Title: The opium poppy genome and morphinan production

Authors: Li Guo[1,2,3,†],Thilo Winzer[4†], Xiaofei Yang[1,5,†], Yi Li[4,†], Zemin Ning[6,†], Zhesi He[4], Roxana Teodor[4], Ying Lu[7], Tim A. Bowser[8], Ian A. Graham[4,*] and Kai Ye[1,2,3,9,*]

Affiliations:

[1]MOE Key Lab for Intelligent Networks & Networks Security, School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049 China.

[2]The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710061, China.

[3]The School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, 710049 China

[4]Centre for Novel Agricultural Products, Department of Biology, University of York, York, YO10 5YW, UK.

[5]Computer Science Department, School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049 China.

[6]The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

[7]Shanghai Ocean University, Shanghai, 201306, China

[8]Sun Pharmaceutical Industries Australia Pty Ltd, Princes Highway, Port Fairy, VIC 3284, Australia

[9]Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, China

*To whom correspondence should be addressed. E-mail: ian.graham@york.ac.uk; kaiye@xjtu.edu.cn

† These authors made an equal contribution

One Sentence Summary: The opium poppy genome reveals gene duplication, rearrangement and fusion events that led to a BIA-gene cluster for noscapine and morphine production.

Abstract: Morphinan-based painkillers are derived from opium poppy. We report a draft of the opium poppy genome, with 2.72Gb assembled into 11 chromosomes with contig N50 and scaffold N50 of 1.77Mb and 204Mb, respectively. Synteny analysis suggests a whole genome duplication at approximately 7.8 million years ago (MYA) and ancient segmental or whole genome duplication(s) that occurred before the Papaveraceae-Ranunculaceae divergence 110 MYA. Syntenic blocks representative of phthalideisoquinoline and morphinan components of a benzylisoquinoline alkaloid cluster of 15 genes provides insight into how it evolved. Paralog analysis identified P450 and oxidoreductase genes that combined to form the *STORR* gene fusion essential for morphinan biosynthesis in opium poppy. Thus gene duplication, rearrangement and fusion events have led to evolution of specialized metabolic products in opium poppy.

Main text: Throughout history opium poppy (*Papaver somniferum* L.) has been both a friend and foe of human civilization. In use since Neolithic times (*1*), the sap, known as opium, contains various alkaloids including morphine and codeine, with effects ranging from pain relief and cough suppression to euphoria, sleepiness and addiction. Opioid based analgesics remain among the most effective and cheap treatment for the relief of severe pain and palliative care but due to their addictive properties careful medical prescription is essential to avoid misuse. Access to morphine equivalents to alleviate serious health-related suffering is unequal: in the USA and Canada over 3000% of estimated need is met, in Western Europe 870%, China 16%, Russia 8%, India 4% and Nigeria 0.2% (*2*). Addressing the lack of access to pain relief or palliative care especially among poor people in low to middle income countries has been recognised as a global health and equity imperative (*2*).

Chemical synthesis or synthetic biology approaches are not as yet commercially viable for any of the morphinan subclass of benzylisoquinoline alkaloids (BIAs) (3, *4*, *5*) and opium poppy remains the only commercial source. Genome rearrangements have been important in the evolution of BIA metabolism in opium poppy. For example, a cluster of 10 genes encode enzymes for production of the antitussive and anticancer compound noscapine, which belongs to the phthalideisoquinoline subclass of BIAs (*6*) and a P450 oxidoreductase gene fusion (*4*, *7*, *8*) is responsible for the key gateway reaction directing metabolites towards the morphinan branch and away from the noscapine branch. Here we present the sequence assembly of the opium poppy genome to aid investigation into the evolution of BIA metabolism and provide a foundation for the further improvement of this medicinal plant.

Large complex plant genomes with an abundance of repeated sequences still pose challenges for genome analysis. Here we combined sequencing technologies (fig. S1), including Illumina Paired-End/Mate-Pair (214X), 10X Genomics linked reads (40X), PacBio (66.8X) and for quality checking, Oxford nanopore and Illumina sequencing of bacterial artificial chromosomes (Table 1; tables S1-S2). The final genome assembly of 2.72Gb covers 94.8% of the estimated genome size (fig. S2-S4; table S3) and 81.6% of sequences have been assigned into individual chromosomes (fig. S5; table S4) using a linkage map generated by sequence-based genotyping of 84 F2 plants (tables S5-S6). We annotated the genome using *MAKER* pipeline (*9*) incorporating protein homolog and transcriptome data from seven tissues (table S7). This predicted 51,213 protein-coding genes (Table 1; fig. S6). The annotation also predicted 9,494 non-coding RNAs (table S8). Benchmarking Universal Single-Copy Orthologs (BUSCO) (*10*) analysis based on plant gene models estimates 95.3% completeness (fig. S7). All predicted protein-coding genes are supported by RNA-seq data or homologs, while 68.8% have significant hits in the InterPro database (Table 1). Repetitive elements make up 70.9% of the genome. Of the repetitive elements, 45.8% are long terminal repeat (LTR) retrotransposons (fig. S8).

Synteny analysis revealed a relatively recent whole genome duplication (WGD) event and traces of what we consider to be ancient segmental duplications although a WGD cannot be ruled out (Fig. 1C; fig. S9-S14). Distribution of synonymous substitutions per synonymous site (*Ks*) for *P. somniferum* paralogous genes and syntenic blocks confirmed a major WGD peak (Fig. 1C; fig. S13A&D and S15; table S9) and a minor segmental duplication peak (Fig. 1C; fig. S13D). Intergenomic co-linearity analysis indicated *P. somniferum* did not experience γ, the hexaploidization event shared in core eudicots, as demonstrated by a 3:2 syntenic relationship between grape (*Vitis vinifera*) (*11*) and *P. somniferum* (fig. S9C&F). Syntenic blocks (greater than 132kb) account for 86% total coverage across the whole genome (Fig. 1A; tables S10-S13). Of the 25,744 genes arising from the WGD, 89.3% are present as two copies and 10.7% present in more than two copies (fig. S10). Gene ontology analysis suggests gene duplicates from the WGD are enriched with terms such as “cell redox homeostasis” and “positive regulation of transcription”(fig. S16). Comparison of *P.*

somniferum with an ancestral eudicot karyotype (AEK) genome (*12*) (Fig. 1B) and with grape also supported the *P. somniferum* WGD (fig. S9 and S11-S12). We used *OrthoFinder* (*13*) to identify 48 single-copy orthologs shared across 11 angiosperm species, and phylogenetic analysis of these using *BEAST* (*14*) indicated that *P. somniferum* diverged from *Aquilegia coerulea* (Ranunculaceae) and *Nelumbo nucifera* (Nelumbonaceae) at around 110 MYA and 125 MYA respectively (Fig. 1D). Using divergence time and mean *Ks* values of syntenic blocks between *P. somniferum* and *A. coerulea*, we estimated the synonymous substitutions per site per year as 6.98e-9 for Ranunculales, which led to the estimated time of the WGD at around 7.8 MYA (Fig. 1C; fig. S13 and S17; table S9). Applying a phylogenomic approach that traces the history of paralog pairs using phylogenetic trees (*15*), we constructed 95 rooted maximum likelihood trees containing a pair of opium poppy paralogs from under the *Ks* peak around 1.5 as anchors, and found 65% of the trees support segmental duplications originating from multiple events occurring before and 35% after the Papaveraceae-Ranunculaceae divergence at 110 MYA (table S14; Database S2 and S3). After applying a more stringent approach with a bootstrap threshold cutoff at 50% for ancient gene pairs (*15*), 21% of trees support duplication events occurring before the Papaveraceae-Ranunculaceae divergence. The *Ks* distributions of *A. coerulea* paralogs and syntenic genes support a WGD event in this representative of the Ranunculaceae at 111.3±35.6 MYA (Fig. 1C, table S9). A synteny dot plot of the opium poppy genome assembly with the *A. coerulea* genome assembly (16) revealed a 2:2 syntenic relationship (fig. S9D) suggesting both the opium poppy and *A. coerulea* WGD events happened after the Papaveraceae-Ranunculaceae divergence as shown in Fig 1D.

The genome assembly allowed us to locate all of the functionally characterised genes of BIA metabolism in opium poppy plus their closely related homologs, to either chromosomes or unplaced scaffold positions (table S15). The noscapine gene cluster occurs within a 584kb region on chromosome 11 along with the (*S*)- *to* (*R*)-*reticuline* (*STORR*) gene fusion plus the remaining four genes in the biosynthetic pathway to production of the morphinan alkaloid, thebaine (Fig. 2A&B). These genes are co-expressed in stems (Fig. 2C; fig. S18) and we refer to them as the BIA gene cluster. None of the other genes known to be associated with BIA metabolism, including *BERBERINE BRIDGE ENZYME* (*BBE*), *TETRAHYDROPROTOBERBERINE N-METHYLTRANSFERASE (TNMT)* and the bifurcated morphinan branch pathway genes, *CODEINE 3-O-DEMETHYLASE* (*CODM*), *THEBAINE 6-O-DEMETHYLASE* (*T6ODM*) and *CODEINONE REDUCTASE* (*COR*), are in a biosynthetic gene cluster (table S15). We used the *plantiSMASH* genome mining algorithm (*17*) to search the 11 chromosomes and unplaced scaffolds that encode annotated genes for additional gene clusters predicted to be associated with plant specialized metabolism. This approach detected the BIA gene cluster and a number of functionally uncharacterised genes across the same region, among which a cytochrome P450 (PS1126530.1) and a methyltransferase (PS1126590.1) exhibited a similar expression pattern as the 15 genes of BIA metabolism (Fig. 2A; table S16). Expression of genes immediately outside the 584kb BIA gene cluster boundaries was low in aerial tissue (table S16). These genes include *METHYLSTYLOPINE 14-HYDROXYLASE* (*MSH*), which is involved in the sanguinarine branch of BIA metabolism (*18*). *MSH* is expressed in root tissue together with other sanguinarine pathway genes, which are dispersed across the genome (table S15). The *plantiSMASH* algorithm also found 49 other possible gene clusters across the 11 chromosomes and 34 on unplaced scaffolds several of which show tissue specific expression patterns (table S16). Paralogs of the morphinan pathway genes *SALUTARIDINE SYNTHASE* (*SALSYN*), *SALUTARIDINE REDUCTASE* (*SALR*), *SALUTARIDINOL-7-O-ACETYL TRANSFERASE (SALAT)* and the recently discovered *THEBAINE SYNTHASE (19)* were identified on an unplaced scaffold in synteny with the BIA biosynthesis gene cluster on chromosome 11 (Fig. 2A; fig. S19 and S20). The expression pattern of these genes match those in the BIA cluster (Fig. 2C; table S16).

To investigate the evolutionary history of the BIA gene cluster we performed two rounds of synteny analysis with either the 'all BLASTp' result as input for blocks with distant homology or the default 'top 5 BLASTp' result for blocks with close homology using *MCScanX (20)*. We found the top ranked syntenic block with the noscapine branch genes has distant homology and is on chromosome 2 (E-value=8.1e-15, all BLASTp), while the top ranked syntenic block for morphinan pathway genes has close homology and is on unplaced scaffold 21 (E-value=0, top 5 BLASTp) (Fig. 2A; tables S17-18). *Ks* and amino acid identity of syntenic gene pairs (fig. S21; table S17) demonstrate that the syntenic block associated with the noscapine branch component of the BIA gene cluster is due to an ancient duplication with a median *Ks* value of 3.9 while the syntenic block associated with the morphinan branch component is due to a much more recent duplication occuring in the same timeframe as the WGD event at around 7.8 MYA.

The fusion event resulting in *STORR* was key for evolution of morphinan biosynthesis in the Papaveraceae (Fig. 2B) (*7*). Gene family analysis of P450 and reductase modules of *STORR* revealed their closest paralogs located 865bp apart on chromosome 2 (Fig. 2D; fig. S22). These paralogs have the same gene orientation and exon/intron boundaries as the *STORR* modules and based on this we propose that the *STORR* gene fusion involved an 865bp deletion following a duplication (Fig. 2D; table S19). *STORR* and its closest paralogs show amino acid sequence identity of 75% and 82% for the P450 and oxidoreductase modules respectively, which suggests the duplication leading to the *STORR* gene fusion occurred earlier than the WGD event (fig. S21).

From thebaine, the bifurcated morphinan branch giving rise to codeine and morphine requires three enzymatic reactions, two catalysed by the 2-oxoglutarate/Fe(II)- dependent dioxygenases, codeine O-demethylase (CODM) and thebaine 6-O-demethylase (T6ODM) and a third catalysed by codeinone reductase (COR; Fig. 2B). The genome assembly reveals that *CODM* and *T6ODM* are encoded by co-localised gene copies on chromosomes 1 and 2 respectively (Fig. 2E; table S19). Phylogenetic analysis shows that copies of both CODM and T6ODM share protein sequence identity greater than 97% whereas closest paralogs of CODM and T6ODM share 75.6% and 88.6% respectively (fig. S21 and S22; table S19). There are four copies of *COR*, two dispersed on chromosome 7 and two adjacent on unplaced scaffold 107 (Fig. 2E). Copies of COR share greater than 95% protein sequence identity and closest paralogs share ~74% (fig. S21 and S22; table S19). This closest paralog analysis indicates, as is the case with *STORR*, *T6ODM*, *CODM* and *COR* emerged before the WGD (fig. S21). Since T6ODM, CODM and COR use thebaine and downstream intermediates, we assume the ability to produce thebaine had evolved prior to WGD. The near sequence identity between the copies within each of the *T6ODM*, *CODM* and *COR* gene families indicates that the increase in copy number of these genes occurred more recently than the WGD event. Based on the above timing of events we speculate that the BIA gene cluster was assembled before the WGD event and following duplication underwent deletion of the noscapine component and STORR, leaving the morphinan component on unplaced scaffold 21.

The presence of genes exclusively associated with biosynthesis of both phthalideisoquinolines and morphinans in the BIA gene cluster implies a selection pressure favoring clustering of genes associated with these classes of alkaloids. *BBE* and *TNMT* functions are not exclusive for noscapine biosynthesis: both are required for sanguinarine biosynthesis which occurs predominantly in root rather than aerial tissues where noscapine and morphine biosynthesis occurs. Selective pressure on *BBE* and *TNMT* associated with their involvement in the biosynthesis of sanguinarine in root tissue may have kept them from being part of the BIA gene cluster even though they are also both expressed in stem tissue (Fig. 2C). Coordinate regulation of gene expression is considered to be part of the selective pressure resulting in gene cluster formation

(*21*). In opium poppy, the exclusivity of gene function and complexity of the gene expression pattern, could have determined which genes are clustered.

References :

1. S. Jacomet, Plant economy and village life in Neolithic lake dwellings at the time of the Alpine Iceman. Veget. Hist. Archaeobot. 18, 47-59 (2009).

2. F. M. Knaul et al., Alleviating the access abyss in palliative care and pain relief—an imperative of universal health coverage: the Lancet Commission report. Lancet, doi:10.1016/S0140-6736(17)32513-8 (2017).

3. M. Gates, G. Tschudi, The synthesis of morphine. J. Am. Chem. Soc. 78, 1380-1393 (1956).

4. S. Galanie, K. Thodey, I. J. Trenchard, M. F. Interrante, C. D. Smolke, Complete biosynthesis of opioids in yeast. Science 349, 1095-1100 (2015).

5. A. Nakagawa et al., Total biosynthesis of opiates by stepwise fermentation using engineered Escherichia coli. Nature Commun. 7, 10390 (2016), doi: 10.1038/ncomms10390.

6. T. Winzer et al., A Papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noscapine. Science 336, 1704-1708 (2012).

7. T. Winzer et al., Morphinan biosynthesis in opium poppy requires a P450-oxidoreductase fusion protein. Science 349, 309-312 (2015).

8. S. C. Farrow, J. M. Hagel, G. A. Beaudoin, D. C. Burns, P. J. Facchini, Stereochemical inversion of (S)-reticuline by a cytochrome P450 fusion in opium poppy. Nature Chem. Biol. 11, 728-732 (2015).

9. M. S. Campbell, C. Holt, B. Moore, M. Yandell, Genome annotation and curation using MAKER and MAKER-P. Curr. Protoc. Bioinform. 48, 4.11.1-4.11.39. (2014).

10. F. A. Simao, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinform. 31, 3210-3212 (2015).

11. O. Jaillon et.al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463-467 (2007).

12. F. Murat, A. Armero, C. Pont, C. Klopp, J. Salse, Reconstructing the genome of the most recent common ancestor of flowering plants. Nature Genet. 49, 490-496 (2017).

13. D. M. Emms, S. Kelly, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16, 157 (2015).

14 . A. J. Drummond, M. A. Suchard, D. Xie, A. Rambaut, Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29, 1969-1973 (2012).

15. Y. Jiao et al., A genome triplication associated with early diversification of the core eudicots. Genome Biol. 13, R3, doi:10.1186/gb-2012-13-1-r3 (2012).

16. The *Aquilegia coerulea* v3.1 genome data (released August 8th, 2014 http://phytozome.jgi.doe.gov/) was produced by the US Department of Energy Joint Genome Institute.

17. S. A. Kautsar, H. G. Suarez Duran, K. Blin, A. Osbourn, M. H. Medema, plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res. 45, W55–W63 (2017).

18. G. A. W. Beaudoin, P. J. Facchini, Isolation and characterization of a cDNA encoding (S)-cis-N-methylstylopine 14-hydroxylase from opium poppy, a key enzyme in sanguinarine biosynthesis. Biochem. Biophys. Res. Comm. 431, 597-603 (2013).

19. X. Chen et al., A pathogenesis-related 10 protein catalyzes the final step in thebaine biosynthesis. Nature Chem. Biol. 14, 738-743 (2018).

20. Y. Wang et al., MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40, e49 (2012).

21. A. Osbourn, Gene clusters for secondary metabolic pathways: an emerging theme in plant biology. Plant Physiol. 154, 531-535 (2010).

22. R. Magnavaca, C. O. Gardner, R. B. Clark, Evaluation of inbred maize lines for aluminum tolerance in nutrient solution, in Genetic Aspects of Plant Mineral Nutrition. Developments in Plant and Soil Sciences, W. H. Gabelman, B. C. Loughman, Eds. (Springer, Dordrecht, 1987), vol. 27, pp. 255-265.

23. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. Nature Methods. 12, 357-360 (2015).

24. M. Pertea et al., StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 33, 290-295 (2015).

25. A. C. Frazee et al., Flexible analysis of transcriptome assemblies with Ballgown. bioRxiv, (2014), (available at https://www.biorxiv.org/content/early/2014/09/05/003665).

26. M. G. Grabherr et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnol. 29, 644-652 (2011).

27. M. J. Raymond, Isolation and characterization of latex-specific promoters from Papaver somniferum. Master Thesis, Virginia Polytechnic Institute and State University, (2004).

28. S. Koren et al., Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27, 722-736 (2017).

29. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754-1760 (2009).

30. E. P. Murchison et al., Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. Cell 148, 780-791 (2012).

31. N. I. Weisenfeld, V. Kumar, P. Shah, D. M. Church, D. B. Jaffe, Direct determination of diploid genome sequences. Genome Res. 27, 757-767 (2017).

32. J.-S. Seo et al., De novo assembly and phasing of a Korean human genome. Nature 538, 243-247 (2016).

33. A. M. Hulse-Kemp et al., Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. Hortic. Res. 5, 4 (2018), doi: 10.1038/s41438-017-0011-0. eCollection 2018..

34. Y. Mostovoy et al., A hybrid approach for de novo human genome sequence assembly and phasing. Nature methods 13, 587-590 (2016).

35. R. Avni et al., Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357, 93-97 (2017).

36. M. C. Luo et al., Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. Nature 551, 498-502 (2017).

37. G. Zhao et al., The Aegilops tauschii genome reveals multiple impacts of transposons. Nature Plants. 3, 946-955 (2017).

38. C. N. Hirsch et al., Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell. 28, 2700-2714 (2016).

39. F. Lu et al., High-resolution genetic mapping of maize pan-genome sequence anchors. Nature Commun. 6, 6914 (2015), doi: 10.1038/ncomms7914.

40. H. Tang et al., ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. 16, 3 (2015), doi: 10.1186/s13059-014-0573-1.

41. C.-S. Chin et al., Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods. 13, 1050-1054 (2016).

42. K.-P. Koepfli, B. Paten, G. K. C. o. Scientists, S. J. O'Brien, The Genome 10K Project: a way forward. Annu. Rev. Anim. Biosci. 3, 57-111 (2015).

43. C.-L. Xiao et al., MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nature Methods. 14, 1072-1074 (2017).

44. A. McKenna et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).

45. J. K. Bonfield, A. Whitwham, Gap5—editing the billion fragment sequence assembly. Bioinformatics 26, 1699-1703 (2010).

46. W. Bao, K. K. Kojima, O. Kohany, Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6, 11 (2015), doi: 10.1186/s13100-015-0041-9.

47. G. S. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. BMC bioinformatics 6, 31 (2005), doi: 10.1186/1471-2105-6-31.

48. O. Keller, M. Kollmar, M. Stanke, S. Waack, A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics 27, 757-763 (2011).

49. I. Korf, Gene finding in novel genomes. BMC bioinformatics 5, 59 (2004).

50. A. Lomsadze, V. Ter-Hovhannisyan, Y. O. Chernoff, M. Borodovsky, Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33, 6494-6506 (2005).

51. P. Lamesch et al., The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40, D1202-1210 (2012).

52. J. C. Dohm et al., The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature 505, 546-549 (2014).

53. P. Jones et al., InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236-1240 (2014).

54. A. Conesa, S. Gotz, Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int. J Plant. Genomics. 2008, 619832 (2008), doi: 10.1155/2008/619832.

55. T. M. Lowe, S. R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955-964 (1997).

56. P. P. Gardner et al., Rfam: updates to the RNA families database. Nucleic Acids Res. 37, D136-140 (2009).

57. E. P. Nawrocki, S. R. Eddy, Query-dependent banding (QDB) for faster RNA similarity searches. PLoS Comput. Biol. 3, e56 (2007), doi: 10.1371/journal.pcbi.0030056.

58. J. A. Bailey, D. M. Church, M. Ventura, M. Rocchi, E. E. Eichler, Analysis of segmental duplications and genome assembly in the mouse. Genome Res. 14, 789-801 (2004).

59. N.H. Putnam et al., The amphioxus genome and the evolution of the chordate karyotype. Nature 453, 1064-1071 (2008).

60. D. Wang, Y. Zhang, Z. Zhang, J. Zhu, J. Yu, KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genom. Proteom. Bioinformatic. 8, 77-80 (2010).

61. R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792-1797 (2004).

62. S. Kumar, G. Stecher, K. Tamura, MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870-1874 (2016).

63. A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690 (2006).

64. L. Scrucca, M. Fop, T. B. Murphy, A. E. Raftery, Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R Journal 8, 289-317 (2016).

65. R. Ming et al., Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.), Genome Biol. 14, R41 (2013).

66. Y. Jiao, J. Li, H. Tang, A. H. Paterson, Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in Monocots. The Plant Cell 26, 2792-2802 (2014).

67. Y. Van de Peer, E. Mizrachi, K. Marchal, The evolutionary significance of polyploidy. Nature Reviews Genetics 18, 411 (2017).

68. H. Badouin et al., The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature 546, 148 (2017).

69. S. Sato et al., The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635 (2012).

70. L. Li, C.J. Stoeckert, D.S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178-2189 (2003).

71. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059-3066 (2002).

72. S. Capella-Gutierrez, J.M. Silla-Martinez, T. Gabaldon, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972-1973 (2009).

73. J. D. Thompson, T. Gibson, D. G. Higgins, Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinformatics. Chapter 2:Unit 2.3. (2002).

74. J. Castresana, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol.17, 540-552 (2000).

75. A.M. Bolger, M. Lohse, B. Usadel. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).

76. M. Bastian, S. Heymann, M. Jacomy. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media (2009), http://gephi.org.

Supplementary Materials

Materials and Methods

Tables S1 – S19

Figs S1 – S23

Databases S1 – S3

References (22 – 76)

Acknowledgements: The authors thank Yuannian Jiao and Lingling Chen for helpful discussions on genome analysis and Jing Hai, Judith Mitchell, Samantha Donninger and Jennifer Hudson for administrative and technical support. This work involved the use of the York Bioscience Technology Facility. Funding: KY, LG, XY are supported by National Science Foundation of China (31671372, 31701739, 61702406), by National Key R&D Program of China (2017YFC0907500), by "the Thousand Talents Plan" and by Key Construction Program of the National "985" Project. ZN is funded by the Wellcome Trust (WT098051); IAG received support from the UK Biotechnology and Biological Sciences Research Council (BB/K018809/1) and The Garfield Weston Foundation. **Author contributions:** KY, ZN and IAG designed and supervised research. TW, RT, TB and YL collected materials for sequencing and generated transcriptome data and linkage map. ZN performed the genome assembly. XY, LG, Ying Lu, TW and YL performed genome annotation and synteny analysis. LG, XY, YL and TW performed transcriptome and phylogenomic analysis. TW, YL, ZH, XY, LG, KY and IAG interpreted evolution of BIA metabolism. XY, LG, ZN, YL, TW prepared figures and tables. IAG, LG, TW, XY, YL, ZN and KY wrote the paper. **Competing interests:** None. **Data and materials availability:** All raw reads and assembled sequence data have been uploaded to NCBI under BioProjectID: PRJNA435796. The Whole Genome Shotgun project has been deposited under the accession number PUWZ00000000. The nine transcriptome datasets have accession numbers GSM3022361-9 and the fifteen BAC sequences accession numbers MG995579-93. HN1 is available from the University of York on behalf of Sun Pharmaceutical Industries Australia Pty Ltd for experimental purposes, subject to a material transfer agreement.

Table 1. Assembly and annotation statistics of opium poppy genome

Contig Assembly	Total Number of Contigs	65,578
	Assembly size	2.71Gb
	N50	1.77Mb
	N90	590kb
	Largest Contig	13.8Mb
Scaffold Assembly	Total Number of Scaffolds	34,388
	Assembly size	2.72Gb
	N50	204.5Mb
	N90	9.9Mb
	Largest Scaffold	270.4Mb
Annotation	GC content	30.5%
	Repeat density	70.9%
	Number of protein-coding genes	51,213
	Average length of protein-coding genes	3,454bp
	Supported by RNA-seq or homologs	100%
	Supported by Protein families	68.8%

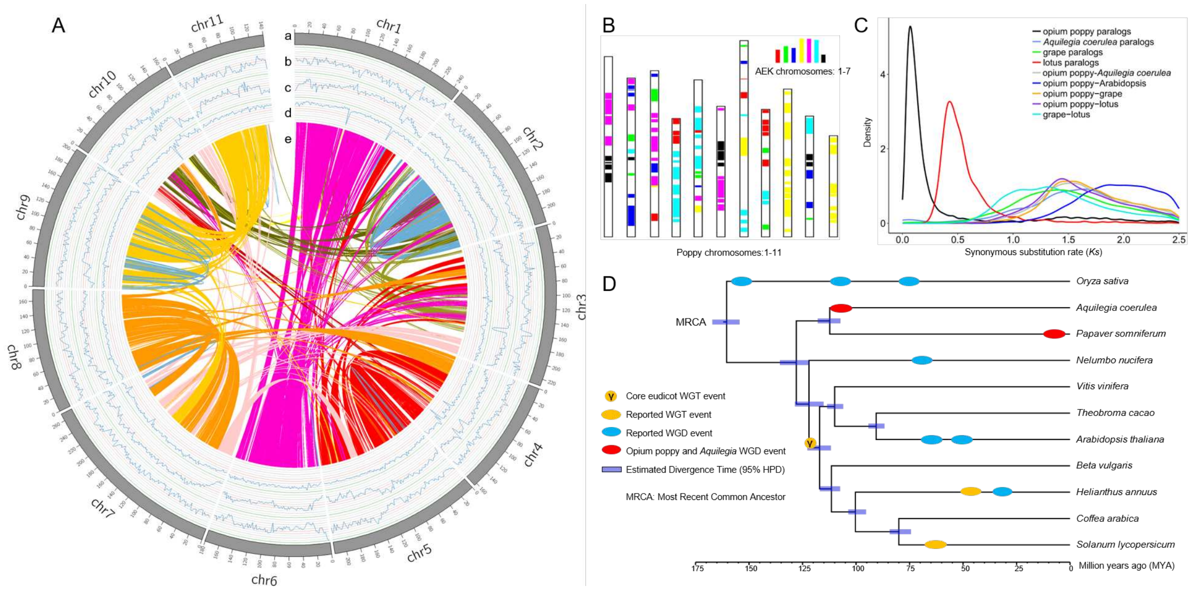


Fig. 1. Opium poppy genome features and whole genome duplication.

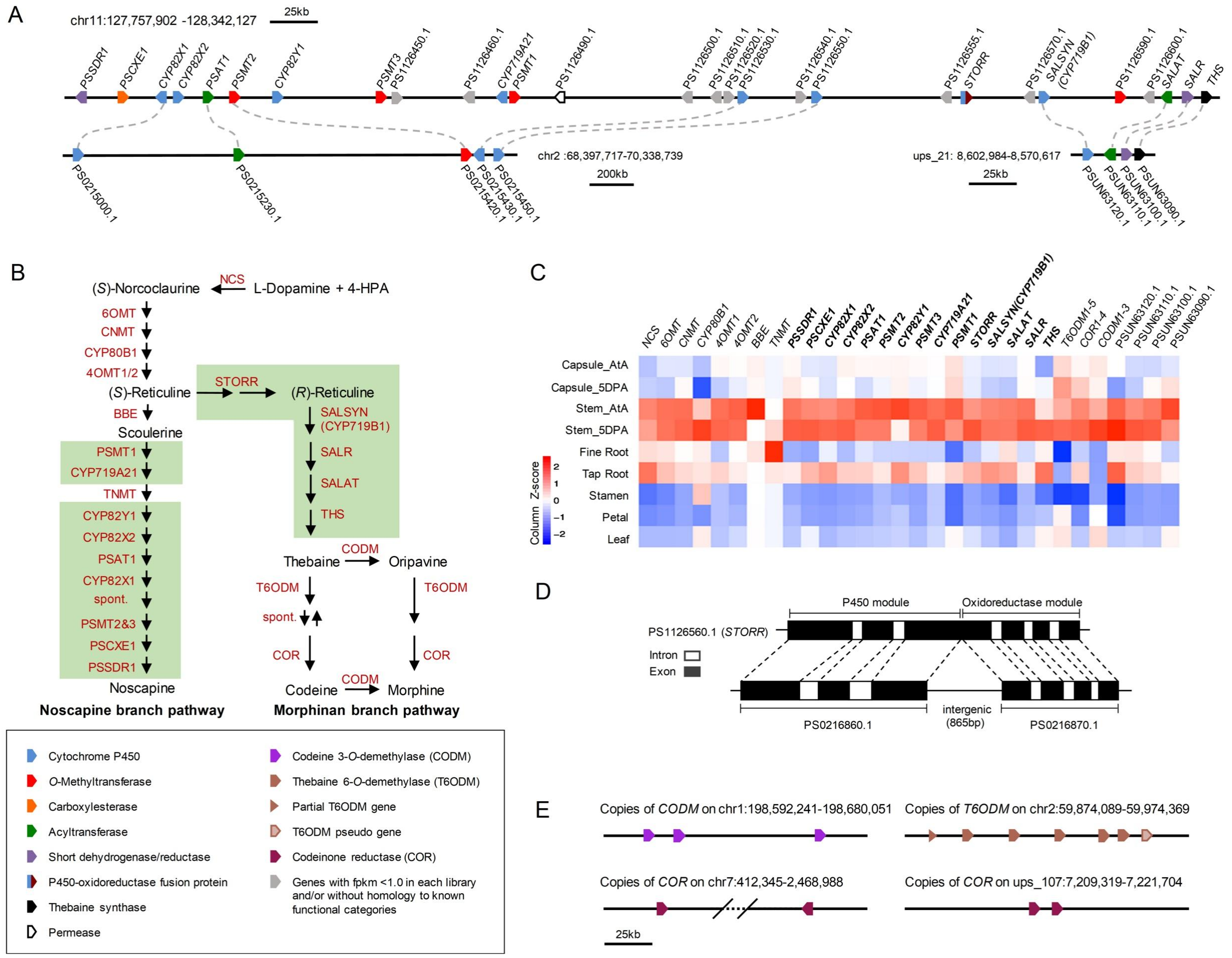


Fig. 2. Genomic arrangement of key genes of BIA metabolism.

Figure Legends

Fig. 1. Opium poppy genome features and whole genome duplication. (**A**) Characteristics of the eleven chromosomes of *Papaver somniferum*. Track **a-c** are the distribution of gene density, repeat density and GC density respectively, with densities calculated in 2Mb windows. Track **d** shows syntenic blocks. Band width is proportional to syntenic block size. (**B**) Comparison with ancestral eudicot karyotype (AEK) chromosomes reveals synteny. The syntenic AEK blocks are painted onto *P. somniferum* chromosomes. (**C**) Synonymous substitution rate (*Ks*) distributions of syntenic blocks for *P. somniferum* paralogs and orthologs with other eudicots are shown in colored lines as indicated. (**D**) Inferred phylogenetic tree with 48 single-copy orthologs of eleven species identified by *OrthoFinder* (*13*). Posterior probabilities for all branches exceed 0.99. Timing of *P. somniferum* whole genome duplication (WGD) was estimated in this study and other reported whole genome triplication (WGT)/WGD timings are superimposed on the tree. Divergence timings are estimated using *BEAST* (*14*) and indicated by light blue bars at the internodes with 95% highest posterior density (HPD).

Fig. 2. Genomic arrangement of key genes of BIA metabolism. (**A**) Arrangement and chromosomal position as indicated of the 584kb BIA gene cluster on chromosome 11 (chr11) encoding fifteen co-expressed genes involved in noscapine and morphine biosynthesis (cluster 49, table S16). Below the BIA gene cluster is shown a syntenic block from chr2 (clusters 11 and 12, table S16) associated with the noscapine component of the cluster and a syntenic block from unplaced scaffold 21 associated with the morphinan component (cluster 70, table S16). Syntenic gene pairs are indicated by dashed lines. (**B**) Schematic representation of noscapine and morphinan branch pathways with the reactions associated with BIA gene cluster highlighted in green boxes. ‘spont.’ indicates spontaneous reactions. 4-HPA: 4-hydroxyphenylacetaldehyde. Gene abbreviations are defined in table S15. (**C**) Tissue-specific expression of noscapine and morphinan biosynthesis genes across different tissues. AtA – At Anthesis, 5DPA – 5 Days Post Anthesis. BIA pathway genes and their expression values (converted to *Z*-scores) across different tissues are visualized as a heatmap. Genes located in the BIA gene cluster are shown in bold. (**D**) Schematic structure of *STORR* on chr11 and the genomic region on chr2 containing its closest paralogs corresponding to the P450 and reductase modules. Dashed lines denote exon/intron boundaries. (**E**) Arrangement of locally duplicated copies of *CODEINE 3-O-DEMETHYLASE* (*CODM*), *THEBAINE 6-O-DEMETHYLASE* (*T6ODM*) and *CODEINONE REDUCTASE* (*COR*) genes (other annotated genes in the associated genomic regions are not shown).

Science
AAAS

Supplementary Materials for

The opium poppy genome and morphinan production

Li Guo, Thilo Winzer, Xiaofei Yang, Yi Li, Zemin Ning, Zhesi He, Roxana Teodor, Ying Lu, Tim A. Bowser, Ian A. Graham,* and Kai Ye*

correspondence to: ian.graham@york.ac.uk; kaiye@xjtu.edu.cn

This PDF file includes:

Other Supplementary Materials for this manuscript includes the following:

Materials and Methods

1. Species variety and plant materials

For sequencing and assembly of the opium poppy reference genome, the proprietary Sun Pharmaceutical Industries Australia Pty Ltd opium poppy variety High Noscapine 1 (HN1) was chosen, which accumulates noscapine in addition to morphine (*6*). HN1 material was grown in Maxi (Fleet) Rootrainers^{TM} (Haxnicks, Mere, UK) under glass in 16 hour days at the University of York horticulture facilities. The growth substrate consisted of 4 parts John Innes No. 2, 1 part Perlite and 2 parts Vermiculite.

After two cycles of self-pollination one individual plant was selected to prepare DNA from leaves for Illumina paired-end and mate-pair sequencing. Subsequently, this selected plant was self-pollinated and its progeny grown to obtain fine leaf material for Ultra High Molecular Weight grade and Next Generation Sequencing grade DNA preparations performed by Amplicon Express (Pullman, WA, USA - see section 2.2). This DNA was used for 10X Genomics and single molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio). For this the seed was sown into trays filled with the growth substrate described above. Two weeks after sowing, the germinating seedlings were transferred into the dark for 24 hours prior to harvesting young emerging leaves.

For RNA sampling, self-pollinated progeny of the HN1 plant selected for genome sequencing was grown in Maxi Rootrainers^{TM} as described above except for the growth substrate, which consisted of 50% sand and 50% Terragreen (Oil-Dri Ltd, Wisbech, UK) to allow easy access to clean root material.The plants were watered daily with a modified Magnavaca nutrient solution (*22*). Concentrations of nutrients in the solution were 1 mM KCl, 1.5 mM NH_4NO_3, 1.5 mM $CaCl_2$,, 50 µM KH_2PO_4, 200 µM $MgSO_4$, 500 µM $Mg(NO_3)_2$, 155 µM $MgCl_2$, 8.26 µM $MnCl_2$, 23.1 µM H_3BO_3, 2.14 µM $ZnSO_4$, 0.56 µM $CuSO_4$, 0.75 µM Na_2MoO_2, 77 µM Fe-HEDTA (ferric hydroxyethyl-ethylene diaminetriacetate). The pH of the nutrient solution was adjusted to 5.5 using sodium hydroxide.

On the first day of anthesis material was sampled from the following seven tissue types using a subset of plants: fine roots, tap root, leaves (the two uppermost ones), stem (the 2 cm long part just underneath the capsule), capsule, petals and stamens. Five days after the onset of anthesis stem and capsule materials were collected from another subset of plants. These plants had been manually pollinated on the first day of anthesis and had shed their petals at the time of sampling.

2. DNA and RNA isolation

2.1 Preparation of genomic DNA for Illumina paired-end read and mate-pair sequencing

A number of young leaves (30-50 mgs each) were collected from the same HN1 individual selected for sequencing of the genome. Genomic DNA was extracted using the BioSprint 96 Plant Kit on the BioSprint 96 Workstation (Qiagen, Crawley, UK) according to the manufacturer's protocol. Extracted DNA was quantified using the Qubit 3.0 Fluorometer (Thermo Fisher Scientific) according to the manufacturer's protocol.

2.2 Preparation of High Molecular Weight grade genomic DNA for 10X Genomics linked reads and PacBio long-read sequencing

High Molecular Weight (HMW) DNA was prepared from young seedling material described above by Amplicon Express (Pullman, WA, USA) using their proprietary protocol for HMW grade (megabase size) DNA preparation. This protocol involves isolation of plant nuclei and yielded pure HMW DNA (fig. S23A). This DNA was used to obtain 10X Genomics linked reads described below. Genomic DNA suitable for PacBio long-read sequencing was prepared from young seedling material described above by Amplicon Express using their proprietary Next Generation Sequencing grade DNA isolation protocol (fig. S23B).

2.3 RNA isolation for transcriptome sequencing

Samples were harvested into liquid nitrogen, then stored at -80°C until extraction. Grinding was performed in jars chilled in liquid nitrogen on the Qiagen TissueLyser as follows: 15 seconds at 20Hz, followed by re-chilling in liquid nitrogen, then a further 15 seconds at 20Hz. 100mg of ground material was used per RNA extraction. RNA was prepared using the Direct-zol RNA Miniprep Kit (Zymo Research, USA) according to the manufacturer's instructions.

The RNA was quantified using a Nanodrop® 2000 spectrophotometer (Thermo Fisher Scientific, UK). Equal amounts of RNA from 3-4 samples were pooled by tissue type to yield pooled samples of 1μg total RNA. RNA quality was assessed by running 1 μl of each pooled sample on a RNA Nano Chip on an Agilent Bioanalyzer 2100 (Agilent Technologies, INC).

3. Genome sequencing

To achieve a high-quality opium poppy genome assembly, we adopted a combination of sequencing methods including Illumina paired-end and mate-pair sequencing, 10X Genomics linked reads, single molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio), and Oxford Nanopore sequencing Technology (ONT). As details in table S1, a total of ~685Gb sequencing data (equivalent to 239× genomic coverage, based on an estimated genome size of 2.87Gb) was generated.

Five size-selected genomic DNA libraries ranging from 470bp to 10kb were constructed for each material. One shotgun library (Paired-End or PE) was made using DNA template fragments size-selected at ~470bp with no PCR amplification (PCR-free). This fragment size was designed to produce a sequencing overlap of the fragments to be sequenced on the Hiseq2500 v2 Rapid mode as 2×265bp, thus creating an opportunity to generate 'stitched' reads of approximately 265bp to 520bp in length. One genomic library of 800bp DNA fragment sizes was prepared using the TruSeq DNA Sample Preparation Kit version 2 with no PCR amplification (PCR-free) following the manufacturer's protocol (Illumina, San Diego, CA). To increase sequence diversity and genome coverage, three separate Mate-Pair (MP) libraries were constructed with 2-5kb, 5-7kb and 7-10kb jumps using the Illumina Nextera Mate-Pair Sample Preparation Kit (Illumina, San Diego, CA). The 800bp shotgun library was sequenced on an Illumina HiSeq2500 as 2x160bp reads (using the v4 Illumina chemistry) while the MP libraries were sequenced on HiSeq4000 as 2X150bp reads. PE and MP libraries construction and sequencing were conducted at Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign.

In addition, DNA fragments longer than 50kb were used to construct one Gemcode library using the Chromium instrument (10X Genomics, Pleasanton, CA). This library was sequenced on HiSeqX platform to produce 2X150bp reads, producing a total of ~128Gb of 10X Chromium library sequencing data. The 10X Chromium library construction and sequencing were conducted at HudsonAlpha Institute for Biotechnology, Huntsville, Alabama. Furthermore, we constructed the 20kb PacBio libraries using BluePippin^{TM} Size-Selection System recommended by Pacific Biosciences. In total, 9µg DNA was sheared to ~20kb targeted size using ultrasonication (Covaris, Woburn, Massachusetts, USA), and finally 6µg DNA was retained to construct the libraries. The quality of shearing processed DNA was examined by FEMTO Pulse pulse field capillary electrophoresis (Advanced Analytical Technologies, Inc.). The sheared DNA was filtered by AMPure PB paramagnetic beads (Beckman Coulter Inc.) with a recovery rate of 80%. The constructed libraries were sequenced by Sequel system in Novogene (Tianjing, China), and a total of 52 SMRT cells were used to yield ~192Gb sequencing data, including 24.1 million clean subreads with an average length of 7.98kb and an N50 of 11.84kb (table S1).

To facilitate genome annotation, we performed RNA sequencing of seven different opium poppy tissues (leaf, petal, stamen, capsule, stem, fine root, tap root). We used 400ng high quality total RNA per pooled sample for mRNA sequencing library preparation. NEBNext® RNA Ultra Directional Library preparation kit for Illumina, NEBNext® Poly(A) mRNA Magnetic Isolation Module (New England BioLabs Inc.,Ipswich, MA), and NEBNext® single 6bp indexing primers, were used according to the manufacturer's instructions. Libraries were pooled at equimolar ratios, and the pool was sent for 2 x 150 base paired-end sequencing on one lane of a HiSeq 3000 system at the University of Leeds Next Generation Sequencing Facility (Leeds, UK). An average of ~26Gb PE reads sequencing data were generated for each tissues. RNAseq analysis followed an in-house pipeline chaining *Hisat2 (23), Stringtie (24)* and *Ballgown (25)* software. Basically, quality-checked RNA-seq reads were aligned to the genome assembly using *Hisat2*, followed by transcript discovery and transcript abundance estimation using *StringTie* and *Ballgown*. In addition, we used *Trinity* v2.1.1 *(26)* for *de nov*o transcriptome assembly and generated EST evidence for gene prediction.

4. BAC library screening, BAC clone sequencing and assembly

The preparation of the HN1 Bacterial Artificial Chromosome (BAC) library was described previously *(6)*. The library was screened for *CODM* containing clones using a 703bp fragment located in the CODM promoter region (*27*). This fragment was amplified with primers AAAATCCGCCCTCCATGC (forward) and CCGACTTTGGCCCACTTGT (reverse) using a PCR digoxigenin (DIG) synthesis kit (Roche Applied Science, Indianapolis, IN) according to the manufacturer's instruction to obtain the DIG-labelled screening probe. Screening of the BAC library was performed by Amplicon Express (Pullman, WA, USA) as described previously (*6*), and resulted in the identification of 10 CODM containing BAC clones, namely BAC33_D07, BAC86_F04, BAC89_C05, BAC109_H06, BAC129_K11, BAC152_G13, BAC158_A11, BAC185_L02, BAC195_N12 and BAC230_D02.

For screening the BAC library for *T6ODM* containing clones, a screening probe was generated as described above using primers CCGAGATTAAGGGTATGTCAGAGG (forward) and CACAAGATCCCCATATGTATATCCAC (reverse). Amplification with these primers generate screening probe fragments between 502 to 526bp (depending on the T6ODM gene copy amplified) corresponding to the 3' end of the gene copies. Five T6ODM containing BAC clones were identified, namely BAC30_E04, BAC70_J09, BAC70_P15, BAC81_K11, BAC127_B22.

Each BAC was sequenced on two sequencing platforms: Paired-end sequencing was performed on the Illumina HiSeq platform. In addition, each BAC clone was sequenced using a MinION sequencer (Oxford Nanopore Technologies Ltd, Oxford, UK). Purified BACs were minimally fragmented using 20 second treatments with NEBNext® dsDNA Fragmentase® (New England Biolabs Inc., Ipswich, MA), and sequencing libraries prepared using Oxford Nanopore Technologies sequencing kit SQK007 with native barcoding expansion pack EXP-NDB002. Briefly, single-stranded nicks in DNA were repaired using NEBNext® FFPE DNA repair mix prior to end-repair and dA tailing using the NEBNext® Ultra™ II End Repair/dA-Tailing Module. Unique barcode sequences were ligated onto fragments for each BAC, before pooling 3-4 BACs per library. Sequencing adapters (including a hairpin adapter to allow for 2D sequencing) were ligated onto the ends of fragments, along with an adapter-binding tether protein. Fragments where tether protein was bound were purified using MyOne™ C1 streptavidin beads (Invitrogen/Thermo Fisher Scientific, UK). Libraries were then run on MinION R9 flow-cells using a 48 hour sequencing protocol, and base calling and demultiplexing was performed using Metrichor's EPI2ME platform (Oxford Nanopore Technologies, Oxford UK). Reads that passed 2D quality checks were split according to barcode for further analysis.

Raw MinION reads of each BAC clone were assembled with CANU v1.3 software (*28*) and insert boundaries were identified with the positions of vector sequences and cross-reference of overlapping BAC clones. These initial assemblies were further corrected with the NANOPOLISH software (https://github.com/jts/nanopolish). The high quality Illumina paired-end short reads were then mapped with BWA software (*29*) to the resulting MinION assemblies, which were used as the BAC scaffolding reference. Consensus references were generated with the mapped alignment and indels were corrected according to the variation analyses of the alignment. The corrected consensus was then used as a new BAC reference in an iterative process until no further improvement could be achieved. This assembly method gave a 99.2 % base identity to a previously assembled and published BAC sequence (BAC164_F07, gene bank accession: JQ659012). The overlapping BACs were then combined to give a continuous 426kb genomic fragment containing the *CODM* copies and a 227kb genomic fragment containing the *T6ODM* copies.

5. Genome assembly

5.1 Genome size estimation

We estimated the genome size on opium poppy using kmer frequency analysis with a kmer size of 61, following the method described in Murchison *et al.* (2012) (*30*).

Fig. S2 shows the frequency against kmer occurrence and a peak value of 17 is observed. Based on this set of frequency data, the genome size is estimated as 2.87Gb.

$G_s = (K_n – K_s)/D = 2.87x10^9$

where $K_n = 48.74x10^9$ – Total number of kmer words;

$K_s = 185x10^6$ - Number of single copy kmer words;

$D = 17$ - Depth of kmer occurrence.

5.2 Genome assembly strategy

Given the challenges in this large and complex plant genome, we adopted a hybrid assembly strategy in the project. The third generation sequencing platform PacBio provides long reads to span repeat-rich genomic regions and ensures longer sequence continuity. In the downstream analysis such as genome annotations, contiguous scaffolding is also essential to capture the whole gene structure. Genome scaffolding relies on long DNA fragments and in recent years application of barcoded linked reads from the 10X genomics platform have begun to replace earlier scaffolding methods that relied on mate-pair data. The high molecular weight (HMW) DNAs offer long fragments up to 1Mb length (*30)*, which help in producing a number of high-quality and contiguous assemblies (*31-39*). Here we present two independent *de novo* assemblies: NRgene 10X and PacBio *Falcon*. We then describe a method to merge the assemblies in order to achieve a high level of sequence continuity for both contigs and scaffolds.

5.3 The NRgene 10X assembly

Genome assembly was conducted using DeNovoMAGIC^{TM} software platform (NRGene, Nes Ziona, Israel). This is a *De Bruijn*-graph-based assembler, designed to efficiently extract the underlying information in the raw Illumina paired-end and mate-pair reads to solve the complexity of the *De Bruijn* graph due to genome polyploidy, heterozygosity and repetitiveness. This task is accomplished using accurate-reads-based traveling in the graph that iteratively connected consecutive phased contigs over local repeats to generate long phased scaffolds (*31-39*). The additional raw Chromium 10X data was utilized to phase polyploidy/heterozygosity, support scaffolds validation and further elongation of the phased scaffolds. Assembly results were summarized in table S1.

In brief, the algorithm is composed of the following steps:

1) *Read pre-processing*. PCR duplicates, Illumina adaptor AGATCGGAAGAGC and Nextera linkers (for MP libraries) were removed. The overlapping reads of the paired-end 450bp 2×265bp libraries were merged with minimal required overlap of 10bp to create the stitched reads.

2) *Error correction*. Following pre-processing, merged paired-end reads were scanned to detect and filter reads with putative sequencing error (contain a sub-sequence that does not re-appear several times in other reads).

3) *Contigs assembly*. The first step of the assembly consists of building a *De Bruijn* graph (kmer=127bp) of contigs from the all paired-end and mate-pair reads. Next, paired-end reads were used to find reliable paths in the graph between contigs for repeat resolving and contigs extension. 10X barcoded reads were mapped to contigs ensure that adjacent contigs were connected only in case there is an evidence that those contigs originate from a single stretch of genomic sequence (reads from the same two or more barcodes were mapped to both contigs).

4) *Scaffolds assembly*. Later, contigs were linked into scaffolds with paired-end and mate-pair information, estimating gaps between the contigs according to the distance of paired-end and mate-pair links. In addition, 10X data was used to validate and support correct phasing during scaffolding.

5) *Fill Gaps*. A final fill-gap step used paired-end and mate-pair links and *De Bruijn* graph information to detect a unique path connecting the gap edges.

6) *Scaffolds elongation and refinement*. 10X barcoded reads were mapped to the assembled scaffolds and clusters of reads with the same barcode mapped to adjacent contigs in the scaffolds were identified to be part of a single long molecule. Next, each

scaffold was scanned with a 20kb length window to ensure that the number of distinct clusters that cover the entire window (indicating a support for this 20kb connection by several long molecules) was statistically significant with respect to the number of clusters that span the left and the right edge of the window. In case where a potential scaffold assembly error was detected the scaffold was broken at the two edges of the suspicious 20kb window. Finally, the barcodes that were mapped to the scaffold edges were compared (first and last 20kb sequences) to generate a scaffolds graph with a link connecting two scaffolds with more than two common barcodes. Linear scaffolds paths in the scaffolds graph were composed into the final scaffolds output of the assembly.

The 10X assembly consists of 2.73Gb and is highly contiguous with scaffold N50 at 15.6Mb and contig N50 at 121kb.

5.4 Chromosome assignment using linkage map

Scaffolds were ordered and oriented to chromosomes using *ALLMAPS* (*40*) based on a linkage group map generated by Keygene N.V. (Wageningen, Netherlands) using their sequence-based genotyping technology.

The F2 population for linkage mapping was set up between varieties Shyama (obtained from www.nickys-nursery.co.uk) and a proprietary Sun Pharmaceutical Industries Australia Pty Ltd variety, HT5. Libraries were prepared from DNA of the 84 F2 individuals and the parental lines, which were sequenced on the Illumina HiSeq and the reads were used for single nucleotide polymorphism (SNP) discovery and genotyping. High quality SNP markers were identified by mapping to a reference assembly from the filtered high quality reads of all samples, and further selected according to Chi-Square test on expected segregation ratio of 1:2:1 for each marker. Eight hundred and eighty seven markers were used for the construction of a linkage group map consisting of 11 major linkage groups and 5 small groups (table S5). The 11 major linkage groups are in accordance with the haploid number of chromosomes in opium poppy (n = 11).

For the 887 markers 723 unique marker sequences were identified and of these 677 were mapped to the 10X scaffold assembly (table S6). The mapped scaffolds spanned a total of 2.4Gb over the entire assembly (87% coverage). Information on the markers was collected and used to generate an input file, which included positions on the linkage map, the scaffolds where the marker sequences were mapped to and the mapped position on the scaffold. This input file was then used for the *ALLMAPS* analyses, resulting in 81.6% of sequences in the 10X assembly being assigned to individual chromosomes (fig. S5; table S4).

5.5 The PacBio Falcon assembly

Based on its reliability and contig continuity we chose *Falcon* (*41*) to assemble PacBio long reads into contiguous sequences. *Falcon* is also the choice for VGP (Vertebrate Genome Project) which aims to produce hundreds of whole genome assemblies with near-reference quality (*42*). We have produced 192Gb raw sequencing reads (table S1) and this means a read coverage of 66.8X given the genome size of 2.87Gb. In the assembly pipeline, the first step is base error correction for all reads. The alignment for candidate read matches takes up to 70% computational time in pairwise and reference genome alignment of long sequencing reads (*43*). After the process of base error correction, overlap graphs are built and consensus contigs are constructed. At this stage, we did not use Illumina reads to correct errors and consensus polishing was performed only after assembly merge with the NRgene 10X assembly. With PacBio reads alone, we obtained an assembly with 2.61Gb and N50 = 1.06Mb.

5.6 Assembly merge

As described previously, the PacBio assembly has long contigs while the 10X barcode reads provide excellent long range linking information for genome scaffolding. Our next step was to merge the two assemblies obtained from different sequencing platforms. Fig. S3 illustrates the method how to merge the assemblies. Our strategy was to maintain the long scaffold structure of the 10X assembly and use long PacBio contigs to replace the 10X sequences which may contain gaps. We first shredded the PacBio contigs into 1kb fragments and then aligned the shredded fragments to the 10X assembly. Here the “shred-and-align” method ensured an end-to-end match for the entire contig mapped to the target scaffold, which normally had long or short gaps. The disadvantage was that short repeats may be placed to different locations. We used a modified version of replace pipeline (https://sourceforge.net/projects/phusion2/files/replace/) for sequence replacement. In the pipeline, small repeats were filtered out as "noise" if the fragment matching location is different to the coordinates of the majority fragments from the same contig. Coordinates of the start and end matching locations were identified and scaffolded sequences from the 10X assembly were replaced with PacBio contiguous sequences, as highlighted in the pink circles in fig. S3B. After assembly merge, we finally carried out three runs of consensus polishing using Illumina paired-end reads. As a complex plant genome, opium poppy can be expected to have features such as a certain degree of heterozygosity (to limit this the inbred HN1 variety was used), whole genome duplication and high repeat content which all pose technical challenges for achieving high quality consensus bases. To monitor base error correction at each step, we used the *GATK* (*44*) pipeline with multiple iterations. Variations (SNPs/indels) were called first and heterozygous variants were filtered out with a minor allele fraction cutoff value of 0.75. Final base changes were made from the VCF (Variant Call Format) file with information of variants and scaffold locations. The assembly statistics at different stages can be seen in table S3.

5.7 Assembly QC

To assess the assembly base quality and genome coverage, we used the previously published BAC164_F07 sequence which includes part of the noscapine gene cluster (*6*). We first aligned this BAC to the final base error corrected assembly (fig. S4A). With even a repeat sequence of ~7kb, the total 11,3261 bases can be completely aligned to one contig with matching identify > 99.99%. There are 5 single base mismatches, 3 single base indels and one indel of 4 bases (fig. S4B). In total, this accounts for 12 base differences compared to our whole genome assembly, or equivalently the error rate indicates a base quality at Q40. To get an even more accurate figure on base quality, we processed the aligned reads with *Gap5* (*45*), which is capable of building and visualizing read pileup in the examined region. This indicates that all remaining mismatch and single indels are due to heterozygous bases (fig. S4C).

We sequenced a genomic region containing the T6ODM gene family (Fig. 2E) as described above (section 4) and found it to be highly repetitive across 227kb, thus suitable for QC. It should be noted that in this highly repetitive region the longest repeat sequence spans over 44kb, which is longer than most PacBio SMRT sequencing reads. For QC purposes we evaluated accuracy of the genome assembly in this highly repetitive region. The dot plot of the ONT sequence against the same region from our PacBio reference genome assembly is shown in fig. S4D. We found that the 227kb ONT sequence could be placed on a single contig in the assembly. However, there was an error in the whole genome assembly in the form of a deletion of ~25kb within the 44kb repetitive region. Thus, while the quality of the assembly in this highly repetitive region is high it is still possible to have errors in long repeat sequences such as collapsed repeats. Overall, the whole genome assembly produced in this project shows high level accuracy at both consensus base level and scaffold structures even in those cases of whole genome duplication where sequence mapping identity ranges from 80%-90%. Our experiences here highlight the tremendous challenges in assemblies for complex and large plant genomes which may be overcome as new sequencing technologies deliver ever longer single-molecule reads.

Chromosome 11 in the opium poppy genome has an assembled length of 140Mb and the BIA gene cluster is located near one of the ends. To confirm the assembly accuracy of region 105Mb – 140Mb containing the BIA gene cluster both at the base pair and structural level we carried out a number of quality checks. First, barcode coverage on the contig using the 10X Genomics Chromium data was investigated using the Scaff10X scaffolding tool (https://github.com/wtsi-hpag/Scaff10X). This confirmed that there are no points with zero coverage indicating there are no breakpoints associated with mis-assembly errors across the 105Mb – 140Mb region (fig. S19A). To further examine assembled sequence accuracy, we checked alignments of long PacBio reads to the BIA gene cluster region. BWA aligner was used to align the PacBio reads against the whole opium poppy assembly and genome visualization tool Gap5 (*45*) was applied to display alignments near the end of chromosome 11 (127.5 - 128.5Mb). The average PacBio read coverage is about 60X in line with our planned sequencing. It is clear that no inconsistencies were present in this region (fig. S19B).

We next compared the released genome assembly (Accession number PRJNA435796) across a 1.6Mb region containing the BIA gene cluster with the NRgene 10X and PacBio assemblies used in production of the released assembly and both are in full agreement (sections 5.3 and 5.5, table S3, fig. S20A-B). We also compared the 1.6Mb region to a 1Mb scaffold with total contig bases of 809kb (N50 = 5622bp; Genbank accession number MH011344) that contains the morphinan component of the BIA gene cluster (*19*). This pairwise comparison showed a number of structural differences between PRJNA435796 and MH011344 (fig. S20C).

6. Annotation of repeats, protein-coding genes and function analysis

6.1 Annotation of repeat DNA sequences

We used Repbase (*46*) and a de novo repeat library to annotate DNA sequences in the opium poppy genome. Repbase was downloaded from http://www.girinst.org/repbase/ and a de novo repeat library from the assembled opium poppy genome was generated using RepeatModeler (version open-1.0.8, http://repeatmasker.org/RepeatModeler/). The repetitive elements in Repbase and the opium poppy de novo repeat library were annotated by RepeatMasker. About 71% of the *Papaver somniferum* genome was identified as repetitive (fig. S8) based on RepeatMasker output. The length of the repetitive elements ranged from 6 to 10^5 bp (fig. S8). The most abundant repetitive element repeat type is long terminal repeat (LTR), making up 45.85% of the genome, including 52.51% Gypsy LTRs, 47.11% Copia LTRs and 0.38% other types of LTRs (fig. S8).

6.2 Protein-coding gene prediction and functional annotation

Gene models of the opium poppy genome were predicted using the *MAKER* pipeline (version 2.31.8) (*9*). In short, *MAKER* first masks repetitive elements in the opium poppy reference genome using *RepeatMasker* (http://repeatmasker.org/). It then applies both evidence-based and *ab initio* gene prediction strategies. For the evidence-based method, *MAKER* uses *BLAST* algorithms to align protein and EST data to the genome. The alignments are further polished by Exonerate to produce gene models (*47*). *MAKER* performs the *ab initio* gene prediction based on the assembly sequence itself and then compares predicted gene models to those determined by EST and protein alignment to revise the gene predictions. The confidence of each predicted gene model is then measured using the Annotation Edit Distance (AED) method, which quantifies the normalized distance between gene model and its supporting evidence.

For gene model prediction, three *ab initio* gene prediction tools were used: *AUGUSTUS* (version 3.3) (*48*), *SNAP* (version 2006-07-28) (*49*) and *GeneMark_ES* (version 3.48) (*50*). Tomato (*Solanum lycopersicum*) was used as species model for the *AUGUSTUS* gene prediction, and the pre-trained model of *Arabidopsis thaliana* was used for the Hidden Markov Models of *SNAP* and *GeneMark_ES*. Swiss-Prot (September 2017) was downloaded and protein sequences of three species, *A. thaliana* (*51*), *Beta vulgaris* (*52*) and *Vitis vinifera* (*11*) were obtained from the Ensembl Plants database (http://plants.ensembl.org/index.html). EST data came from the transcriptome assembly by *Trinity* (version 2.1.1) (*26*) using RNA-seq data generated in this work (table S7). *MAKER* pipeline initially predicted 172,347 candidate gene models. We removed genes lacking transcript support or having an AED > 0.5 to produce a high-confidence annotated gene set of 51,224 genes. Genes encoding 10 or fewer amino acids were dropped and manual checking against functionally characterised genes of benzylisoquinoline alkaloid (BIA) metabolism was carried out. This yielded 51,209 protein-coding genes supported by either EST or protein evidence, of which 41,766 are on 11 chromosomes, and 9,443 are on unplaced-scaffolds. Closer inspection of the region spanning gene cluster 49 containing the BIA cluster on chromosome 11 (table S16) revealed 4 additional expressed open reading frames giving a total of 51,213 protein-coding genes. Annotation features such as length distribution of gene, transcript, protein sequence and exon number distribution are shown in fig. S6.

The Benchmarking Universal Single-Copy Orthologs (BUSCO) test was used to determine the sensitivity of our annotation of the 51,213 protein-coding genes, using the plant early release version (v1.1b1, release May 2015) (*10*). The BUSCO test reported 95.3% of complete gene models (38% complete single-copy and 62% duplicated genes, respectively) plus 1.4% additional fragmented gene models (fig. S7), suggesting a high degree of completeness of gene annotation. We also validated our annotation by searching for a list of 25 known BIA biosynthesis genes of *P. somniferum*. The protein sequences of these genes were retrieved from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/protein/), and aligned to our annotated genes by *BLASTp* program. We confirmed that all the genes were included in our annotated results (table S15).

We annotated the functions of predicted protein-coding genes using *InterProScan* (version 5.25-64.0) (*53*) with default parameters, and Gene Ontology (GO) terms for each gene were assigned using *Blast2GO* (version 4.1) (*54*). In total, about 68.8% (35,216) predicted genes have functional domains or GO annotations. GO enrichment analysis of gene sets was performed in *Blast2GO* against opium poppy genome as reference. Statistical significance was tested by Fisher's exact test corrected in multiple tests using *Bonferroni* method under false discovery rate (FDR) threshold of 0.05.

7. Non-coding RNA annotation

Non-coding RNAs (ncRNAs) were annotated using various databases and software packages. The result is summarized in table S8. First, tRNAs and their secondary structures were annotated using *tRNAscan-SE* (version 1.3.1) (*55*) with default parameters. In total, 5,467 tRNAs were predicted in *P. somniferum* genome with sum of lengths of about 403kb. Ribosomal RNAs (rRNAs) were annotated based on sequence homology with various plant rRNAs (the GenBank IDs: AJ307354 for 5S rRNA, AJ232900 for 5.8S rRNA, X16077 for 18S rRNA, and AH001750 for 28S rRNA) using *BLASTn* program (version 2.2.26, E-value cutoff 1e-5). This resulted in detection of 2,283 copies of rRNA with a total length of about 362.6kb, including 338 18S, 219 28S, 61 5.8S and 1665 5S rRNAs. To annotate microRNAs (miRNAs) and small nuclear RNA (snRNA), we searched the Rfam database (version 9.1) (*56*) using *BLASTn* (version 2.2.26, with parameter -W 7 -e 1 -v 10000 -b 10000) and *INFERNAL* (version 0.81, with default parameters) (*57*). We detected 266 miRNAs and 1,478 snRNAs, with sums of lengths about 31.6kb and 128.8kb, respectively.

8. Genome synteny analysis

8.1 Whole genome duplication in opium poppy genome

To study opium poppy genome evolution, we searched for genome wide duplications in our assembled opium poppy genome. First, we performed self-alignment of the assembled genome sequence using megablast as described previously (*58*). The analysis revealed long stretches of duplications within the *P. somniferum* genome that are either inter-chromosomal (between chromosome 1 and 6, 4 and 5, 7 and 8, and 9 and 11) or intra-chromosomal (chr2) (fig. S9A&B). Secondly, we performed all-vs-all paralog analysis in *P. somniferum* genome using reciprocal best hits (RBH) from primary protein sequences by self-*BLASTp* in opium poppy. RBHs are defined as reciprocal best *BLASTp matches* with e-value threshold of 1e-5, c-score (BLAST score / best BLAST score) threshold of 0.3 (*59*), and alignment length threshold of 100 amino acids. The synonymous substitution rate (*Ks*) of RBH gene pairs was calculated based on YN model in *KaKs_Calculator* v2.0 (*60*). We detected 13,377 RBH paralogous gene pairs in the opium poppy genome, and the RBH paralog *Ks* distribution shows a single *Ks* peak at around 0.1 (fig. S13 and S16).

To distinguish whether this peak represents a whole genome duplication event or background small-scale duplications (fig. S13), we performed synteny analysis on *P. somniferum* genes using *MCScanX* (*20*) with default parameters from top five self-*BLASTp* hits. We detected 645 syntenic blocks across the whole genome including 25,744 genes. The total length of syntenic blocks is about 2.34Gb (~86% of whole genome), and the maximum and minimum block size are ~41Mb and ~132Kb, respectively. We found that the majority (93.9%) of the paralogous gene pairs are located inter-chromosomally, i.e. between chromosome 1 and 6 (1,959), 4 and 5 (1,006), 7 and 8 (983), and 9 and 11 (1,458) (Fig. 1A, table S12). In addition, we found that several segmental duplication blocks spanned a large proportion of the corresponding chromosomes (Fig. 1A; table S13). For example, ~ 70% of chr1 were duplicated with ~ 87.4% of chr6, while ~ 69.4% of chr9 were duplicated with ~ 78.8% of chr11 (table S13). This is consistent with the finding from whole-genome DNA alignment analysis. The widespread and well-maintained one-versus-one syntenic blocks (Fig. 1A) indicates that a whole genome duplication (WGD) event has occurred in the *P. somniferum* genome.

Indeed, analysis of duplication types of the *P. somniferum* paralogs by *MCScanX* (fig. S10) indicate that WGD/segmental duplication is the dominant type (50.3%) compared to three other types: dispersed (30.1%), proximal (4.6%) and tandem (6.7%). The synonymous substitution rate (*Ks*) was calculated for opium poppy syntenic block gene pairs and *Ks* distribution clearly showed a major peak at around 0.1 (Fig. 1C, fig. S13), suggesting the presence of a recent whole genome duplication. That this syntelog *Ks* peak is close to the RBH *Ks* peak suggests opium poppy has a whole genome duplication mixed with background gene duplications (fig. S13). Taken together, our analysis provides convincing evidence for a single whole genome duplication event in the opium poppy genome. In addition, the syntenic *Ks* distribution revealed a minor peak at around 1.5, indicating the opium poppy genome has underwent additional segmental duplications (Fig. 1C; fig. S13D).

8.2. Intergenomic comparison

To investigate the evolution of opium poppy, we compared its genome with five other eudicots: *Vitis vinifera* (grape), *Arabidopsis thaliana (Arabidopsis)*, *Coffea arabica (coffee), Nelumbo nucifera (lotus)* and *Aquilegia coerulea.* The orthologs between opium poppy and these species were identified using both RBH and syntenic block analysis described above (Fig. 1C; fig. S9 and S14-S15) with primary protein sequences. For core eudicots such as grape, Arabidopsis and coffee, a γ hexaploidization event occurred before divergence of Rosids and Asterids. Grape is often used as a reference genome for investigating the evolutionary history of eudicot genomes since its genome underwent minimal rearrangement following the γ event. Syntenic analysis using opium poppy and grape genomes suggested that opium poppy did not experience the γ event as suggested by a 3:2 syntenic relationship between grape and opium poppy (fig. S9C&F). Murat *et al.* (*12*) constructed the genome of the most recent ancestor of flowering plants, referred to as the ancestral eudicot karyotype (AEK). We compared the opium poppy genome to AEK in addition to the grape genome (*11*). The synteny dot plot (fig. S9) and genome painter image (Fig. 1B and fig. S9E) both illustrate that most AEK and grape segments have two syntenic copies in *P. somniferum*, suggesting that opium poppy clearly underwent a whole genome duplication event. Moreover, we calculated the ortholog depth of *P. somniferum* per AEK and grape genes from *BLASTp* analysis (sequence identity >= 0.5, E-value <= 1e-40), and a number of genes showed a high peak (1,859 and 5,243 collinear genes in AEK and grape, respectively) at depth of two (fig. S12). The genome comparisons also revealed signs of genome rearrangement events having occurred in the opium poppy genome following whole genome duplication as shown in dual synteny plots (fig. S11).

8.3 Phylogenetic analysis and estimation of divergence time

The assembled and annotated opium poppy genome allowed us to investigate its evolutionary history. Single-copy orthologs among taxa are commonly used to achieve robust phylogenetic reconstruction with high confidence and concordance. Using *OrthoFinder* v2.0 (*13*) we identified a set of 48 single-copy orthologs from 11 angiosperm species including the monocot *Oryza sativa,* opium poppy, *A. coerulea, N. nucifera, A. thaliana, B. vulgaris, Coffea arabica, Theobroma cacao, V. vinifera, S. lycopersicum* and *Helianthus annuus*. Based on this ortholog set, a phylogenetic tree of the eleven plant species was constructed as follows: for each single-copy gene a coding sequence alignment was created using *MUSCLE* v3.8.31 (*61*) and then all coding sequence alignments were concatenated in *MEGA* (*62*). The concatenated alignment was then used to construct a maximum likelihood phylogenetic tree using *RAxML* v7.2.6 (*63*) and the maximum likelihood tree was then used as a starting tree to estimate species divergence time using *BEAST* v2.1.2 (Bayesian Evolutionary Analysis Sampling Trees) (*14*). For the divergence time estimation, we used a calibrated Yule model with a strict clock rate, and gamma hyperparameter of prior distribution. To calibrate the divergence time, a Log Normal model was chosen for monocot-dicot split time (mean: 150 MYA. Std dev: 4MYA) and grape-cacao split time (mean: 110 MYA. Std dev: 4MYA). The Markov chain Monte Carlo was repeated 10,000,000 times with 1000 steps.

8.4 Estimate of whole genome duplication timing

To estimate the timing of the whole genome duplication event in opium poppy, *Ks* values of opium poppy syntenic block genes were calculated using YN model in *KaKs_Calculator* v2.0 (*58*). The *Ks* values were then fitted to a mixture model of

Gaussian distribution to determine the number of components in the *Ks* distribution (Fig. 1C; fig. S13; and table S9) using the *Mclust* R package (*64*). We identified components associated with WGD and segmental duplications with their mean *Ks* value and standard deviations. We then plotted the ortholog *Ks* distributions between opium poppy and other eudicots to compare the relative substitution rates in different species (Fig. 1C and fig. S14-S15). First of all, we observed faster substitution rates for Arabidopsis than grape, lotus and Aquilegia, as shown by a larger *Ks* for opium poppy-Arabidopsis syntelogs than for syntelogs between opium poppy and grape, lotus or Aquilegia (Fig. 1C and fig. S14). Therefore, we conclude that Arabidopsis is not appropriate for estimating substitution rate in the opium poppy lineage. We then observed that opium poppy has a faster substitution rate than grape, because the *Ks* between genome-wide opium poppy-grape syntelog pairs are smaller than those among triplicated grape genes. Ming *et al* reported that lotus substitution rate is slower than grape (*65*). Because opium poppy has the fastest substitution rate among the three species, neither grape nor lotus is suitable for estimating substitution rate for opium poppy. To time the opium poppy WGD, we estimated the average evolutionary rate for Ranunculales using *P. somniferum*, a Papaveraceae and *Aquilegia coerulea*, a Ranunculaceae. Divergence time of 110 million years ago (MYA) between *P. somniferum* and *A. coerulea* was obtained based on our divergence estimation using *BEAST* (Fig. 1D). Given the mean *Ks* value (1.53) of *P. somniferum-A. coerulea* and their divergence date *T* (110 MYA), we calculated the synonymous substitutions per site per year (*r*) for Ranunculales equaling 6.98e-9 ($T = Ks / 2r$). The *r* value was applied to time the *P. somniferum* WGD. We dated the opium poppy WGD ($Ks = 0.11 \pm 0.061$) around 7.8±4.35MYA (Fig. 1C and table S9). To better understand the relationship between polyploidy events in the Papaveraceae and Ranunculaceae we also performed reciprocal best hit and syntenic analysis on a high quality whole genome assembly of *A. coerulea* (*16*) to identify potential whole genome duplications. Overall we detected 5,630 RBH paralogous gene pairs and 82 syntenic blocks containing 895 gene pairs. The *Ks* was calculated for both the RBH paralogous genes and syntenic block gene pairs of *A. coerulea*. Comparison of the two *Ks* distributions showed a major peak at 1.55±0.50 (Fig. 1C; fig. S13; table S9) representing the *A. coerulea* WGD. Using the *r* value for Ranunculales, we dated this *A. coerulea* WGD at 111.3±35.6 MYA (Fig. 1D; table S9).

Given the overall 2:2 syntenic relationship between opium poppy and *A. coerulea* (fig. S9D), the WGD in both species appears to be lineage specific, indicating that the *A. coerulea* WGD may have occurred soon after its divergence from opium poppy.

The estimated timing for WGD events of opium poppy and *A. coerulea* as well as previously reported WGD/WGT (whole genome triplication) events in five other angiosperm species (*N. nucifera* (65), *O. sativa* (66), *A. thaliana* (67), *H. annuus* (68) and *S. lycopersicum* (69)) are displayed in the phylogenetic tree (Fig. 1D).

9. Phylogenomic analysis

Phylogenomic analysis was conducted to determine the timing of segmental duplication events in opium poppy as described by Jiao *et al.* (*15*). Each pair of opium poppy paralogs under the segmental duplication *Ks* peak (1.4~1.6) were used as anchor genes in searching for homologous genes in a public database for 22 different land plant species (http://fgp.huck.psu.edu/planttribes_data/22Gv1.0.tar.bz2). *OrthoMCL* (*70*) was implemented to identify 261 orthogroups for the 23 land plant species, from which 95 orthogroups were obtained by including orthogroups with just two opium poppy syntenic paralogs. Protein sequence alignments were created for each orthogroup using *MAFFT* (*71*) L-INS-i iterative refinement method and automatically trimmed by *trimAl* (*72*). Each alignment was then used to construct maximum likelihood phylogenetic tree using *RAxML* v7.2.6 (*63*), searching for the best maximum likelihood tree with the PROTGAMMAJTT model by conducting 100 bootstrap replicates. All gene trees were rooted using the outmost taxon in the reference species tree (www.timetrees.org) as the outgroup (Database S2). Each tree was examined to determine the likely placement of paleo-duplication event(s) throughout angiosperm evolution following procedures described in Jiao *et al.* (*15*).

10. Gene family analysis

To investigate nucleotide identity level of coding sequences of the STORR P450 module, STORR reductase module/COR and CODM/T6ODM to their closest paralogs we firstly searched and retrieved the gene family members in the annotated proteins of the opium poppy genome, then performed phylogenetic analyses. The P450 and reductase sequences in Winzer *et al.* 2015 (*7*) were also included in the analyses of STORR P450 module and STORR reductase module/*COR* gene families.

Both CODM (ADD85331.1) and T6ODM (ADD85329.1) protein sequences were used as query sequence in a *BLASTp* search in the curated Swissprot database via the NCBI webpage (http://www.ncbi.nlm.nih.gov/). Two hits with significant similarity (Q39224.1 and A2A1A0.1) were added to the subsequent analyses of the CODM/T6ODM gene family.

Species-specific identifiers have been assigned as follows: *Ammi majus* (_AMMMJ), *Arabidopsis lyrata* (_ARALY), *Arabidopsis thaliana* (_ARATH), *Digitalis purpurea* (_DIGPU), *Erythroxylum coca* (_ERYCB), *Eschscholzia californica* (_ESCCA), Fragaria x ananassa (_FRAAN), *Glycyrrhiza glabra* (_GLYGL), *Gossypium hirsutum* (_GOSHI), *Glycine max* (_SOYBN), *Hordeum vulgare* (_HORVU), *Malus domestica* (_MALDO), *Medicago sativa* (_MEDSA), *Nicotiana tabacum* (_TOBAC), *Nicotiana tomentosiformis* (_NICTO), *Oryza sativa Japonica* (_ORYSJ), *Panax ginseng* (_PANGI), *Papaver rhoeas* (_PAPRH), *Pisum sativum* (_PEA), *Sesbania rostrata* (_SESRO), and *Coptis japonica* (_COPJA).

GenBank accession numbers for the protein sequences are as follows:

Cytochrome P450s: CYP82A3_SOYBN (O49858.1), CYP82A1_PEA (Q43068.2), CYP82A4_SOYBN (O49859.1), CYP82A2_SOYBN (O81972.1), CYP82D1_GOSHI (AII31758.1), CYP82D2_GOSHI (AII31759.1), CYP82D3_GOSHI (AII31760.1), CYP82D47_PANGI (H2DH24.1), CYP82E4v1_TOBAC (ABA07805.1), CYP82E4_NICTO (ABM46920.1), CYP82E4v2_TOBAC (ABA07804.1), CYP82E3_NICTO (ABM46919.1), CYP82G1_ARALY (EFH61953.1), CYP82G1_ARATH (Q9LSF8.1), NMCH_ESCCA (AAC39454.1), CYP82C4_ARATH (Q9SZ46.1), CYP82C3_ARATH (O49396.3), CYP82C2_ARATH (O49394.2), CYP82H1_AMMMJ (AAS90126.1), CYP82F1_ARALY (EFH56916.1), AFB74614 (AFB74614.1, CYP82X1), AFB74616 (AFB74616.1, CYP82X2), AFB74617 (AFB74617.1,CYP82Y1), P6H_ESCCA (F2Z9C1.1), L7X0L7.1 (L7X0L7.1, P6H), and

L7X3S1.1 (L7X3S1.1, MSH), AKO60176_PAPRH (AKO60176.1), and STORR_CYP82Y2 (AKN63431.1).

Oxidoreductases: Q9SQ68.1 (Q9SQ68.1, COR1.3), Q9SQ67.2 (Q9SQ67.2, COR1.4), Q9SQ69.1 (Q9SQ69.1, COR1.2), Q9SQ70.1 (Q9SQ70.1, COR1.1), B9VRJ2.1 (B9VRJ2.1, COR1.5), Q9SQ64.1 (Q9SQ64.1, COR2), PKR1_GLYGL (BAA13113.1), CR_MEDSA (Q40333), 6DCS_SOYBN (P26690.1), GALUR_FRAAN (O49133.1), NADO2_ORYSJ (Q7G765.1), NADO1_ORYSJ (Q7G764.1), MER_ERYCB (E7C196.1), CR_SESRO (CAA11226.1), AKRCA_ARATH (Q84TF0.1), AKRC9_ARATH (Q0PGJ6.1), AKRCB_ARATH (Q9M338.1), AR1_DIGPU (CAC32834.1), AR2_DIGPU (CAC32835.1), AKRC8_ARATH (O80944.2), S6PD_MALDO (P28475.1), ALDR_HORVU (P23901.1), and STORR_oxired (AKN63431.1).

2-oxoglutarate/Fe(II)-dependent dioxygenase: Q39224_ARATH (Q39224.1), A2A1A0_COPJA (A2A1A0.1), ADD85329.1 (ADD85329.1, T6ODM), ADD85330.1 (ADD85330.1), and ADD85331.1 (ADD85329.1, CODM)

Protein sequence alignments were made firstly with *ClustalX* (*73*), then conserved blocks were evaluated and selected with Gblocks v0.91b by allowing gap positions within final blocks (*74*) and used in the subsequent phylogenetic analyses. The best-scoring maximum likelihood tree of a thorough maximum likelihood analyses in conjunction with bootstrap analyses of 100 replicates was carried out with *RAxML* (*63*). Groups with above 70% bootstrap value were considered as strongly supported (fig. S22).

Closest paralogs of STORR P450 module, STORR reductase module/COR and CODM/T6ODM were identified from the phylogentic branches (fig. S22) and level of pairwise nucleotide sequence identity between all pairs were then calculated with *EMBOSS* Stretcher (https://www.ebi.ac.uk/Tools/psa/emboss_stretcher/nucleotide.html) and the amino acid sequence identities were calculated using *BLASTp* software. All the results were summarised in table S19.

11. Transcriptomic analysis

The RNA sequencing reads were first checked for quality using *FastQC* (https://github.com/s-andrews/FastQC). Illumina sequencing adapters and poor-quality reads (quality score < 30) were trimmed using *Trimommatic* v0.32 (*75*). The cleaned high-quality RNA reads were used for *de novo* assembly of transcripts using *Trinity* v2.1.1 (*26*), providing EST evidence for genome annotation. To estimate the transcript abundance for annotated opium poppy genes, the trimmed RNA reads were aligned against reference genome using *Hisat2* (*23*) and transcripts were discovered and quantified by *Stringtie* (*24*) and *Ballgown* (*25*) respectively using default parameters. The processed transcriptome data from different opium poppy tissues was analyzed by K-means clustering in using in-house R scripts, identifying 20 co-expression gene modules (fig. S18B). Each module is categorized based on types of tissues where genes have higher average expression levels. Co-expression gene and GO networks were then constructed and visualized in *Gephi* v0.9.2 (*76*) based on the clustering results (fig. S18C).

12. Genome mining for gene clusters of plant specialized metabolism

To search for potential gene clusters that are associated with plant specialized metabolism, *plantiSMASH* version 3.0.5-a04b4cd (*17*) was used to mine the sequences of the 11 chromosomes along with their GFF (General Feature Format) annotation files. Default parameters were used, and *plantiSMASH* ClusterFinder function predicted a total of 50 gene clusters across all 11 chromosomes. The same analysis was also extended to the 426 unplaced scaffolds that contain annotated genes and this resulted in identification of a further 34 clusters (table S16). The results were parsed and summarized with additional Pfam (version 31.0) entries and gene expression patterns across 7 tissue types.

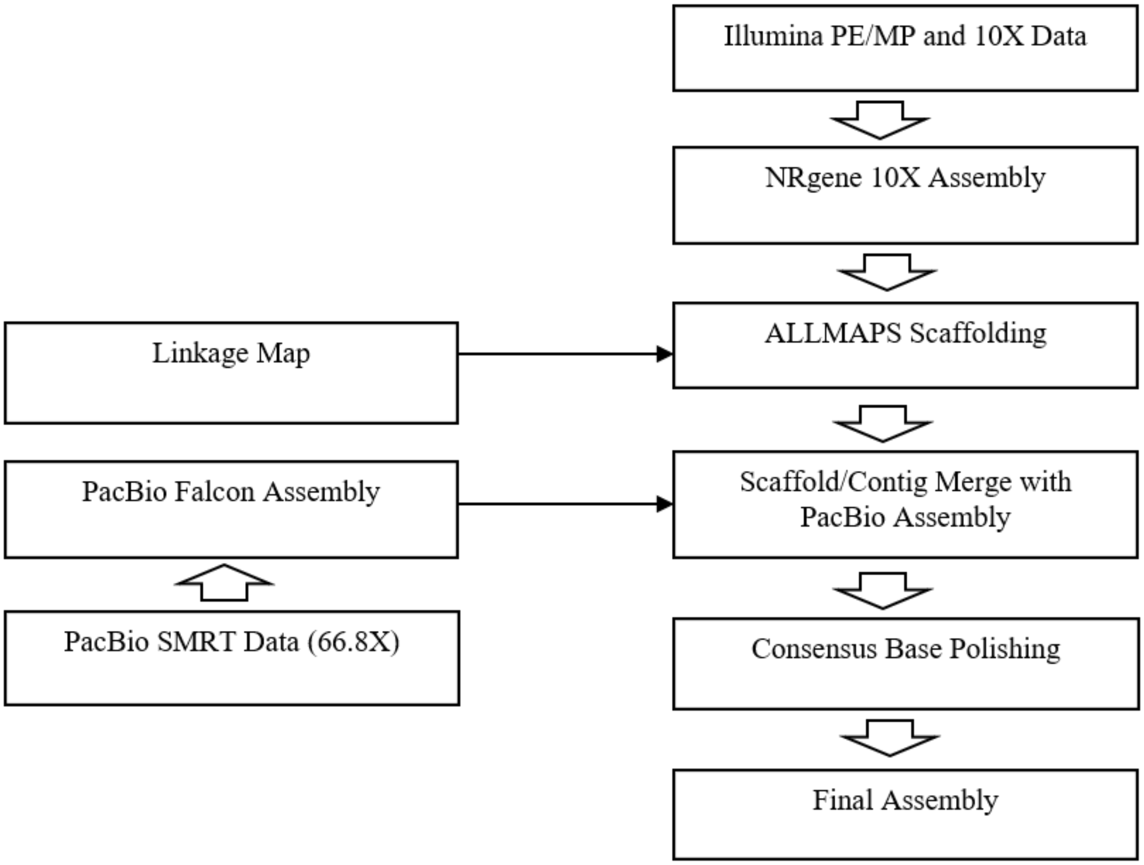


Fig. S1.
Genome assembly flowchart demonstrating assembly merge and data integration.

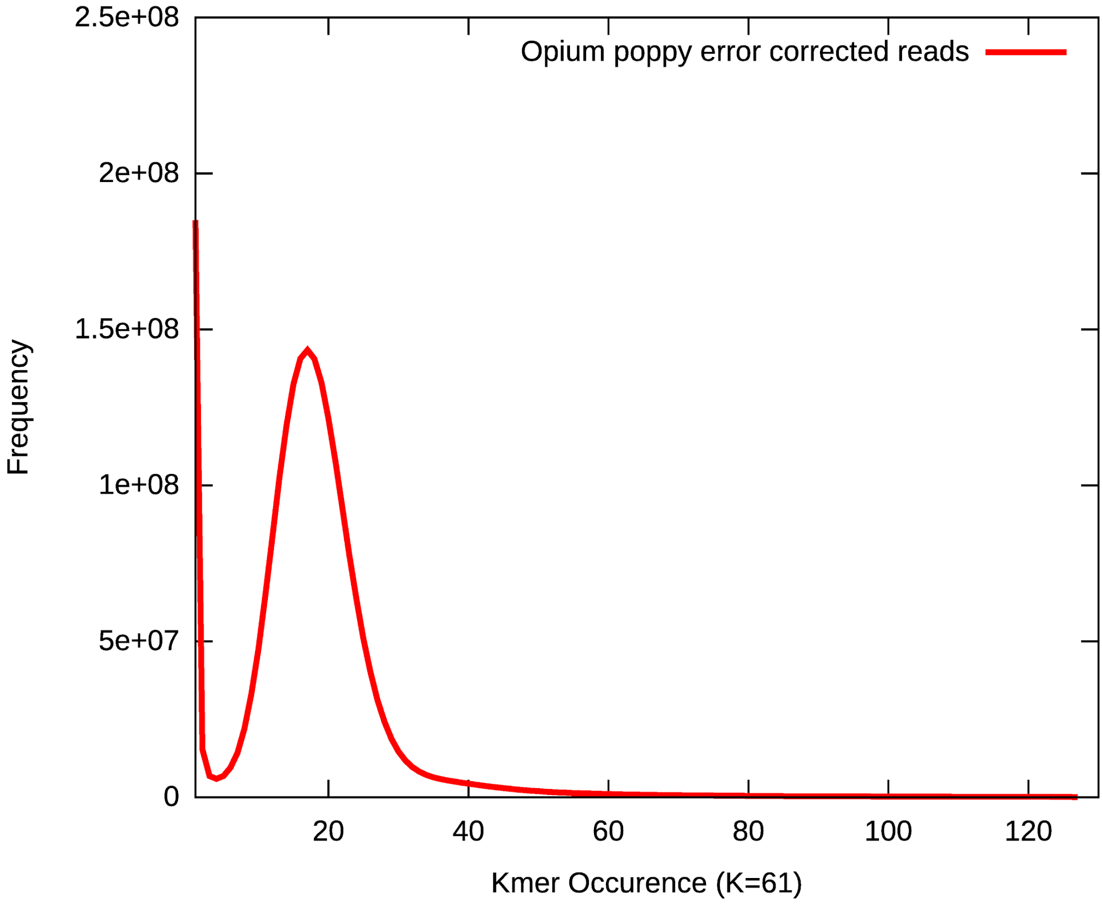


Fig. S2.

Kmer frequency distributions from base error corrected reads. With *K*=61, there is a frequency peak value at 17 which is used for genome size estimation.

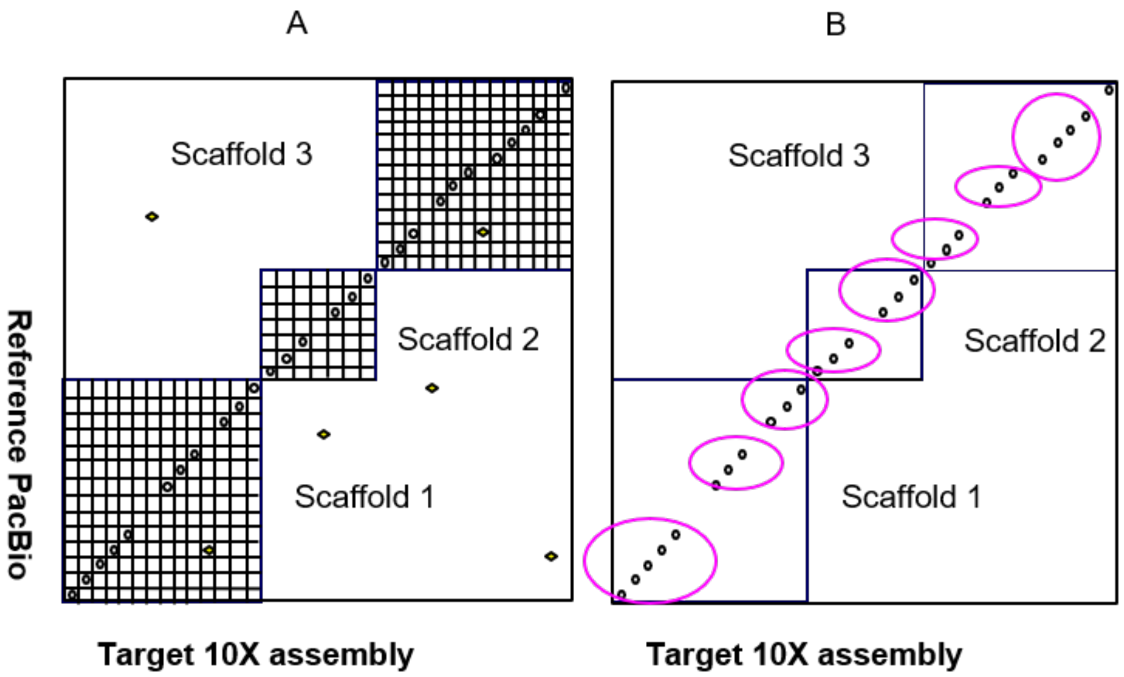
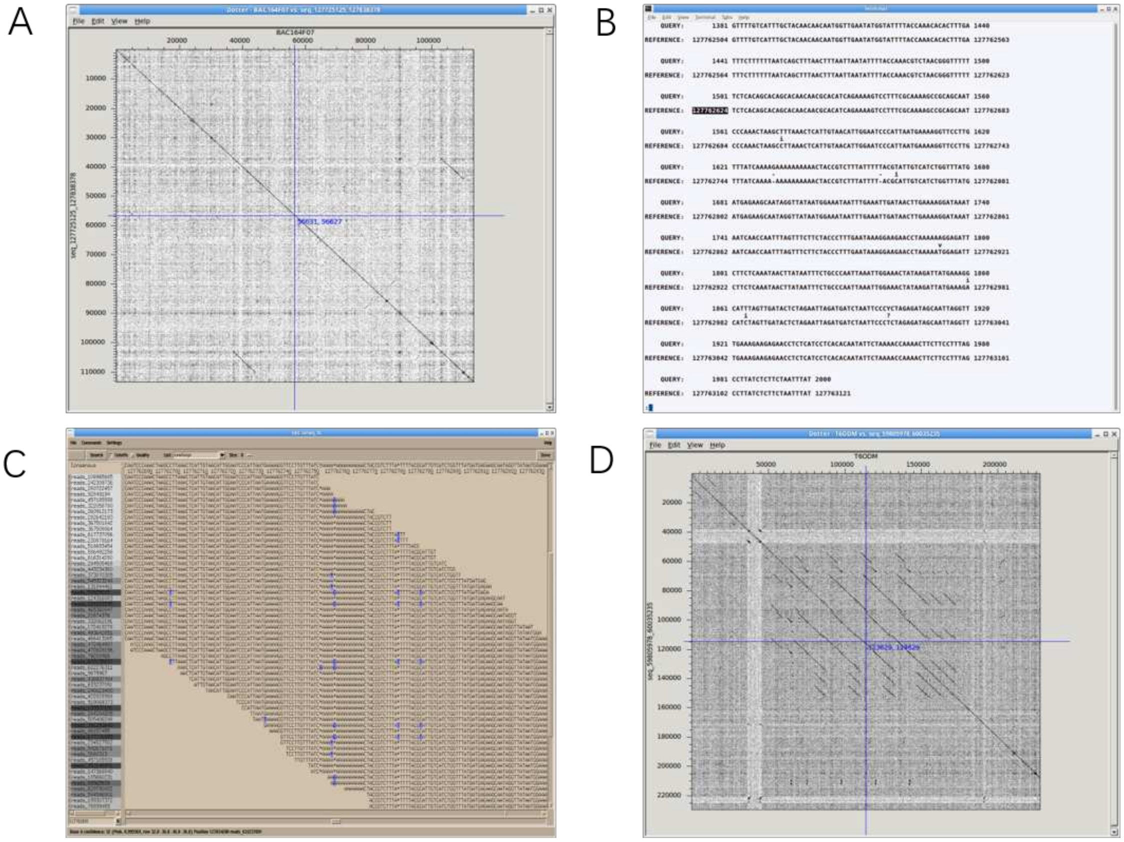


Fig. S3.
Assembly merge in which longer PacBio contig sequences are merged into an assembly (NRgene 10X) with much longer scaffolds, but shorter contigs. **(A)** PacBio contigs are shredded into 1kb fragments and then aligned to the target NRgene 10X assembly. After the alignment, repeats are filtered out as noise; **(B)** Coordinates of start and end matching locations are identified and target sequences likely with gaps are replaced with PacBio contiguous sequences.



```
    QUERY:         1381 GTTTTGTCATTTGCTACAACAACAATGGTTGAATATGGTATTTTACCAAACACACTTTGA 1440
REFERENCE:    127762504 GTTTTGTCATTTGCTACAACAACAATGGTTGAATATGGTATTTTACCAAACACACTTTGA 127762563

    QUERY:         1441 TTTCTTTTTTAATCAGCTTTAACTTTAATTAATATTTTACCAAACGTCTAACGGGTTTTT 1500
REFERENCE:    127762564 TTTCTTTTTTAATCAGCTTTAACTTTAATTAATATTTTACCAAACGTCTAACGGGTTTTT 127762623

    QUERY:         1501 TCTCACAGCACAGCACAACAACGCACATCAGAAAAGTCCTTTCGCAAAAGCCGCAGCAAT 1560
REFERENCE:    127762624 TCTCACAGCACAGCACAACAACGCACATCAGAAAAGTCCTTTCGCAAAAGCCGCAGCAAT 127762683

    QUERY:         1561 CCCAAACTAAGCTTTAAACTCATTGTAACATTGGAATCCCATTAATGAAAAGGTTCCTTG 1620
                                    i
REFERENCE:    127762684 CCCAAACTAAGCCTTAAACTCATTGTAACATTGGAATCCCATTAATGAAAAGGTTCCTTG 127762743

    QUERY:         1621 TTTATCAAAAGAAAAAAAAAACTACCGTCTTTATTTTTACGTATTGTCATCTGGTTTATG 1680
                                  .                             -  i
REFERENCE:    127762744 TTTATCAAAA-AAAAAAAAAACTACCGTCTTTATTTT-ACGCATTGTCATCTGGTTTATG 127762801

    QUERY:         1681 ATGAGAAGCAATAGGTTATAATGGAAATAATTTGAAATTGATAACTTGAAAAGGATAAAT 1740
REFERENCE:    127762802 ATGAGAAGCAATAGGTTATAATGGAAATAATTTGAAATTGATAACTTGAAAAGGATAAAT 127762861

    QUERY:         1741 AATCAACCAATTTAGTTTCTTCTACCCTTTGAATAAAGGAAGAACCTAAAAAAGGAGATT 1800
                                                                          v
REFERENCE:    127762862 AATCAACCAATTTAGTTTCTTCTACCCTTTGAATAAAGGAAGAACCTAAAAATGGAGATT 127762921

    QUERY:         1801 CTTCTCAAATAACTTATAATTTCTGCCCAATTAAATTGGAAACTATAAGATTATGAAAGG 1860
                                                                                   i
REFERENCE:    127762922 CTTCTCAAATAACTTATAATTTCTGCCCAATTAAATTGGAAACTATAAGATTATGAAAGA 127762981

    QUERY:         1861 CATTTAGTTGATACTCTAGAATTAGATGATCTAATTCCCYCTAGAGATAGCAATTAGGTT 1920
                                i                                   ?
REFERENCE:    127762982 CATCTAGTTGATACTCTAGAATTAGATGATCTAATTCCCTCTAGAGATAGCAATTAGGTT 127763041

    QUERY:         1921 TGAAAGAAGAGAACCTCTCATCCTCACACAATATTCTAAAACCAAAACTTCTTCCTTTAG 1980
REFERENCE:    127763042 TGAAAGAAGAGAACCTCTCATCCTCACACAATATTCTAAAACCAAAACTTCTTCCTTTAG 127763101

    QUERY:         1981 CCTTATCTCTTCTAATTTAT 2000
REFERENCE:    127763102 CCTTATCTCTTCTAATTTAT 127763121
```

Fig. S4.

Finished or unfinished BACs against the whole genome assembly for QC. (**A)** Start to end match between BAC164_F07 and whole genome assembly; (**B)** Only a few base differences (mismatch or single base indels) are observed; (**C)** Read pileup from Gap5 (*45*) indicates that the base differences are due to heterozygosity; (**D)** Collapsed repeats are present in cases where long PacBio reads cannot span across repetitive regions.

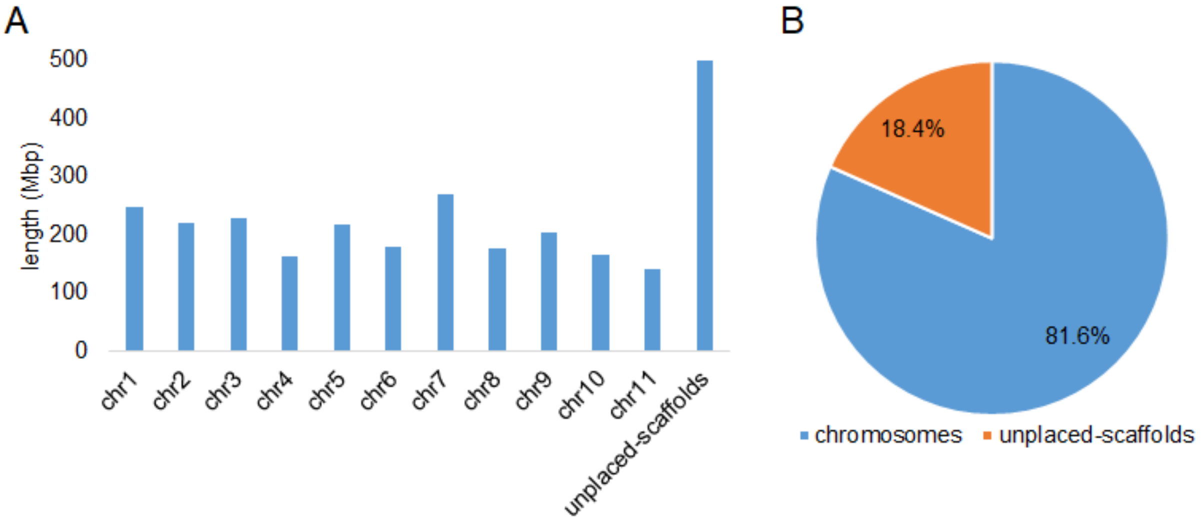


Fig. S5.
Summary of the assembled opium poppy genome. **(A)** The size (Mbp) of 11 assembled chromosomes and unplaced scaffolds (sum of 34,377 scaffolds). **(B)** The proportions of 11 chromosomes and unplaced scaffolds.

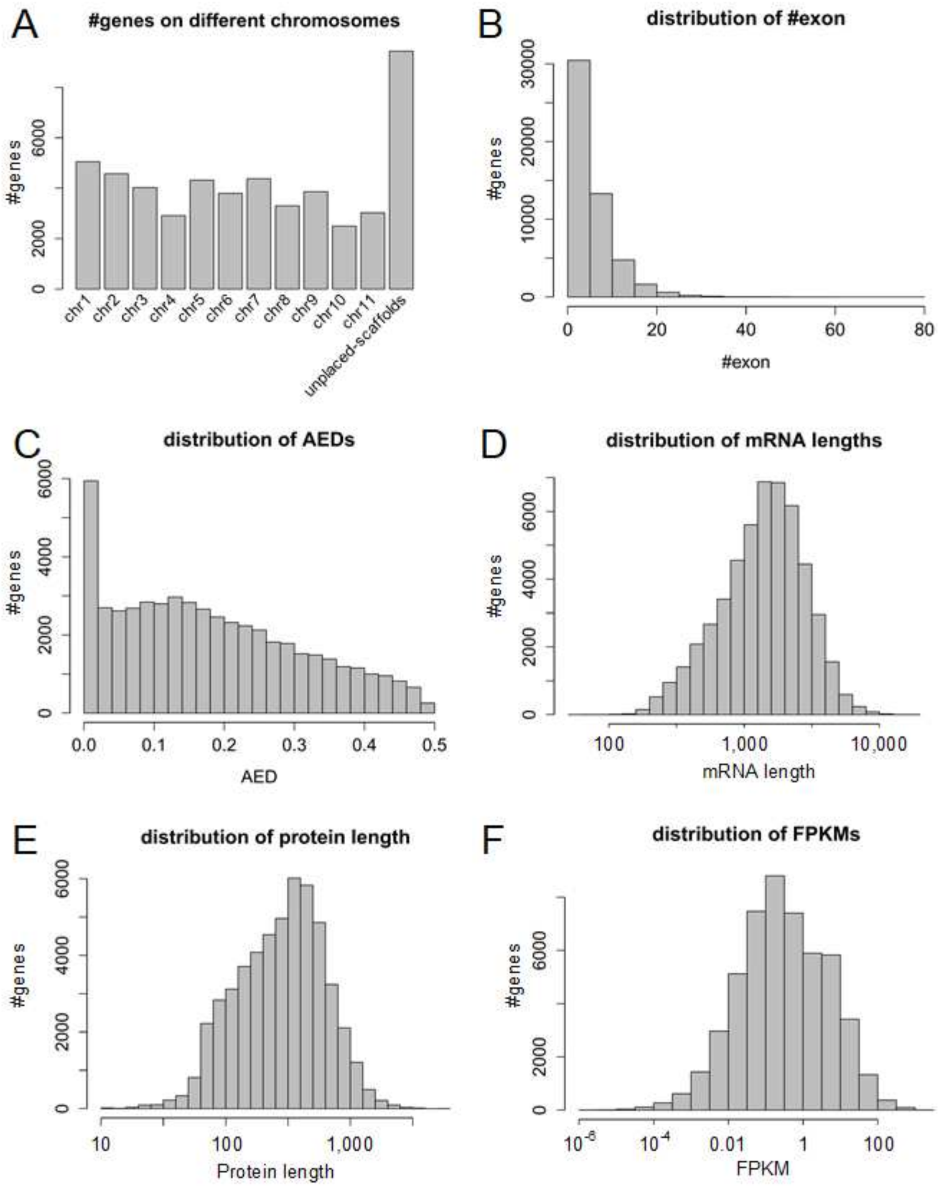


Fig. S6.

Characteristics of predicted opium poppy protein-coding genes. **(A)** gene numbers on 11 chromosomes and unplaced scaffolds (sum of 34,377 scaffolds). **(B)** Distribution of exon numbers. **(C)** Distribution of annotation edit distance (AED) of each gene. **(D)** Distribution of mRNA sequence length. **(E)** Distribution of protein sequence length. **(F)** Distribution of transcript abundance measured by FPKM (Fragments per Kilobase per Million mapped reads).

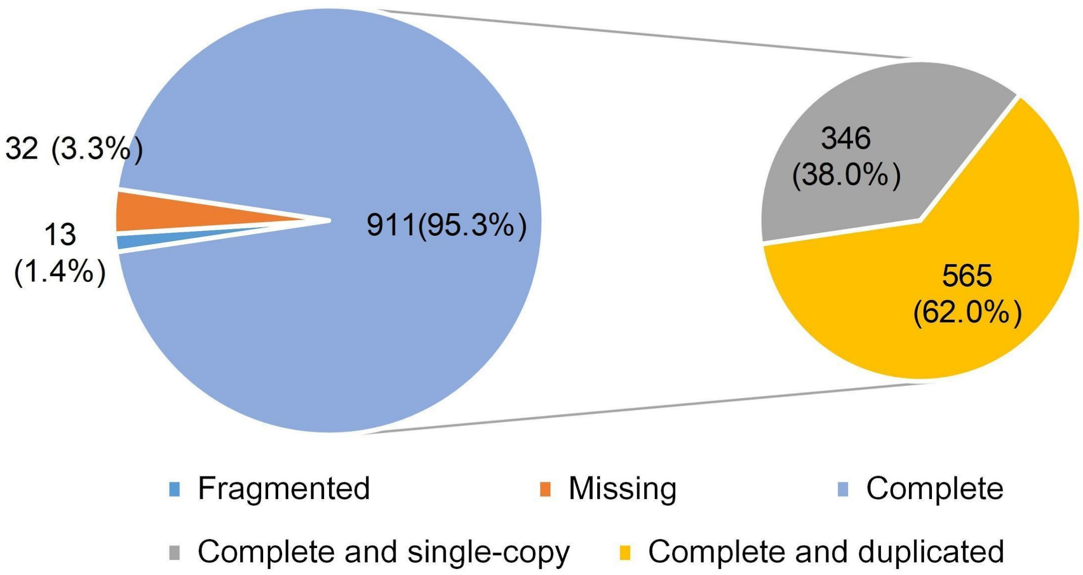


Fig. S7.

Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis of opium poppy annotated genes. The predicted protein-coding genes in the opium poppy genome gave 95.3% of the plant early release version (v1.1b1, release May 2015) database (*10*). Of these 38% were single-copy and 62% were duplicated.

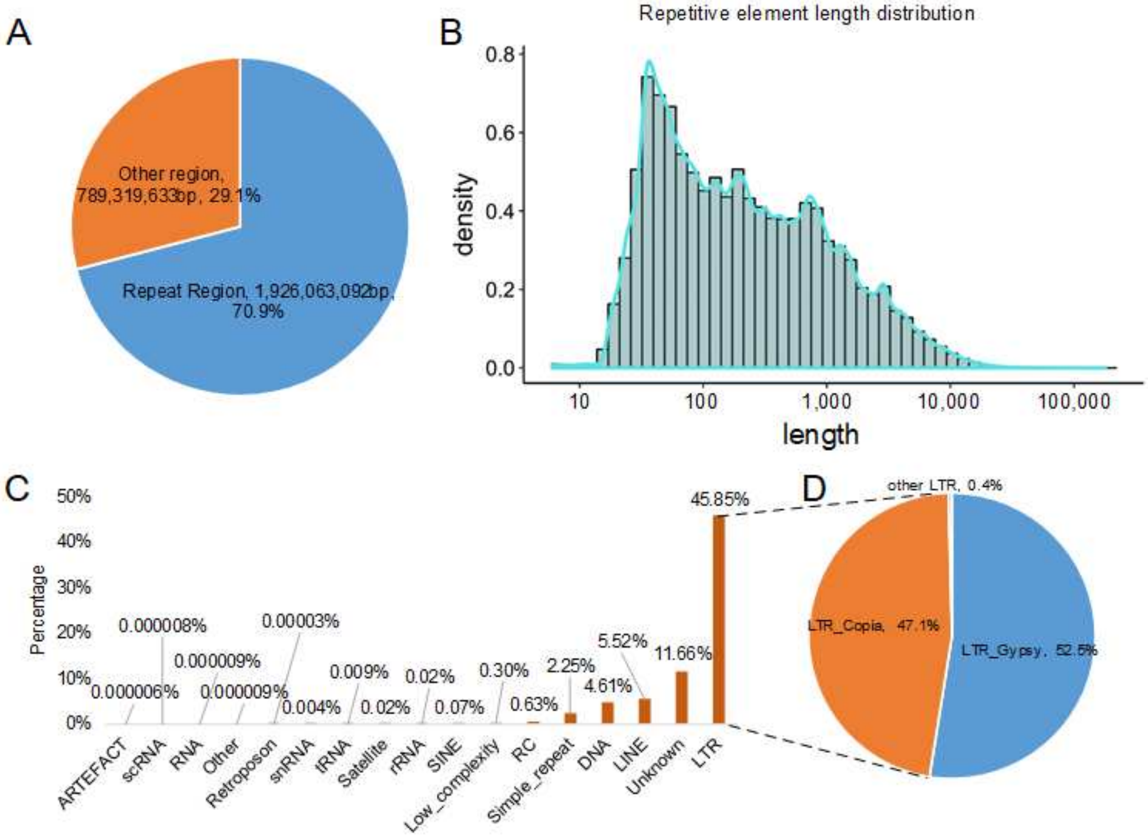


Fig. S8.

Characteristics of repetitive elements in the opium poppy genome. **(A)** The proportions of repetitive elements in the opium poppy genome. **(B)** The length distribution of repetitive elements. **(C)** The proportions of different classes of repetitive elements in the opium poppy genome. The LTRs (long terminal repeats) are the most abundant repetitive elements. **(D)** The proportions of different LTR species.

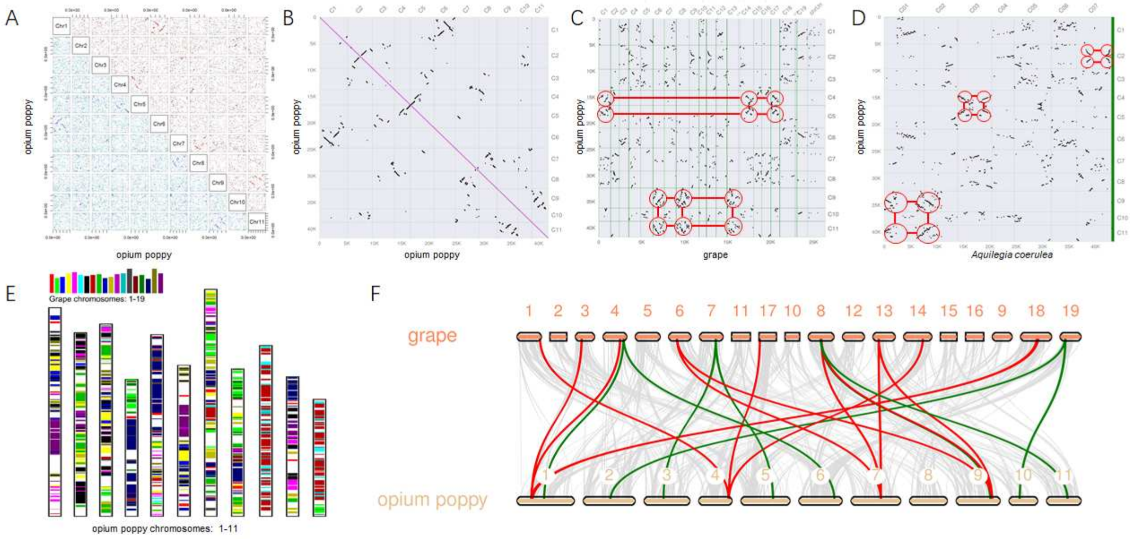


Fig. S9.

Synteny analysis within the opium poppy genome and between the opium poppy and grape (*Vitis vinifera*) genomes. **(A)** Dot plot matrix displaying the DNA sequence alignment of 11 chromosomes in opium poppy. (**B)** Dot plot of paralogs in opium poppy to show the segmental duplication events. **(C)** Dot plot illustrating the comparative analysis of the opium poppy and grape genomes, the red circles highlight several major duplication events, the dots represent the synteny gene pairs. **(D).** Dot plot illustrating the comparative analysis of the opium poppy genome assembly and the *Aquilegia coerulea* genome assembly (*16*). The red circles highlight several major duplication events, the dots represent the synteny gene pairs. **(E)** Genome painter image displays gene collinearity between the grape and opium poppy genomes. Synteny from paralogs and orthologs was detected by *MCScanX* (*20*). **(F).** Macrosynteny between grape and opium poppy karyotypes. Green lines highlight the two copies of opium poppy syntenic blocks per corresponding grape block. Red lines highlight the three copies of grape syntenic block per corresponding opium poppy block.

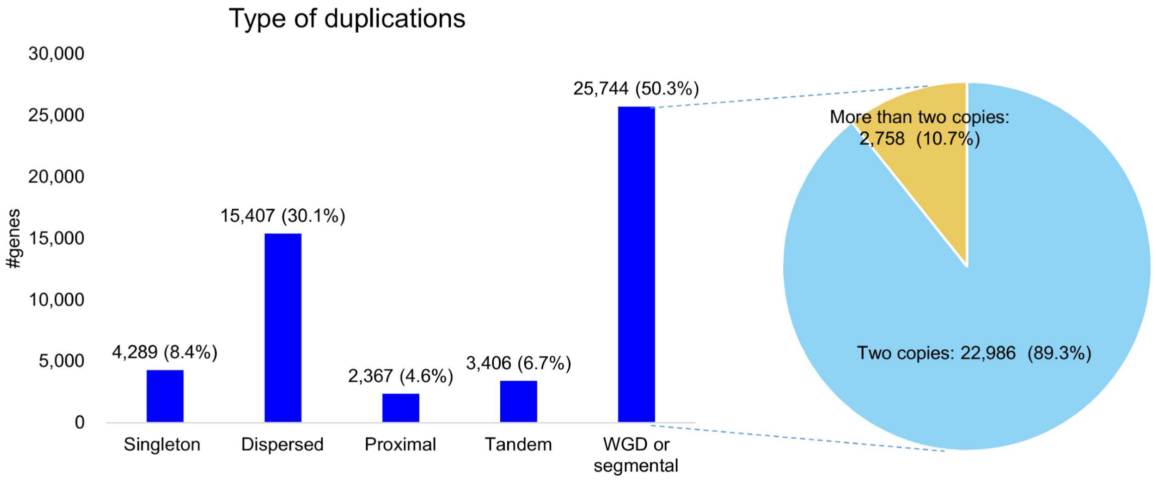


Fig. S10.

Types of gene duplication in the opium poppy genome. **(A)** Distribution of different duplication types classified by *MCScanX* (*20*) as follows: Singleton: no duplication; WGD/segmental: whole genome or segmental duplications (collinear genes in collinear blocks); Tandem: consecutive duplication; Proximal: duplications in nearby chromosomal region but not adjacent; Dispersed: duplications of modes other than tandem, proximal or WGD/segmental. **(B)** Pie-chart showing 89.3% of collinear genes in collinear blocks are present as two copies and the remainder are present in more than two copies.

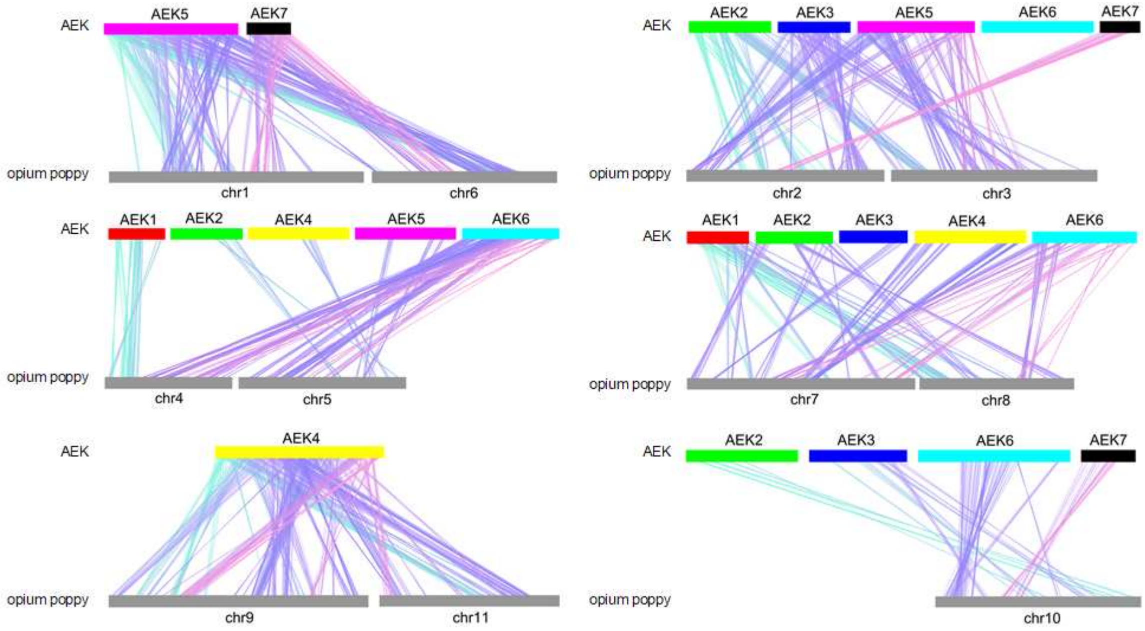


Fig. S11.

Dual synteny plots showing the opium poppy genome rearrangement events. The AEK (Ancestral Eudicot Karyotype) chromosomes (1~7) are colored consistently with Fig. 1B. The synteny blocks were detected by *MCScanX* (*20*) using top five *BLASTp* hits.

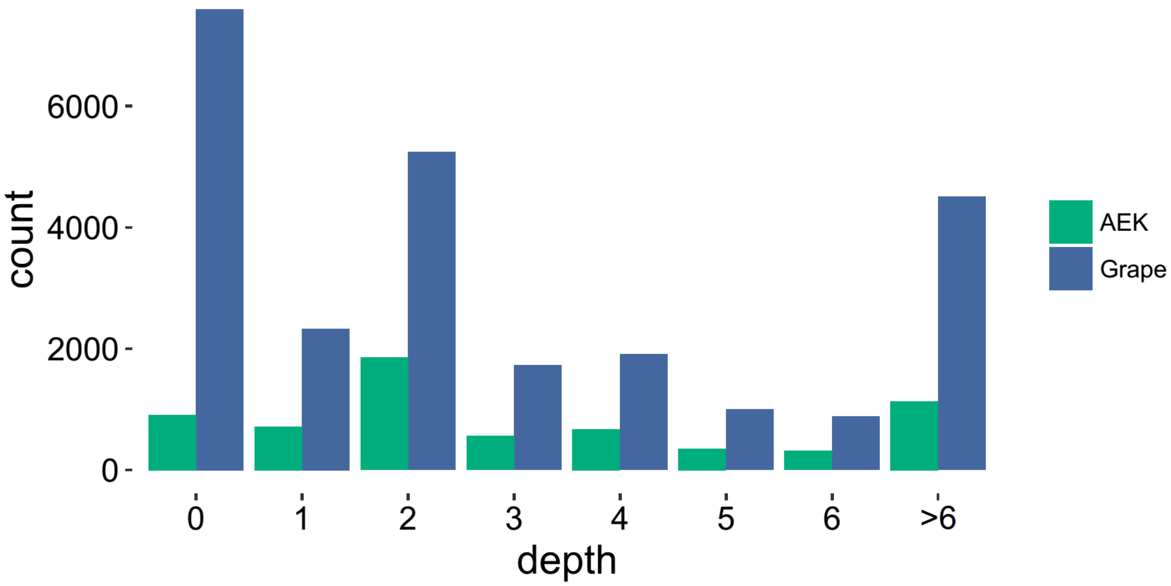


Fig. S12.

Ortholog depth density plot of opium poppy genes vs. AEK (Ancestral Eudicot Karyotype) and grape (*Vitis vinifera*) genes. Ortholog depth refers to the number of opium poppy genes orthologous to each of equivalent AEK and grape genes. Orthologs were detected by *BLASTp* with e-value ≤1e-40 and sequence identity ≥ 0.5.

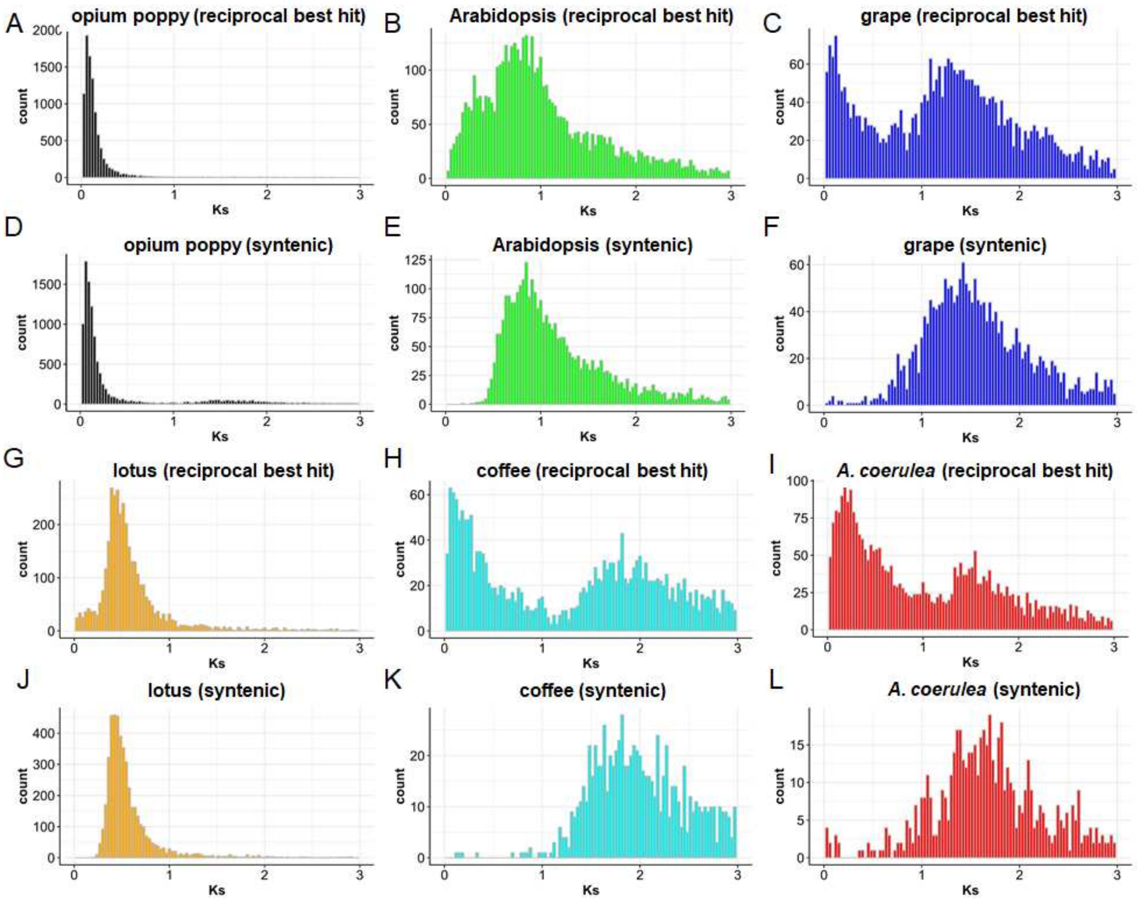


Fig. S13.

Histogram distribution of *Ks* (synonymous substitution rate) for paralogous gene pairs identified through reciprocal best hit analysis (ABCGHI) and syntenic block gene pairs identified through *MCScanX* (DEFJKL) in different eudicot species: opium poppy (*Papaver somniferum*), Arabidopsis (*Arabidopsis thaliana*), grape (*Vitis vinifera*), lotus (*Nelumbo nucifera*), coffee (*Coffea arabica*) and Aquilegia (*Aquilegia coerulea*).

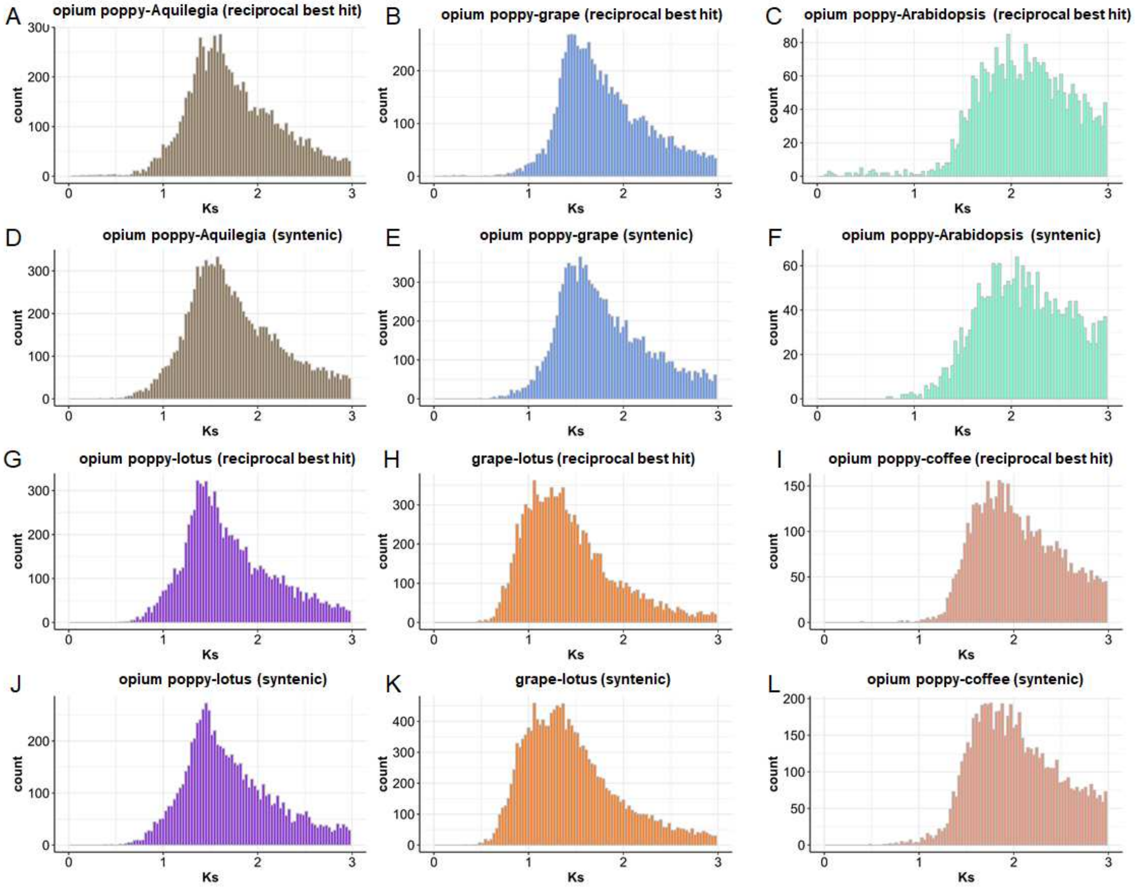


Fig. S14.

Histogram distribution of *Ks* (synonymous substitution rate) for orthologous gene pairs between opium poppy (*Papaver somniferum*) and five different eudicot species: Aquilegia (*Aquilegia coerulea*), Arabidopsis (*Arabidopsis thaliana*), grape (*Vitis vinifera*), lotus (*Nelumbo nucifera*) and coffee (*Coffea arabica*) identified through reciprocal best hit analysis (ABCGHI) and syntenic block analysis through *MCScanX* (DEFJKL) .

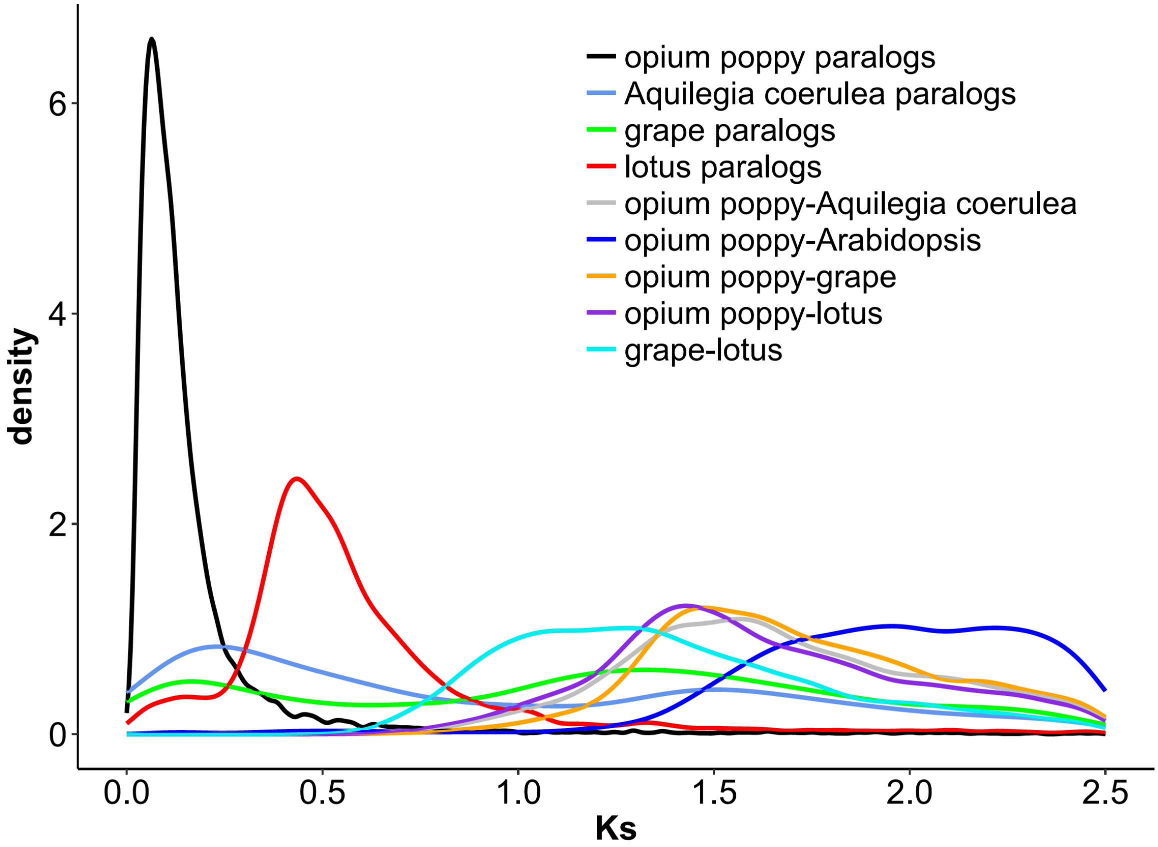


Fig. S15.

Ks (Synonymous substitution rate) distributions for opium poppy RBH (reciprocal best hit) paralogs and orthologs with other eudicots: Arabidopsis (*Arabidopsis thaliana*), grape (*Vitis vinifera*), lotus (*Nelumbo nucifera*) and Aquilegia (*Aquilegia coerulea*).

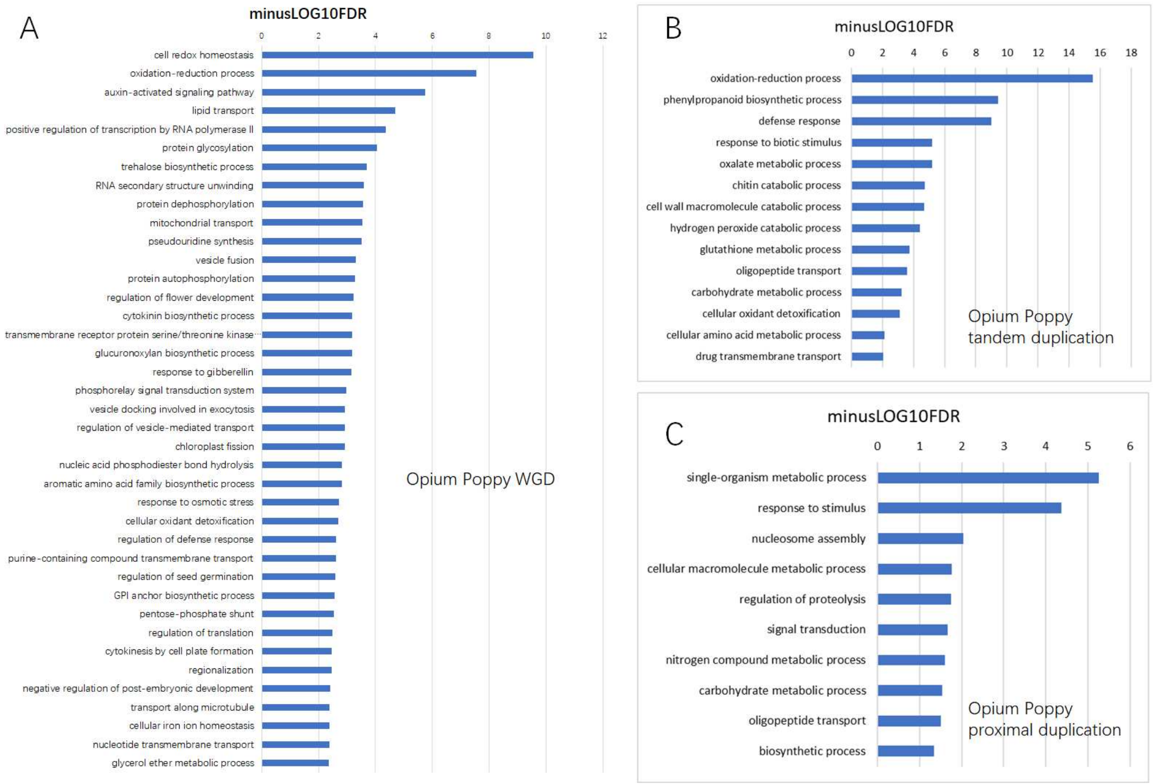


Fig. S16.

Gene ontology enrichment analysis of opium poppy WGD (whole genome duplication) (A) and local gene duplications (B. Tandem duplications; C. Proximal duplications). Number on X-axis represent the minus value of log10 transformed FDR (false discovery rate) in Fisher’s exact tests corrected in multiple tests using *Bonferroni* method.

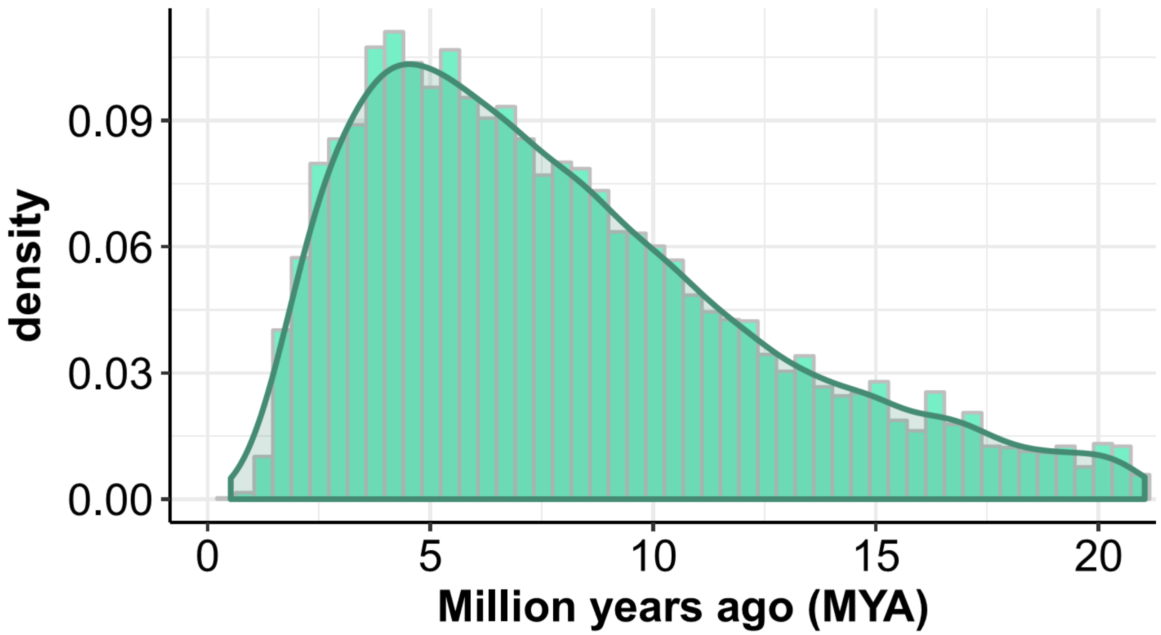


Fig. S17.

Age distribution of opium poppy WGD. Based on the estimated divergence time between *P. somniferum* and *Aquilegia coerulea* (110 MYA) using *BEAST* and the mean *Ks* value (1.53) of *P. somniferum*-*A. coerulea* , we calculated the number of substitution per synonymous site per year for Ranunculales with *r* = 6.98E-9 (divergence date = *Ks* / 2*r*). The same *r* was applied to calculate the age distribution of *P. somniferum* WGD as (7.8±4.35 MYA) based on the *Ks* values (0.108±0.0037).

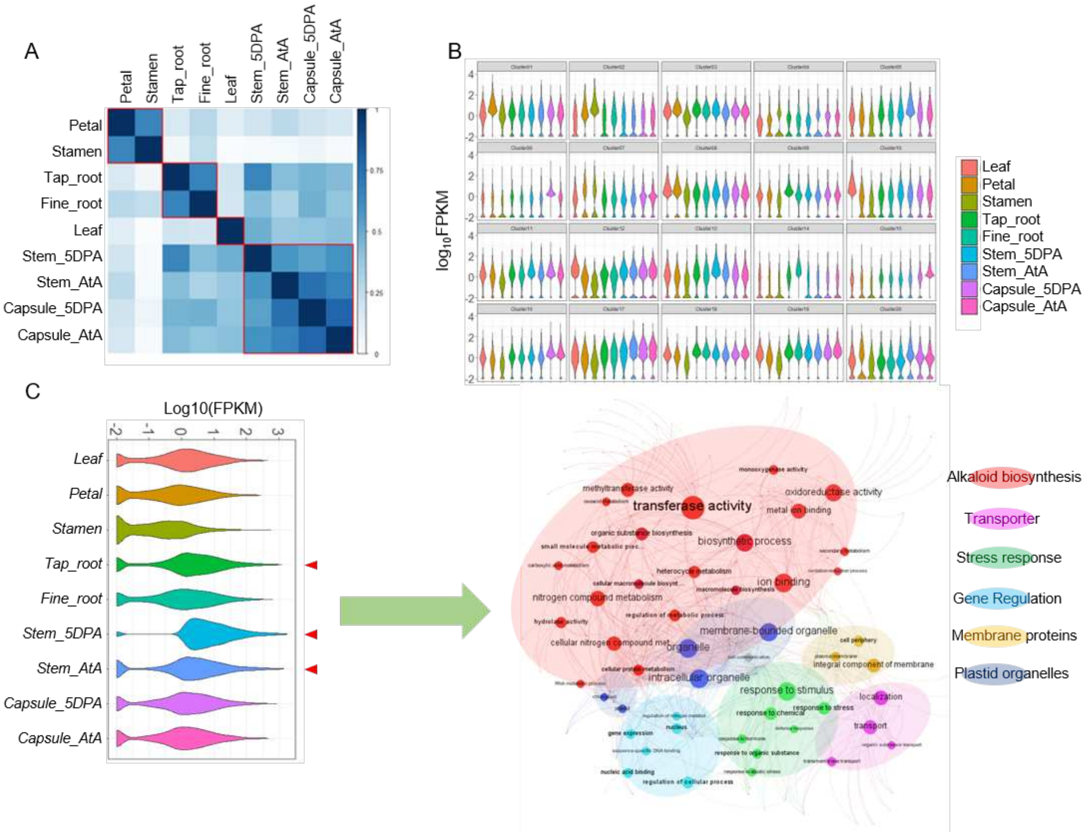


Fig. S18.

Transcriptome analysis reveals tissue-specificity of opium poppy gene expression and identifies a co-regulation network of benzylisoquinoline alkaloid (BIA) and stress response genes. **(A)** Correlation plot of transcriptome in different opium poppy tissues : capsule at anthesis (Capsule_AtA), capsule at 5 days post onset of anthesis (Capsule_5DPA), stem at anthesis (Stem_AtA), stem at 5 days post onset of anthesis (Stem_5DPA), Tap_root, Fine_root, Leaf, Petal and Stamen. Pearson correlation coefficients are colored in scale (0 ~1). Red rectangles delineates the four hierarchical clusters of different tissues. **(B)** Violin plots of different co-expression gene clusters identified using K-means clustering of tissue-specific transcriptomes. In each cluster, Y-axis and X-axis represent Log10-transformed FPKM and tissue types respectively. **(C)** BIA biosynthesis co-regulation network of opium poppy. Left: A violin plot of gene expression levels across nine different tissue types in the BIA super gene cluster co-expression module. Red arrows indicate the three tissue types where gene expression levels are significantly higher than other tissue types. Right: GO ontology Network visualization of the BIA super gene cluster co-expression module. Submodules and nodes are colored correspondingly based on their involvement in specific biological processes.

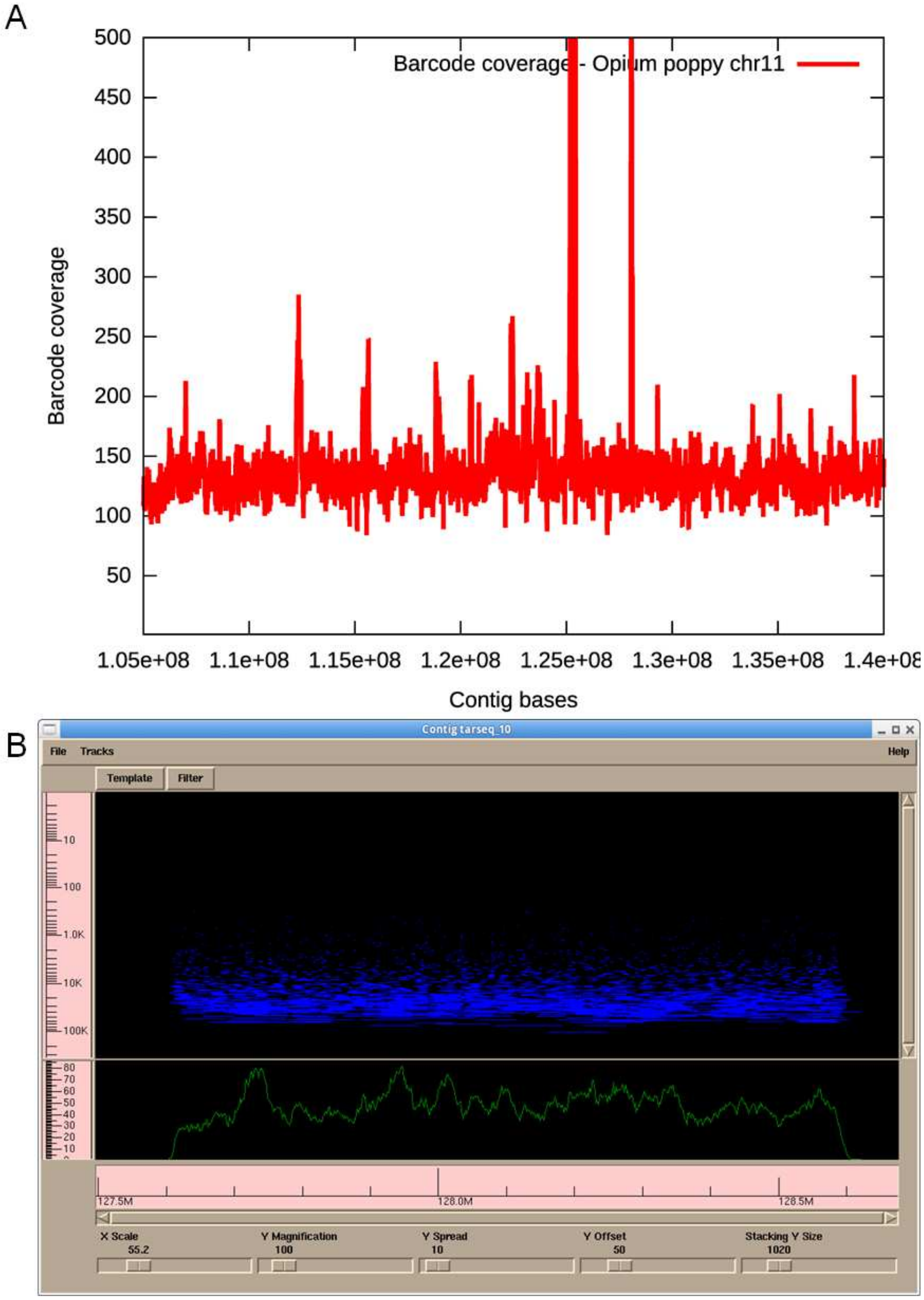


Fig. S19.

Assessment of the assembly accuracy on that region of Chromosome 11 containing the BIA gene cluster. **(A)** Barcode coverage near the end of Chromosome 11 (105Mb – 140Mb). **(B)** Alignment of PacBio long reads against the opium poppy released assembly in the 127.5-128.5Mb region of chromosome 11 containing the BIA gene cluster. Here blue lines display mapping location for each long PacBio read while read coverage is shown in green lines. This confirms continuous read coverage across the 127.5-128.5Mb region of chromosome 11.

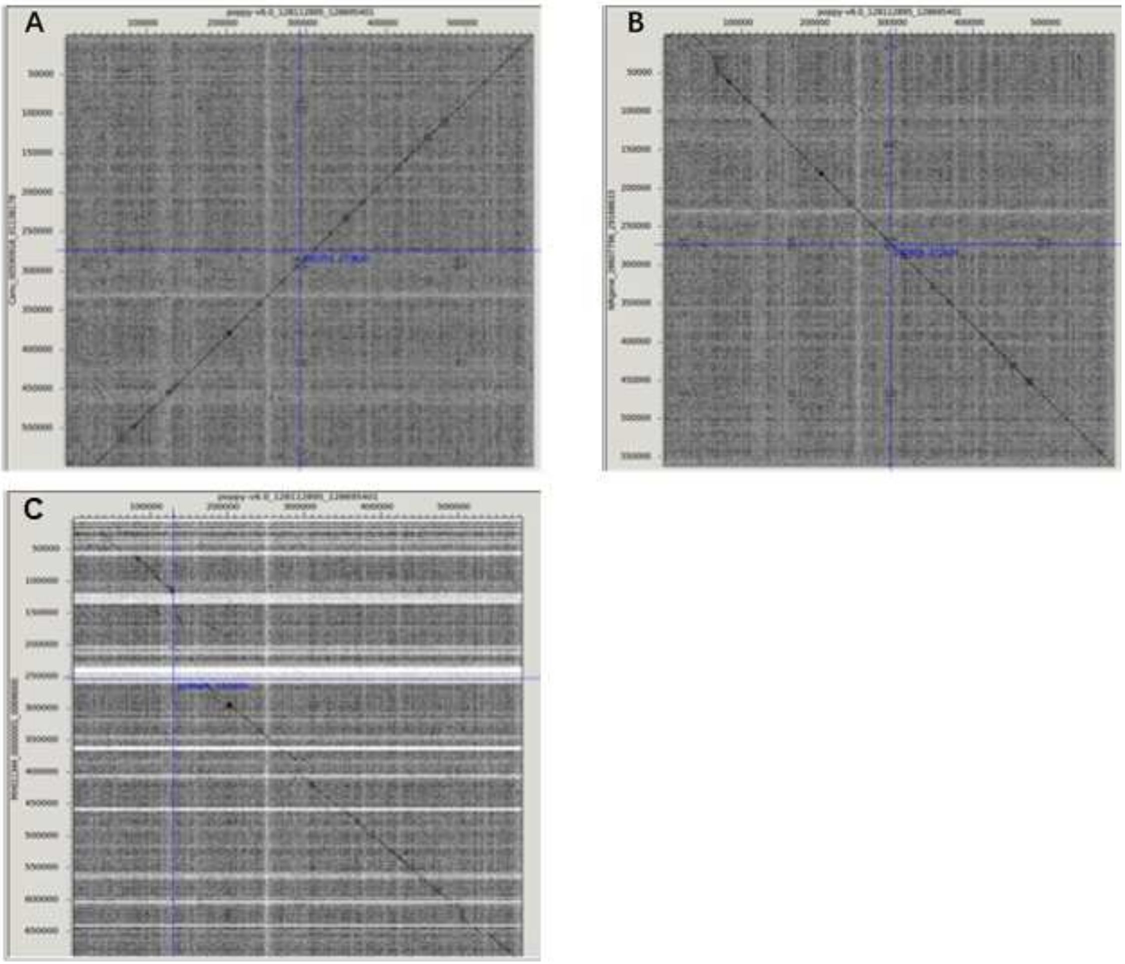


Fig. S20.
Comparison of released opium poppy assembly (accession number PRJNA435796) with: **(A)** the PacBio assembly, **(B)** the NRgene 10X assembly, **(C)** scaffold MH011344 from Chen *et al* (*19*).

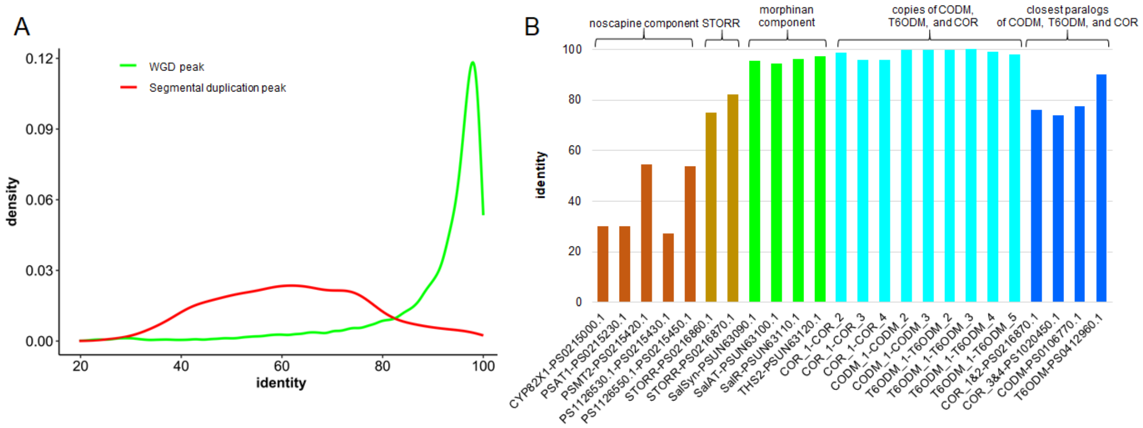


Fig. S21.

Amino acid identity distribution of the two *Ks* peaks for opium poppy as presented in Fig. 1C and the amino acid identities of gene pairs associated with BIA metabolism. **(A)**. Amino acid identity distributions of syntenic gene pairs involved in the opium poppy *Ks* peaks. **(B)**. Amino acid identity of syntenic gene pairs involved in the noscapine and morphinan branch components of the BIA gene cluster (Fig. 2A), STORR with its closest paralogs corresponding to the P450 and reductase modules and local duplicated copies and closest paralogs of COR, CODM, and T6ODM.

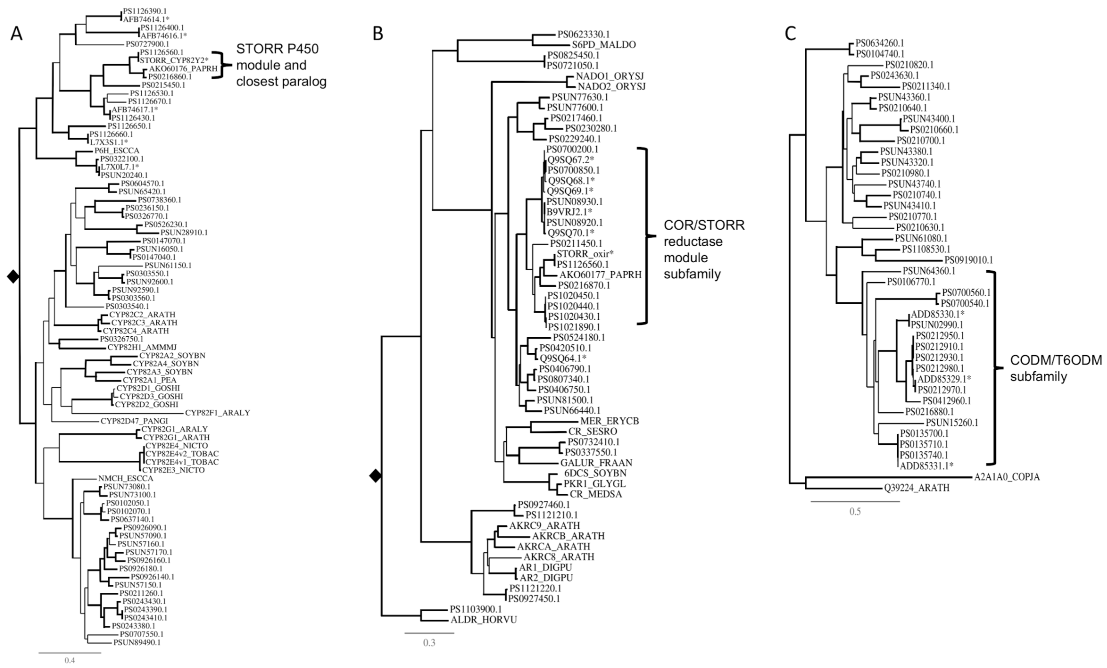


Fig. S22.

Phylogenetic analysis of the cytochrome P450 CYP82, oxidoreductase and CODM/T6ODM gene subfamilies from *P. somniferum*. **(A)** cytochrome P450 CYP82 subfamily together with the N-terminal cytochrome P450 module (STORR_CYP82Y2). **(B)** aldo/keto-reductase 4 subfamily together with the C-terminal oxidoreductase module (STORR_oxired). **(C)** CODM/T6ODM subfamily. All strongly supported subgroups of STORR modules, COR and CODM/T6ODM are highlighted. The Asterisk following a taxon name in the trees indicates a protein sequence derived from a reported opium poppy cDNA sequence. Apart from the annotated opium poppy proteins and two field poppy sequences (AKO60176_PARRH and AKO60177_PARRH), the remaining sequences used in (A) and (B) were reported previously in Winzer *et al.* 2015 (*7*). Species-specific identifiers: *Ammi majus* (_AMMMJ), *Arabidopsis lyrata* (_ARALY), *Arabidopsis thaliana* (_ARATH), *Digitalis purpurea* (_DIGPU), *Erythroxylum coca* (_ERYCB), *Eschscholzia californica* (_ESCCA), *Fragaria x ananassa* (_FRAAN), *Glycyrrhiza glabra* (_GLYGL), Gossypium hirsutum (_GOSHI), Glycine max (_SOYBN), Hordeum vulgare (_HORVU), *Malus domestica* (_MALDO), *Medicago sativa* (_MEDSA), *Nicotiana tabacum* (_TOBAC), *Nicotiana tomentosiformis* (_NICTO), *Oryza sativa Japonica* (_ORYSJ), *Panax ginseng* (_PANGI), *Papaver rhoeas* (_PAPRH), *Pisum sativum* (_PEA), *Sesbania rostrata* (_SESRO), and *Coptis japonica* (_COPJA).

Abbreviations: *N*-methylcoclaurine 3'-hydroxylase (NMCH), *N*-methylstylopine 14-hydroxylase (MSH), protopine 6-hydroxylase (P6H), codeinone reductase (COR), aldose reductase (AR), polyketide reductase (PKR), chalcone reductase (CR), 6'-deoxychalcone synthase (6DCS), galacturonate reductase (GALUR), NAD(P)H- dependent oxidoreductase (NADO), Methylecgonone reductase (MER), aldo-keto reductase (AKR), sorbitol-6-phosphate dehydrogenase (S6PD), aldehyde reductase (ALDR).

All branches are drawn to scale as indicated by the scale bar (substitutions/site). The solid diamonds indicate the root of the phylogenetic trees. Strongly supported nodes with above 70% bootstrap values are highlighted with thickened lines.

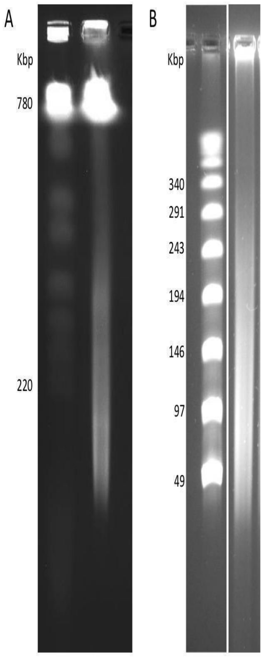


Fig. S23.

Pulsed-field gel electrophoresis images of high molecular weight (HMW) DNA prepared from young opium poppy seedling material. **(A)** HMW DNA prepared by Amplicon Express using their protocol for HMW grade (megabase size) DNA preparations. **(B)** HMW DNA prepared by Amplicon Express using their protocol for NGS grade DNA preparations. Running conditions: 1% agarose, 0.5X TBE. 6 V/cm, 120° included angle, initial switch time: 5 sec, final switch time: 25 sec, run time: 16 hours, 200 ng of genomic DNA loaded, ethidium bromide staining. Lambda PFG Ladder from New England Biolabs (Ipswich, MA, USA) was used as size marker.

Table S1. Raw sequencing data

Table S2. Evaluation of PacBio data

Table S3. Assembly statistics at different stages

Table S4. Summary of Chromosome length in the opium poppy genome

Table S5. The linkage map of opium poppy

Table S6. Unique marker sequences used to construct the opium poppy linkage map in table S5

Table S7. Summary of RNA-seq data

Table S8. ncRNA annotation results

Table S9. Summary of the peaks in Ks distribution of opium poppy paralogs plus opium poppy orthologs and other species

Table S10. Summary of accumulated syntenic block coverage for each scaffold

Table S11. The syntenic blocks detected by *MCScanX* with default parameters

Table S12. The number of paralogous gene pairs in different scaffold pairs

Table S13. Summary of size and proportion of syntenic blocks in each scaffold pair

Table S14. Summary of orthogroup phylogenetic trees supporting the timing of segmental duplications in opium poppy

Table S15. The benzylisoquinoline alkaloid metabolism genes

Table S16. Gene clusters predicted by the *plantiSMASH* method on the opium poppy genome assembly

Table S17. Syntenic blocks on chromosome 2 and unplaced scaffold 21 associated with the BIA gene cluster genes on chromosome 11

Table S18. Details of syntenic blocks across the whole genome - *MCScanX* output file

Table S19. Pairwise sequence comparisons of STORR, CODM, T6ODM, and COR with their corresponding closest paralogs

Captions for databases S1 to S3

Supplementary file 1: All the supplementary tables (table S1 – S19, Supplementary_tables.xlsx).

Supplementary file 2: Phylogenetic trees of 95 orthogroups each containing opium poppy paralog pairs and their homologous genes in 22 land plant species (phylogenomic_trees.pdf).

Supplementary file 3: Multiple sequence alignment (.aln) and phylogenetic tree files (newick format) generated by the phylogenomic analysis on 95 orthogroups each containing opium poppy paralog pairs and their homologous genes in 22 land plant species. (Alignment&Trees.tar.gz)

References:

22. R. Magnavaca, C. O. Gardner, R. B. Clark, Evaluation of inbred maize lines for aluminum tolerance in nutrient solution, in Genetic Aspects of Plant Mineral Nutrition. Developments in Plant and Soil Sciences, W. H. Gabelman, B. C. Loughman, Eds. (Springer, Dordrecht, 1987), vol. 27, pp. 255-265.
23. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. Nature Methods. 12, 357-360 (2015).
24. M. Pertea et al., StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 33, 290-295 (2015).
25. A. C. Frazee et al., Flexible analysis of transcriptome assemblies with Ballgown. bioRxiv, (2014), (available at https://www.biorxiv.org/content/early/2014/09/05/003665).
26. M. G. Grabherr et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnol. 29, 644-652 (2011).
27. M. J. Raymond, Isolation and characterization of latex-specific promoters from Papaver somniferum. Master Thesis, Virginia Polytechnic Institute and State University, (2004).
28. S. Koren et al., Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27, 722-736 (2017).
29. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754-1760 (2009).
30. E. P. Murchison et al., Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. Cell 148, 780-791 (2012).
31. N. I. Weisenfeld, V. Kumar, P. Shah, D. M. Church, D. B. Jaffe, Direct determination of diploid genome sequences. Genome Res. 27, 757-767 (2017).
32. J.-S. Seo et al., De novo assembly and phasing of a Korean human genome. Nature 538, 243-247 (2016).
33. A. M. Hulse-Kemp et al., Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. Hortic. Res. 5, 4 (2018), doi: 10.1038/s41438-017-0011-0. eCollection 2018..

34. Y. Mostovoy et al., A hybrid approach for de novo human genome sequence assembly and phasing. Nature methods 13, 587-590 (2016).

35. R. Avni et al., Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357, 93-97 (2017).

36. M. C. Luo et al., Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. Nature 551, 498-502 (2017).

37. G. Zhao et al., The Aegilops tauschii genome reveals multiple impacts of transposons. Nature Plants. 3, 946-955 (2017).

38. C. N. Hirsch et al., Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell. 28, 2700-2714 (2016).

39. F. Lu et al., High-resolution genetic mapping of maize pan-genome sequence anchors. Nature Commun. 6, 6914 (2015), doi: 10.1038/ncomms7914.

40. H. Tang et al., ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. 16, 3 (2015), doi: 10.1186/s13059-014-0573-1.

41. C.-S. Chin et al., Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods. 13, 1050-1054 (2016).

42. K.-P. Koepfli, B. Paten, G. K. C. o. Scientists, S. J. O'Brien, The Genome 10K Project: a way forward. Annu. Rev. Anim. Biosci. 3, 57-111 (2015).

43. C.-L. Xiao et al., MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nature Methods. 14, 1072-1074 (2017).

44. A. McKenna et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).

45. J. K. Bonfield, A. Whitwham, Gap5—editing the billion fragment sequence assembly. Bioinformatics 26, 1699-1703 (2010).

46. W. Bao, K. K. Kojima, O. Kohany, Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6, 11 (2015), doi: 10.1186/s13100-015-0041-9.

47. G. S. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. BMC bioinformatics 6, 31 (2005), doi: 10.1186/1471-2105-6-31.

48. O. Keller, M. Kollmar, M. Stanke, S. Waack, A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics 27, 757-763 (2011).

49. I. Korf, Gene finding in novel genomes. BMC bioinformatics 5, 59 (2004).

50. A. Lomsadze, V. Ter-Hovhannisyan, Y. O. Chernoff, M. Borodovsky, Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33, 6494-6506 (2005).

51. P. Lamesch et al., The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40, D1202-1210 (2012).

52. J. C. Dohm et al., The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature 505, 546-549 (2014).

53. P. Jones et al., InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236-1240 (2014).

54. A. Conesa, S. Gotz, Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int. J Plant. Genomics. 2008, 619832 (2008), doi: 10.1155/2008/619832.

55. T. M. Lowe, S. R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955-964 (1997).

56. P. P. Gardner et al., Rfam: updates to the RNA families database. Nucleic Acids Res. 37, D136-140 (2009).

57. E. P. Nawrocki, S. R. Eddy, Query-dependent banding (QDB) for faster RNA similarity searches. PLoS Comput. Biol. 3, e56 (2007), doi: 10.1371/journal.pcbi.0030056.

58. J. A. Bailey, D. M. Church, M. Ventura, M. Rocchi, E. E. Eichler, Analysis of segmental duplications and genome assembly in the mouse. Genome Res. 14, 789-801 (2004).

59. N.H. Putnam et al., The amphioxus genome and the evolution of the chordate karyotype. Nature 453, 1064-1071 (2008).

60. D. Wang, Y. Zhang, Z. Zhang, J. Zhu, J. Yu, KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genom. Proteom. Bioinformatic. 8, 77-80 (2010).

61. R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792-1797 (2004).

62. S. Kumar, G. Stecher, K. Tamura, MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870-1874 (2016).

63. A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690 (2006).

64. L. Scrucca, M. Fop, T. B. Murphy, A. E. Raftery, Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R Journal 8, 289-317 (2016).

65. R. Ming et al., Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.), Genome Biol. 14, R41 (2013).

66. Y. Jiao, J. Li, H. Tang, A. H. Paterson, Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in Monocots. The Plant Cell 26, 2792-2802 (2014).

67. Y. Van de Peer, E. Mizrachi, K. Marchal, The evolutionary significance of polyploidy. Nature Reviews Genetics 18, 411 (2017).

68. H. Badouin et al., The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature 546, 148 (2017).

69. S. Sato et al., The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635 (2012).

70. L. Li, C.J. Stoeckert, D.S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178-2189 (2003).

71. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059-3066 (2002).

72. S. Capella-Gutierrez, J.M. Silla-Martinez, T. Gabaldon, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972-1973 (2009).

73. J. D. Thompson, T. Gibson, D. G. Higgins, Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinformatics. Chapter 2:Unit 2.3. (2002).

74. J. Castresana, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol.17, 540-552 (2000).

75. A.M. Bolger, M. Lohse, B. Usadel. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).

76. M. Bastian, S. Heymann, M. Jacomy. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media (2009), http://gephi.org.