contig N50 values of 65.6 and 5.3 Mb (Table 1 and Supplementary Tables 3–6, Supplementary Data 2, Supplementary Method 6). Facilitated by Hi-C scaffolding, 92.4% of original P. somniferum sequences were anchored on 11 chromosomes with a scaffold N50 of 249.6 Mb, much-improved assembly contiguity.
over the original version24 (Table 1 and Supplementary Fig. 7, Supplementary Data 2, Supplementary Method 6). After reassignment of chromosome IDs, we found the BIA gene cluster was located at chr4 rather than chr11. Benchmarking Universal Single-Copy Orthologs (BUSCO)45 evaluation revealed the high genome completeness (92.8–95.3%) (Table 1, Supplementary Fig. 8). We annotated 41,407 to 106,517 protein-coding genes and 12,429–23,109 non-coding RNAs for the three Papaver species (Table 1 and Supplementary Figs. 9, 10, Supplementary Method 7). Repetitive elements make up more than 70% for each Papaver genome, and over 50% of them are long terminal repeat retrotransposons (LTRs) (Table 1 and Supplementary Fig. 11, Supplementary Method 7).

### Whole-genome duplication events in three Papaver species

Synteny analysis revealed two rounds of WGDs in *P. setigerum* but no WGD in *P. rhoes*, and confirmed the reported single WGD in *P. somniferum*24 (Fig. 1b, c and Supplementary Figs. 12–16, Supplementary Method 8). In addition, a collinearity analysis of three *Papaver* species with grape (*Vitis vinifera*)46 and ancestral eudicot karyotype genome47 (Supplementary Figs. 17 and 18) confirmed a lack of the whole-genome triplication (y event) occurred in core eudicots in *Papaver* genus.

We next investigated whether *P. somniferum* and *P. setigerum* shared any WGD. Using the estimated divergence time of around 7.7 million years ago (Mya), which is consistent with TimeTree48, and mean synonymous substitutions per synonymous site (Ks) of syntenic gene pairs between *P. setigerum* and *P. rhoes*, we estimated the mutation rate as $8.1 \times 10^{-9}$ synonymous substitutions per site per year for *Papaver*. Based on this rate and Ks distribution, the *P. somniferum* WGD occurred at around 7.2 Mya, within the range of previous estimation24 (Fig. 1d and Supplementary Data 3, Supplementary Method 8). Phylogenomic analysis of eight angiosperms using 48 single-copy orthologs identified by OrthoFinder49 indicated *P. setigerum* and *P. somniferum* diverged at around 4.9 Mya (Fig. 1e and Supplementary Method 8), following the divergence of their common ancestor from *P. rhoes* at around 7.7 Mya. Since the *P. somniferum* WGD predated the divergence of *P. somniferum* and...
Table 1 The statistics of assemblies and annotations of three *Papaver* genomes.

|               | *P. setigerum* | *P. rhoes* | *P. somniferum* |
|---------------|---------------|------------|-----------------|
| **Contig**    |               |            |                 |
| Total number of contigs | 553          | 3,273      | 61,801          |
| Assembly size (MB)   | 4590.03      | 2541.84    | 2709.32         |
| **Contig N50 (MB)** | 65.57        | 5.29       | 1.74            |
| **Contig N90 (MB)** | 13.37        | 0.61       | 0.12            |
| Largest contig (MB)| 178.78       | 39.02      | 13.77           |
| **Scaffold**      |               |            |                 |
| Total number of scaffolds | 381          | 237        | 55,380          |
| Assembly size (MB) | 4590.12      | 2542.27    | 2712.53         |
| Scaffold N50 (MB)  | 211.16       | 329.41     | 249.6           |
| Scaffold N90 (MB)  | 155.08       | 33.61      | 172.47          |
| Largest scaffold (MB) | 329.11      | 361.76     | 328.07          |
| BUSCO for genome   | 94.50%       | 92.80%     | 95.30%          |
| GC content         | 36.88%       | 37.94%     | 37.28%          |
| Number of gaps     | 172          | 3,036      | 6,421           |
| **Annotation**     |               |            |                 |
| Number of protein-coding genes | 106,517   | 41,470     | 55,316          |
| Supported by RNA-seq or homologs | 100%       | 100%       | 100%            |
| Supported by protein family | 70.56%   | 70.14%     | 70.05%          |
| Repeat density     | 71.55%       | 73.92%     | 76.68%          |
| Number of ncRNA    | 23,109       | 12,429     | 12,636          |

*P. setigerum* (4.9 Mya), this WGD is essentially the same event as the first WGD (WGD-1) in *P. setigerum* which has further undergone a lineage-specific WGD (WGD-2) dated around 4.0 Mya (Fig. 1d, e and Supplementary Data 3).

Ancestral genomes and accelerated non-random post-WGD rearrangements. Rearrangements following WGD events are common in plant genome evolution. We are curious about what genome structural changes led to the present-day karyotypes of three *Papaver* species. We adopted the computational strategy proposed by Sankoff et al. and developed a bottom-up workflow to reconstruct pre- and post-WGD-1 ancestors for three *Papaver* species (Supplementary Figs. 19, 20 and Supplementary Method 9). Ancestral genome reconstruction relies on accurate genome assembly, and potential misassembly confounds the reconstruction and downstream analysis. Based on the high-quality genome assemblies of three *Papaver* species, the reconstructed pre- and post-WGD-1 ancestral genome had six and eleven protochromosomes, respectively (Fig. 2a and Supplementary Fig. 21). Compared to pre-WGD-1 ancestral genome, *P. rhoes* likely needed at least five chromosomal fissions and four chromosomal fusions to reach its current structure of seven chromosomes (Fig. 2a and Supplementary Fig. 21). By contrast, *P. setigerum* experienced a much more complex evolutionary history involving two rounds of WGDs and post-WGD rearrangements that finally shaped its karyotype of 22 chromosomes. Shared with *P. somniferum*, *P. setigerum* likely underwent at least 11 chromosomal fissions and 12 chromosomal fusions after WGD-1, and then at least 20 chromosomal fissions and 20 chromosomal fusions following a lineage-specific WGD-2 (Fig. 2a).

We observed a non-random distribution of chromosomal fissions and chromosomal fusions associated with the transformation of the reconstructed pre-WGD-1 protochromosomes to modern chromosomes (Fig. 2). Three chromosomes (chr7, chr16, and chr19) in *P. setigerum* were significantly enriched with fusion events, while chr21 was significantly depleted of them (p-value < 0.05, z-test, Fig. 2a). Similarly, in pre-WGD-1 ancestor, protochromosome 4 was significantly enriched with fusion events, whereas protochromosome 2 was significantly depleted of them (p-value < 0.05, z-test, Fig. 2b). In addition, the numbers of WGDs and fission events needed to shape the current karyotypes in three *Papaver* species have a superlinear correlation (Fig. 2b), indicating that post-WGD genome rearrangements might have been accelerated. WGD plays a significant role in the plant secondary metabolism evolution, inspiring us to investigate the function of post-WGD diploidization in *Papaver* species. The genes around the shuffling breakpoints in *P. somniferum* were enriched in KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways related to isoquinoline and indole alkaloid biosynthesis, as well as the metabolism of amino acids, such as tyrosine and phenylalanine, whereas breakpoint-vicinity genes in *P. setigerum* were enriched with plant-pathogen interactions and environmental adaptation (Supplementary Fig. 22). These results suggest a post-WGD diploidization might have played a part in the optimization of the biosynthetic pathways of alkaloids such as morphinans in the ancestor of *P. somniferum* and *P. setigerum*.

Recruitment of new genes to BIA gene cluster locus. Although the chromosomal shuffling events affected isoquinoline alkaloid metabolism, we did not observe their impact on the BIA gene cluster (Fig. 2a and Supplementary Data 4, 5, Supplementary Method 10). Therefore, we compared three *Papaver* genomes to identify additional structural variation events potentially leading to the formation of the BIA gene cluster. *STORR* encodes a fusion protein of a cytochrome P450 and an oxidoreductase (Supplementary Fig. 23), and enables the gateway reaction towards morphinan biosynthesis. If the deletion of the intergenic region between pre-fusion modules is the sole event causing *STORR* fusion, one should expect collinearity between loci encompassing *STORR* and its pre-fusion modules. However, such collinearity was observed in neither *P. somniferum* nor *P. setigerum* (Supplementary Figs. 12 and 24), intriguing us to hypothesize that *STORR* evolution may involve additional translocation besides the fusion event with unknown order, the so-called “fusion, translocation” (FT) event.

We then systematically examined two genomic loci implicated in the translocation event of the three species and traced the evolutionary history of *STORR*. We defined the donor loci as genomic regions syntenic with the *P. somniferum* regions containing the pre-fusion modules, and the recipient loci as genomic regions collinear with the *P. somniferum* regions harboring the *STORR* locus (Fig. 3a and Supplementary Figs. 25, 26). In addition, donor loci carrying the pre-fusion modules represent the prior state of the FT event, while other donor loci denote the post-state. As for recipient loci, the ones carrying *STORR* represent the post-state, while the others represent the prior state (Fig. 3a and Supplementary Figs. 25, 26). We found all four types of loci appeared exactly once in *P. somniferum*, but twice in *P. setigerum*, while only the prior states of donor and recipient loci were observed once in *P. rhoes* (Fig. 3a and Supplementary Figs. 25, 26, Supplementary Methods 11), supporting our hypothesis.

Considering the phylogeny and WGD history among the three species, we proposed a parsimonious evolutionary model to illustrate a burst of genomic rearrangements giving the birth of *STORR* to the current BIA cluster (Fig. 3b). The ancestor of three species contains the pre-fusion modules at donor loci and an empty recipient locus, which is preserved in *P. rhoes*. After divergence from *P. rhoes*, WGD-1 at around 7.2 Mya resulted in two copies of pre-fusion modules, of which one copy was converted to *STORR* by a FT event and inherited by *P. somniferum*. After divergence from *P. somniferum*, *P. setigerum*
underwent a lineage-specific WGD-2 at around 4.0 Mya, yielding two copies of STORR at recipient loci and two copies of pre-fusion module at donor loci (Fig. 3b). Based on comparative analysis of three Papaver genomes, current computational strategies, and parsimonious assumption, our STORR evolutionary model represents a working hypothesis. The inclusion of sequencing data from additional species may lead to an updated evolutionary model or alternative hypotheses.

We next investigated the evolutionary history of the remaining genes in the BIA gene cluster. Unlike previous analyses focusing on gene trees53, we integrated evidence of multiple sources from synteny, phylogeny, protein sequence alignments, and WGD (Fig. 4a and Supplementary Figs. 27–36, Supplementary Methods 11, 12) and found that four genes (PSSDR1, CYP82X1, CYP719A21, and PSMT1) in noscapine branch and two (SALSYN and SALR) in morphinan branch were already present in MRCA of three Papaver species (Fig. 4b). Three genes (STORR, SALAT, and THS) in the morphinan branch were assembled before P. somniferum diverged from P. setigerum, while the assembly of noscapine branch was completed in P. somniferum via adding six new genes (PSCXE1, CYP82X2, PSAT1, PSMT2, CYP82Y1, and PSMT3) through lineage-specific duplications (Fig. 4b and Supplementary Figs. 27–35, Supplementary Method 11). Specifically, we inferred five genes including PSCXE1, PSAT1, PSMT3, SALAT, and THS may be assembled into the BIA gene cluster by putative dispersed duplications (the gene and their inferred original copy are non-syntenic and not adjacent) while CYP82X2 was likely generated from a tandem duplication of CYP82X1 (Fig. 4b and Supplementary Figs. 27–35). Considering the difficulty of distinguishing dispersed duplication from old tandem duplication with gene deletions, we do not rule out the possibility that PSCXE1, PSAT1, PSMT3, SALAT, and THS were assembled into BIA gene cluster by ancient tandem duplications with follow-up gene deletions (Supplementary Fig. 37).

Clustered BIA genes are co-regulated and evolved in a coordinated manner. Assembly of 15 genes encoding two distinct pathways into a compact gene cluster is striking. The genetic components of the BIA pathway should allow the biosynthesis of morphinan and noscapine albeit at their original loci before clustering, raising questions on the necessity of gene clustering in evolution. In fact, besides the BIA gene cluster, the phenomenon of gene clustering has been reported for multiple plant secondary metabolic pathways54. This increasingly common theme suggests that plants, with a sessile lifestyle, are under selection pressure to evolve a special genetic architecture to facilitate their adaptation to environmental stimuli55, consistent with our gene family analysis (Supplementary Data 6–8). Recently, we and others have reported that gene clustering likely enables the coordinated...
expression of metabolic enzymes for BIA biosynthesis in opium poppy stem and root tissues\textsuperscript{24,36}. However, it remains unclear how \textit{Papaver} species achieve the coordinated BIA gene expression by evolving this gene cluster. To approach this question, we dissected the tissue-specific expression profiles of the BIA gene cluster and its closest paralogs within \textit{P. somniferum} and \textit{P. setigerum} (Supplementary Method 13). For STORR gene-related loci, we found that the donor locus had low expression levels across all tissues among three species while the recipient locus exhibited high expression in the stem where morphinan biosynthesis primarily occurs\textsuperscript{24} (Fig. 5a and Supplementary Fig. 38). Expression pattern of the other eight new genes at the BIA gene cluster raised from a series of whole-genome duplications (WGD), “fusion, translocation” events. Purple and blue colors denote the cytochrome P450 and oxidoreductase modules. MRCA most recent common ancestor, WGD whole-genome duplication, Mya million years ago. Source data underlying \textbf{a} are provided as a Source Data file.}

\section*{Fig. 3 Evolutionary history of STORR.} \textbf{a} The collinearity of genes in the donor loci and the recipient loci showing both prior and post states of FT event in three \textit{Papaver} species supporting the “fusion, translocation” (FT) event leading to STORR formation at BIA gene cluster. The directions of arrows indicated the chromosome from 5' to 3'. The dash lines indicate the boundary between donor loci and recipient loci, and the boundary between prior states and post states. The detailed diagram with the genomic context was shown in Supplementary Fig. 25. \textbf{b} A proposed evolutionary model for the birth of STORR at BIA gene cluster formed by a series of whole-genome duplications (WGD), “fusion, translocation” events. Purple and blue colors denote the cytochrome P450 and oxidoreductase modules. MRCA most recent common ancestor, WGD whole-genome duplication, Mya million years ago. Source data underlying \textbf{a} are provided as a Source Data file.

Total expression of all gene copies contributes to morphinan production. Copy number variation of BIA biosynthetic genes
Fig. 4 Evolutionary history of benzylisoquinoline alkaloid (BIA) gene cluster. a The syntenic relations at BIA gene cluster in three Papaver species. Solid lines denote the syntenic relations detected by MCScanX, and the dash lines denote the top BlastP hits. Different colors denote different genes. The gene name or ID is labeled for each gene. The numbers above the genes represent sequence identity levels from BlastP. The directions of arrows indicate the chromosome from 5′ to 3′. BH best hit. b Summary of BIA gene cluster evolution. Putative tandem duplication as a structural variation event producing identical adjacent segments while the other non-locally duplicated ones are "putative dispersed duplication". MRCA most recent common ancestor, WGD whole-genome duplication, Mya million years ago. Source data underlying a are provided as a Source Data file.
Source data are provided as a Source Data STORR copy or copies of cating a potential gene dosage effect. However, we found the extra production of morphinans compared to *P. setigerum* due to WGD-2 did not lead to more abundant pro-

*P. somniferum* sequences between *P. setigerum* median; box limits, upper and lower quartiles; whiskers, data range.

8 in *P. somniferum* has been positively correlated with the production of morphinans the donor and the recipient loci related with BIA genes were shown in Supplementary Fig. 38. By contrast, compared to *P. somniferum* cultivars in a previous genomic study, indi-

Fig. 5 Clustered BIA genes are co-regulated and evolved in a coordinated manner. a The heatmap of the average normalized expression level of genes at the donor and the recipient loci related with STORR in six tissues of three *Papaver* species. Details of the gene expressions related to BIA genes were shown in Supplementary Fig. 38. b Hi-C interaction heatmap of regions including two copies of morphinan gene cluster on chr15 (top) and chr8 (down) of *P. setigerum*, shown at a resolution of 10 kb. The morphinan branch gene regions are marked with colored boxes. c The comparison of the interactions between two morphinan gene copies in *P. setigerum*. The p-value is calculated by a two-sided Wilcoxon rank-sum test. For the boxplot, the centerline, median; box limits, upper and lower quartiles; whiskers, data range. n is the number of bins. d Dotplots of STORR, SALSYN, SALAT, SALR, and THS promoter sequences between *P. somniferum* (one copy) and *P. setigerum* (two copies). Promoter is defined as 2 kb upstream region of the transcription start site. Source data are provided as a Source Data file.

has been positively correlated with the production of morphinans in *P. somniferum* cultivars in a previous genomic study, indicating a potential gene dosage effect. However, we found the extra copy or copies of STORR, SALSYN, SALAT, and SALR in *P. setigerum* due to WGD-2 did not lead to more abundant production of morphinans compared to *P. somniferum* (Fig. 6 and Supplementary Fig. 2). Moreover, the eroded promoters of one copy of the morphinan gene cluster in *P. setigerum* caused low expression, while the intact promoter of the second copy maintained high expression of genes (Supplementary Method 13). Therefore, the copy number is not the only factor affecting BIA productions, and gene expression may also play an important role. We systematically compared the copy number and total expression of all copies of BIA genes encoding enzymes for the morphinan biosynthetic pathway in the three species. We observed that four genes encoding key enzymes, including STORR, SALAT, THS, and T6ODM, were missing in the corresponding syntenic blocks of *P. rhoes* YM1 (Fig. 4, Supplementary Fig. 42), probably leading to its negligible amount of three morphinans compared to *P. somniferum* HN1 (Fig. 6 and Supplementary Fig. 2). By contrast, compared to *P. somniferum* HN1, *P. setigerum* DCW1 carries twice as many copies of STORR, SALSYN, SALR, and T6ODM, but fewer copies of COR and CODM catalyzing the final step to codeine and morphine production (Fig. 6 and Supplementary Fig. 42, Supplementary Data 5). The pure difference of gene copy numbers between the two species could not explain the large gap in their morphinan productions. Thus, we compared the total expression of all copies of each morphinan branch gene in the stem where morphinan biosynthesis primarily takes place, and found their expression in *P. somniferum* HN1 was at least twice as high as that in *P. setigerum* DCW1 (Fig. 6 and Supplementary Fig. 42, Supplementary Data 5). As the morphinan biosynthesis occurs in a cascade fashion, we expect the total expression of all copies of each gene and its copies along the pathway to yield an amplified difference in the gross production of the final products, which is exactly what we observed in the two *Papaver* relatives (Fig. 6). As such, these results suggest the morphinan production is attributed to the total expression of all copies of key genes besides their copy numbers. Finally, distinct copy numbers for CODM, T6ODM, and COR have been observed among three *Papaver* species (Fig. 6 and Supplementary Fig. 42) and among cultivars of *P. somniferum*, suggesting that dosage of these genes is possibly under positive selection.
Discussion
Gene clustering, a frequent genetic paradigm in microbial metabolic pathways, has long been thought as a rarity in higher eukaryotes such as land plants. Until recently, with the burst of plant genome sequencing projects and advancement of bioinformatic tools, biosynthetic pathways for over 30 plant natural products are known to be encoded by gene clusters, including A. thaliana (e.g., thalianol)\textsuperscript{23}, opium poppy (e.g., morphinan)\textsuperscript{2,24}, maize (e.g., DIMBOA)\textsuperscript{57}, rice (e.g., mimilactones)\textsuperscript{4}, and tomato (e.g., α-Tomatine)\textsuperscript{3}, etc. Our studies on opium poppy, and recent examples shown by other groups\textsuperscript{56} have demonstrated that the formation and diversification of biosynthetic gene clusters in plants have been made possible largely by gene gain and loss through structural variations\textsuperscript{58}.

In this study, the contribution of large-scale genome variations, e.g., WGD and subsequent genome rearrangements, to the rise and divergence of biosynthetic pathways is well illustrated by the reconstructed evolutionary history of STORR and other key BIA genes. Three high-quality Papaver genomes and their distinct WGD status enabled us to infer the origins and steps leading to the current BIA gene cluster. We showed that duplication was a major factor contributing to the formation of the BIA gene cluster. The origin and putative type of the duplications were inferred based on synteny, WGD, and protein sequence alignments analysis of three Papaver genomes. Alternative explanations of the BIA gene cluster formation such as “old tandem duplications” are equally possible but shall be tested with more sequencing data and genome analyses. Interspecies comparison of the morphinan production revealed that total expression of gene copies of BIA genes contributed to the morphinan production. Consistently, the importance of gene copies for the BIA biosynthetic genes was also shown by an intraspecies comparison of nine P. somniferum cultivars\textsuperscript{56}. Therefore, gene copies are pivotal to both intraspecies and interspecies variation of morphinan production in Papaver.

Papaver species produce diverse types of natural products besides morphinans and noscapine. For example, Menéndez-Perdomo et al. summarized 44 authentic BIA in P. somniferum, including salutaridine, papaverine, and narcotoline\textsuperscript{59}, and Grauso et al. summarized 128 organic compounds isolated from P. rhoas, including leucine and rhoeadine\textsuperscript{60}. The fact that biosynthetic pathways of the non-morphinan and non-noscapine compounds are poorly understood in Papaver species limited our further investigation of their metabolic pathways in this study. Thus, we focused on the evolutionary steps leading to the efficient production of the morphinans and noscapine in this interspecies comparison. Expectedly, P. rhoas which produces a trace amount of morphinans and noscapine lacks a bona fide BIA gene cluster. We found that P. rhoas has a primitive metabolic pathway leading to the formation of morphinan precursors such as dopamine, 4-Hydroxyphenylacetaldehyde, and (S)-Norco-claurine encoded by TYDC (PrhUNG23530.0) and TryAT (Prh05G30560.0), NCS (PrhUNG12800.0), respectively. Given the phylogenetic relationships of the three Papaver spp., it suggests that an ancient repertoire of unregulated enzymes with promiscuous reactivities might be already present in Papaver to allow the production of morphinans at minimal levels. WGD-1 apparently energized the ancestral genome and triggered chromosomal rearrangements for boosting the production levels of beneficial metabolites. As a result, a set of new genes originated from fusions and gene duplications were recruited to a locus pre-equipped with desired regulatory elements, leading to the
formation of the BIA gene cluster. The formation of bifunctional STORR enzyme as a result of gene fusion channels metabolic flow preferentially into the morphinan pathway. Recruitment of subsequent enzymes for both morphinan and noscapine pathways as they were branching included neither forwardly nor backwardly evolving progression of evolutionary steps, rejecting both hypotheses of forward pathway evolution and retrograde evolution of biochemical pathway. Instead, once a relatively large repertoire of enzymes becomes available after WGD-1, duplicated promiscuous enzymes can be readily recruited from various genomic regions into a locus pre-equipped with tissue-specific regulatory elements to maximize flux through morphinan and noscapine pathways, as elegantly explained by the patchwork model, which describes the way of recruitment of enzymes in a pathway like a patchwork manner. There are extensive instances supporting the patchwork model in bacteria. For example, the pathway for degradation of pentachlorophenol, in instances supporting the patchwork model in bacteria. For example, the pathway for degradation of pentachlorophenol, in Sphingomonas chlorophenoica and metabolic pathway for 2CNB degradation in Pseudomonas stutzeri. By contrast, an evolutionary model for metabolic pathways in eukaryotes, as elegantly explained by the patchwork model, which describes the way of recruitment of enzymes in a pathway like a patchwork manner. We showed in this study that post-WGD genome rearrangement played a major role in achieving this tissue-specific co-expression of the morphinan gene cluster for P. somniferum and P. setigerum. We showed that new genes were recruited from lowly expressed genomic loci to the pre-configured locus with a stem-specific expression. It remains a mystery why such a low entropy of tightly packed gene cluster ever exists and evolves coordinate, as it is common to find genes at various genomic loci encoding the same pathway, functionally linked via gene co-regulation. In fact, unclustered biosynthetic genes are common in plants that somehow are co-expressed in specific tissues. One example is the flame lily (Gloriosa superba) colchicine biosynthetic pathway encoded by genes showing no apparent genomic clustering but strongly co-expressed in seeds and corms. For opium poppy, genes such as CDM and T6ODM encoding the tailoring enzymes in morphinan biosynthesis are also unclustered but co-expressed with clustered core genes encoding the morphinan pathway even though located on different chromosomes. Among plant metabolic gene clusters, the genetic architecture of a core gene cluster plus several peripheral or satellite genes is rather common. The gene clusters can be continuous without any intervening genes such as avacinin (oat) gene cluster, while some have genes of unknown function separating the cluster genes such as thaliana. The DIMBOA morphinan gene cluster. There are also biosynthetic pathways such as tomatine, morphine, and cucurbitacin C encoded by a core cluster with satellite genes or cluster. The morphine biosynthetic pathway is composed of a core cluster coupled with several peripheral genes (in trans) or satellite gene groups. Recent studies also show that collinearity exits within a single gene cluster such as the noscapine gene cluster where the cluster is organized in modules that correspond to early, middle and late steps of the pathway. The diverse architecture of metabolic gene clusters shows the evolutionary history of gene clustering for secondary metabolic pathways is complex and species-specific, involving complicated genomic variations and selection processes. Yet how plant biosynthetic genes, either clustered or dispersed in genome, are co-regulated at both epigenetic and transcriptional levels remains largely unknown and needs further investigation. We presented substantial evidence that epigenetic regulations and a non-random erosion of cis-elements within the gene cluster on chr8 in P. setigerum underlies the efficient biosynthesis of BIA. The fusion protein STORR couples two reactions, forming the core of a so-called “metabolon” together with co-expressed subsequent morphinan biosynthetic enzymes, to maximize the efficiency of morphinan production but minimize diffusion of toxic intermediate products. The lately formed noscapine pathway joined the locus, perhaps exploiting the optimized regulatory elements and chromatin-level regulation, further expanding the production line of the secondary metabolite factory.

Finally, as we and others have reported the structural variants such as copy number variants within morphinan biosynthetic genes, opium poppy genome as well as its morphinan gene cluster are not static and probably still evolving given the natural and human artificial selection process. Thus, it would be interesting to observe how biosynthetic pathways continue to evolve due to human selection, and whether biosynthetic genes that are clustered but did not co-express, and genes that co-expressed but non-clustered will have a converged or diverged evolutionary trajectory in future opium poppy crops.

Methods

Plant materials, DNA, and RNA isolation. P. setigerum variety DCW1, P. rhoes variety YMR1, and P. sommiferum variety HN1 were grown in Azalea pots in a growth chamber with 16 h of light located at Xi’an Jiaotong University Laboratory of BioData Sciences. Genomic DNA was extracted from fresh leaves (the uppermost ones) harvested from 6-week-old seedlings of three Papaver species. For Illumina and regular Oxford Nanopore (ONT) sequencing, high molecular weight (HMW) genomic DNA was prepared by the CTAB method and purified with QIAGEN® Genomic kit (Cat#13343, QIAGEN). For ultra-long ONT sequencing, ultra-high molecular weight DNA was extracted by the SDS method handled gently to sustain the length of DNA. For transcriptome sequencing (RNA-seq), Total RNA was extracted from the six tissue types on the first day of anthesis, root, leaves, stem (the 2-cm-long part just underneath the capsule), capsule, petal, and stamens, using TRIzol reagent (TIANGEN). RNA integrity was determined using regular agarose gel electrophoresis, Nanodrop (ThermoFisher Scientific), and Agilent 2100 Bioanalyzer (Agilent Technologies). RNA sample of high quality (OD260/280 within range [1.8, 2.2], OD260/230 ≥ 2.0, RIN ≥ 8, > 1 µg) was used to construct the sequencing library.

Genome and transcriptome sequencing. DNA and RNA sequencing in this study were performed using sequencers at Nectomics Inc. (Wuhan, China). For ONT sequencing, about 4 and 10 µg high molecular weight DNA was used for ultra- and ultra-long library construction, respectively. About 700 ng (regular) and 800 ng (ultra-long) DNA libraries were sequenced on a Nanopore PromethION sequencer. A total of eight ONT cells were used to sequence the P. setigerum genome. For Illumina paired-end sequencing, a total of 1.5 µg DNA was used to construct the library and the libraries were sequenced by using the Illumina NovaSeq platform to generate 150 bp paired-end reads with insert size around 400 bp. Hi-C library construction was performed according to a 3C protocol established for maize. Briefly, 5 g fresh leaves were fixed with 1% formaldehyde solution in MS buffer (10 mM potassium phosphate, pH 7.0, 50 mM NaCl; 0.1 M sucrose), and homogenized with liquid nitrogen, and then subjected to nuclei isolation and enrichment. Chromatin was digested for 16 h with 400 U HindIII restriction enzyme (New England Biolabs) at 37 °C. DNA ends were labeled with biotin and DNA ligation was performed by adding T4 DNA ligase (New England Biolabs, USA) and incubation at 16 °C for 4–6 h, followed by reverse cross-linking using proteinase K (Promega, USA). Purified DNA was fragmented to a size of 300–500 bp, and DNA fragments labeled by biotin were finally separated on Dynabeads® M-280 Streptavidin (Life Technologies, USA). Hi-C libraries were controlled for quality and sequenced on an Illumina Novaseq platform. RNA-seq was performed according to the TruSeq RNA Library Preparation Kit (Illumina, USA). The RNA-seq libraries were sequenced on an Illumina Novaseq platform and paired-end reads of 150 bp were generated.

Genome assembly. To obtain chromosome-scale genome assemblies, all sequencing reads were first subjected to quality control. The ONT reads were assembled into contigs by using NewDenovo (v2.26) software, upon which we next performed three rounds of polishing with ONT reads and another round of polishing using Illumina paired-end reads by Nextpolish (v1.2.0) to yield high-quality contigs. Redundant sequences in the polished contigs were removed by purge_dups. Then, breakhis (v1.1) was used to identify assembly breakpoints of polished contigs and built Hi-C reads. Finally, 3d-DNA (v180922) was used to reorder and anchor contigs into scaffolds and chromosomes. To assess genome completeness, we
applied BUSCO (Benchmarking Universal Single-Copy Orthologs) (v3) analysis using the plant early release version (v1.161, release May 2015)55. In addition, we aligned BlastP and RTR opposition sequences to the Blast database, and detected SNP and Indels by GATK (v4.1.8)72 to evaluate the base accuracy of the assemblies.

**Genome annotation.** We used Repbase73 and the species-specific de novo repeat library constructed by RepeatModeler (v0.1.0.8) to annotate the repeat DNA sequences in three Papaver species. RepeatMasker (v0.4.0.7) was applied to annotated the repeat elements. In addition, we applied LTR-Finder (v1.17)74, LTRharvest (v1.5.2.8)75, and LTR_retriever (v2.8.3.0)76 to detect LTR elements. Protein-coding genes were predicted using the MAKER2 pipeline (v2.3.18)77 integrating evidence-based and ab initio gene predictors. The evidence includes Swiss-Prot (downloaded in Jan. 2020), protein sequences of A. thaliana78, Beta vulgaris79, and Vitis vinifera80, and transcripts assembled by Trinity (v2.1.19)81. Three ab initio gene prediction tools including AUGUSTUS (v3.3.3)82, SNAP (v2006-07-28)83, and GeneMark_EK (v3.48)84. Predicted genes were evaluated in terms of whether they were supported by transcript or protein homolog, the annotation edit distance (AED), and the exon eAED (v2.6.4). We kept the predicted genes with AED < 0.5, eAED < 0.5, and ones supported by transcript or protein homologs. The function domains of the proteins were annotated by InterProScan with default parameters84. Non-coding RNAs (ncRNAs) were annotated using csmRNA from INFERNAL (v1.1.2)85 package based on Rfam database (v14.1)86.

**Whole-genome duplication identification.** The syntenic analysis was performed by MScanX72 with default parameters from top-five BlastP hits. Within each Papaver genome, the proportion of genes with WGD/segmental duplication types and the widespread and well-maintained copy numbers of the syntenic blocks indicate no WGD in P. rhoeas, single WGD in P. somniferum, and two WGDs in P. setigerum. In addition, comparison among three Papaver genomes, between ancestral eudicot karyotype (AEK)47 and each genome, as well as between Vitis vinifera80 and each Papaver genome provides other evidence for confirming the number of WGDs in three Papaver genomes.

**Phylogenomic analysis and divergence time estimation.** We applied Ortho-Finder (v2.3.4)89 to detect single-copy orthologs from the three Papaver genomes and other five angiosperm species including the monocot Oryza sativa88, Aquilegia coerulea89, Maulea cordata90, Arabidopsis thaliana91, and Vitis vinifera92. Then, MAFFT (v7)93 was applied to align the detected single-copy ortholog pairs and the conserved regions were inferred by using Ghodb (v0.16b)94 with default parameters. The maximum likelihood phylogenetic tree was constructed by using RAxML (v8.2.12)94 with 100 bootstraps. Based on the neutral theory and molecular clock95,96,97, the mean divergence time was calculated from the synonymous substitutions per site per year (\(z\)) using the Mclust R package98. We identified pPGs, consisting of 152 Mya, http://timetree.org/100, the constrain time taxon of Aquilegia–Papaver (127.9–139.4 Mya, http://timetree.org/)101, and the constrain time taxon of A. thaliana and V. vinifera (107–135 Mya, http://timetree.org/)101. We estimated the P. rhoeas and P. somniferum divergence time of 7.7 Mya, consisting of timeWEB tree webing (http://timetree.org/)101 reports. Similarly, we estimated the divergence time of P. somniferum and P. setigerum as 4.9 Mya.

**Timing of whole-genome duplications.** To estimate the timing of the WGD event in P. somniferum and P. setigerum, Ks values of P. somniferum syntenic block genes and Ks values of P. setigerum syntenic block genes were calculated respectively using YN model in KaKs_Calculator (v2.0)99. For P. setigerum, we considered the reciprocal best matches among the syntenic gene pairs as the pairs from WGD-2 (the second WGD event) while other syntenic gene pairs were grouped as the pairs from WGD-1 (the first WGD event). The Ks value distributions were then fitted to a mixture model of Gaussian distribution using the Mclust R package98. We identified four gene pairs from P. setigerum, two genes from P. somniferum, and one gene from P. rhoeas, as a core pPG based on the corresponding numbers of WGD. Finally, DRIMM-Synteny99 were performed on the core pPG to identify the non-overlapping (NO) synteny blocks in the three Papaver species. We filtered blocks without satisfying \(2z:1z\) ratio in P. setigerum, P. somniferum, and P. rhoeas. And then, pPGs except for core pPGs were filtered in NO synteny blocks. In stage two, the ancestor protochromosomes were reconstructed by solving GMP (genomic median problem)100,101 and GGHP (guided genomic halving problem)101 in the scenario of genomes with multiple WGD events based on the detected NO synteny blocks. We modeled GMP and GGHP in scenario with multiple WGDs as a block matching optimization (MO) problem based on SCoG (single cut or join) genomic distance102, and solved this problem by an integer programming method. In the third stage, we inferred the possible order of pPGs based on the previously reconstructed ancestor protochromosomes and NO syntenic blocks. We built a directed weighted pPG graph for each syntenic block. A topological sorting method with a greedy strategy was performed on each graph to get possible gene orders. Our work can also be applied to similar evolutionary scenarios of Papaver. We are still exploring the solutions for other complex evolutionary scenarios and eventually will provide a generalized and user-friendly framework for all possible evolutionary scenarios. Our work is based on the accuracy of genome assembly. But now even with the cutting-edge sequencing data and widely used assembly methods, assembly errors are inevitable103. The potential misassembly may affect the reconstruction and downstream analysis. Therefore, computational evaluation and experimentally validation of genome assemblies are important to obtain more reliable results.

**Gene expression analysis.** The cleaned RNA reads were aligned against the assembled genome using Hisat2 (v2.2.1)106, and transcripts were discovered and quantified by Stringtie (v2.1.4) and Ballgown107, respectively, with default parameters. We measured the gene expression level by TPM (transcripts per million). To compare the TPMs between different species, we normalized the TPM by calculating \(z\)-score in each species. For each species, we first calculated the mean and standard deviation values of all TPMs. Then we calculated TPM of each gene as \(TPM = (TPM – mean)/sd\), where the mean is the mean value of all TPMs, and sd is the standard deviation of all TPMs. Furthermore, we calculated the normalized TPM as \(TPM = TPM + min_TPM\), to make the normalized TPM non-negative, where \(min_TPM\), is the minimize value of all \(TPM\)

**Hi-C data analysis.** Juicer (v1.6108) and Tadtools (v0.76)109 were used to calculating Hi-C contact matrix and generating the heatmap. The chromatin loops detected by HTCUP110.

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

**Data availability.** Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this Article is available as a downloadable Document Information File. The raw sequencing data generated in this study are available in PRJNA720042 and the National Genomics Data Center under accession code PRJCA004217. This study used previously published RNA-seq data under accession number GSE11111114, the genome assembly data are available at the Genome Warehouse in National Genomics Data Center under accession number GWHAZP00000008, GWHAZP00000009, and GWHAZP00000000 for P. rhoeas genome, P. setigerum genome and P. somniferum genome, respectively. The genome annotations of three Papaver genomes are available from Github [https://xjtlu-omics.github.io/Papaver-Genomics]. The used genomes and annotations of Oryza sativa [http://plantsensembl.org/Oryza_sativa/Info/Index], Arabidopsis thaliana [http://plants.ensembl.org/Arabidopsis_thaliana/Info/Index], and Vitis vinifera [http://plants.ensembl.org/Vitis_vinifera/Info/Index] are downloaded from EnsemblPlants database. The used genomes and annotations of Aquilegia coerulea are downloaded from Phytozone database [https://phytozone-nxt.jgi.doe.gov/info/Accessv3v1.3]. The used genomes and annotations of Maulea cordata are downloaded from the GenBank database under accession number GCA_002174775.1. Source data are provided with this paper.

**Code availability.** The ancestral genome reconstruction code has been deposited to https://github.com/XJTU-YeLab/TAG111 and https://doi.org/10.5281/zenodo.5528515. The analysis scripts have been deposited to https://github.com/XJTU-YeLab/Papaver-Genomics/tree/main/analysis_scripts112 and https://doi.org/10.5281/zenodo.5528517.

Received: 21 March 2021; Accepted: 29 September 2021; Published online: 15 October 2021
References

1. Field, P. E. & Osbourn, A. E. Metabolic diversification-independent assembly of benzylisoquinoline gene clusters in different plants. Science 320, 543–547 (2008).
2. Winzer, T. et al. A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science* 336, 1704–1708 (2012).
3. Itkin, M. et al. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* 341, 175–179 (2013).
4. Vedel-Persson, P. B., Xu, M., Jin, Y., Coates, R. M. & Peters, R. J. Identification of syn-pimara-7,15-diene synthase reveals functional clustering of terpene synthases involved in rice phytoalexin/aldehyde biosynthesis. *Plant Physiol.* 135, 2098–2105 (2004).
5. Jonczyk, R. et al. Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in maize: characterization of Bx6 and Bx7. *Plant Physiol.* 146, 1053–1063 (2008).
6. Xiong, X. et al. The Taxus genome provides insights into paclitaxel biosynthesis. *Nature Plants* 11, 728–732 (2015).
7. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. PlantSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* 45, W55–W63 (2017).
8. Topfer, M., Fuchs, L. M. & Aharoni, A. The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Res.* 45, 7049–7063 (2017).
9. Nutzmann, H. W., Scazzocchio, C. & Osbourn, A. Metabolic gene clusters in fungi. *Nat. Rev. Genet.* 13, 145–150 (2012).
10. Hurst, L. D., Pal, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* 5, 170–176 (2004).
11. Itkin, M. et al. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* 341, 175–179 (2013).
12. Winzer, T. et al. Metaphase of a metabolic gene cluster in *Papaver somniferum* L. and *Papaver setigerum* DC their hybrid. *Genetics* 31, 1–130 (1966).
13. Choe, S. et al. Species identification of *Papaver* by metabolic profiling. *Forensic Sci. Int.* 211, 51–60 (2011).
14. Asghari-Zarakia, R., Razmi, S., Madadi, R. & Fathi, M. Karyological study of the medicinal plant *Papaver rhoas* from northwest of Iran. *Afr. J. Biotechnol.* 10, 11173–11177 (2011).
15. Clark, M. C., Maldalina, A. & Ion, B. I. The study of mitotic chromosomes at *Papaver rhoas* 1 (2n = 14) species. *Analele titin ifice ale Universit. Cuza din Iaşi (serie nou), Sec iunea I, Genetic Biologie Molecular* 32, 117–118 (1988).
16. van der Steen, T. et al. A multiple GAL pathway genes clusters evolved independently and by different mechanisms in fungi. *Plant Cell* 11, 1705–1717 (1999).

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-021-26330-8 | www.nature.com/naturecommunications

© 2021 Nature Research Ltd. All rights reserved.
67. Louwers, M. et al. Tissue- and expression level-specific chromatin looping at maize b1 epialles. Plant Cell 21, 832–842 (2009).
68. Wick, R. R. & Holt, K. E. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Res 8, 2138 (2019).
69. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics 36, 2233–2255 (2020).
70. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36, 2896–2898 (2020).
71. Dudchenko, O. et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 356, 92–95 (2017).
72. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
73. Rao, W., Kojima, K. K. & Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mol. DNA 6, 11 (2015).
74. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35, W265–W268 (2007).
75. Ellingshaus, D., Kurtz, S. & Willhoft, O. LTRHarvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9, 18 (2008).
76. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 176, 1410–1422 (2018).
77. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. Curr. Protoc. Bioinformatics 48, 4 11–11 (2014).
78. Lamesch, P. et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40, D1202–D1210 (2012).
79. Dohm, J. C. et al. The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature 505, 546–549 (2014).
80. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652 (2011).
81. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics 27, 757–763 (2011).
82. Korf, I. Gene finding in novel genomes. BMC Bioinformatics 5, 59 (2004).
83. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33, 6494–6506 (2005).
84. Jones, P. et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240 (2014).
85. Nawrocki, E. P. Annotating functional RNAs in genomes using Infernal. Methods Mol. Biol. 1097, 163–197 (2014).
86. Nawrocki, E. P. et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 43, D130–D137 (2015).
87. Wang, Y. et al. McScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40, e49 (2012).
88. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature 436, 793–800 (2005).
89. Filiault, D. L. et al. The Aquilegia genus provides insights into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. Elife 7, https://doi.org/10.7554/eLife.36426 (2018).
90. Liu, X. et al. The Genome of medicinal plant Macleaya cordata provides new insights into bentsylosiounline alkaloids metabolism. Mol. Plant 10, 975–989 (2017).
91. Cheng, C. Y. et al. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 89, 789–804 (2017).
92. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780 (2013).
93. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552 (2000).
94. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690 (2006).
95. Kimura, M. Molecular evolutionary clock and the neutral theory. J. Mol. Evol. 26, 24–33 (1987).
96. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19, 301–302 (2003).
97. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs, Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteom. Bioinformatics 8, 77–80 (2010).
98. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J. 19(2), 30 (2017).
99. Pham, S. K. & Pevzner, P. A. DRIMM-Synten: decomposing genomes into evolutionary conserved segments. Bioinformatics 26, 2509–2516 (2010).
100. Tannier, E., Zheng, C. & Sankoff, D. Multichromosomal median and halving problems under different genomic distances. BMC Bioinformatics 10, 120 (2009).
101. Zheng, C., Zhu, Q., Adam, Z. & Sankoff, D. Guided genome halving: hardness, heuristics and the history of the Hemiascomycetes. Bioinformatics 24, i96–i104 (2008).
102. Feijao, P. & Medanis, J. SCI: a breakpoint-like distance that simplifies several rearrangement problems. IEEE/ACM Trans. Comput. Biol. Bioinformatics 8, 1318–1329 (2011).
103. Alekseyev, M. A. & Pevzner, P. A. Breakpoint graphs and ancestral genome reconstructions. Genome Res. 19, 943–957 (2009).
104. Ansalmetti, Y., Luhmann, N., Berard, S., Tannier, E. & Chauve, C. Comparative methods for reconstructing ancient genome organization. Methods Mol. Biol. 1704, 343–362 (2018).
105. Rai, A. et al. Chromosome-level genome assembly of Ophiopilzica punila reveals the evolution of campothecin biosynthesis. Nat. Commun. 12, 405 (2021).
106. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37, 907–915 (2019).
107. Pertea, M., Kim, D., Pertea, G. M., Lee, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. 11, 1650–1667 (2016).
108. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3, 95–98 (2016).
109. Kruse, K., Hug, C. B., Hernandez-Rodriguez, B. & Vaquerizas, J. M. TADtool: visual parameter identification for TAD-calling algorithms. Bioinformatics 32, 3190–3192 (2016).
110. Rao, S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680 (2014).
111. Yang, X. et al. Three chromosome-scale Papaver genomes reveal punctuated patchwork evolution of the morphinan and noscapine biosynthesis pathway. Zenodo https://doi.org/10.5281/zenodo.5528515 (2021).
112. Yang, X. et al. Three chromosome-scale Papaver genomes reveal punctuated patchwork evolution of the morphinan and noscapine biosynthesis pathway. Zenodo https://doi.org/10.5281/zenodo.5528517 (2021).
