Title

SOI: Robust identification of orthologous synteny with the *Orthology Index* and broad applications in evolutionary genomics

Short title: Robust identification of orthologous synteny

Authors

Ren-Gang Zhang[1,2†], Hong-Yun Shang[1,2†], Richard Ian Milne[3], Fabricio Almeida-Silva[4,5], Hengchi Chen[4,5], Min-Jie Zhou[1,2], Heng Shu[1,2], Kai-Hua Jia[8*], Yves Van de Peer[4,5,6,7*], Yong-Peng Ma[1*]

Affiliations

[1] Yunnan Key Laboratory for Integrative Conservation of Plant Species with Extremely Small Populations / Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201 Yunnan, China.

[2] University of the Chinese Academy of Sciences, Beijing 100049 Beijing, China.

[3] Institute of Molecular Plant Sciences, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JH, UK

[4] Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium

[5] VIB Center for Plant Systems Biology, VIB, 9052 Ghent, Belgium

[6] Department of Biochemistry, Genetics and Microbiology, Centre for Microbial Ecology and Genomics, University of Pretoria, Pretoria 0028, South Africa

[7] College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China

[8] Institute of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences, Jinan 250100 Shandong, China.

[†] These authors contributed equally.

[*] Corresponding to: Yong-Peng Ma (mayongpeng@mail.kib.ac.cn)

Yves Van de Peer (yves.vandepeer@psb.vib-ugent.be)

Kai-Hua Jia (kaihuajia_saas@163.com)

Abstract

With the explosive growth of whole-genome datasets, the accurate detection of orthologous synteny has become crucial for the reconstruction of evolutionary history. However, the currently available methods for the identification of orthologous synteny have great limitations: the methods are difficult to scale with varying polyploidy histories, and the accurate removal of out-paralogy is challenging. In this study, we developed a scalable and robust approach, the *Orthology Index* (*OI*), to identify orthologous synteny. Our evaluation of a large-scale dataset with diverse polyploidization events demonstrated that the technique is highly reliable. This discovery highlights *OI* as a potentially unified criterion for the identification of orthologous synteny, and this is further validated using simulation-based benchmarks. In addition, we explore its broad applications in reconstructing the evolutionary histories of plant genomes, including inference of polyploidy, identification of reticulation, and phylogenomics. In conclusion, *OI* offers a robust, interpretable, and scalable approach for identifying orthologous synteny, significantly enhancing our analytical prowess in plant evolutionary genomics.

Keywords: Orthologous synteny, Polyploidy, Reticulation, Phylogenomics, *Orthology Index*

MAIN TEXT

Introduction

Reconstruction of the evolutionary histories of organisms, including inference of the tree/network of life, identification of polyploidy events, and placement of these events on the tree/network, usually relies on the orthologous relationships at gene, chromosomal block, chromosome, subgenome and/or whole genome scales. In general, the greater the number of reliably orthologous genes or loci are involved in the reconstruction of evolutionary history, the more confidence can be placed in the reconstruction. Indeed, phylogenomic reconstructions using genome-scale data have become the gold standard for understanding the evolution of lineages in the tree of life (1). Orthologous synteny has been established as a proxy to generate a maximum number of reliable orthologs, from the chromosomal to the subgenomic/genomic scales, for example, in the reconstruction of evolutionary histories for the octoploid strawberry (2) and the major angiosperm clades (3). The identification of orthologous relationships is also vital for other synteny-based analyses, such as inference of polyploidy (4) and subgenome phasing (2). However, accurate identification of orthologous synteny is still challenging. This is especially true in plants, where recurrent whole-genome duplication (WGD, also known as polyploidization) events have produced substantial numbers of syntenic paralogs. This can significantly complicate the inference of the orthology (1,5) and may mislead the reconstruction of evolutionary history. For example, the overlooked orthologous relationships in the synteny between two *Papaver* species resulted in an incorrect interpretation of the history of polyploidy in these species (6-8).

To date, two main strategies have been employed to identify orthologous synteny. One strategy (the "criterion strategy") is to filter the detected synteny using certain criteria, which are usually specific to each case and are therefore not scalable for large-scale datasets. One widely-used criterion for this is the synonymous substitution rate (*Ks*). Because the WGD event (which produces paralogs) and speciation event (which produces orthologs) occurred at different times in the past, *Ks* between syntenic gene pairs can be used to differentiate the ages of orthologous and paralogous synteny (9,10). However, such events occur at different times in different lineages, and substitution rates also vary between lineages, both of which cause *Ks* values to vary case by case. Hence, *Ks*-based methods are not always effective for distinguishing syntenic blocks from different evolutionary events (11), and cannot be universally applied to the large-scale automated identification of orthologous syntenic blocks. Another criterion, *homo*, has been proposed for use with WGDI, to extract the best homology (orthology) of syntenic blocks (12). Unfortunately, this criterion relies on a parameter, *multiple*, to define the top number of hits as best hits (12), but *multiple* cannot be universally applied to different cases with different syntenic depth ratios. A further tool, QUOTA-ALIGN (11) was developed to screen orthologous syntenic blocks under given constraints on syntenic depths (*QUOTA*). However, the prior *QUOTA* criterion must be set by users, and will differ between cases with different evolutionary histories, thus it is also challenging to robustly distinguish synteny blocks from different events using QUOTA-ALIGN. For example, it is necessary to set *QUOTA* = 1:1 for *Arabidopsis thaliana* and *A. lyrata*, but to set *QUOTA* = 4:2 for *A. thaliana* and poplar (11). This is also similarly required when setting the *multiple* parameter in WGDI (-c option).

An alternative strategy (the "pre-inferred strategy") for the identification of orthologous

synteny is the use of pre-inferred orthologs to call synteny, with tools such as MCScanX_h (13). This strategy is scalable for large-scale datasets. However, hidden out-paralogs (i.e. false positives in ortholog inference) in the pre-inferred orthologs may result in out-paralogous syntenic blocks that need to be further removed with the above mentioned "criterion" strategy. Our previous work, such as that on *Salix* (14) and *Populus* (9), faced this problem. Moreover, the presence of hidden orthologs (i.e. false negatives in ortholog inference) can reduce the efficiency of the subsequent detection of orthologous synteny in practice.

To address these issues and to screen orthologous synteny robustly for large-scale genomic datasets, we developed a scalable approach called the *Orthology Index* (*OI*), which describes the proportion of collinear gene pairs that are pre-inferred as orthologs (see Methods). We evaluated the efficacy of *OI* using a large-scale dataset comprising 91 well-documented cases with diverse WGD and speciation events and a simulated benchmarking dataset. Our results demonstrated the high robustness and accuracy of *OI*, which outperformed the existing methods. We then explored its broad usefulness in evolutionary inference in analyses of polyploidy, reticulation and phylogenomics. We finally integrated the index into a toolkit (freely available from https://github.com/zhangrengang/SOI) to facilitate its use.

Methods

Empirical data collection and pre-processing

The genomic data were obtained from public databases or from the corresponding authors, as detailed in **Table S2**.

An all-versus-all BLAST search of protein sequences for each species was conducted pairwise using DIAMOND v0.9.24 (15). Orthologous relationships were inferred using OrthoFinder v2.3.1 (Emms and Kelly 2019) (parameters: -M msa). Syntenic/collinear blocks were identified with the `-icl` option of WGDI (12) v0.6.2 (default parameters). The synonymous substitution rate (*Ks*) was calculated for homologous gene pairs using the `-ks` option of WGDI (default parameters).

Definition of the *Orthology Index*

We propose an index, named the *Orthology Index*, to distinguish the syntenic blocks of orthology from out-paralogy. The index (*OI*) is defined as:

$$OI = \frac{n}{m},$$

where m represents the total number of collinear gene pairs in a collinearity block, and n denotes the number of collinear gene pairs pre-inferred as orthologs. Thus, the index represents the proportion of orthologous syntenic gene pairs within a syntenic block. It ranges from 0 to 1 and can differentiate orthology from out-paralogy. Employment of the index is expected to result in a polarized pattern, where orthologous synteny is expected to result in a high *OI* value (with an expectation of 1, relying on the recall of pre-inferred orthologs), whereas out-paralogous synteny is expected to result in a relatively low *OI* value (with an expectation of 0, relying on the precision of pre-inferred orthologs). From our experience with many different cases (Figs. S1–90), the major peak with the highest index can be considered to be orthologous in principle.

Implementation of the *Orthology Index*

Using this index as a foundation, we developed a user-friendly all-in-one toolkit (called SOI) for visualization and downstream analyses. The subcommand 'dotplot' enables visualization and evaluation of synteny, with the dots colored by the *OI* or *K*s values. The subcommand 'filter' retrieves orthologous blocks by discarding all blocks with less than a default *OI* value of 0.6. However, users can also apply a stringent standard (higher *OI* and block length cutoffs) to obtain long, highly credible blocks. The subcommand 'cluster' groups orthologous syntenic genes into syntenic orthogroups (SOGs) by constructing an orthologous syntenic graph and applying the Markov Cluster (MCL) algorithm (16) to perform graph clustering with breaking weak links and bridging unexpectedly disrupted links. The clustering algorithm is widely employed by popular tools of orthogroup clustering, such as OrthoFinder2 (17), SonicParanoid2 (18), and OrthoMCL (19). The subcommand 'outgroup' retrieves syntenic orthologs from outgroups that lack WGDs shared with ingroups. The subcommand 'phylo' reconstructs multi-copy or single-copy gene trees, by aligning protein sequences with MAFFT v7.481 (20), converting protein alignments to codon alignments with PAL2NAL v14 (21), trimming alignments with trimAl v1.2 (22) (parameter: -automated1) and reconstructing maximum-likelihood trees with IQ-TREE v2.2.0.3 (23). These gene trees serve as input to infer a species tree with the coalescence-based method ASTRAL-Pro v1.10.1.3 (24). The default threshold for missing taxa is set to 40 %, according to an evaluation for this parameter (25).

This tool is implemented in Python3 and supports synteny outputs from state-of-the-art synteny detectors, including MCscan/JCVI (26), MCscanX (13) and WGDI (12), as well as orthology outputs from OrthoFinder2 (17), Broccoli (27), SonicParanoid2 (18), Proteinortho6 (28), InParanoid (29), OrthoMCL (19) and other tools upon request. These orthology inference methods have competitive accuracy (18,27,30) and thereby should produce similar *OI* values. The tool can be easily installed using the conda environment or the Apptainer/Singularity container system (31). Since both synteny and orthology analyses are generally standard procedures for a genomics project, the *OI* tool can be seamlessly integrated into these and their downstream pipelines. The source code is accessible on GitHub (https://github.com/zhangrengang/SOI).

Simulation-based benchmarks

To validate the feasibility and robustness of *OI*, we performed simulation-based benchmarks. We simulated a shared-WGD scenario similar to the Salicaceae case (Fig. 1A) using Zombi (32). The input species tree was "(((A1:0.2, B1:0.2):ΔT, (A2:0.2, B2:0.2):ΔT):0.2, O:ΔT+0.4):0.1", where A1 and A2, and B1 and B2 is the two subgenomes of taxa A and B, respectively, and were treated as pseduo-species, and O is an outgroup species, while ΔT varied from 0.01 to 1 (scaled in time unit of substitutions per site). 5000 coding sequences were output for each simulation with other parameters (e.g. those regarding substitutions and chromosomal rearrangements) being default (see https://github.com/AADavin/Zombi/tree/master/Parameters/). Then the simulated three-taxon genomic data were input to the aforementioned OrthoFinder2–WGDI–SOI pipeline to calculate the *Ks* and *OI* values and to identify syntenic orthologs between taxa A and B. The true positives (TP), false positives (FP), and false negatives (FN) were counted from the results, and the precision (P = TP / (TP + FP)), recall (R = TP / (TP + FN)), and F1-score ($2 \times P \times R / (P + R)$) were calculated. The simulation was independently replicated 50 times for each ΔT condition.

Results and Discussion

Identification of orthologous synteny using the Orthology Index

We aimed to create a method that would robustly distinguish orthologous synteny from out-paralogy for genomes of any two species. The genomes should have shared a whole genome duplication event (WGD, or polyploidization; producing out-paralogy) and a speciation event (producing orthology), and these events should have occurred within a certain time period (denoted as ΔT, **Fig. 1A**). To evaluate our approach, we selected ninety well-documented plant species pairs, with each pair being descended from a common ancestor that had undergone at least one polyploidy event (**Table S1**, **Figs. S1-90**), along with the typical poplar–willow (Salicaceae) case (**Fig. 1**).

Fig. 1 shows an analysis for the Salicaceae case, represented by *Populus trichocarpa* and *Salix dunnii*. The two genera speciated ~50 million years ago (Mya) (33,34), after a WGD event in their common ancestor ~60 Mya (33-36) (**Fig. 1A**). Additionally, both genera also share a much earlier paleohexaploidization event (the γ event, whole genome triplication or WGT) in the common ancestor of core eudicots ~120 Mya (37), making three detectable evolutionary events in total (**Fig. 1A**).

With the "criterion" strategy using the collinearity detector WGDI (12), syntenic blocks derived from these three evolutionary events could be identified and distinguished visually using their distinct synonymous substitution rate (*Ks*) values and fragmentation extent (the more ancient the event, the more fragmented the blocks or greater the chromosomal rearrangements, and the higher the *Ks* value) (**Fig. 1B**). These three categories of syntenic blocks were denoted WGT-SBs ($Ks \approx 1.5$, most fragmented, out-paralogous), WGD-SBs ($Ks \approx 0.27$, moderately fragmented, out-paralogous), and S-SBs ($Ks \approx 0.13$, least fragmented, orthologous) (**Fig. 1B**) according to the evolutionary events (i.e. WGT, WGD and speciation) (**Fig. 1A**). However, the recent two *Ks* peaks derived from WGD-SBs (out-paralogy) and S-SBs (orthology) largely overlapped (**Fig. 1B**) and were difficult to split with a hard cutoff. This is likely to be because the evolutionary rates in different genes also vary, resulting in similar accumulated substitutions for genes from different events. Indeed, only a few empirical cases (e.g. **Figs. S4, 30**) can employ a hard *Ks* cutoff to distinguish orthology and out-paralogy. For some extreme cases where the ΔT was quite small, *Ks* peaks from two events (shared WGD and speciation) overlapped completely, and showed only one peak (e.g. **Figs. S1, 8, 16, 34**) with no distinction between orthology and out-paralogy.

With the "pre-inferred" strategy using the orthology inference method, OrthoFinder2 (17), the peak ($Ks \approx 0.27$) of out-paralogs from WGD-SBs was significantly reduced, while out-paralogs from the older WGT event (WGT-SBs, $Ks \approx 0.13$) were almost entirely eliminated (**Fig. 1C**). This suggests that the level of accuracy of orthology inference by OrthoFinder2 (~ 85 % orthologs located in the orthologous chromosomal segments for the Salicaceae case) is relatively high. Nevertheless, in the Salicaceae case, this method inferred a substantial number of "orthologous" genes (~ 15 %) that exhibit synteny in the out-paralogous blocks ($Ks \approx 0.27$; **Fig. 1C**), suggesting that they are hidden paralogs and would introduce out-paralogous synteny if they were imported into tools such as MCScanX_h (13). This problem was also observed in numerous instances where the ΔT was not substantial (e.g. **Figs. S1, 5, 6, 8, 10**), which could be attributed to systematic errors or biased gene loss, and which could have a detrimental effect on downstream analyses, such as phylogenomics (1).

Consequently, we combined the algorithmic advances of the two methods described above by introducing a straightforward index, referred to as the *Orthology Index* (*OI*), to determine the orthology of a syntenic block. *OI* represents the proportion of gene pairs that are pre-inferred to be orthologs in a syntenic block. Thus, orthologous synteny expects an *OI* of 1 (i.e., all syntenic gene pairs are inferred to be orthologs), while out-paralogous synteny expects an *OI* of 0 (i.e., none of the syntenic gene pairs are inferred to be orthologs). However, for real datasets, due to the limited recall and precision from orthology inference methods, the *OI* distribution could be biased from the expectations. For the Salicaceae case, the WGT-SBs and WGD-SBs (both out-paralogous but with different ages) exhibit an *OI* of nearly 0 and approximately 0.1, respectively, whereas the S-SBs (orthologous) display an *OI* peak of approximately 0.9 (**Fig. 1D**), close to the expectations. The *OI* distribution reflected the same orthologous and out-paralogous relationships as *Ks* in this case, but was able to much better distinguish between them than was *Ks* (**Fig. 1 B** and **D**). Remarkably, the *OI* peaks derived from orthology and out-paralogy did not overlap, but showed a polarized pattern with a very clear, wide dividing range (*OI* = ~0.3–0.7) between the orthology and out-paralogy peaks (**Fig. 1D**). Therefore, in this case, the *OI* was easily able to distinguish orthology from out-paralogy, with a hard cutoff between 0.3 and 0.7 (**Fig. 1D**).

We next applied the index analysis to ninety study cases with diverse polyploidy histories and varying ΔT (**Table S1, Figs. S1-90**). We found that there were also clear divides around *OI* = ~0.5–0.6 in nearly all instances (although minor noise around *OI* = ~0.5–0.6 can be observed in some cases) (**Figs. 2A, S1-90D**), suggesting that a potentially unified *OI* criterion could be employed for the identification of orthologous synteny. The *OI* cutoff was further adjusted to 0.6 empirically to identify orthologous synteny according to the *OI* distributions in these cases (**Figs. 1D and 2A, Figs. S1–90D**), in order to balance the visible false positives (out-paralogy mis-identified as orthology) and false negatives (excessive removal of true orthologous synteny). Interestingly, we observed that the index performed exceptionally and consistently well in identifying orthologous synteny in these instances (**Fig. 1E, Figs. S1–90E**).

For example, in the Salicaceae case, a cutoff value of *OI* = 0.6 resulted in retrieval of a clean 1:1 orthology relationship without visible false positives or false negatives (**Fig. 1E**), as expected. In this case, 12.2 % of the syntenic blocks and 47.6 % of the syntenic pairs were retained. Of the retained syntenic gene pairs, 90.4 % were pre-inferred as orthologs using OrthoFinder2, while the remaining 9.6 % were presumed to be orthologs but were not identified using OrthoFinder2. Moreover, although some visible false positives or false negatives were observed in a few study cases, there were many fewer false positives than when using the above two strategies (**Figs. S50, 68, 74, 86E**). This finding highlights the efficiency and robustness of *OI* as a unified criterion for identifying orthologous synteny.

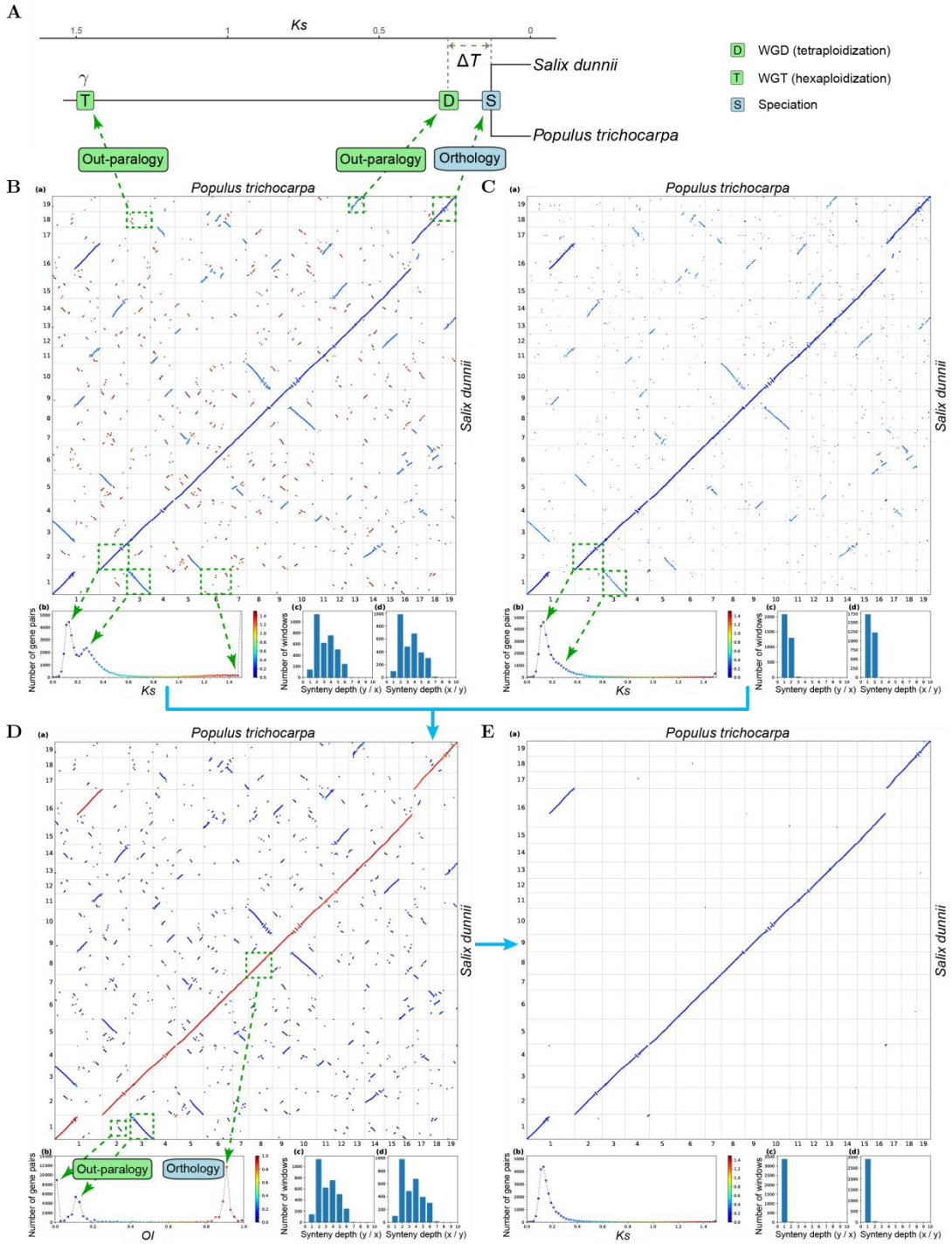


Fig. 1. The illustration of the *Orthology Index* when identifying orthologous synteny for the typical Salicaceae case. A) Schematic of the evolutionary history of the poplar and willow genomes, adapted from the literature (33-36). **B**) *Ks*-colored dot plots showing synteny detected by WGDI, with an observable distinction between the three categories of syntenic blocks derived from three evolutionary events (three peaks: *Ks* ≈ 1.5, *Ks* ≈ 0.27, and *Ks* ≈ 0.13). **C**) *Ks*-colored dot plots illustrating the orthology as inferred by OrthoFinder2, with an observable high proportion (~15 %) of hidden out-paralogs (*Ks* ≈ 0.27). **D**) *Orthology Index* (*OI*)-colored dot plots: integrating synteny (B) and orthology (C), showing polarized distinction of the three categories of syntenic blocks (three peaks: *OI* ≈ 0, *OI* ≈ 0.1, and *OI* ≈ 0.9). **E**) *Ks*-colored dot plots of synteny after application of an *OI* cutoff of

0.6, with clean one-to-one orthology as expected from the evolutionary history. B–E are plotted using the 'dotplot' subcommand with four subplots: a) dot plots colored by *Ks* or *OI* (x-axis and y-axis, chromosomes of the two genomes; a dot indicates a homologous gene pair between the two genomes), b) histogram of *Ks* or *OI* (x-axis, *Ks* or *OI*; y-axis, number of homologous gene pairs), using the same color map as the dot plots, c-d) synteny depth (relative ploidy) derived from 50-gene windows (x-axis, synteny depth; y-axis, number of windows). Examples of the syntenic blocks from three evolutionary events [referred to as WGT-SBs ($Ks \approx 1.5$, $OI \approx 0$), WGD-SBs ($Ks \approx 0.27$, $OI \approx 0.1$), and S-SBs ($Ks \approx 0.13$, $OI \approx 0.9$)] are highlighted with dashed squares. These are associated with the evolutionary events and peaks of *Ks* or *OI*, indicated by arrows, and labeled as 'Out-paralogy' or 'Orthology'. Additional cases illustrating other lineages can be found in **Figs. S1–90** (summarized in **Table S1**).

To validate the feasibility and robustness of *OI*, we further performed benchmarking by simulating a shared-WGD scenario similar to that of the Salicaceae case (**Fig. 1A**). ΔT was correlated to the unexpected noises around $OI = 0.5$ in the empirical datasets (**Fig. 2B**), and was therefore assumed to be a key parameter affecting the accuracy in identifying orthologous synteny. With a varied ΔT ($\Delta T \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$, in a time unit of expected substitutions per site), *OI* consistently showed much higher ability to distinguish between orthology and out-paralogy than did *Ks* (**Fig. 2C**). At values of ΔT between 0.01 and 0.1, *Ks* values from both orthology and out-paralogy showed only one peak without any observable distinction. When ΔT reached 0.2, the two *Ks* peaks still overlapped significantly and could not be efficiently distinguished. In striking contrast, *OI* showed clear dividing ranges around $OI = \sim 0.5–0.6$ for all values of ΔT (**Fig. 2C**). By applying a hard cutoff of *OI* around ~0.5–0.6, both the precision and recall consistently approached to 1.0 under different values of ΔT (**Fig. 2D**). Moreover, when ΔT was between 0.01 and 0.05, where OrthoFinder2 indicated a relatively low accuracy in ortholog inference (precision = 0.62–0.85, recall = 0.66–0.83, and F1 score = 0.64–0.84), the *OI* method with a hard cutoff (cutoff = 0.5 or 0.6) highly improved both the precision (0.992–1.0) and recall (0.992–1.0) of ortholog detection (**Fig. 2D–E**). This suggests that *OI* could significantly improve ortholog inference under these conditions and could therefore facilitate syntelog-based analyses, such as phylogenomics and pangenomics. Overall, these simulation-based benchmarking results were highly consistent with those above based on empirical datasets, demonstrating that the *OI* method outperforms the existing methods (*Ks*-based and "pre-inferred") and is able to produce accurate results when identifying orthologous synteny.

OI can therefore be applied in automated pipelines for large-scale datasets (ranging from dozens to hundreds of genomes at present) with a unified threshold of *OI*. We integrated the *OI* method into a toolkit (https://github.com/zhangrengang/SOI) to identify orthologous synteny and to facilitate downstream applications.

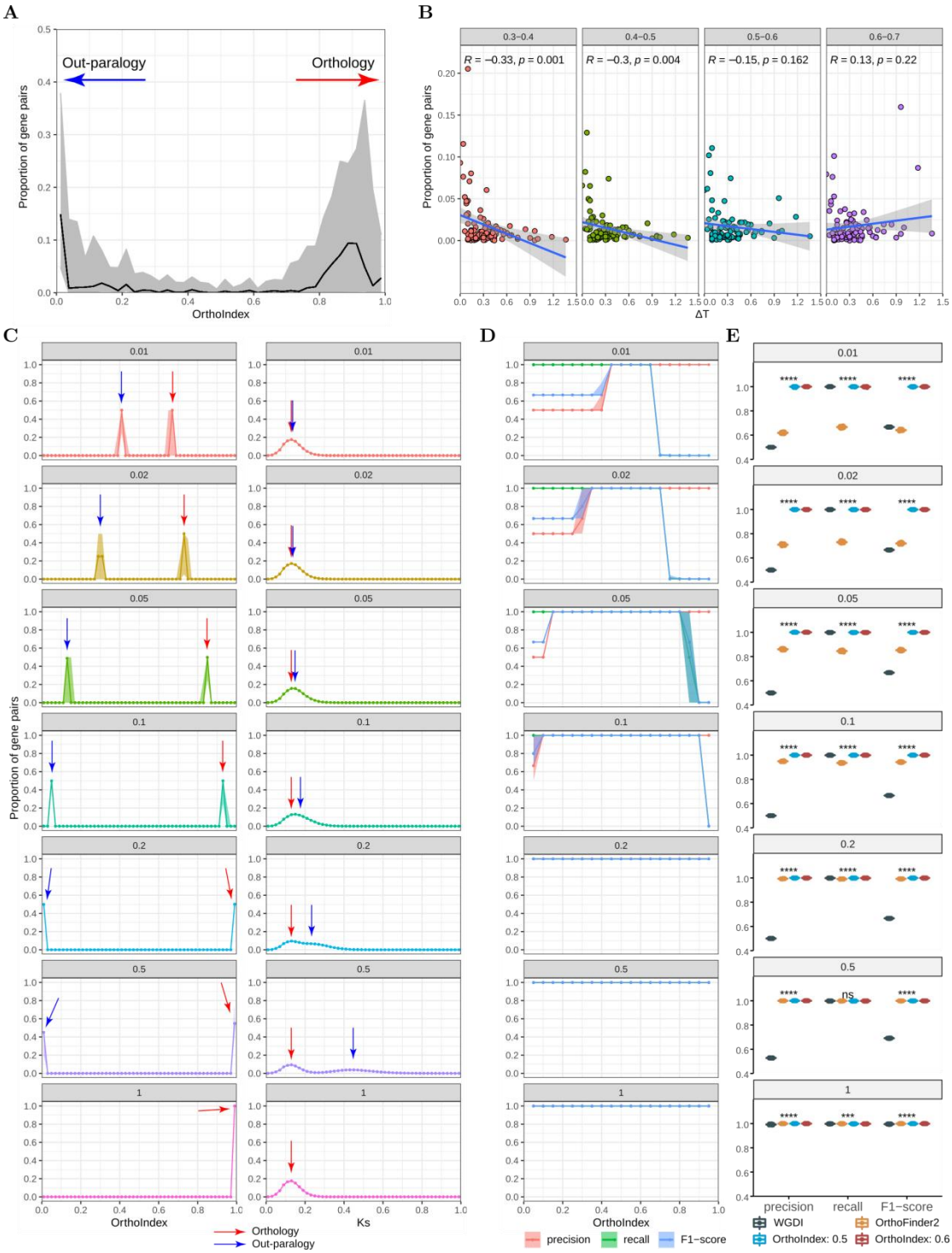


Fig. 2. The performance of the *Orthology Index* in identifying orthologous synteny in empirical and simulated datasets. A) Summary of *OI* distributions in the 91 empirical test cases. The black line and gray shadow represent the median and percentile-based 95 % confidence interval (CI) values, respectively. **B)** The correlation between ΔT and the noise around OI = 0.5 in the empirical datasets. The ΔT values were roughly obtained from the *Ks* peak values of shared WGD and speciation events. The noise was indicated by the accumulated proportion of syntenic gene pairs within 0.3–0.4, 0.4–0.5, 0.5–0.6 and 0.6–0.7 intervals of the *OI*. **C)** Comparisons of inter-genomic *Ks* and *OI* distributions in the simulated datasets at different values of ΔT ($\Delta T \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$, in time units of expected substitutions per site). The line/point and shadow represent the median and 95 % CI values

from 50 repeated simulations, respectively. **D)** Comparisons of the precision, recall and F1 scores of orthology identification using different *OI* cutoffs (0.05–0.95) from the simulated benchmarks. The line/point and shadow represent the median and percentile-based 95 % CI values from 50 repeated simulations, respectively. **E)** Comparisons of the precision, recall and F1 score of orthology identification using WGDI, OrthoFinder2 and *OI* with cutoffs 0.5 and 0.6, based on the simulated benchmarks. The boxplot represents the values from 50 repeated simulations. ns, $P > 0.05$; ***, $P <= 0.001$; ****, $P <= 0.0001$; Kruskal-Wallis test.

Applications of *OI* in polyploidy inference

Polyploidization (or WGD) has been identified as a critical mechanism in eukaryotic evolution and is pervasive in ancient and recent plant lineages (38). Both synteny and *Ks*-based methods have been widely used in the inference of polyploidy, including identification of occurrence and resulting ploidy and the phylogenetic placement of WGDs (12,39). However, inter-species orthologous synteny patterns have been previously overlooked in polyploidy inference, potentially leading to misunderstandings regarding the evolutionary history of WGD events (see discussions in (40)). We argue that distinguishing synteny resulting from orthology from that resulting from out-paralogy is not only straightforward but also vital for the inference of the evolutionary history of the polyploidy. We illustrate a straightforward model to explain the process of polyploidy inference using the patterns of orthologous synteny (**Fig. 3A**). When two genomes exhibit a 2:2 ratio of synteny depth, there are two main hypotheses to consider: either the genomes share a tetraploidization event (WGD) that occurred in a common ancestor, or they each underwent a lineage-specific tetraploidization event (WGD) independently, after they diverged (**Fig. 3A**). The *OI* offers a straightforward and visual method to test these two hypotheses with separated orthology and out-paralogy. If the genomes display a clear 1:1 orthology + 1:1 out-paralogy, similar to the Salicaceae case (**Fig. 1D**; ignoring the very ancient out-paralogy from the γ event), the first hypothesis is supported (**Fig. 3A, left panel**). In contrast, if they exhibit a 2:2 orthology, the second hypothesis is supported (**Fig. 3A, right panel**). Applying this to 90 known cases (**Figs. S1–90**) suggests that these inferences of polyploidy history from orthologous synteny patterns are reasonable and accurate.

Apiaceae and Araliaceae (both within Apiales) provide the perfect test case for this. Previous synteny and *Ks*-based analysis indicates either a WGD event in their common ancestor (41), or independent WGD events in the two families (42-44). Therefore, we reexamined the polyploidy of the available genomes from the two families using the newly developed *OI* method.

We first determined the occurrence of polyploidization and the resulting ploidy using the orthologous synteny patterns and a summarized term, relative ploidy (p, i.e, the overall orthologous synteny depth relative to a given reference genome), for the Apiales genomes, referring to the *Vitis vinifera* (Vitaceae) genome, which has not undergone lineage-specific polyploidy since the paleohexaploid γ event shared by the core eudicots (37). Specifically, $p > 1$ indicates one or more independent polyploidization events since the divergence with the reference genome, which resulted in a ploidy of $2p$ for a haploid genome assembly (e.g. $p = 2$ suggesting tetraploidy and $p = 3$ suggesting hexaploidy). Using *Centella asiatica* (Apiaceae) and *Aralia elata* (Araliaceae) as representatives, we observed a clear 2:1 orthologous synteny depth ($p = 2$) with the *Vitis vinifera* (Vitaceae) genome (**Fig. S91**), suggesting that both species have undergone a

tetraploidization event following their divergence from Vitaceae, consistent with previous research (43,44). However, upon comparison of the two genomes, a clear 1:1 orthology (p = 1) + 1:1 out-paralogy pattern was highlighted by the *OI* (**Fig. 3B**), aligning with the first model (**Fig. 3A, left panel**) and thus supporting the hypothesis of shared WGD (ω, **Fig. 3C**) in their common ancestor (see Song et al (41)). This inference was corroborated with macro-synteny phylogenies (**Fig. S92**), demonstrating the reliability of this technique. Similar patterns were also observed in other combinations of genomes from the two families (e.g. *C. asiatica* : *Eleutherococcus senticosus* = 1:2 orthology) (**Figs. S93–94**), further supporting the inference. Notably, we observed a considerable disparity in branch length (or substitution rate) between the two families from their most recent common ancestor (MRCA), with the mean branch length of Apiaceae (*L1* = 0.205) nearly three times than that of Araliaceae (*L2* = 0.074) (**Fig. 3C**). This discrepancy might explain the unreasonable phylogenetic placements drawn from the traditional *Ks*-based methods (42-44), which assumed equal substitution rates for different lineages when estimating the relative timing of polyploidization and speciation events. In fact, the assumption of equal substitution rates is often false or uncertain in practice.

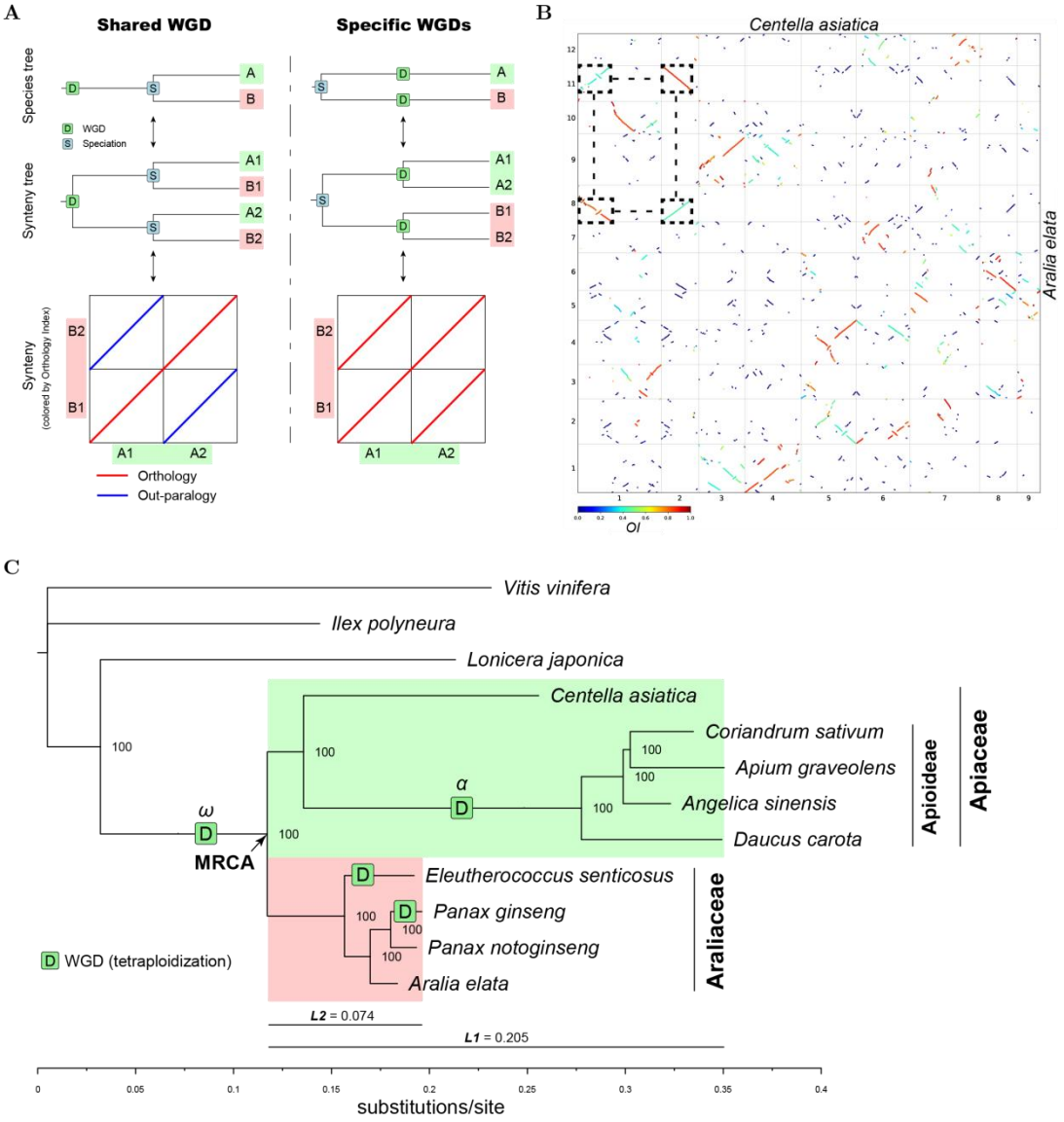


Fig. 3. Inference of polyploidy in Apiales (Apiaceae + Araliaceae) genomes using the *Orthology*

Index. **A)** Schematic illustration of determination of shared or lineage-specific polyploidy events using the orthologous synteny patterns identified with the *Orthology Index*. Despite a similar 2:2 ratio of synteny depth, the two scenarios have distinct patterns of orthologous synteny (1:1 orthology vs. 2:2 orthology). The labels A and B indicate two species, and A1, A2 and B1, B2 indicate duplicated chromosomes or blocks from WGD event(s). **B)** *Orthology Index*-colored dot plots indicating orthologous and out-paralogous synteny between the genomes of *Centella asiatica* (Apiaceae) and *Aralia elata* (Araliaceae). A typical 1:1 orthology + 1:1 out-paralogy synteny pattern is highlighted by the dashed squares. **C)** Phylogeny reconstructed from the genomes of certain species in the Apiaceae and the Araliaceae (Apiales) with labels indicating polyploidization events. *L1* and *L2* represent the average branch length (substitution rate) of the Apiaceae and the Araliaceae, respectively, from their most recent common ancestor (MRCA). Numbers at the nodes denote bootstrap values. The maximum-likelihood phylogenetic tree was reconstructed using IQ-TREE2, based on concatenated codon alignments of 2,363 single-copy genes (with at most 20 % taxa missing). Additional evidence supporting the inferred polyploidization events can be found in **Figs. S91–97**.

Furthermore, we deduced additional lineage-specific polyploidy events within the two families using the *OI*-based approach **(Fig 3C)**. We observed that all studied genomes from species in the Apioideae subfamily (including *Daucus carota*, *Angelica sinensis*, *Apium graveolens* and *Coriandrum sativum*) exhibit a 2:1 orthologous synteny depth (p = 2) when compared to the *C. asiatica* genome (**Fig. S95**), suggesting that they each experienced a tetraploidization event subsequent to their divergence from *C. asiatica*. Moreover, any two genomes within this subfamily display distinct 1:1 orthology (p = 1) + 1:1 out-paralogy patterns (**Figs. S56, 96**), suggesting a shared tetraploidization event (α), which is in line with previous research (41,44). Thus the α event was placed between the stem and crown of the Apioideae subfamily (**Fig. 3C**). In addition, employing the *OI*, we also inferred one species-specific tetraploidization event each in the *Panax ginseng* and *Eleutherococcus senticosus* genomes (**Fig. 3C, Figs. S57–59, 97**), consistent with some previous research (43,44) but in conflict with the inference of two independent WGD events in *P. ginseng* in Song et al (41), which may be misled by other types of gene duplications (non-WGD).

In summary, each previous inference for polyploidy in the Apiales genomes (41-44) was partly correct but did not represent the whole picture. *OI* can now unravel the history of polyploidy with strong evidence of orthologous synteny patterns. These findings suggest that misinterpretation may be difficult to avoid if inter-genomic orthologous synteny patterns are not considered, potentially leading to the confusion with out-paralogs and other types of gene duplications. Traditional *Ks*-based methods can be only used with great uncertainty in the identification of occurrence and phylogenetic placement of WGDs (45), while traditional synteny-based methods can not be used in phylogenetic placement of WGDs when showing potential out-paralogous synteny. In contrast, orthologous synteny inferred using the *OI,* which combines both synteny and orthology information, is straightforward to use in the identification of the occurrence of polyploidization and the resulting ploidy, as well as in the high-confidence phylogenetic placement of WGDs.

Applications in identification of reticulation

Reticulation, driven by allopolyploidization or hybridization, is a significant factor in eukaryotic evolution, yielding novel phenotypes that facilitate ecological diversification and the occupation of new niches (46). Numerous genomes originating from recent reticulation events have been documented (summarized in Jia et al (47)), encompassing critical cereals (48), fruits (49), vegetables (50), trees (9) and fish (51). We evaluated the application of the *Orthology Index* in several representative, well-documented cases, ranging from simple to complex evolutionary scenarios (**Fig. 4**).

The cases of *Arabidopsis* (52) and *Arachis* (53) represent simple hybridization or allopolyploidization scenarios. From the *OI*-colored dot plots, the two subgenomes of both the hybrid (*Arabidopsis thaliana* × *Arabidopsis renosa*) and tetraploid (*Arachis hypogaea*) show clear and separate orthologous relationships with their diploid progenitors (**Fig. 4A and 4C, Fig. S98-99**). Based on the orthologous relationships revealed by the *OI*, the hybridization events can be easily inferred with straightforward visualization (**Fig. 4B and 4D**).

The orthologous relationships between the genomes in the complicated polyploid species complexes *Triticum* and *Papaver* are also clear from the *OI*-colored dot plots (**Fig. 4E and 4G, Fig. S100–101**). The two subgenomes of the tetraploid *T. turgidum* are orthologous to the two subgenomes of the hexaploid *T. aestivum* (**Fig. 4E, Fig. S100**), leading the inference that *T. turgidum* is the intermediate tetraploid progenitor of the allohexaploid *T. aestivum* (**Fig. 4F**), in line with previous results (48). Similarly, a reticulate allopolyploidization origin in two *Papaver* genomes (**Fig. 4H**) can be inferred from the *OI*-colored dot plots (**Fig. 4G, Fig. S101**). This inference agrees with our previous work (7); however, the phylogenetic relationships between homoeologous subgenomes cannot be resolved directly from the *OI*-colored dot plots and require further evidence, such as chromosome or subgenome-scale phylogenies (7).

Therefore, we conclude that reticulate speciation with intact subgenomic structure can be simply and directly inferred from the *OI*-colored dot plots. Although phylogeny-based methods could provide a deeper insight into the reticulation (2), the *OI* method is more straightforward and allows visualization that at least suggests a hypothesis to test further. Indeed, the straightforward visualization of the patterns of orthologous synteny could make it easy to recognize the signals of reticulation and thus to reasonablely elucidate the evolutionary history (e.g. (6-8)).

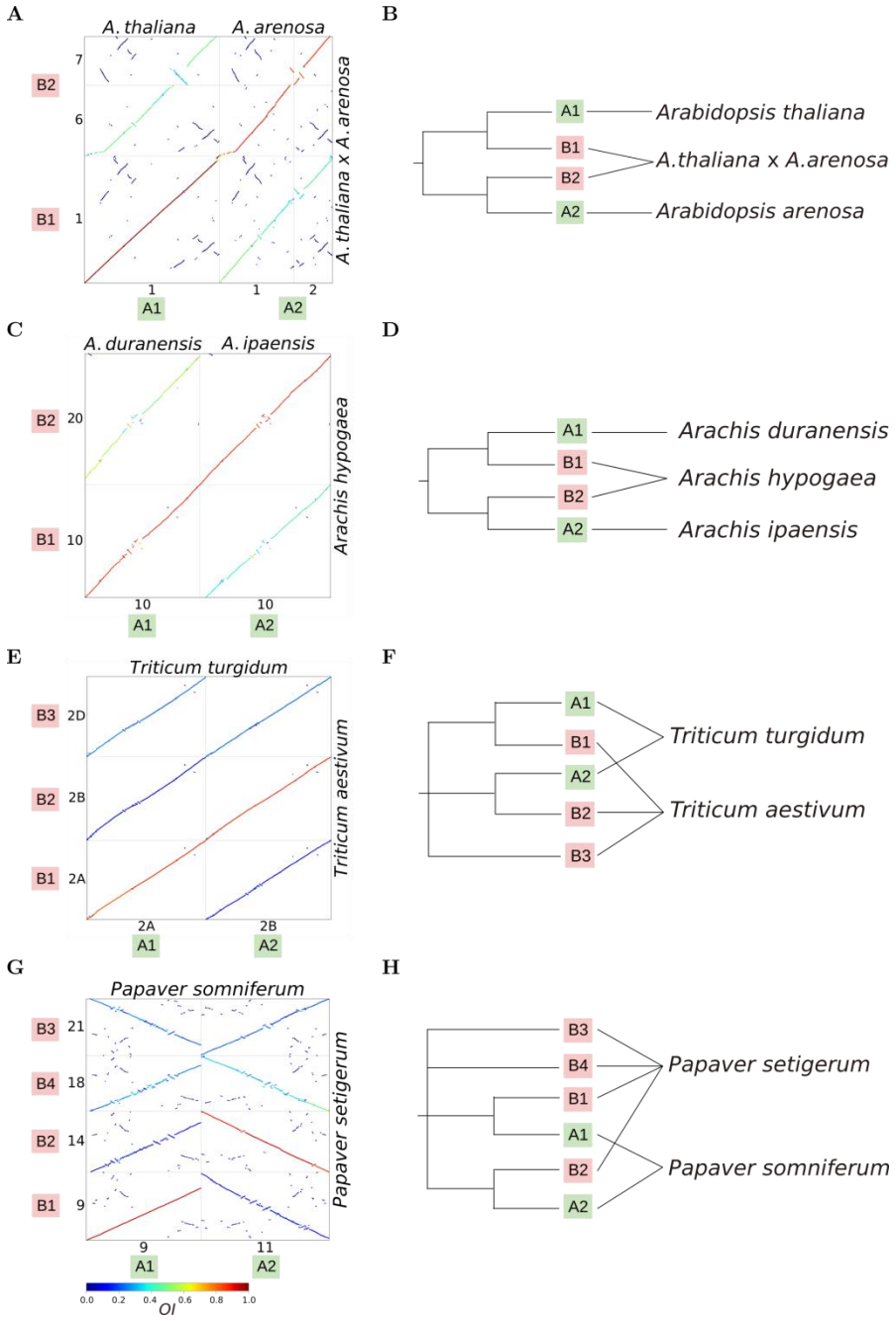


Fig. 4. **Examples of reticulation inferences based on the *Orthology Index*. A–B**) *Orthology Index*-colored dot plots (**A**) of the genomes of *Arabidopsis thaliana* + *Arabidopsis arenosa* and their hybrid *Arabidopsis thaliana* × *Arabidopsis arenosa*, and the inference (**B**; in dendrogram form) from the orthologous relationships. **C–D**) *Orthology Index*-colored dot plots (**C**) of the genomes of the tretaploid *Arachis hypogaea* and its diploid progenitors, and the inference (**D**; in dendrogram form) from the orthologous relationships. **E–F**) *Orthology Index*-colored dot plots (**E**) of the genomes of the hexaploid *T. aestivum* and its intermediate tretaploid *T. turgidum*, and the inference (**F**; in dendrogram form) from the orthologous relationships. **G–H**) *Orthology Index*-colored dot plots (**G**) of the genomes of the neo-octoploid *P. setigerum* and its intermediate tretaploid *P. somniferum*, and the inference (**H**; in dendrogram form) from the orthologous relationships. Only one set of representative homoeologous chromosomes is shown in the dot plots; dot plots with the full set of chromosomes can be found in **Figs. S98-101**.

464
465
466 **Applications in phylogenomics**
467 Accurate inference of orthology plays a crucial role in the estimation of species trees, and the
468 pseudo-orthologs (i.e. hidden paralogs) derived from WGD and gene loss can mislead species tree
469 inference greatly under some circumstances (54). As demonstrated above (**Fig. 1–2, Figs. S1–90**),
470 the *OI* exhibited a high level of accuracy in identifying syntenic orthologs (or syntelogs) and can
471 therefore minimize the detrimental influence of pseudo-orthologs. Consequently, the resulting
472 knowledge of syntenic orthologs can be directly applied to species tree reconstruction. We used
473 the example of the core eudicots to showcase this application. It is accepted that all core eudicots
474 share a paleohexaploidization event (γ event, WGT) around 120 Mya, whereas no two orders
475 within the core eudicots share an additional polyploidy event (38). Therefore, it is appropriate to
476 use the *OI* to remove the out-paralogy produced from the γ event, and to better resolve the
477 phylogenetic relationships among the orders of the core eudicots, many of which remain poorly
478 resolved, such as the Celastrales–Oxalidales–Malpighiales (COM) clade (55). Here we utilized the
479 genome-scale syntenic orthologs inferred by the *OI* to reconstruct a backbone phylogeny of the
480 core eudicots, aiming to minimize the detrimental influence of the γ event.
481 We compiled a high-quality genomic dataset that covers 28 (70 %) of the 40 orders and 98
482 (33 %) of the 298 families of core eudicots treated in APG IV (55). We then applied the *OI* to this
483 dataset to identify syntenic orthologs. This resulted in the identification of 54,322 syntenic
484 orthogroups (SOGs). After filtering, 12,277 multi-copy and 5,154 single-copy SOGs were
485 retrieved, allowing for up to 40 % missing taxa (**Fig. 5A**). This imbalance between the numbers of
486 multi-copy and single-copy SOGs is attributable to lineage-specific polyploidy events within the
487 core eudicots. As a result of these polyploidy events, the occupancy of single-copy SOGs showed
488 significant decreases in species with high relative ploidy (i.e. orthologous syntenic depth relative
489 to the *Vitis vinifera* genome) (**Fig 5B**). Nevertheless, the order-level species tree topologies based
490 on the two gene sets were identical and both were strongly supported with high posterior
491 probabilities (**Fig. 5C)**, although the tree based on multi-copy SOGs was more robust with equal
492 or higher posterior probabilities at nearly all nodes (**Fig. 5C)** and there were slight differences in
493 the positions of a few of the species (**Figs. S102–103**).
494 We observed large incongruences between our phylogeny and the APG IV (55), as nearly half
495 of the orders covered in this study have inconsistent phylogenetic positions (**Fig. 5C)**. For
496 example, in our analysis, the COM clade was not monophyletic and was placed with the malvids,
497 in contrast to that from APG IV (**Fig. 5C)**. The incongruences are likely to be due to the fact that
498 the APG IV analysis was conducted using a very limited number of phylogenetic markers (55). In
499 contrast, recent phylogenomics/phylotranscriptomic studies are consistent with most of our
500 findings (56-63). For example, the phylogenetic positions of the Fagales, Rosales, Celastrales–
501 Oxalidales–Malpighiales, Myrtales, Cornales and Ericales in our study (**Fig. 5C)** are consistent
502 with the genome-scale phylogenomics based on coalescent-based analysis of 482 single-copy
503 nuclear orthologous sequences (60). In addition, the phylogenetic positions of the Santalales
504 (sister to the superrosids), Aquifoliales (sister to the Garryales), and Crossosomatales (sister to the
505 fabids + remaining malvids) in our study (**Fig. 5C)** are consistent with the phylogenetic inference
506 based on coalescent tree analysis of 410 single-copy gene families extracted from transcriptome
507 and genome data (57).

508 Our results further provide interesting phylogenomic insights into the core eudicots. In our
509 study, the Vitales was found to be sister to the Saxifragales with high support (**Fig. 5C**), and this
510 clade was sister to the remaining rosids (**Fig. 5C**). These data are inconsistent with the hypothesis
511 that the Vitales are sister to the Saxifragales and all other taxa in the superrosids clade (56-
512 58,60,61,63). We also found that the Zygophyllales were sister to the malvids clade with high
513 support (**Fig. 5C**), inconsistent with its placement as sister of the Myrtales (57,61,63). These
514 discrepancies can be mainly attributed to incomplete lineage sorting and ancient hybridization (64).
515 Considering that our analyses involved ten thousand syntenic orthologous gene families, many
516 more than than the hundreds or thousands of loci used in previous studies (56-63), and also that
517 we have minimized the detrimental effects of the shared γ event, our results could represent a
518 more accurate depiction of the real tree of life of the core eudicots than the other mentioned
519 studies. Previous studies have used ortholog-based methods, which are prone to misidentify out-
520 paralogs as orthologs, and we argue that our orthologous synteny-based method should be
521 emphasized in plant phylogenomic studies. Nevertheless, our method is limited by taxon sampling
522 (lack of high-quality genomes for some orders and families). However, this is expected to be
523 resolved in the near future with the continuing development and effort in the field of genome
524 sequencing (65).

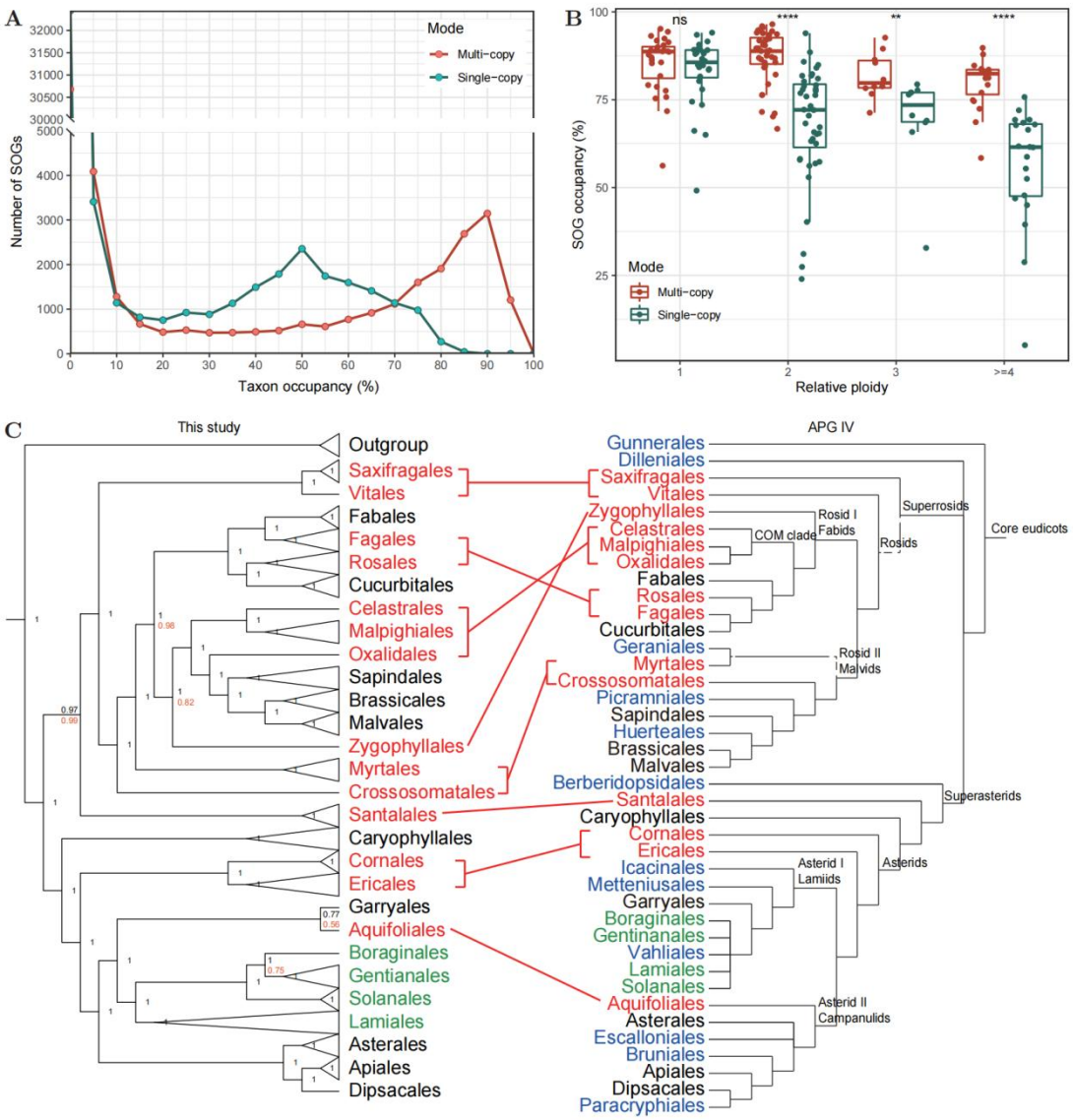


Fig. 5. An example (core eudicots) of phylogenomics based on the *Orthology Index*. **A**) The number of multi-copy and single-copy syntenic orthogroups (SOGs) with different taxon occupancy. **B**) The SOG occupancy in species with different relative ploidy (i.e. orthologous syntenic depth to the grape genome) with at most 40 % taxa missing. Each point represents one species. ns, $P > 0.05$; **, $P <= 0.01$; ****, $P <= 0.0001$; Wilcoxon test. **C**) Phylogenetic relationships within the core eudicots reconstructed in this study versus in APG IV. Conflicting positions are marked in red; unresolved relationships in APG IV are marked in green, and orders not covered in this study are marked in blue. The numbers at the nodes are posterior probabilities from ASTRAL, with the black representing those from the multi-copy SOGs and orange representing those from the single-copy SOGs (omitted for equal values). Further details of the two trees reconstructed in this study can be found in **Figs. S102–103**.

Limitations

The *Orthology Index* may not perform well in extremely complex cases. For instance, when ΔT is notably small (e.g. radiation following polyploidization in a few generations), similar to other

methods (e.g. *Ks*-based), *OI* may also find it difficult to distinguish between out-paralogy and orthology (for a difficult case also see Fig. S68). Therefore, although we set a unified *OI* cutoff (0.6) in our pipeline, users should manually inspect the results (mainly dot plots) for confirmation, and the extremely complex cases showing unexpected patterns should be investigated on a case-by-case basis. This method is also limited in some scenarios where orthology inference and/or synteny detection is limited. For example, synteny is known not to be conserved in distantly-related or fast-evolving lineages (66), suggesting that this method should not be applied in cases without conserved synteny (e.g. angiosperms–ferns). Additionally, fragmented assemblies, as well as mis-assemblies, can disrupt synteny and subsequently reduce the efficiency of the method. However, this is likely to cease to be a concern in the near future with the fast development of sequencing and assembly techniques.

Conclusions

In summary, we present a human-interpretable and machine-actionable approach to distinguish orthology from out-paralogy for syntenic blocks. The approach can identify orthologous synteny robustly, as validated with nearly 100 representative cases. We have demonstrated the broad and valuable applications of this approach to the reconstruction of evolutionary history in plant genomes, including reconstruction of the tree/network of life, and identification of and placing of polyploidy events on the tree/network. This approach will extend our analytical capacity in evolutionary genomics and might reduce the misleading data generated using some traditional methods.

Code availability

The codes of the SOI tool and pipeline, and typical examples can be found on GitHub (https://github.com/zhangrengang/SOI and https://github.com/zhangrengang/evolution_example).

Data availability

The codon alignments and gene trees of the core eudicots are available from Figshare (https://doi.org/10.6084/m9.figshare.24174930).

Acknowledgments

We thank Professors Ying-Xiong Qiu, Xiao-Ming Song, Quan-Jun Hu, Tao Zhou, and Zhao-Ying Liu for generously sharing their genomic data, and the other researchers who have already released their genomic data publicly.

Funding

National Key Research and Development Program (2022YFF1301702); Key Basic Research Programs of Yunnan Province (202101BC070003 and 202302AE090018); Conservation grant for PSESP inYunnan Province (2022SJ07X-03).

Author contributions

RGZ and YPM conceived and designed the study; RGZ, HYS, MJZ and HS programmed and/or tested the code; RGZ collected and analyzed the data; RGZ and HYS prepared figures; RGZ, KHJ

and HYS drafted the manuscript; YPM, RIM, FA, HC, KHJ and YVdP revised the manuscript; all authors approved the final manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Figures and Tables

Figs. S1–90. *Orthology Index* in the identification of orthologous synteny in species X and species Y. Refer to Fig. 1 for detailed descriptions.

Fig. S91. *Orthology Index*-colored dot plots showing orthologous syntenic relationships between *Centella asiatica* : *Vitis vinifera* (2:1 orthologous synteny depth ratio), *Aralia elata* : *Vitis vinifera* (2:1), *Angelica sinensis* : *Vitis vinifera* (4:1), and *Panax ginseng* : *Vitis vinifera* (4:1).

Fig. S92. Macro-synteny phylogenies inferred based on concatenated 1:2:2 (*Vitis vinifera* : *Aralia* elata : *Centella* asiatica) orthologous syntenic genes. The leaf labels on the trees are chromosome numbers. *n*, number of orthologous syntenic genes involved. Numbers at nodes represent the bootstrap values (percentage).

Fig. S93. *Orthology Index*-colored dot plots showing orthologous syntenic relationships between *Centella* asiatica and the Araliaceae (*Centella asiatica* : *Aralia* elata = 1:1, *Centella asiatica* : *Eleutherococcus senticosus* = 1:2, *Centella asiatica* : *Panax ginseng* = 1:2, *Centella asiatica* : *Panax notoginseng* = 1:1).

Fig. S94. *Orthology Index*-colored dot plots showing orthologous syntenic relationships between *Aralia* elata and the Apioideae (all 1:2).

Fig. S95. *Orthology Index*-colored dot plots showing orthologous syntenic relationships between *Centella asiatica* and the Apioideae (all 1:2).

Fig. S96. *Orthology Index*-colored dot plots showing orthologous syntenic relationships within the Apioideae (all 1:1).

Fig. S97. *Orthology Index*-colored dot plots showing orthologous syntenic relationships within the Araliaceae (*Panax ginseng* : *Panax notoginseng* = 2:1, *Aralia elata* : *Panax notoginseng* = 1:1, *Eleutherococcus senticosus* : *Panax notoginseng* = 2:1). The other orthologous syntenic depth ratios (*Aralia elata* : *Eleutherococcus senticosus* = 1:2, *Aralia elata* : *Panax ginseng* = 1:2, *Eleutherococcus senticosus* : *Panax ginseng* = 2:2) can be found in Fig. S57-59.

Fig. S98. *Orthology Index*-colored dot plots showing orthologous syntenic relationships between *Arabidopsis thaliana* × *A. arenosa* and *A. thaliana* + *A. arenosa*.

Fig. S99. *Orthology Index*-colored dot plots showing orthologous syntenic relationships between *Arachis hypogaea* and *A. duranensis* + *A. ipaensis*.

Fig. S100. *Orthology Index*-colored dot plots showing orthologous syntenic relationships between *Triticum turgidum* and *T. aestivum*.

Fig. S101. *Orthology Index*-colored dot plots showing orthologous syntenic relationships between *Papaver somniferum* and *P. setigerum*.

Fig. S102. Phylogenetic relationships within the core eudicots based on 12,277 multi-copy SOGs. The numbers at the nodes are posterior probabilities from ASTRAL. Bar, 3.0 coalescent units.

Fig. S103. Phylogenetic relationships within the core eudicots based on 5,154 single-copy SOGs. The numbers at the nodes are posterior probabilities from ASTRAL. Bar, 3.0 coalescent

units.

Table S1. Cases with shared polyploidization event(s) used in this study.

Table S2. Genomic data used in this study (accessed before 2023-06-29).

References

1. Steenwyk, J.L., Li, Y., Zhou, X., Shen, X. and Rokas, A. (2023) Incongruence in the phylogenomics era. *Nat. Rev. Genet.*, **24**, 834–850.
2. Zhang, R.G., Shang, H., Jia, K. and Ma, Y. (2024) Subgenome phasing for complex allopolyploidy: case-based benchmarking and recommendations. *Brief. Bioinform.*, **25**, bbad513.
3. Ma, J., Sun, P., Wang, D., Wang, Z., Yang, J., Li, Y., Mu, W., Xu, R., Wu, Y. and Dong, C. *et al.*. (2021) The *Chloranthus sessilifolius* genome provides insight into early diversification of angiosperms. *Nat. Commun.*, **12**, 6929.
4. Sun, P., Lu, Z., Wang, Z., Wang, S., Zhao, K., Mei, D., Yang, J., Yang, Y., Renner, S.S. and Liu, J. (2024) Subgenome-aware analyses reveal the genomic consequences of ancient allopolyploid hybridizations throughout the cotton family. *Proc. Natl. Acad. Sci. U. S. A.*, **121**, e1981046175.
5. Guo, C., Luo, Y., Gao, L.M., Yi, T.S., Li, H.T., Yang, J.B. and Li, D.Z. (2023) Phylogenomics and the flowering plant tree of life. *J. Integr. Plant Biol.*, **65**, 299–323.
6. Yang, X., Gao, S., Guo, L., Wang, B., Jia, Y., Zhou, J., Che, Y., Jia, P., Lin, J. and Xu, T. *et al.*. (2021) Three chromosome-scale *Papaver* genomes reveal punctuated patchwork evolution of the morphinan and noscapine biosynthesis pathway. *Nat. Commun.*, **12**, 6030.
7. Zhang, R.G., Lu, C., Li, G., Lv, J., Wang, L., Wang, Z., Chen, Z., Liu, D., Zhao, Y. and Shi, T. *et al.*. (2023) Subgenome-aware analyses suggest a reticulate allopolyploidization origin in three *Papaver* genomes. *Nat. Commun.*, **14**, 2204.
8. Yang, X., Gao, S., Xu, T., Wang, B., Jia, Y. and Ye, K. (2023) Reply to "Subgenome-aware analyses suggest a reticulate allopolyploidization origin in three *Papaver* genomes". *Nat. Commun.*, **14**, 2203.
9. An, X., Gao, K., Chen, Z., Li, J., Yang, X., Yang, X., Zhou, J., Guo, T., Zhao, T. and Huang, S. *et al.*. (2021) High quality haplotype-resolved genome assemblies of *Populus tomentosa* Carr., a stabilized interspecific hybrid species that is widespread in Asia. *Mol. Ecol. Resour.*, **22**, 786–802.
10. Zhang, T., Qiao, Q., Du, X., Zhang, X., Hou, Y., Wei, X., Sun, C., Zhang, R., Yun, Q. and Crabbe, M.J.C. *et al.*. (2022) Cultivated hawthorn (*Crataegus pinnatifida* var. *major*) genome sheds light on the evolution of Maleae (apple tribe). *J. Integr. Plant Biol.*, **64**, 1487–1501.
11. Tang, H., Lyons, E., Pedersen, B., Schnable, J.C., Paterson, A.H. and Freeling, M. (2011) Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics*, **12**, 102.
12. Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., Xi, Z., Wang, X. and Liu, J. (2022) WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant.*, **15**, 1841–1851.
13. Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B. and Guo, H. *et al.*. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic. Acids. Res.*, **40**, e49.

14. Chen, J., Huang, Y., Brachi, B., Yun, Q., Zhang, W., Lu, W., Li, H., Li, W., Sun, X. and Wang, G. *et al.*. (2019) Genome-wide analysis of Cushion willow provides insights into alpine plant divergence in a biodiversity hotspot. *Nat. Commun.*, **10**, 5230.

15. Buchfink, B., Xie, C. and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.

16. van Dongen, S. (2008) Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.*, **30**, 121–141.

17. Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.*, **20**, 238.

18. Cosentino, S., Sriswasdi, S. and Iwasaki, W. (2024) SonicParanoid2: fast, accurate, and comprehensive orthology inference with machine learning and language models. *Genome Biol.*, **25**, 195.

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

20. Standley, D.M. and Katoh, K. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

21. Suyama, M., Torrents, D. and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic. Acids. Res.*, **34**, W609–W612.

22. Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

23. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A. and Lanfear, R. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, **37**, 1530–1534.

24. Zhang, C. and Mirarab, S. (2022) ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics*, **38**, 4949–4950.

25. Morel, B., Williams, T.A., Stamatakis, A. and Schwartz, R. (2023) Asteroid: a new algorithm to infer species trees from gene trees under high proportions of missing data. *Bioinformatics*, **39**, btac832.

26. Tang, H., Krishnakumar, V., Zeng, X., Xu, Z., Taranto, A., Lomas, J.S., Zhang, Y., Huang, Y., Wang, Y. and Yim, W.C. *et al.*. (2024) JCVI: A versatile toolkit for comparative genomics analysis. *iMeta*, **3**, e211.

27. Derelle, R., Philippe, H. and Colbourne, J.K. (2020) Broccoli: Combining phylogenetic and network analyses for orthology assignment. *Mol. Biol. Evol.*, **37**, 3389–3396.

28. Klemm, P., Stadler, P.F. and Lechner, M. (2023) Proteinortho6: pseudo-reciprocal best alignment heuristic for graph-based detection of (co-)orthologs. *Frontiers in Bioinformatics*, **3**.

29. Persson, E. and Sonnhammer, E.L.L. (2022) InParanoid-DIAMOND: faster orthology analysis with the InParanoid algorithm. *Bioinformatics*, **38**, 2918–2919.

30. Deutekom, E.S., Snel, B. and van Dam, T.J.P. (2021) Benchmarking orthology methods using phylogenetic patterns defined at the base of Eukaryotes. *Brief. Bioinform.*, **22**, bbaa206.

31. Kurtzer, G.M., Sochat, V. and Bauer, M.W. (2017) Singularity: Scientific containers for mobility of compute. *PLoS One*, **12**, e177459.

32. Davín, A.A., Tricou, T., Tannier, E., de Vienne, D.M. and Szöllősi, G.J. (2020) Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages. *Bioinformatics*, **36**, 1286–1288.

33. Dai, X., Hu, Q., Cai, Q., Feng, K., Ye, N., Tuskan, G.A., Milne, R., Chen, Y., Wan, Z. and Wang,

Z. (2014) The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Res.*, **24**, 1274.

34. He, L., Jia, K.H., Zhang, R.G., Wang, Y., Shi, T.L., Li, Z.C., Zeng, S.W., Cai, X.J., Wagner, N.D. and Hörandl, E. *et al.*. (2021) Chromosome-scale assembly of the genome of *Salix dunnii* reveals a male-heterogametic sex determination system on chromosome 7. *Mol. Ecol. Resour.*, **21**, 1966–1982.

35. Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S. and Salamov, A. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

36. Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H. (2008) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.

37. Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N. and Jubin, C. *et al.*. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

38. Van de Peer, Y., Mizrachi, E. and Marchal, K. (2017) The evolutionary significance of polyploidy. *Nat. Rev. Genet.*, **18**, 411–424.

39. Chen, H., Zwaenepoel, A., Van de Peer, Y. and Robinson, P. (2024) wgd v2: a suite of tools to uncover and date ancient polyploidy and whole-genome duplication. *Bioinformatics*, **40**, btae272.

40. Chen, B., Li, D., Luo, K., Jiu, S., Dong, X., Wang, W., Li, X., Hao, T., Lei, Y. and Guo, D. *et al.*. (2024) Chromosome-level assembly of *Lindenbergia philippensis* and comparative genomic analyses shed light on genome evolution in Lamiales. *Front. Plant Sci.*, **15**, 1444234.

41. Song, X., Sun, P., Yuan, J., Gong, K., Li, N., Meng, F., Zhang, Z., Li, X., Hu, J. and Wang, J. *et al.*. (2021) The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in apiales. *Plant Biotechnol. J.*, **19**, 731–744.

42. Jiang, Z., Tu, L., Yang, W., Zhang, Y., Hu, T., Ma, B., Lu, Y., Cui, X., Gao, J. and Wu, X. *et al.*. (2021) The chromosome-level reference genome assembly for *Panax notoginseng* and insights into ginsenoside biosynthesis. *Plant Commun.*, **2**, 100113.

43. Wang, Y., Zhang, H., Ri, H.C., An, Z., Wang, X., Zhou, J., Zheng, D., Wu, H., Wang, P. and Yang, J. *et al.*. (2022) Deletion and tandem duplications of biosynthetic genes drive the diversity of triterpenoids in *Aralia elata*. *Nat. Commun.*, **13**, 2224.

44. Yang, Z., Chen, S., Wang, S., Hu, Y., Zhang, G., Dong, Y., Yang, S., Miao, J., Chen, W. and Sheng, J. (2021) Chromosomal-scale genome assembly of *Eleutherococcus senticosus* provides insights into chromosome evolution in Araliaceae. *Mol. Ecol. Resour.*, **21**, 2204–2220.

45. McKibben, M.T.W., Finch, G. and Barker, M.S. (2024) Species-tree topology impacts the inference of ancient whole-genome duplications across the angiosperm phylogeny. *Am. J. Bot.*, **111**, e16378.

46. Otto, S.P. and Whitton, J. (2000) Polyploid incidence and evolution. *Annu. Rev. Genet.*, **34**, 401–437.

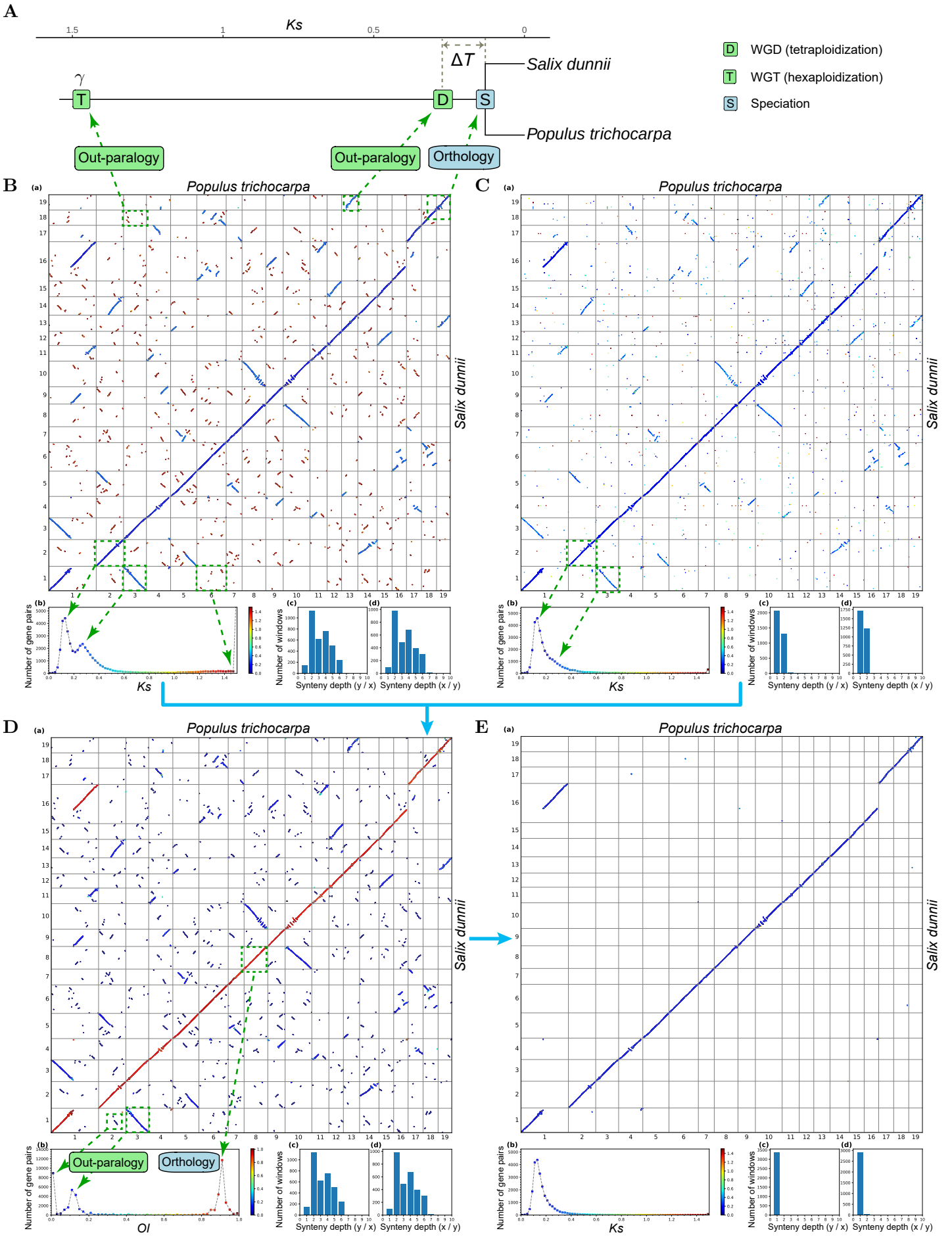
47. Jia, K., Wang, Z., Wang, L., Li, G., Zhang, W., Wang, X., Xu, F., Jiao, S., Zhou, S. and Liu, H. *et al.*. (2022) SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific *k*-mers. *New Phytol.*, **235**, 801–809.

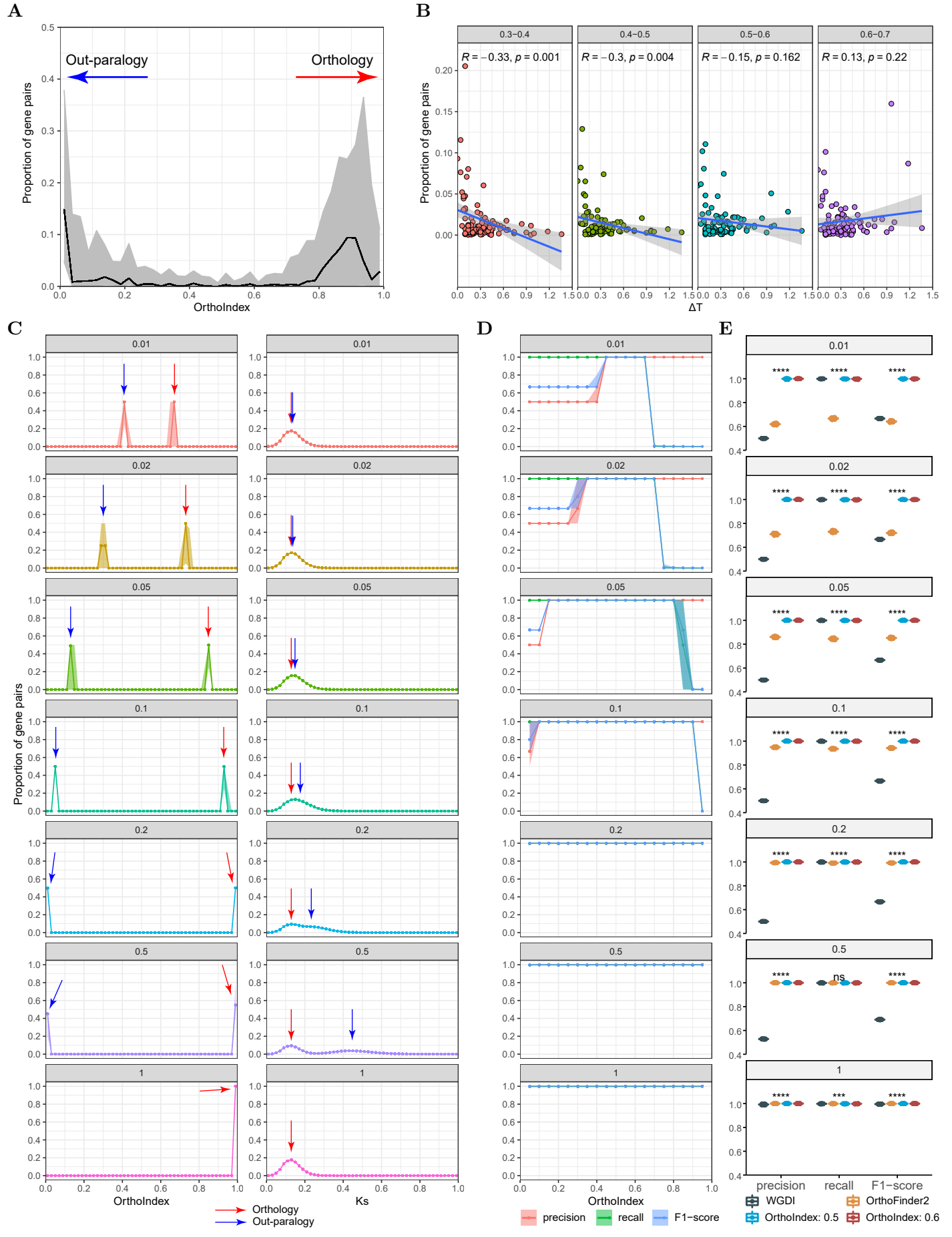
48. Marcussen, T., Sandve, S.R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K.S., Wulff, B.B.H., Steuernagel, B., Mayer, K.F.X. and Olsen, O.A. *et al.*. (2014) Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, **345**, 1250092.

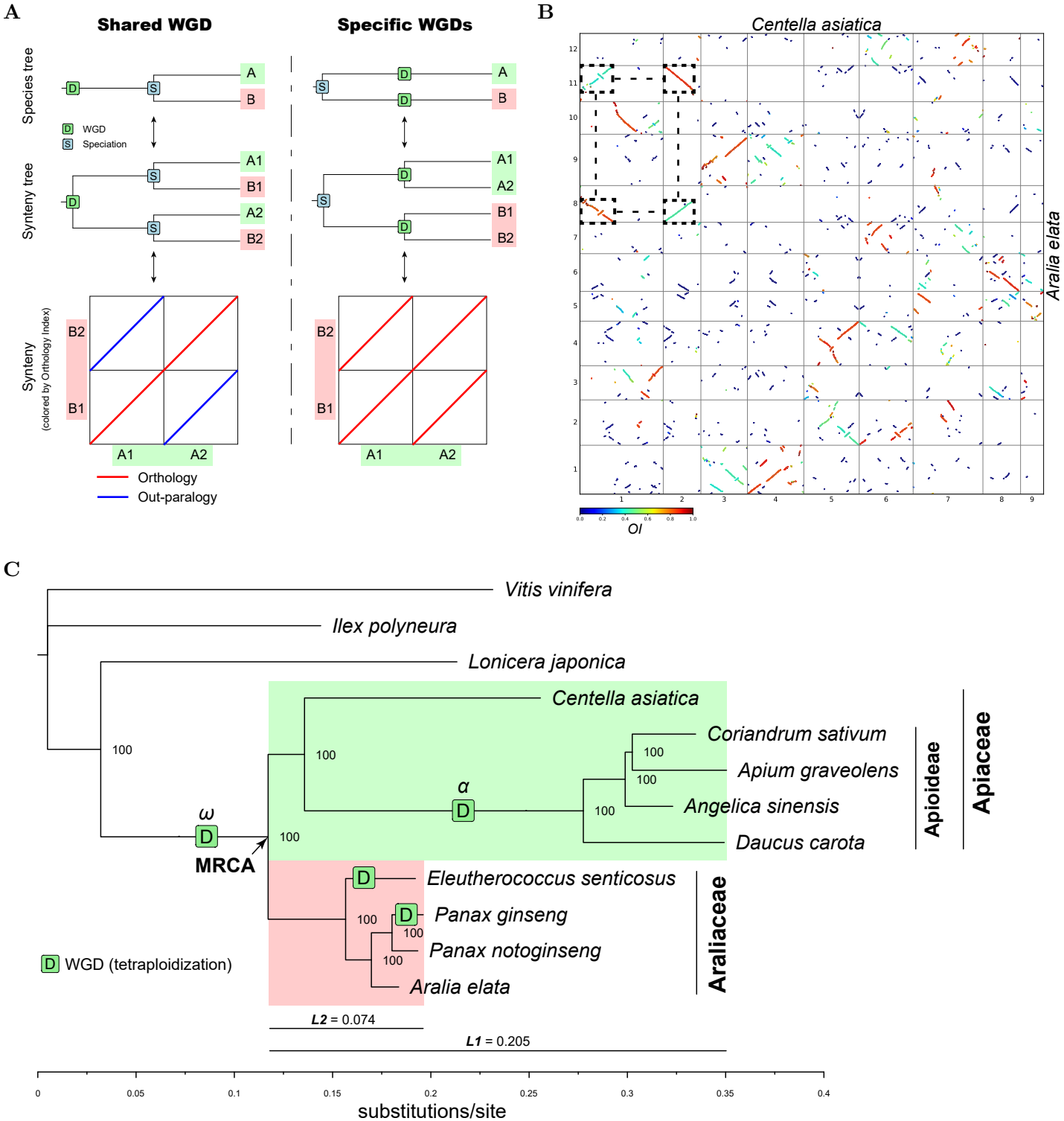
49. Edger, P.P., Poorten, T.J., VanBuren, R., Hardigan, M.A., Colle, M., McKain, M.R., Smith, R.D.,

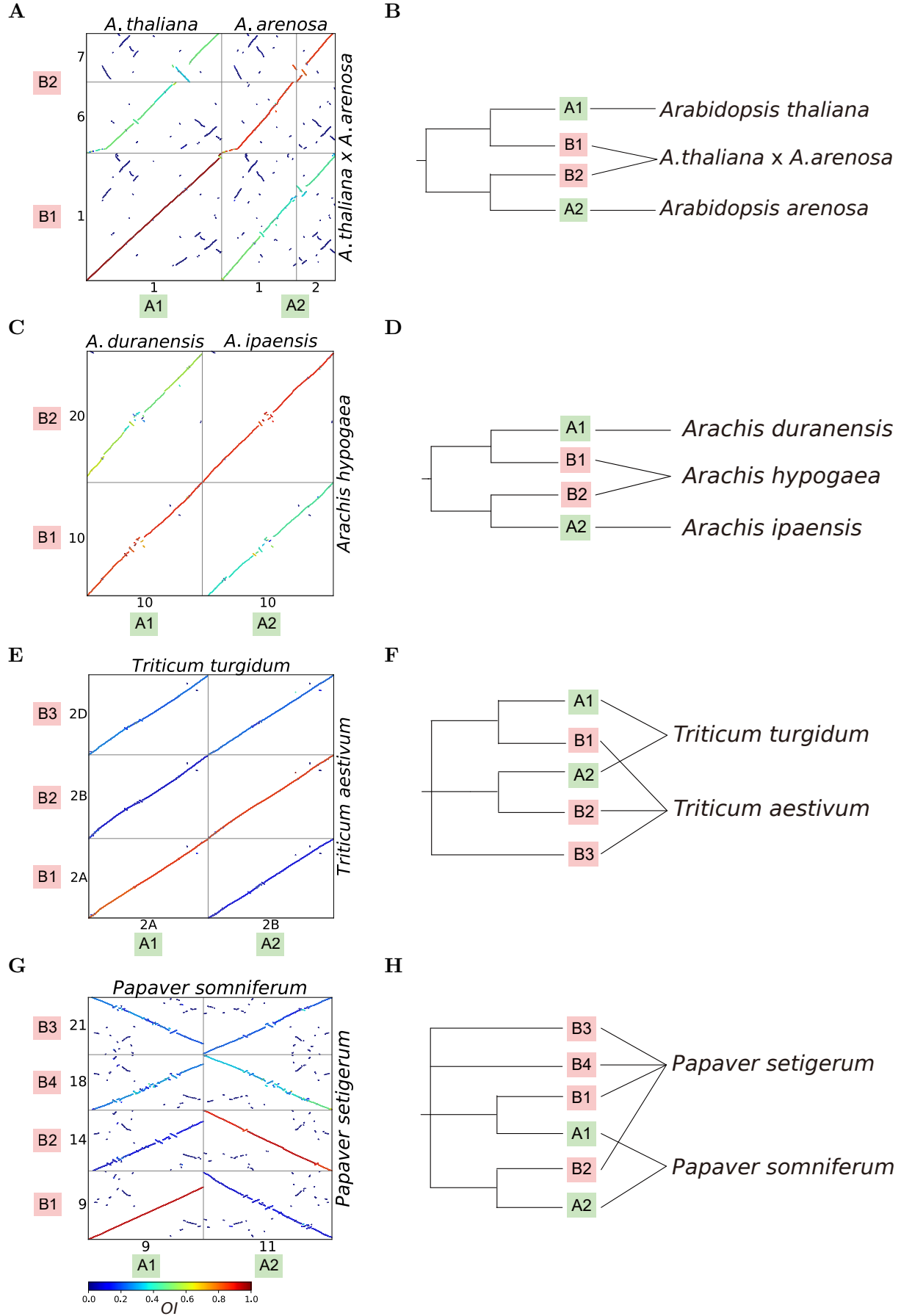
762 Teresi, S.J., Nelson, A.D.L. and Wai, C.M. *et al.*. (2019) Origin and evolution of the octoploid
763 strawberry genome. *Nat. Genet.*, **51**, 541–547.
764 50. Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A., Tang, H., Wang, X., Chiquet, J., Belcram, H.,
765 Tong, C. and Samans, B. *et al.*. (2014) Early allopolyploid evolution in the post-Neolithic *Brassica*
766 *napus* oilseed genome. *Science*, **345**, 950–953.
767 51. Li, J.T., Wang, Q., Huang, Y.M., Li, Q.S., Cui, M.S., Dong, Z.J., Wang, H.W., Yu, J.H., Zhao,
768 Y.J. and Yang, C.R. *et al.*. (2021) Parallel subgenome structure and divergent expression evolution of
769 allo-tetraploid common carp and goldfish. *Nat. Genet.*, **53**, 1493–1503.
770 52. Jiang, X., Song, Q., Ye, W. and Chen, Z.J. (2021) Concerted genomic and epigenomic changes
771 accompany stabilization of *Arabidopsis* allopolyploids. *Nat. Ecol. Evol.*, **5**, 1382–1393.
772 53. Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M.K., Zhang, C., Chang, W., Zhang, L.,
773 Zhang, X. and Tang, R. *et al.*. (2019) The genome of cultivated peanut provides insight into legume
774 karyotypes, polyploid evolution and crop domestication. *Nat. Genet.*, **51**, 865–876.
775 54. Xiong, H., Wang, D., Shao, C., Yang, X., Yang, J., Ma, T., Davis, C.C., Liu, L. and Xi, Z. (2022)
776 Species tree estimation and the impact of gene loss following whole-genome duplication. *Syst. Biol.*, **71**,
777 1348–1361.
778 55. The Angiosperm Phylogeny Group. (2016) An update of the Angiosperm Phylogeny Group
779 classification for the orders and families of flowering plants: APG IV. *Bot. J. Linnean Soc.*, **181**, 1–20.
780 56. Zeng, L., Zhang, N., Zhang, Q., Endress, P.K., Huang, J. and Ma, H. (2017) Resolution of deep
781 eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and
782 genomic datasets. *New Phytol.*, **214**, 1338.
783 57. Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A.,
784 Graham, S.W., Grosse, I., Li, Z., Melkonian, M. and Mirarab, S. *et al.*. (2019) One thousand plant
785 transcriptomes and the phylogenomics of green plants. *Nature*, **574**, 679–685.
786 58. Yang, L., Su, D., Chang, X., Foster, C.S.P., Sun, L., Huang, C., Zhou, X., Zeng, L., Ma, H. and
787 Zhong, B. (2020) Phylogenomic insights into deep phylogeny of angiosperms based on broad nuclear
788 gene sampling. *Plant Commun.*, **1**, 100027.
789 59. Zhang, C., Zhang, T., Luebert, F., Xiang, Y., Huang, C., Hu, Y., Rees, M., Frohlich, M.W., Qi, J.
790 and Weigend, M. *et al.*. (2020) Asterid phylogenomics/phylotranscriptomics uncover morphological
791 evolutionary histories and support phylogenetic placement for numerous whole genome duplications.
792 *Mol. Biol. Evol.*, **37**, msaa160.
793 60. Hu, H., Sun, P., Yang, Y., Ma, J. and Liu, J. (2023) Genome-scale angiosperm phylogenies based
794 on nuclear, plastome, and mitochondrial datasets. *J. Integr. Plant Biol.*, **65**, 1479–1489.
795 61. Liu, L., Chen, M., Folk, R.A., Wang, M., Zhao, T., Shang, F., Soltis, D.E. and Li, P. (2023)
796 Phylogenomic and syntenic data demonstrate complex evolutionary processes in early radiation of the
797 rosids. *Mol. Ecol. Resour.*, **23**, 1673–1688.
798 62. Baker, W.J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., Botigué, L.R., Brewer, G.,
799 Carruthers, T. and Clarkson, J.J. *et al.*. (2022) A comprehensive phylogenomic platform for exploring
800 the angiosperm tree of life. *Syst. Biol.*, **71**, 301–319.
801 63. Maurin, O., Epitawalage, N., Françoso, E., Roy, S.R., Simpson, L., Toledo Romero, E., Cowan,
802 R.S., Kim, J.T., DeBolt, L. and Hendriks, K.P. *et al.*. (2024) Phylogenomics and the rise of the
803 angiosperms. *Nature*, **629**, 843–850.
804 64. Shang, H., Jia, K., Zhou, M., Yang, H., Ma, Y. and Zhang, R. (2024) Phytop: A tool for
805 visualizing and recognizing signals of incomplete lineage sorting and hybridization using species trees

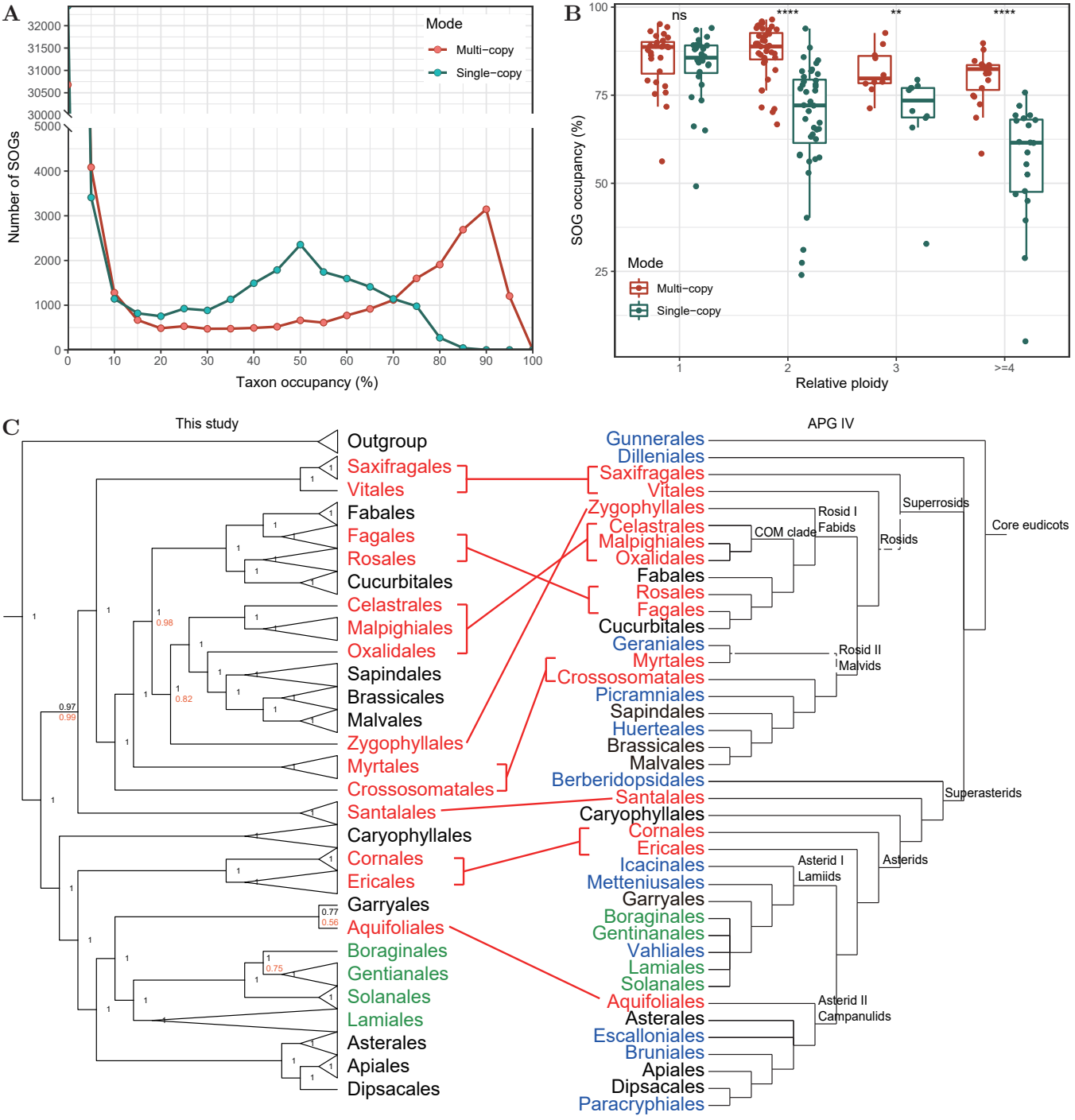
806 output from ASTRAL. *bioRxiv*.
807 65. Lewin, H.A., Richards, S., Lieberman Aiden, E., Allende, M.L., Archibald, J.M., Bálint, M.,
808 Barker, K.B., Baumgartner, B., Belov, K. and Bertorelle, G. *et al.*. (2022) The Earth BioGenome
809 Project 2020: Starting the clock. *Proc. Natl. Acad. Sci. U. S. A.*, **119**, e2115635118.
810 66. Kristensen, D.M., Wolf, Y.I., Mushegian, A.R. and Koonin, E.V. (2011) Computational methods
811 for Gene Orthology inference. *Brief. Bioinform.*, **12**, 379–391.
812


A
Ks
1.5
1
0.5
0
ΔT
γ
T
D
S
Salix dunnii
Populus trichocarpa
D WGD (tetraploidization)
T WGT (hexaploidization)
S Speciation
Out-paralogy
Out-paralogy
Orthology
B
C
D
E
(a)
Populus trichocarpa
Salix dunnii
(b)
Number of gene pairs
Ks
OI
(c)
Number of windows
Synteny depth (y / x)
(d)
Synteny depth (x / y)


A
Out-paralogy
Orthology
Proportion of gene pairs
OrthoIndex
B
0.3–0.4
0.4–0.5
0.5–0.6
0.6–0.7
R = −0.33, p = 0.001
R = −0.3, p = 0.004
R = −0.15, p = 0.162
R = 0.13, p = 0.22
Proportion of gene pairs
ΔT
C
Proportion of gene pairs
OrthoIndex
Ks
Orthology
Out-paralogy
D
OrthoIndex
precision
recall
F1−score
E
precision
recall
F1−score
WGDI
OrthoFinder2
OrthoIndex: 0.5
OrthoIndex: 0.6
ns


A
Shared WGD
Specific WGDs
Species tree
Synteny tree
Synteny
(colored by Orthology Index)
WGD
Speciation
A
B
A1
B1
A2
B2
Orthology
Out-paralogy
B
Centella asiatica
Aralia elata
OI
C
Vitis vinifera
Ilex polyneura
Lonicera japonica
Centella asiatica
Coriandrum sativum
Apium graveolens
Angelica sinensis
Daucus carota
Eleutherococcus senticosus
Panax ginseng
Panax notoginseng
Aralia elata
Apioideae
Apiaceae
Araliaceae
ω
α
MRCA
WGD (tetraploidization)
L2 = 0.074
L1 = 0.205
substitutions/site


A
A. thaliana
A. arenosa
A. thaliana x A. arenosa
B2
B1
7
6
1
1
1
2
A1
A2
B
A1
B1
B2
A2
Arabidopsis thaliana
A.thaliana x A.arenosa
Arabidopsis arenosa
C
A. duranensis
A. ipaensis
Arachis hypogaea
B2
B1
20
10
10
10
A1
A2
D
A1
B1
B2
A2
Arachis duranensis
Arachis hypogaea
Arachis ipaensis
E
Triticum turgidum
Triticum aestivum
B3
B2
B1
2D
2B
2A
2A
2B
A1
A2
F
A1
B1
A2
B2
B3
Triticum turgidum
Triticum aestivum
G
Papaver somniferum
Papaver setigerum
B3
B4
B2
B1
21
18
14
9
9
11
A1
A2
0.0
0.2
0.4
0.6
0.8
1.0
OI
H
B3
B4
B1
A1
B2
A2
Papaver setigerum
Papaver somniferum


A
Mode
Multi−copy
Single−copy
Number of SOGs
Taxon occupancy (%)
B
ns
****
**
****
SOG occupancy (%)
Relative ploidy
>=4
C
This study
Outgroup
Saxifragales
Vitales
Fabales
Fagales
Rosales
Cucurbitales
Celastrales
Malpighiales
Oxalidales
Sapindales
Brassicales
Malvales
Zygophyllales
Myrtales
Crossosomatales
Santalales
Caryophyllales
Cornales
Ericales
Garryales
Aquifoliales
Boraginales
Gentianales
Solanales
Lamiales
Asterales
Apiales
Dipsacales
APG IV
Gunnerales
Dilleniales
Saxifragales
Vitales
Zygophyllales
Celastrales
Malpighiales
Oxalidales
Fabales
Rosales
Fagales
Cucurbitales
Geraniales
Myrtales
Crossosomatales
Picramniales
Sapindales
Huerteales
Brassicales
Malvales
Berberidopsidales
Santalales
Caryophyllales
Cornales
Ericales
Icacinales
Metteniusales
Garryales
Boraginales
Gentinanales
Vahliales
Lamiales
Solanales
Aquifoliales
Asterales
Escalloniales
Bruniales
Apiales
Dipsacales
Paracryphiales
Rosid I
Fabids
COM clade
Rosid II
Malvids
Superrosids
Rosids
Core eudicots
Superasterids
Asterids
Asterid I
Lamiids
Asterid II
Campanulids