Whole-genome microsynteny-based phylogeny of angiosperms

Tao Zhao (tao.zhao@nwafu.edu.cn)
Ghent University

Jiayu Xue
Institute of Botany, Jiangsu Province and Chinese Academy of Sciences

Arthur Zwaenepoel
Ghent University

Shu-min Kao
Ghent University

Zhen Li
Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Gent, Belgium.

M. Schranz
Wageningen University https://orcid.org/0000-0001-6777-6565

Yves Van de Peer
Ghent University https://orcid.org/0000-0003-4327-3730

Article

Keywords: Synteny network, gene order, matrix representation, phylogeny, angiosperms, maximum-likelihood

Posted Date: August 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-51378/v1

Version of Record: A version of this preprint was published at Nature Communications on June 9th, 2021.
See the published version at https://doi.org/10.1038/s41467-021-23665-0.

Whole-genome microsynteny-based phylogeny of angiosperms

Tao Zhao[1,2,3], Jia-Yu Xue[4], Arthur Zwaenepoel[1,2], Shu-Min Kao[1,2], Zhen Li[1,2], M. Eric Schranz[5], Yves Van de Peer[1,2,6,7]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium.

[2]Center for Plant Systems Biology, VIB, Ghent, Belgium.

[3]State Key Laboratory of Crop Stress Biology for Arid Areas/Shaanxi Key Laboratory of Apple, College of Horticulture, Northwest A & F University, Yangling, 712100, China.

[4]Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing, China.

[5]Biosystematics Group, Wageningen University and Research, Wageningen, The Netherlands.

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

[7]College of Horticulture, Nanjing Agricultural University, Nanjing, China.

To whom correspondence should be addressed:

yves.vandepeer@psb.ugent.be

tao.zhao@nwafu.edu.cn

Abstract

Plant genomes are generally very complex and dynamic structures, and vary greatly in size, organization, and architecture. This is mainly due to the often-excessive numbers of transposable and repetitive elements, as well as to the fact that many plants are ancient or recent polyploids. Such (recurrent) whole-genome duplications are usually followed by genomic rearrangements, gene transpositions and gene loss, making local gene order-based phylogenetic inference particularly challenging. Nevertheless, microsynteny, i.e. the conservation of local gene content and order, has been recognized as a valuable and alternative phylogenetic character to sequence-based characters (nucleotides or amino acids) for the inference of phylogenetic trees, but to date its application for reconstructing larger phylogenies has been, for several reasons, limited. Here, by combining synteny network analysis, matrix representation, and maximum likelihood, we have reconstructed a microsynteny-based phylogenetic tree for more than 120 available high-quality plant genomes, representing more than 50 different plant families and 30 plant orders within the angiosperms. Comparisons with sequence alignment-based trees and current phylogenetic classifications show that we reconstruct very accurate and robust phylogenies, albeit with sometimes important alternative sister-group relationships. For instance, our synteny-based tree positioned Vitales as early-diverging eudicots, Saxifragales belongs to superasterids, and magnoliids as sister to monocots. We discuss how synteny-based phylogeny can be complementary to traditional methods and could provide additional insights into some long-standing controversial phylogenetic relationships.

Key words

Synteny network, gene order, matrix representation, phylogeny, angiosperms, maximum-likelihood

Introduction

Microsynteny (hereafter also simply referred to as synteny), or the local conservation of gene order and content, provides a valuable means to infer the shared ancestry of groups of genes and is commonly used to infer the occurrence of ancient polyploidy events[1,2], to identify genomic rearrangements[3], and to establish gene orthology relationships[4-6], particularly for large gene families where sequence-based phylogenetic methods may be inconclusive[7,8]. In addition, microsynteny has also been used for the inference of phylogenetic relationships. For instance, recently, Drillon et al. (2020)[9] developed a bottom-up pairwise (partial splits) distance-based tree reconstruction approach that starts with the identification of breakpoints between synteny blocks, followed by the identification of partial splits. This method was successfully applied to 13 vertebrate genomes and 21 yeast genomes[9]. However, in general, application of synteny data for phylogenetic inference in complex eukaryotic genomes has been very limited[9-11]. Compared to sequence-based phylogenetic approaches, methods based on gene order analyses are usually computationally costly and ancestral genome reconstruction and inference of gene order rearrangements is a highly challenging combinatorial and algorithmic problem[12-14]. Most methods hitherto developed can only handle pairs of genomes, or single-chromosome organisms and organelles, and such methods have therefore only been applied to simpler genomes such as plastid genomes or bacterial genomes, or simplified genome datasets or simulated datasets[15,16].

Plant genomes are highly diverse[17,18] and affected by a complex interplay of small- and large-scale duplication events[19,20], hybridization[21], transposable element (TE) activity[22], and gene loss[23,24]. Consequently, employing synteny information for phylogenetic inference from large plant genome datasets has been notoriously difficult. Ideally, inferring phylogenetic relationships from synteny data should involve dealing with (1) multiple chromosomes; (2) genome duplications; and (3) multiple genomes at the same time, which greatly increases the complexity of the problem and computational cost. Recently, some of us developed a novel approach in which microsynteny information is converted into a network data structure, and which has proven to be well suited for evolutionary synteny comparisons among many eukaryotic nuclear genomes[7,18]. Using this approach, conservation or divergence of genome structure can be conveniently summarized and reflected by synteny cluster

sizes and composition. Importantly, the network representation of synteny relationships provides an abstraction of structural homology across genomes that is in principle amenable to tree inference using standard approaches from phylogenetics, however this possibility has hitherto not been explored.

Here, we combined synteny networks and their matrix representation with standard maximum-likelihood based statistical phylogenetics to reconstruct phylogenies for real-life plant whole genome datasets. We have constructed a well-resolved ‘synteny tree’ using currently available representative genomes of flowering plants. The obtained tree is highly comparable with current phylogenetic classifications, although some notable differences concerning the phylogenetic positions of magnoliids, Vitales, Caryophyllales, Saxifragales, and Santalales were observed. We believe that our approach provides a noteworthy complement to more classical approaches of tree inference and could help to solve some longstanding problems that may remain difficult to solve with sequence (alignment) based methods.

Results

Highly-resolved microsynteny-based phylogeny for angiosperm genomes

After quality control, 123 fully-sequenced plant genomes were used for synteny-based phylogenetic analysis, which includes synteny network construction and clustering, matrix representation of synteny followed by maximum likelihood estimation (Fig. 1, see Methods for details). The size of the matrix used for ML tree inference is 123 x 137,833, which contains a binary presence/absence coding for each cluster in the synteny network. The resulting best ML tree demonstrated that the overall monophyly of most clades was strongly supported, and 110 of the 122 nodes had ≥ 95% bootstrap support (Fig. 2). *Amborella* was used as the sister group to other angiosperms in order to root the obtained phylogenies. Nymphaeales was resolved as a successive lineage right after *Amborella,* again sister to all other angiosperm genomes (Fig. 2). The monophyly of mesangiosperms was also strongly supported (Fig. 2). Interestingly, magnoliids (including Laurales and Magnoliales) form a sister group to monocots (BS = 95%), with the resulting clade sister to the eudicots (BS = 100%). Notably, our synteny tree strongly favors Vitales as sister to the rest of the core eudicots, or as another clade of early-diverging eudicots, branching off after Proteales (*Nelumbo*) (both BS = 100%) (Fig. 2). Moreover, in the synteny tree, we find Santalales as sister to Saxifragales (BS = 99%), and with both sister to Caryophyllales (BS = 99%), which in turn was found as sister to all other asterids (BS = 94%). Synteny trees positioned Myrtales (*Eucalyptus* and *Punica*) as early-diverging rosids (BS = 100%), and Malpighiales as early-diverging Malvids (BS = 100%). Almost all the nodes within Brassicales were fully resolved (Fig. 2). The monophyly of the Nitrogen-Fixing Clade was fully recovered, and supports a relationship of ((Fagales, Fabales), (Rosales, Cucurbitales)), however with lower support (46% and 64% BS support, respectively) (Fig. 2).

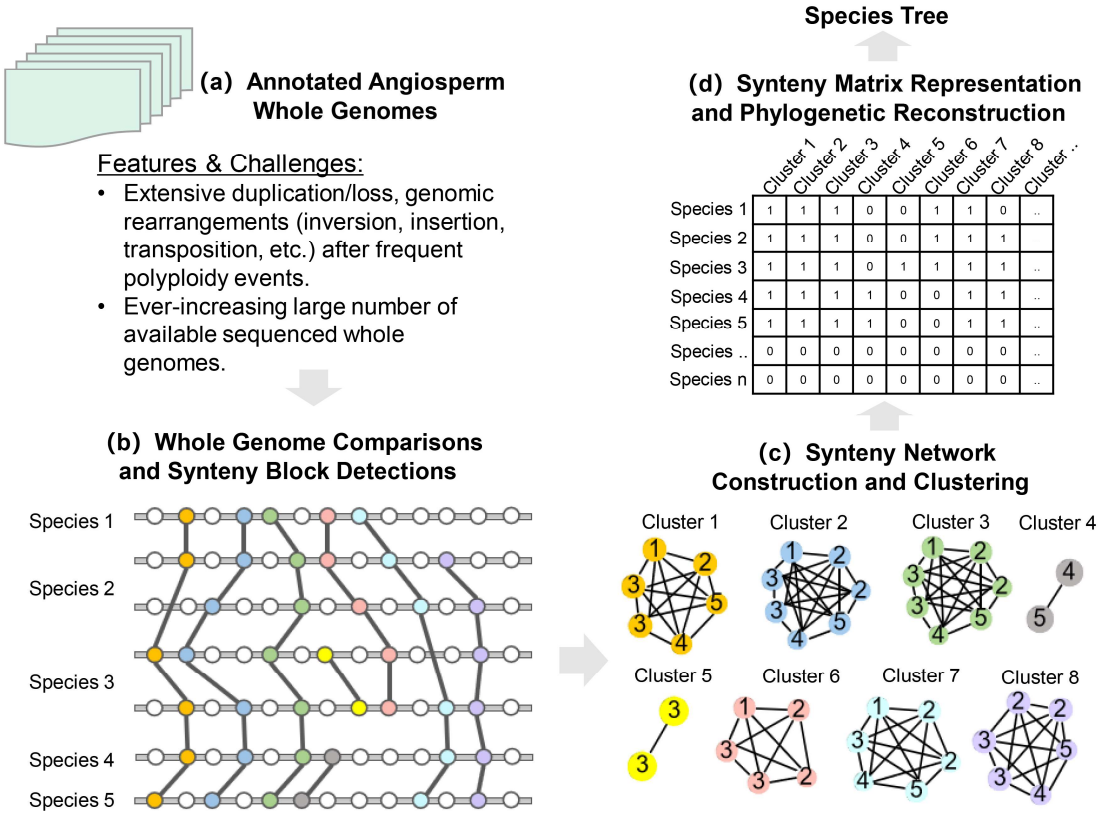


Fig. 1 Whole-genome microsynteny-based species tree inference. (**a**) Whole-genome datasets with all predicted genes are used for phylogeny reconstruction. (**b**) The synteny network approach first conducts all pairwise reciprocal genome comparisons, followed by synteny block detection. All syntenic blocks constitute the synteny network database (see Methods for details). (**c**) We analyzed all synteny clusters after clustering the entire network database. Synteny clusters vary in size and node compositions. Shared genomic rearrangements are reflected by cluster compositions. Specific anchor pairs shared by a lineage/species form specific clusters (e.g. Clusters 4-6). We account for the presence or absence of the same recurring anchors for multiple blocks derived from whole-genome or segmental duplications (e.g. for Species 2 and 3 in Clusters 2, 3, 5, and 8). (**d**) The phylogenomic profiling of all clusters constructs a binary matrix, where rows represent species and columns represent clusters. The synteny matrix comprehensively represents phylogenomic gene order dynamics. It transforms the concept of synteny comparisons from analyzing massive parallel coordinates plots into analyzing profiles of individual clusters/networks. Each cluster stands for a shared homologous 'context'. For example, genes (driven by transposable elements) can be transposed as insertions into new contexts or can be lost from the original context (e.g. genes in Clusters 4-6). As long as such transpositions are shared by different genomes (e.g. genes in Clusters 4 and 6) or within the same genome because of whole genome duplication (e.g. genes in Cluster 5), specific clusters will emerge and corresponding signals will be added to the matrix. This synteny matrix is used as the input for species tree inference by maximum likelihood (referred to as Syn-MRL).

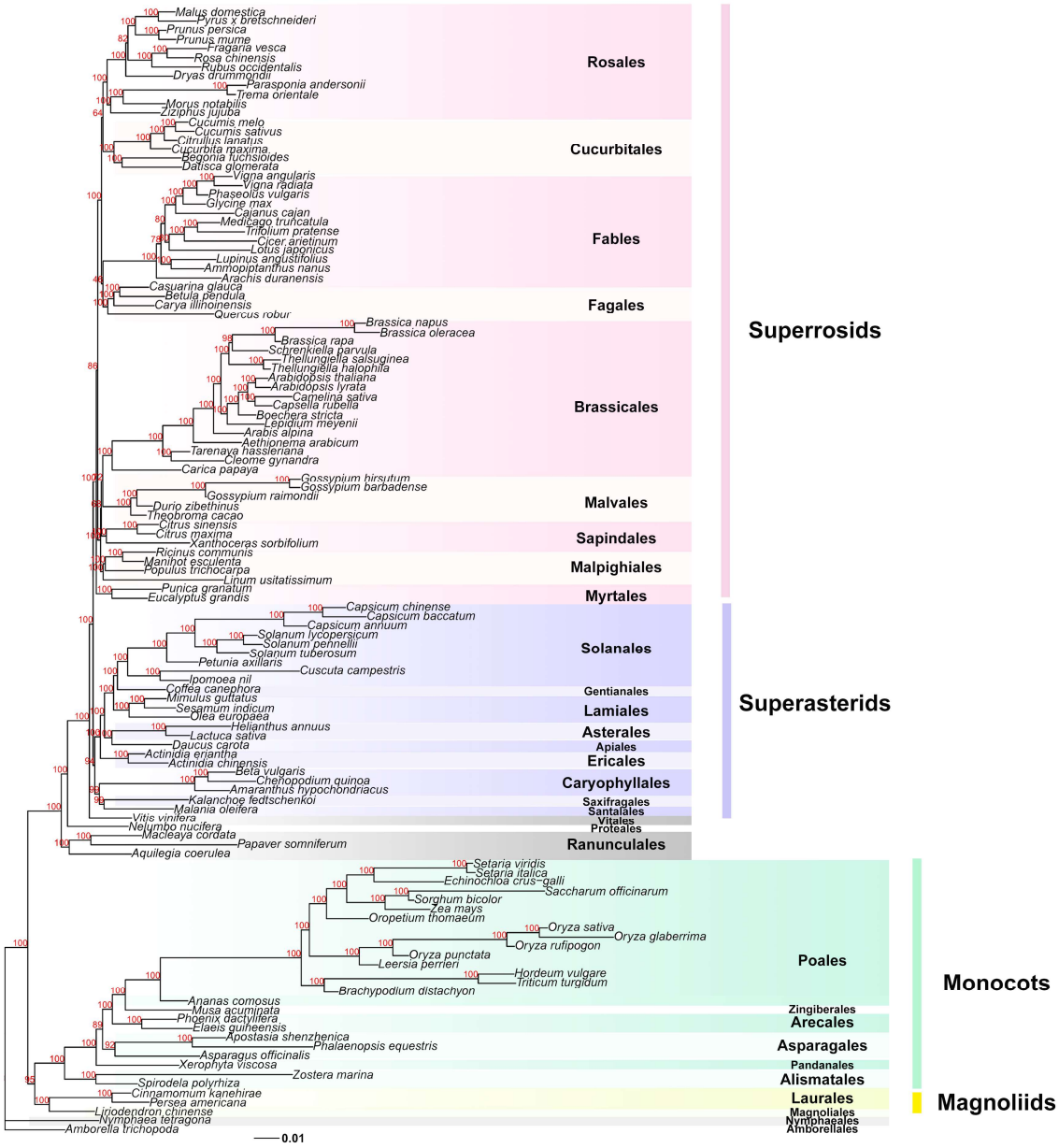


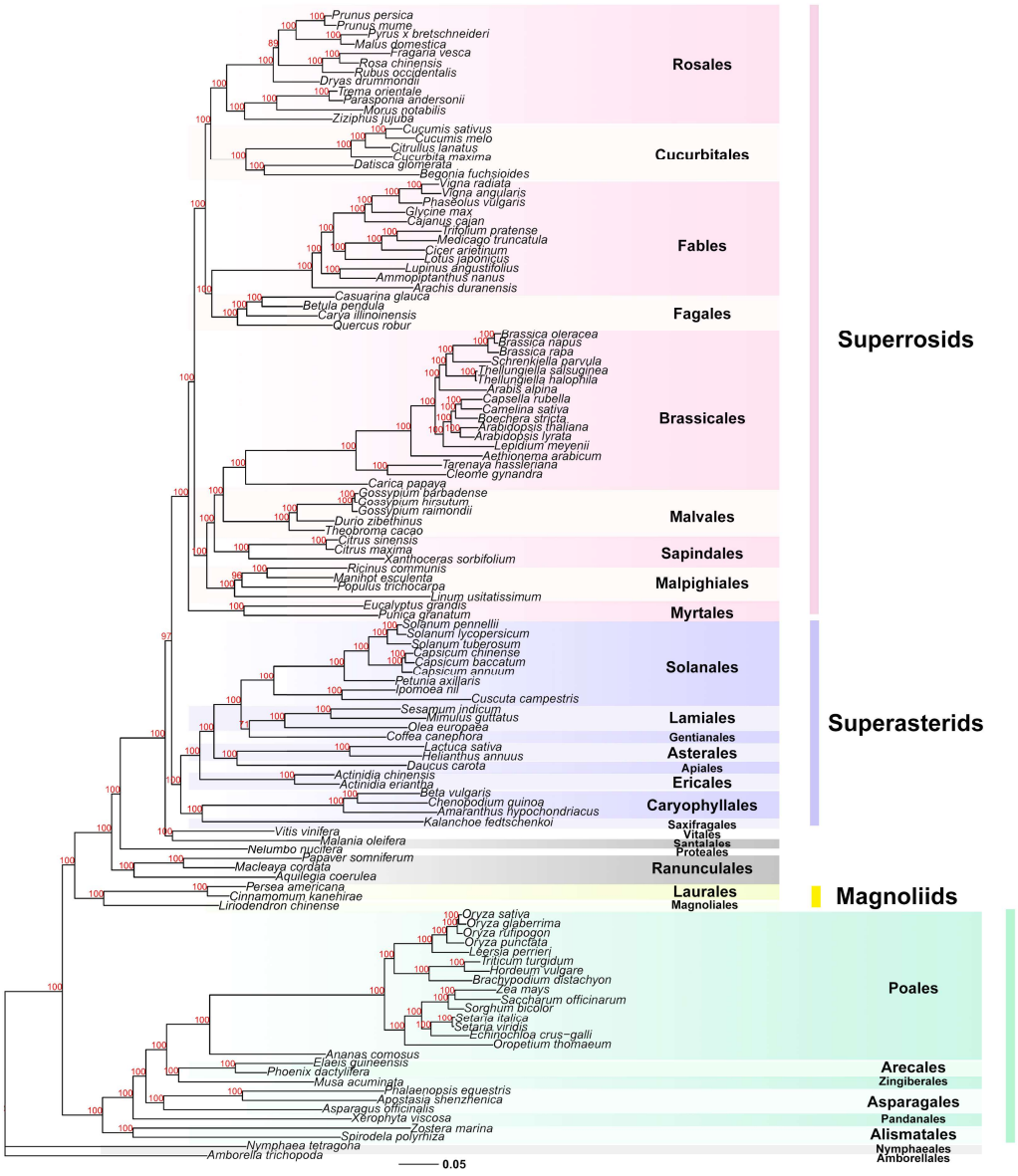
Fig. 2 Maximum likelihood (ML) tree for 123 fully-sequenced flowering plant genomes based on the microsynteny approach. The tree is rooted by *Amborella*, and four main clades, i.e. superrosids, superasterids, monocots, and magnoliids are shaded in light-red, light-purple, light-green, and light-yellow, respectively. Ultrafast bootstrapping values are denoted for all the nodes. Names for the different plant orders follow the APG IV classification[25].

Comparison of the synteny tree with sequence alignment-based phylogenies (using whole-genome derived gene markers), and with current phylogenetic classifications

Taking advantage of the large whole-genome plant dataset used in this study, we also reconstructed phylogenies using widely-adopted sequence alignment-based approaches for comparison with our microsynteny-based tree. We used BUSCO gene sets, which are widely adopted in benchmarking genome assembly and annotation quality. Besides BUSCO, we developed a set of CSSC (Conserved Single-copy Synteny Clusters) gene markers from a profiling and screening of all synteny clusters (see Methods). We used supermatrix, supertree, and multispecies coalescent as representative sequence-based approaches (Supplementary Fig. 1) for these two sets of gene markers, and generated six phylogenetic trees (Supplementary Figs. 2-7). Overall, the six sequence alignment-based trees (hereafter referred to as SA trees) are highly similar, and only differ by a few bootstrap support values and a few minor sister-group relationships (Supplementary Figs. 2-7). We used the supermatrix-BUSCO tree as the representative of the SA trees and compared it with the synteny tree (Fig. 3).

Comparing both trees shows that the synteny tree we have obtained is highly congruent with the SA tree (Figs. 2-3, Supplementary Fig. 8). Interestingly, some notable differences were found between the two trees, such as the positioning of magnoliids, Santalales, Zingiberales, and Gentianales (Figs. 2-3, Supplementary Fig. 8). Additional subtle differences were found within the orders, such as in Poales and Brassicales (Figs. 2-3, Supplementary Fig. 8). We further compared our synteny tree with the current Angiosperm Phylogeny Group (APG) phylogeny (version IV)[25], and the recent 1000 plant transcriptomes (1KP) phylogeny of green plants[20] at the order level (Supplementary Figs. 9-11). Apart from the orders that do not have representative genomes yet, our synteny tree generally shows strong congruence to both the APG and 1KP trees (note however that a certain degree of incongruence exists between the two latter trees (Supplementary Fig. 9)) (Supplementary Figs. 10-11). Besides the relationship of magnoliids and monocots and dicots, all discrepancies are confined to the positions and relationships of Vitales, Santalales, Saxifragales, Caryophyllales (Supplementary Figs. 9-11), which reflect long-time controversies in plant systematics and have been well acknowledged in the

189 literature[20,25-33]. Our SA trees agree with the 1KP tree on magnoliids being sister to
190 eudicots, which is different from the synteny tree (magnoliids sister to monocots) and
191 APG tree (magnoliids sister to both monocots and eudicots) (Supplementary Figs. 8-
192 11); the APG tree and the 1KP tree resolve Vitales and Saxifragales as early-
193 diverging lineages within Rosids, whereas our SA trees (except the ASTRAL-
194 BUSCO tree) and synteny tree favor Vitales as sister to the rest of the core eudicots,
195 and Saxifragales cluster within Superasterids (Supplementary Figs. 2-11). We tested
196 whether our synteny matrix (under the current synteny network construction
197 parameters) would reject representative alternative topologies. To this end, we used
198 the 'approximately unbiased' (AU) test[34] to evaluate the support for alternative
199 topologies regarding the positions of magnoliids, Vitales, Saxifragales,
200 Caryophyllales, and Myrtales (Supplementary Fig. 12). The test shows that our ML
201 synteny tree was found to be significantly better than the alternative topologies with
202 all alternative topologies rejected at the $p = 0.05$ level, except for the scenario where
203 magnoliids are sister to eudicots ($p = 0.146$, Supplementary Fig. 12a). This means
204 that, at least based on our synteny matrix and the (admittedly *ad hoc*) Markov model
205 of character evolution, alternative scenarios such as magnoliids sister to both
206 monocots and eudicots (Supplementary Fig. 12b), Vitales (Supplementary Fig. 12c),
207 or Saxifragales (Supplementary Fig. 12d) as early-diverging rosids, and so on, were
208 significantly less well supported.



209

210 **Fig. 3** Maximum likelihood (ML) tree based on the concatenation of protein
211 alignments of BUSCO genes. This tree was used as the representative SA
212 (sequence-alignment based) tree.

Discussion

Phylogenetic trees serve many purposes and are an indispensable tool for the interpretation of evolutionary trends and changes. So far, abundant tools and sophisticated substitution models have been developed for molecular sequence alignment-based phylogenetic inference. In contrast, although it is well acknowledged that there is phylogenetic signal in gene order dynamics[4,35-39], efficient tools for analyzing phylogenomic synteny properties, reconstructing ancestral genomes and inference of genome rearrangements, particularly for large datasets, remain scarce. In this study, we present an approach that bypasses the challenging combinatorial problem of inferring genome rearrangements and ancestral genome organization by integrating information on pairwise homologous genomic organization across many plant genomes in a network representation. As in classical orthogroup inference, we perform similarity searches across multiple genomes and cluster a network representation thereof, but our graph representation now also includes the genomic context (i.e. synteny information), and the resulting clusters are therefore defined by sequence level homology and structural homology at a micro-synteny level. The idea developed in the present paper is that each cluster, no matter how conserved, family- or lineage-specific, contains a phylogenetic signal that can readily be used for tree inference. By reducing whole-genome syntenic comparisons to the synteny network representation, information with regard to the actual syntenic contexts of the genes is ignored while only relations of shared syntenic contexts between genes are retained. By identifying clusters in the resulting synteny network, evolutionarily relevant homologous features of genome structure across species are efficiently captured while we are able to abstract from much of the complex features of genome evolution.

Accuracy and reliability of the synteny tree

We exploited the total information residing in all synteny clusters by encoding their phylogenomic synteny pattern profiles into a large binary matrix (synteny matrix) (Fig. 1). Standard ML-based tree inference was then applied to reconstruct a species tree (Fig. 1). The process is similar to the MRL (Matrix Representation with Likelihood) supertree method (which uses the same data matrix as Matrix Representation with

Parsimony, but with ML-based inference, yielding higher accuracy[40,41]), except that MRL is based on a set of input trees, in contrast with our synteny matrix based on synteny clusters. To avoid misunderstanding, in this study we confine the usage of 'MRL' to the supertree method; and we refer to our 'supercluster' approach as Syn-MRL (Synteny Matrix Representation with Likelihood).

For the Syn-MRL approach, we use the Mk model, which is a Jukes-Cantor type model for discrete morphological data[42]. Each column of the data matrix is regarded as an independent character. The two-state binary encoding for each column represents the two groups of related species regarding that specific character. There is no further ordering or special weighing for the elements in the data matrix. The Mk model may arouse some concern regarding the symmetric two-state model (i.e., there is an equal probability of changing from state 0 to state 1 and from state 1 to state 0), as generally, time-reversible Markov models of evolution may not be ideal for the Syn-MRL approach. However, a first and important observation was that these models do seem to result in very reasonable well-supported phylogenies, in a similar vein as the MRL approach[40]. We hypothesize this is because, firstly, we may reasonably suspect that for our data, state 0 or 1 has the same probability of being the ancestral or derived state, as we can hypothesize that the emergence of a new lineage-specific synteny cluster and loss of another, are due to the same processes, e.g. transposition. Secondly, although the Mk model allows numerous state changes, in practice it yields trees with identical likelihood scores compared to a modified Mk model where character states can change only once or not at all[42]. This suggests that more biologically plausible non-reversible models, where for instance the re-emergence of a synteny cluster (a secondary 0→1 transition after the *initial* emergence and subsequent loss of the cluster) occurs at a different rate than the initial 0→1 transition, might not result in a substantially better fit. Nevertheless, we believe developing other probabilistic models for synteny cluster evolution is a fruitful avenue for further research.

Despite much progress in the last two decades, it is currently well-acknowledged that some phylogenetic relationships in plants remain especially controversial. For example, the relationships among monocots, eudicots, magnoliids, Ceratophyllaceae, and Chloranthales remain unsolved[20,25,43], while also the relationships among core rosids, asterids, Saxifragales, Vitales, Santalales, and Caryophyllales have been

enigmatic[20,25-29]. We have compared our synteny tree with SA trees and state-of-art classifications (represented by APG (IV) and 1KP) in a general way (Supplementary Figs. 8-11). Overall, our synteny tree showed great accuracy and congruence on the classification of the major lineages and clades. On the other hand, the synteny tree also provides alternative sister-group relationships for some of the recalcitrant clades mentioned higher. It should be noted that it is not our intention to argue that our tree is the 'true' tree, finally resolving the contentious relationships between some of the plant clades, but rather to provide a novel way to reconstruct and consider large-scale gene order-based phylogenetic trees.

The phylogenetic position of magnoliids has long been discussed[43], and even recently, based on whole-genome information of new magnoliid genomes, different sister-group relationships for magnoliids have been proposed[31-33,44]. Also, a sister-group relationship for magnoliids and monocots, as suggested in the current study, has been suggested before, based on different approaches and data[45-49]. Based on microsynteny, we have explored the relationships of magnoliids, monocots, and eudicots in more detail. Our synteny-based approach provides a means to investigate phylogenetic signal in the data matrix and trace those back to the genomic regions where differential gene order arrangements are located. To this end, we focused on the 'submatrix' of synteny profiles for magnoliids and extracted 15,424 magnoliid-associated synteny clusters (Fig. 4a). A hierarchical clustering of the phylogenomic profile of these synteny clusters showed 1,107 synteny clusters may related to the grouping of magnoliids and monocots (Fig. 4a). To validate their contribution to the final ML tree, we first removed these signals from the entire matrix and reconstructed the phylogenetic tree, after which the species tree obtained favors magnoliids as sister to eudicots (BS = 100%) (Supplementary Fig. 13). To further understand the genomic distribution of specific genes in these clusters joining magnoliids and monocots, we reorganized the cluster profiles according to the chromosome gene arrangement of a magnoliid representative (*Cinnamomum kanehirae*)[31](Supplementary Table S2). In doing so, we observed a number of 'signature' blocks consisting of specific anchor pairs that are shared by monocots and magnoliids with exclusion of eudicots (Supplemental Table S2-sheet 1, contexts with highlighted yellow rows). As an example, we highlight a synteny context of 15 genes where 8 genes (highlighted red) are only found in synteny between magnoliid

and monocot genomes (Fig. 4b), with flanking genes generally conserved across angiosperms (highlighted in blue) (Fig. 4b, Supplemental Table S2-sheet2). However, an alternative explanation could be that the synteny context is lost in eudicots, from which it might be 'wrongly' concluded that monocots and magnoliids share some 'derived' characters and therefore share a common ancestry. Also, some synteny contexts are shared with basal eudicots (Ranunculales) from the 1,107 clusters (Supplementary Table S2). Nevertheless, with more representative genomes of Chloranthales, Ceratophyllales, and early-diverging angiosperms to be added into the analysis, a better resolution of the genomic rearrangements for magnoliids and related lineages could be obtained.

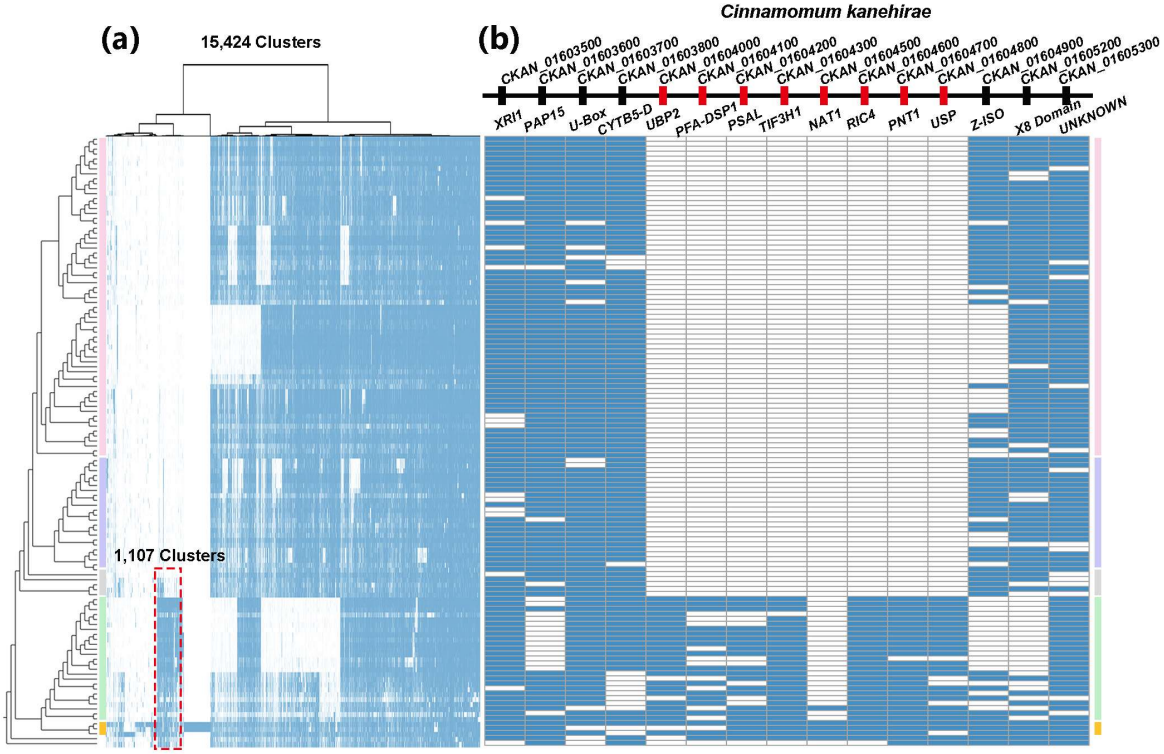


Fig. 4 Magnoliids-associated signals and a representative example of phylogenetically informative microsynteny. (**a**) Hierarchical clustering (ward.D) of 15,424 magnoliids-associate cluster profiles based on jaccard distance. On the far-left, the synteny-based species tree is displayed (same as Fig. 1). Superrosids, superasterids, early-diverging eudicots, monocots, and magnoliids are shaded in light-red, light-purple, light-grey, light-green, and light-yellow, respectively. 1,107 clusters supporting a grouping of magnoliids and monocots (Supplemental Table S2). (**b**) One example from all supporting signals. A fifteen-gene context in the genome of *Cinnamomum kanehirae* (a magnoliid) shows eight neighboring genes (highlighted in orange) only present in magnoliids and monocot genomes, while the flanking genes (colored blue) are generally conserved angiosperm-wide.

Apart from the contentious phylogenetic relationship of magnoliids with related lineages, evolutionary relationships of early-diverging core eudicots, such as Vitales, Saxifragales, Santalales, and Caryophyllales as suggested by microsynteny are strongly supported compared to competing hypotheses under the Syn-MRL approach (as reflected by the AU test statistic). For example, our synteny-based phylogenies strongly support Vitales as early-diverging, right after Proteales (*Nelumbo*). Trees based on concatenation, MRL, and ASTRAL-CSSC trees in general also support Vitales as early-diverging but also forming a sister-group relationship with Santalales (Supplementary Figs. 2-5, and 7). Several studies have already reported the positioning of Vitales as sister to the core-eudicots or as early-diverging eudicots, based on mitochondrial genes or genomic data[50,51], assembly of multi-nuclear genes[52], and large-scale transcriptome data[53,54]. The recent 1KP study also observed substantial gene-tree discordance for Vitales from analyses employing coalescent and supermatrix approaches, as well as plastid genomes[20]. Moreover, a recent analysis involving syntenic comparisons of the columbine (Aquilegia) and grapevine genomes revealed that the gamma palaeohexaploidy at the root of the core-eudicots is possibly a result of hybridization between a tetraploid and a diploid species (Aköz and Nordborg, 2019). The hybrid origin of core-eudicots could help us better understand the phylogenetic incongruence of the early-diverging groups within core-eudicots including Vitales[55].

Insights towards ancestral introgressive hybridization

Within the same plant family or order, our synteny tree also revealed some subtle differences with the SA trees (Figs 2-3, Supplementary Fig. 8). For example, the position of *Boechera* and *Arabis* in Brassicaceae, and the position of the Pooideae clade (including wheat, barley, and *Brachypodium*) in Poales (Supplementary Fig. 8).

Very interestingly, a study has comprehensively investigated the particular species branching order of three monophyletic groups in the Brassicaceae: *Boechera* (Clade B), *Capsella* and *Camelina* (Clade C), and *Arabidopsis* (Clade A)[56]. Forsyth et al. (2018) explicitly proved the existence of massive nuclear introgression between Clades B and C, which largely reduced the sequence divergence between them. As a result, the majority of single-copy gene trees strongly support the branching order

of (A,(B,C)), while the true branching order is supposed to be (B,(A,C))[56]. Their study employed multiple approaches and measurements to show that neither gene duplication and loss nor ILS, nor phylogenetic noise can adequately explain such incongruence, except introgressive hybridization[56]. Interestingly, our results are congruent with the findings of Forsythe et al. (2018). Alignment-based trees (supermatrix, MRL, and ASTRAL) all support (A,(B,C)) (Supplementary Fig.s 2-7), whereas only the synteny tree supports the 'true' species branching order (B,(A,C)).

Regarding *Arabis* in Brassicaceae, another recent study identified shared genomic block associations between *Aethionema* (which is sister to all other Brassicaceae) and *Arabis* by extending the analysis of Brassicaceae genomic blocks to *Aethionema*[57]. Their study of macrosynteny supports Arabideae as the next diverging clade following *Aethionema* in Brassicaceae[57], which is also consistent with our synteny tree.

For Pooideae in Poales, in order to test whether potential bias exists caused by specific genome(s), we further compared the clustering of Pooideae (wheat, barley, and Brachypodium) with the rest of the Poaceae family by removing one or two genomes from Pooideae from the synteny matrix and rebuilding phylogenetic trees. We obtained consistent results and topologies regarding the positioning of the clade of Pooideae (Supplementary Figs. 14-17). To our knowledge, similar relationships regarding the PACMAD and the BEP clades have rarely been reported, however, the comparative phylogenomic large-scale gene family expansion and contraction analysis from the wheat genome project showed a similar clustering pattern of species based on gene family profiles, thus provided some evidence in support of our result (Figure 4A of International Wheat Genome Sequencing Consortium, 2018[58]). If the synteny tree is indeed true, it would reshape our understanding of the evolution of some of the most important crops including wheat, rice, and maize, as well as the origin of the C_4 lineages[59].

The notion that species boundaries can be obscured by introgressive hybridization is increasingly accepted[56,60-62]. Thus, it is plausible that the incongruences discussed above may indicate an effect from introgression caused by (recent or more ancestral) hybridization. Since the focal point of synteny information is to reflect genomic structure variation instead of gene sequence changes, we believe it is plausible that genome-level synteny-based phylogeny inference may avoid the bias caused by

(extensive) introgression for sequence-based species tree reconstruction and can reflect the 'true' (or perhaps better, 'main') species branching order. Also, TE mobilization often follows hybrid speciation or introgression[63-65] which leads to gene transpositions. However, our approach can capture potential gene transpositions and rearrangements as novel anchor pairs within synteny blocks and convert those into phylogenetic signals (explained in Fig. 1). Nevertheless, in the case of true hybrid speciation, where a bifurcating species tree does not exist, our method can be similarly unsuitable as other phylogenetic inference methods that do not allow phylogenetic network inference. In any case, the whole-genome synteny approach could provide unique opportunities towards developing phylogenetic inference methods that are robust to introgressive hybridization[66].

Caveats and conclusion

Here, we have presented a methodological roadmap to reconstruct species trees based on synteny information from large volumes of available whole-genome data, which can be applied to any set of genomes. However, it should be noted that our approach depends on the quality of genome assemblies and their gene annotations, which is the basis for synteny detection. Also, parameter settings for synteny detection should be tested and compared beforehand (here we adopt the parameterization found to be most appropriate for the analysis of angiosperm genomes in a previous study[18]). One should consider the evolutionary distance and genomic properties of the added genomes, as synteny conservation can be scarce for distantly related species. For example, caution should be taken when comparing a gymnosperm or fern genome to an angiosperm genome because in such comparisons, microsynteny might be inadequate due to extensive genome rearrangements. On the contrary, for highly similar genomes, our approach might simply not provide enough resolution due to the lack of informative rearrangements. However, in such cases, stricter synteny detection parameter settings, as well as consideration of gene orientation might help to increase the number of informative signals for resolving the tree topology.

To conclude, using a character matrix derived from a network representation of pairwise microsynteny relations, we here explored a maximum likelihood approach to

reconstruct phylogenies using genome structure data across a large set of angiosperm genomes. Our resulting synteny-based species tree showed high resolution and strong consistency with phylogenies of angiosperms using more classical methods to infer tree topologies and current classifications, although some notable differences were identified. We hope that our approach might offer a complementary way to consider and evaluate ambiguous phylogenetic relationships. Furthermore, as more and more high-quality genomes from underrepresented plant phyla are becoming available at increasing rates, we expect our approach to become more sensitive and informative in future applications.

Methods

Genome resources

Reference genomes were obtained from public repositories, including Phytozome, CoGe, GigaDB, and NCBI. For each genome, we downloaded FASTA format files containing protein sequences of all predicted gene models and the genome annotation files (GFF/BED) containing the positions of all the genes. We modified all peptide sequence files and genome annotation GFF/BED files with corresponding species abbreviation identifiers. After constructing our synteny network database and clustering (see further), poor quality genomes could be relatively easily identified (see further), and were removed from the database for further analysis. After quality control, the final list of genomes used in the current study and related information for each genome can be found in Supplemental Table S1-Sheet1, genomes that were filtered out due to low contiguity were listed in Supplemental Table S1-Sheet2.

We acknowledge that our taxon sampling is limited to currently available high-quality genomes and thus many important lineages could not be included.In our current study, we have included genome sequences of as many representative angiosperm families and orders as possible (https://www.plabipd.de/portal/web/guest/angiosperm-phylogenetic-view). However, some important plant lineages still lack well-assembled genomes (such as for Dilleniales and Chloranthales), while others only have few representatives (such as Santalales and Saxifragales), and still others are relatively 'over-represented' (such as Brassicales, Fabales, and Poales). For 'over-represented' orders, we kept all qualified genomes because this provides an opportunity to test the resolution and robustness of our approach at different levels (e.g. order-level, family-level, and species-level).

We manually downloaded each genome and checked the completeness of the annotation files. Selecting genomes to be used for the analysis initially not based on certain cutoffs, such as N50 or BUSCO completeness. First, candidate genomes were all used for the synteny network construction. Next, 'genome quality' is checked manually in the phylogenomic profiling plot/matrix (of all synteny clusters, where rows represent species/genomes and columns are clusters). Genomes of poor completeness and contiguity - indicated by lighter rows (for an example, see Figure 5

of Zhao and Schranz, 2019[18], rows indicated by black arrows) - are removed from the microsynteny network. After this step, 123 fully sequenced plant genomes were used for further analysis (Supplemental Table 1). The overall sampling covered 31 orders and 52 different families of angiosperms. We are aware that for some orders there is an overrepresentation (Brassicales, Fabales, and Poales), while for orders such as Santalales, Saxifragales, and Gentianales, there is only one representative. Moreover, for some genera, several genomes are included, for example for *Oryza*, *Solanum*, *Gossypium*, and *Brassica*. Among those, *B. napus*, *G. herbaceum,* and *G. arboreum* are polyploid (allopolyploid) genomes.

Pipeline for whole-genome microsynteny-based phylogenetic inference

Our synteny-based phylogenetic reconstruction approach includes four main steps, in turn namely phylogenomic synteny network construction, network clustering, matrix representation, and maximum-likelihood estimation. Together we call our approach 'Syn-MRL' for short.

The synteny network construction consists of two main steps: first, all-vs-all reciprocal annotated-protein comparisons of the whole genome using DIAMOND was performed[67], followed by MCScanX[68], which was used for pairwise synteny block detection. Parameter settings for MCScanX have been tested and compared before[18]; here we adopt 'b5s5m25' (b: number of top homologous pairs, s: number of minimum matched syntenic anchors, m: number of max gene gaps), which has proven to be appropriate by various studies for the evolutionary distances among angiosperm genomes. To avoid large numbers of local collinear gene pairs due to tandem arrays, if consecutive homologs (up to five genes apart) share a common gene, homologs are collapsed to one representative pair (with the smallest E-value). Further details regarding phylogenomic synteny network construction can be found in a tutorial available in the associated GitHub repository (https://github.com/zhaotao1987/SynNet-Pipeline). Each pairwise synteny block represents pairs of connected nodes (syntenic genes), all pairwise identified synteny blocks together form a comprehensive synteny network with millions of nodes and edges. In this synteny network, nodes are genes (from the synteny blocks), while edges connect syntenic genes. For our work, the entire synteny network summarizes

503 information from 7,435,502 pairwise syntenic blocks, and contains 3,098,333 nodes
504 (genes) and 94,980,088 edges (syntenic connections).

505 The entire synteny network (database) is clustered for further analysis. We used the
506 Infomap algorithm for detecting synteny clusters within the map equation
507 framework[69] (https://github.com/mapequation/infomap). We have discussed before
508 why Infomap is more appropriate for clustering phylogenomic synteny networks[18].
509 We used the two-level partitioning mode with ten trials (--clu -N 10 --map -2). The
510 network was treated as undirected and unweighted. Resulting synteny clusters vary
511 in size and composition, which is associated with synteny either being well-
512 conserved or rather lineage-/species-specific. A typical synteny cluster comprises of
513 syntenic genes shared by groups of species, which precisely represent phylogenetic
514 relatedness of genomic architecture among species (Fig. 1). Here, we classified the
515 entire synteny network into 137,833 synteny clusters.

516 A cluster phylogenomic profile shows its composition by the number of nodes in
517 each species. We summarize the total information residing in all synteny clusters as
518 a data matrix for tree inference. Phylogenomic profiles of all clusters construct a
519 large data matrix, where rows represent species, and columns as clusters (Fig. 1).
520 The matrix was then reduced to a binary presence-absence matrix to obtain the final
521 synteny matrix (Fig. 1). Here, the dimension of our input synteny matrix is 123 ×
522 137,833.

523 Tree estimation was based on maximum-likelihood as implemented in IQ-TREE
524 (version 1.7-beta7) (Nguyen et al., 2014), using the MK+R+FO model. (where “M”
525 stands for “Markov” and “k” refers to the number of states observed, in our case, k =
526 2). The +R (FreeRate) model was used to account for site-heterogeneity, and
527 typically fits data better than the Gamma model for large datasets[70,71]. State
528 frequencies were optimized by maximum-likelihood (by using ‘+FO’). We generated
529 1000 bootstrap replicates for the SH-like approximate likelihood ratio test (SH-aLRT),
530 and 1000 ultrafast bootstrap (UFBoot) replicates (-alrt 1000 -bb 1000)[72].

531

532 **Sequence alignment-based phylogenetic reconstruction**

533 For sequence-based phylogenetic inference, we employed three commonly used
534 approaches, namely a supermatrix (also called superalignment or concatenation)

535 approach, a reconciliation approach based on the multispecies coalescent (MSC),
536 and a supertree approach using matrix representation with likelihood (MRL). For
537 each of these, we used two sets of whole-genome derived gene markers
538 independently, namely BUSCO genes (Benchmarking Universal Single-Copy
539 Orthologs)[73] and CSSC genes (Conserved Single-copy Synteny Clusters).

540 The criterion for characterizing of CSSC genes was: median number of nodes across
541 species < 2, present in ≥ 90% genomes, and presence within Poaceae, monocots
542 (except Poaceae), Asterids, Rosales, Brassicaceae, and Fabaceae must ≥ 50%.
543 BUSCO analysis (v3.0, embryophyta_odb9, with 1440 profiles) identified a total of
544 1438 conserved single-copy genes from the 123 angiosperm genome sequences,
545 compared to 883 identified as CSSC. Multiple sequence alignments were performed
546 using MAFFT (version 7.187)[74] . Two rounds of alignment trimming and filtering were
547 conducted by trimAl[75]. First, the alignments were trimmed through heuristic selection
548 of the automatic method (-automated1). Second, sequences with less than 50%
549 residues that pass the minimum residue overlap threshold (overlap score: 0.5) were
550 removed (-resoverlap 0.5 -seqoverlap 50). Alignment concatenation was conducted
551 by catfasta2phyml (https://github.com/nylander/catfasta2phyml). The length of our
552 concatenated gene sequence alignments of BUSCO and CSSC genes were 591,196
553 and 341,431 amino acids, respectively.

554 Maximum-likelihood analyses were conducted using IQ-TREE[76]. For sequence-
555 based tree construction, we used the JTT+R model for protein alignments, both for
556 the construction of trees based on single alignments, as well as for the concatenated
557 sequence alignments. For all trees, we performed bootstrap analysis with 1000
558 bootstrap replicates (SH-aLRT and UFBoot (-alrt 1000 -bb 1000)).

559 ASTRAL-Pro[77] was used for the tree summary approach based on multi-species
560 coalescence to estimate the species trees from 1438 BUSCO gene trees and 883
561 CSSC gene trees, respectively. ASTRAL-Pro is the latest update of ASTRAL, which
562 can now account for multi-copy trees.

563 For the MRL supertree analysis, 1438 BUSCO gene trees and 883 CSSC gene trees
564 were used as two independent data sets. For each tree of the two sets, branch
565 support from bootstrap analysis has been included as a standard output of IQ-TREE
566 (suffixed as '*.phy.splits.nex'). We then encoded all splits (bipartitions) with ≥ 85%

UFBoot support of all the trees into the data matrix. Thus, each column (matrix element) in the matrix represents a well-supported bipartition. Coding is similar to the "Baum-Ragan" coding method (0, 1, ?)[78], but without question marks because '?' was originally designed for missing taxa as trees were multi-sourced. The dimensions of the matrices are 123 × 139,538 for the BUSCO gene trees, and 123 × 102,617 for the CSSC gene trees. We used the same binary model (MK+R+FO) in IQ-TREE, and parameter settings were as the one described earlier in Syn-MRL supercluster analysis.

To assess whether alternative sister-group relationships of certain plant clades could be statistically rejected given the synteny matrix, we performed approximate unbiased (AU) tests[34], as implemented in IQ-TREE[76], under the 'MK+R+FO' model, with 10,000 replicates.

Acknowledgements

Z. L. is supported by a postdoctoral fellowship from the Special Research Fund of Ghent University (BOFPDO2018001701). A. Z. is supported by a PhD fellowship of the Research Foundation Flanders (FWO). Yves Van de Peer acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 833522).

Supplemental documents

Supplementary Fig. 1 Representative sequence-based methods for reconstructing species trees used in this study, we used superaligment concatenation/supermatrix, multispecies coalescent, and MRL-based supertree methods for two sets of whole-genome-derived markers (BUSCO and CSSC) (see text for details). First, multiple sequence alignments (MSA) were first build for each of the whole-genome-derived markers (BUSCO and CSSC) and used as input for the Supermatrix method inferring species trees based on the concatenation of gene alignments. Second, independent gene trees can be inferred for each alignment, after which a species tree can be inferred from the set of obtained gene trees under the multispecies coalescent model, or by using a supertree method. In the latter case, we used a MRL-based method (see text for details). For example, Clade 1+2 is well-supported (BS ≥ 85%) in Tree 1, then this branching order (phylogenetic grouping) is coded as the first column of the matrix, similarly, Clade 1+2+3 of Tree 1 is coded as the second column. Leaves from well-supported nodes of all trees construct a binary branch matrix, which is then used for phylogenetic analysis by maximum likelihood.

Supplementary Fig. 2 Concatenation-BUSCO tree.

Supplementary Fig. 3 Concatenation-CSSC tree.

Supplementary Fig. 4 MRL-BUSCO tree.

Supplementary Fig. 5 MRL-CSSC tree.

Supplementary Fig. 6 ASTRAL-BUSCO tree.

Supplementary Fig. 7 ASTRAL-CSSC tree.

Supplementary Fig. 8 Comparison of synteny tree and the SA tree. Both trees are rooted by *Amborella*, and three main clades, i.e. superrosids, superasterids, and monocots are shaded in light-red, light-purple, and light-green, respectively. Eight differences are indicated by indexed black dots. Branches are not drawn to scale. Ultrafast bootstrapping values (see text for details) were marked for nodes with less than 100% support.

Supplementary Fig. 9 Comparison of phylogenetic relationships between the phylogeny of Angiosperm Phylogeny Group IV and 1KP.

Supplementary Fig. 10 Comparison of phylogenetic relationships between APG IV and the synteny tree.

Supplementary Fig. 11 Comparison of phylogenetic relationships between the phylogeny of 1KP study (angiosperms part) and the synteny tree.

Supplementary Fig. 12 Approximate unbiased (AU) test for alternative topologies, with resulting *p* values indicated under the trees. Alternative (tested) topologies include (**a**) magnoliids as sister to eudicots, which is the only scenario tested that cannot be significantly rejected (see text for details). (**b**) magnoliids as sister to both monocots and eudicots, (**c**) Vitales as early-diverging rosids, (**d**) Saxifragales as early-diverging rosids, (**e**) Vitales and Saxifragales as early-diverging rosids, (**f**) Malpighiales as early-diverging Fabids, and (**g**) Malpighiales as early-diverging Fabids plus Myrtales as early-diverging Malvids.

Supplementary Fig. 13 Syn-MRL tree based on the synteny matrix without 1107 (Figure 4a) specific signals.

Supplementary Fig. 14 Syn-MRL tree based on the synteny matrix without *Triticum turgidum* (wheat).

Supplementary Fig. 15 Syn-MRL tree based on the synteny matrix without *Hordeum vulgare* (barley).

Supplementary Fig. 16 Syn-MRL tree based on the synteny matrix without *Brachypodium distachyon* (Brachypodium).

Supplementary Fig. 17 Syn-MRL tree based on the synteny matrix without *Triticum turgidum* (wheat) and *Hordeum vulgare* (barley).

Supplementary Table S1 List of genomes used in this study.

Supplementary Table S2 Phylogenomic profiling rearranged by *Cinnamomum kanehirae* chromosomes and gene orders.

Data availability

Datasets used in this study are available at DataVerse (https://doi.org/10.7910/DVN/7ZZWIH). This includes all annotated protein sequences in FASTA format of each genome, entire synteny network database

(edgelist), network clustering result, trimmed alignments and corresponding phylogenetic trees of BUSCO and CSSC genes, bipartitions with support values for each tree, and binary data matrices. Related codes and software parameters are available at Github (https://github.com/zhaotao1987/Syn-MRL).

Author contributions

Y.V.d.P., M.E.S, and J.X. conceived the idea. T.Z. and Y.V.d.P. designed the study. T.Z. and S.K. performed the analysis. T.Z., J.X., Y.V.d.P, Z.L., M.E.S., and A.Z. analyzed data. T.Z. and Y.V.d.P. wrote the manuscript. All authors discussed the results and commented on the manuscript.

References

1. Van de Peer, Y. Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet* **5**, 752-763 (2004).
2. Bowers, J.E., Chapman, B.A., Rong, J. & Paterson, A.H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433-8 (2003).
3. Pevzner, P. & Tesler, G. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res* **13**, 37-45 (2003).
4. Dewey, C.N. Positional orthology: putting genomic evolutionary relationships into context. *Brief Bioinform* **12**, 401-412 (2011).
5. Koonin, E.V. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**, 309-338 (2005).
6. Van Bel, M. *et al.* Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol* **158**, 590-600 (2012).
7. Zhao, T. *et al.* Phylogenomic synteny network analysis of MADS-box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation. *Plant Cell* **29**, 1278-1292 (2017).
8. Sampedro, J., Lee, Y., Carey, R.E., Depamphilis, C.W. & Cosgrove, D.J. Use of genomic history to improve phylogeny and understanding of births and deaths in a gene family. *Plant J* **44**, 409-419 (2005).
9. Drillon, G., Champeimont, R., Oteri, F., Fischer, G. & Carbone, A. Phylogenetic reconstruction based on synteny block and gene adjacencies. *Mol Biol Evol* (2020).
10. Luo, H. *et al.* Phylogenetic analysis of genome rearrangements among five mammalian orders. *Mol Phylogen Evol* **65**, 871-882 (2012).
11. Feng, B. *et al.* Reconstructing yeasts phylogenies and ancestors from whole genome data. *Sci Rep* **7**, 1-12 (2017).
12. Watterson, G.A., Ewens, W.J., Hall, T.E. & Morgan, A. The chromosome inversion problem. *J Theor Biol* **99**, 1-7 (1982).
13. Blanchette, M., Bourque, G. & Sankoff, D. Breakpoint phylogenies. *Genome informatics* **8**, 25-34 (1997).
14. Moret, B.M., Bader, D.A., Wyman, S., Warnow, T. & Yan, M. A new implementation and detailed study of breakpoint analysis. in *Biocomputing 2001* 583-594 (World Scientific, 2000).
15. Bourque, G. & Pevzner, P.A. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Res* **12**, 26-36 (2002).

16. Luo, H., Shi, J., Arndt, W., Tang, J. & Friedman, R. Gene order phylogeny of the genus Prochlorococcus. *PloS one* **3**, e3837-e3837 (2008).
17. Murat, F., Peer, Y.V.d. & Salse, J. Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biol Evol* **4**, 917-928 (2012).
18. Zhao, T. & Schranz, M.E. Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proc Natl Acad Sci U S A* **116**, 2165-2174 (2019).
19. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat Rev Genet* **18**, 411 (2017).
20. Leebens-Mack, J.H. *et al.* One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679-685 (2019).
21. Payseur, B.A. & Rieseberg, L.H. A genomic perspective on hybridization and speciation. *Mol Ecol* **25**, 2337-60 (2016).
22. Fedoroff, N. Transposons and genome evolution in plants. *Proc Natl Acad Sci USA* **97**, 7002-7007 (2000).
23. Freeling, M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu Rev Plant Biol* **60**, 433-453 (2009).
24. Bowles, A.M., Bechtold, U. & Paps, J. The origin of land plants is rooted in two bursts of genomic novelty. *Curr Biol* (2020).
25. Chase, M.W. *et al.* An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot J Linn Soc* **181**, 1-20 (2016).
26. Zeng, L. *et al.* Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytol* **214**, 1338-1354 (2017).
27. Moore, M.J., Soltis, P.S., Bell, C.D., Burleigh, J.G. & Soltis, D.E. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci USA* **107**, 4623 (2010).
28. Worberg, A. *et al.* Phylogeny of basal eudicots: Insights from non-coding and rapidly evolving DNA. *Org Divers Evol* **7**, 55-77 (2007).
29. Soltis, D.E. *et al.* Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Bot J Linn Soc* **133**, 381-461 (2000).
30. Soltis, D.E. & Soltis, P.S. Nuclear genomes of two magnoliids. *Nat Plants* **5**, 6 (2019).
31. Chaw, S.M. *et al.* Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat Plants* **5**, 63-73 (2019).

32. Chen, J. *et al. Liriodendron* genome sheds light on angiosperm phylogeny and species–pair differentiation. *Nat Plants* **5**, 18 (2019).
33. Rendón-Anaya, M. *et al.* The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *Proc Natl Acad Sci USA* **116**, 17081-17089 (2019).
34. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst Biol* **51**, 492-508 (2002).
35. Boore, J.L. The use of genome-level characters for phylogenetic reconstruction. *Trends Ecol Evol* **21**, 439-46 (2006).
36. Rokas, A. & Holland, P.W. Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol* **15**, 454-459 (2000).
37. Nadeau, J.H. & Taylor, B.A. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci USA* **81**, 814-818 (1984).
38. Sankoff, D. *et al.* Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc Natl Acad Sci USA* **89**, 6575-6579 (1992).
39. Sankoff, D. & Nadeau, J.H. Chromosome rearrangements in evolution: From gene order to genome sequence and back. *Proc Natl Acad Sci USA* **100**, 11188-11189 (2003).
40. Nguyen, N., Mirarab, S. & Warnow, T. MRL and SuperFine+MRL: new supertree methods. *Algorithms Mol Biol* **7**, 3 (2012).
41. Mirarab, S., Bayzid, M.S. & Warnow, T. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst Biol* **65**, 366-80 (2016).
42. Lewis, P.O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst Biol* **50**, 913-25 (2001).
43. Soltis, D.E. & Soltis, P.S. Nuclear genomes of two magnoliids. *Nat Plants* **5**, 6-7 (2019).
44. Hu, L. *et al.* The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nat Commun* **10**, 1-11 (2019).
45. Soltis, P.S., Soltis, D.E. & Chase, M.W. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* **402**, 402-4 (1999).
46. Sun, G. *et al.* Archaefructaceae, a new basal angiosperm family. *Science* **296**, 899-904 (2002).
47. Endress, P.K. & Doyle, J.A. Reconstructing the ancestral angiosperm flower and its initial specializations. *Am J Bot* **96**, 22-66 (2009).

763 48. Zhang, N., Zeng, L., Shan, H. & Ma, H. Highly conserved low-copy nuclear genes as
764 effective markers for phylogenetic analyses in angiosperms. *New Phytol* **195**, 923-
765 937 (2012).
766 49. Sun, M. *et al.* Deep phylogenetic incongruence in the angiosperm clade Rosidae. *Mol*
767 *Phylogen Evol* **83**, 156-166 (2015).
768 50. Qiu, Y.L. *et al.* Angiosperm phylogeny inferred from sequences of four mitochondrial
769 genes. *J Syst Evol* **48**, 391-425 (2010).
770 51. Liu, Y., Cox, C.J., Wang, W. & Goffinet, B. Mitochondrial phylogenomics of early land
771 plants: mitigating the effects of saturation, compositional heterogeneity, and codon-
772 usage bias. *Syst Biol* **63**, 862-878 (2014).
773 52. Finet, C., Timme, R.E., Delwiche, C.F. & Marlétaz, F. Multigene phylogeny of the
774 green lineage reveals the origin and diversification of land plants. *Curr Biol* **20**, 2217-
775 2222 (2010).
776 53. Wickett, N.J. *et al.* Phylotranscriptomic analysis of the origin and early diversification
777 of land plants. *Proc Natl Acad Sci U S A* **111**, E4859-68 (2014).
778 54. Zeng, L. *et al.* Resolution of deep angiosperm phylogeny using conserved nuclear
779 genes and estimates of early divergence times. *Nat Commun* **5**, 4956 (2014).
780 55. Aköz, G. & Nordborg, M. The *Aquilegia* genome reveals a hybrid origin of core
781 eudicots. *Genome Biol* **20**, 256 (2019).
782 56. Forsythe, E.S., Nelson, A.D. & Beilstein, M.A. Biased gene retention in the face of
783 massive nuclear introgression obscures species relationships. *bioRxiv*, 197087
784 (2018).
785 57. Walden, N., Nguyen, T.-P., Mandáková, T., Lysak, M.A. & Schranz, M.E. Genomic
786 blocks in *Aethionema arabicum* support *Arabideae* as next diverging clade in
787 Brassicaceae. *Front Plant Sci* **11**, 719 (2020).
788 58. International Wheat Genome Sequencing, C. *et al.* Shifting the limits in wheat
789 research and breeding using a fully annotated reference genome. *Science* **361**(2018).
790 59. Sage, R.F., Sage, T.L. & Kocacinar, F. Photorespiration and the evolution of C_4
791 photosynthesis. *Annu Rev Plant Biol* **63**, 19-47 (2012).
792 60. Dasmahapatra, K.K. *et al.* Butterfly genome reveals promiscuous exchange of
793 mimicry adaptations among species. *Nature* **487**, 94 (2012).
794 61. Fontaine, M.C. *et al.* Mosquito genomics. Extensive introgression in a malaria vector
795 species complex revealed by phylogenomics. *Science* **347**, 1258524 (2015).
796 62. Edelman, N.B. *et al.* Genomic architecture and introgression shape a butterfly
797 radiation. *Science* **366**, 594-599 (2019).

63. Shan, X. *et al.* Mobilization of the active MITE transposons *mPing* and *Pong* in rice by introgression from wild rice (*Zizania latifolia* Griseb.). *Mol Biol Evol* **22**, 976-990 (2005).
64. Wang, N. *et al.* Transpositional reactivation of the *Dart* transposon family in rice lines derived from introgressive hybridization with *Zizania latifolia*. *BMC Plant Biol* **10**, 190 (2010).
65. Serrato-Capuchina, A. & Matute, D.R. The role of transposable elements in speciation. *Genes* **9**, 254 (2018).
66. Folk, R.A., Soltis, P.S., Soltis, D.E. & Guralnick, R. New prospects in the detection and comparative analysis of hybridization in the tree of life. *Am J Bot* **105**, 364-375 (2018).
67. Buchfink, B., Xie, C. & Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60 (2015).
68. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**, e49-e49 (2012).
69. Rosvall, M. & Bergstrom, C.T. Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci USA* **105**, 1118-1123 (2008).
70. Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 993-1005 (1995).
71. Soubrier, J. *et al.* The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Mol Biol Evol* **29**, 3345-3358 (2012).
72. Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q. & Vinh, L.S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol* **35**, 518-522 (2017).
73. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
74. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).
75. Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
76. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274 (2014).
77. Zhang, C., Scornavacca, C., Molloy, E. & Mirarab, S. ASTRAL-Pro: quartet-based species tree inference despite paralogy. *bioRxiv* (2019).

78. Baum, B.R. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* **41**, 3-10 (1992).

Figures

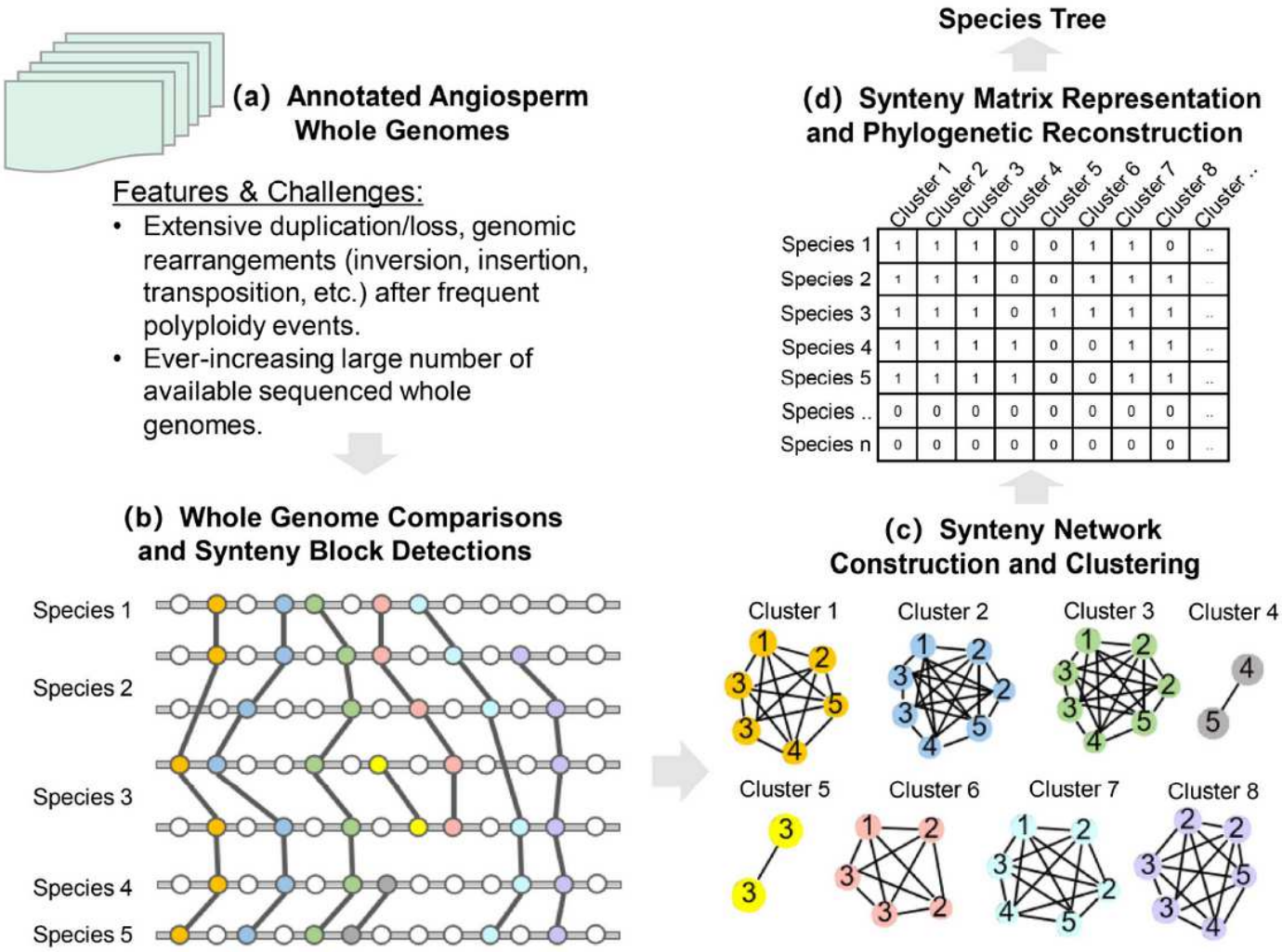


Figure 1

Whole-genome microsynteny-based species tree inference. (a) Whole genome datasets with all predicted genes are used for phylogeny reconstruction. (b) The synteny network approach first conducts all pairwise reciprocal genome comparisons, followed by synteny block detection. All syntenic blocks constitute the synteny network database (see Methods for details). (c) We analyzed all synteny clusters after clustering the entire network database. Synteny clusters vary in size and node compositions. Shared genomic rearrangements are reflected by cluster compositions. Specific anchor pairs shared by a lineage/species form specific clusters (e.g. Clusters 4-6). We account for the presence or absence of the same recurring anchors for multiple blocks derived from whole-genome or segmental duplications (e.g. for Species 2 and 3 in Clusters 2, 3, 5, and 8). (d) The phylogenomic profiling of all clusters constructs a binary matrix, where rows represent species and columns represent clusters. The synteny matrix comprehensively represents phylogenomic gene order dynamics. It transforms the concept of synteny comparisons from analyzing massive parallel coordinates plots into analyzing profiles of individual clusters/networks. Each cluster stands for a shared homologous ‘context’. For example, genes (driven by

transposable elements) can be transposed as insertions into new contexts or can be lost from the original context (e.g. genes in Clusters 4-6). As long as such transpositions are shared by different genomes (e.g. genes in Clusters 4 and 6) or within the same genome because of whole genome duplication (e.g. genes in Cluster 5), specific clusters will emerge and corresponding signals will be added to the matrix. This synteny matrix is used as the input for species tree inference by maximum likelihood (referred to as Syn-MRL).

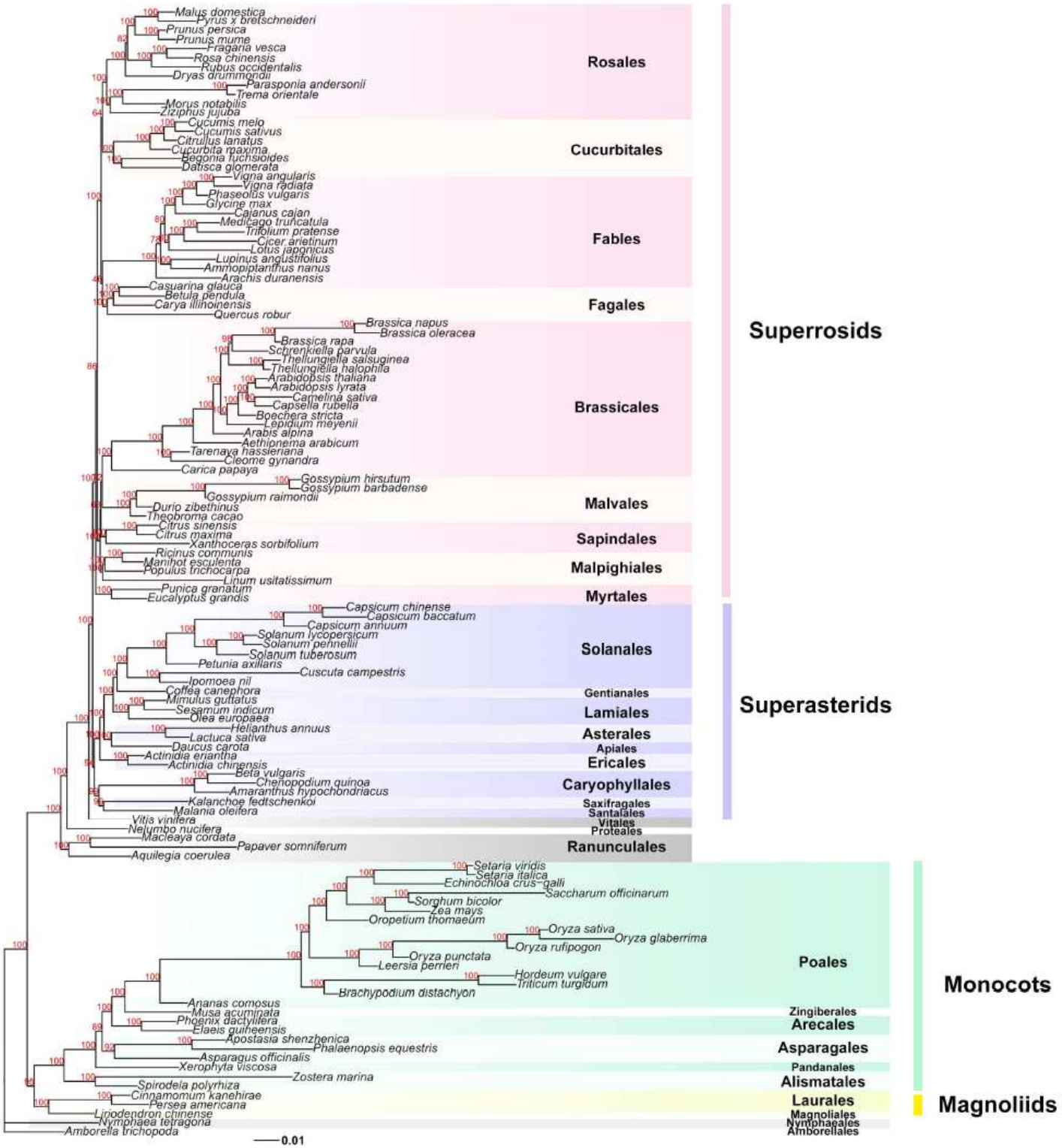


Figure 2

Maximum likelihood (ML) tree for 123 fully-sequenced flowering plant genomes based on the microsynteny approach. The tree is rooted by Amborella, and four main clades, i.e. superrosids, superasterids, monocots, and magnoliids are shaded in light-red, light-purple, light-green, and light-yellow, respectively. Ultrafast bootstrapping values are denoted for all the nodes. Names for the different plant orders follow the APG IV classification25.

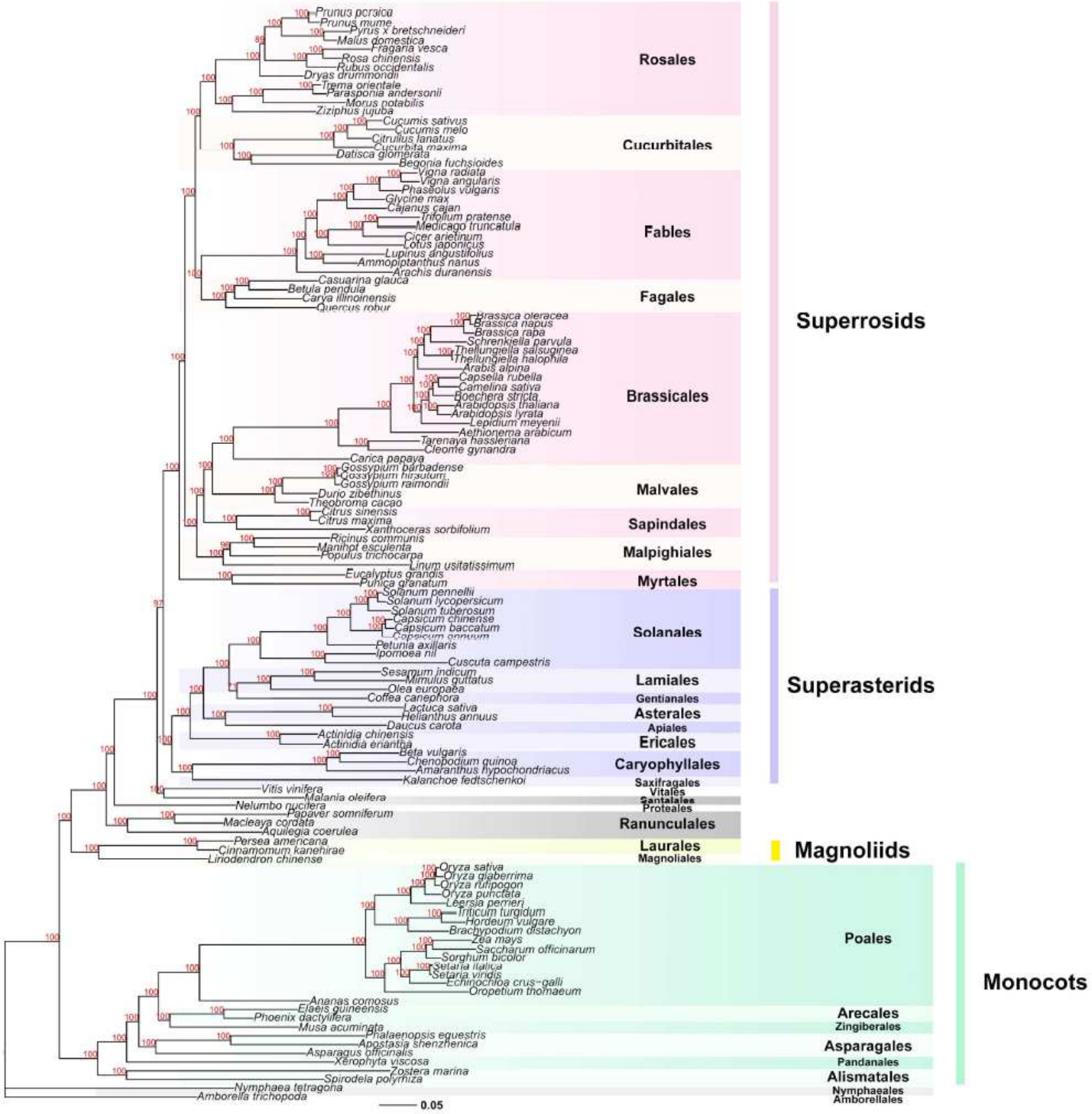


Figure 3

Maximum likelihood (ML) tree based on the concatenation of protein alignments of BUSCO genes. This tree was used as the representative SA (sequence-alignment based) tree.

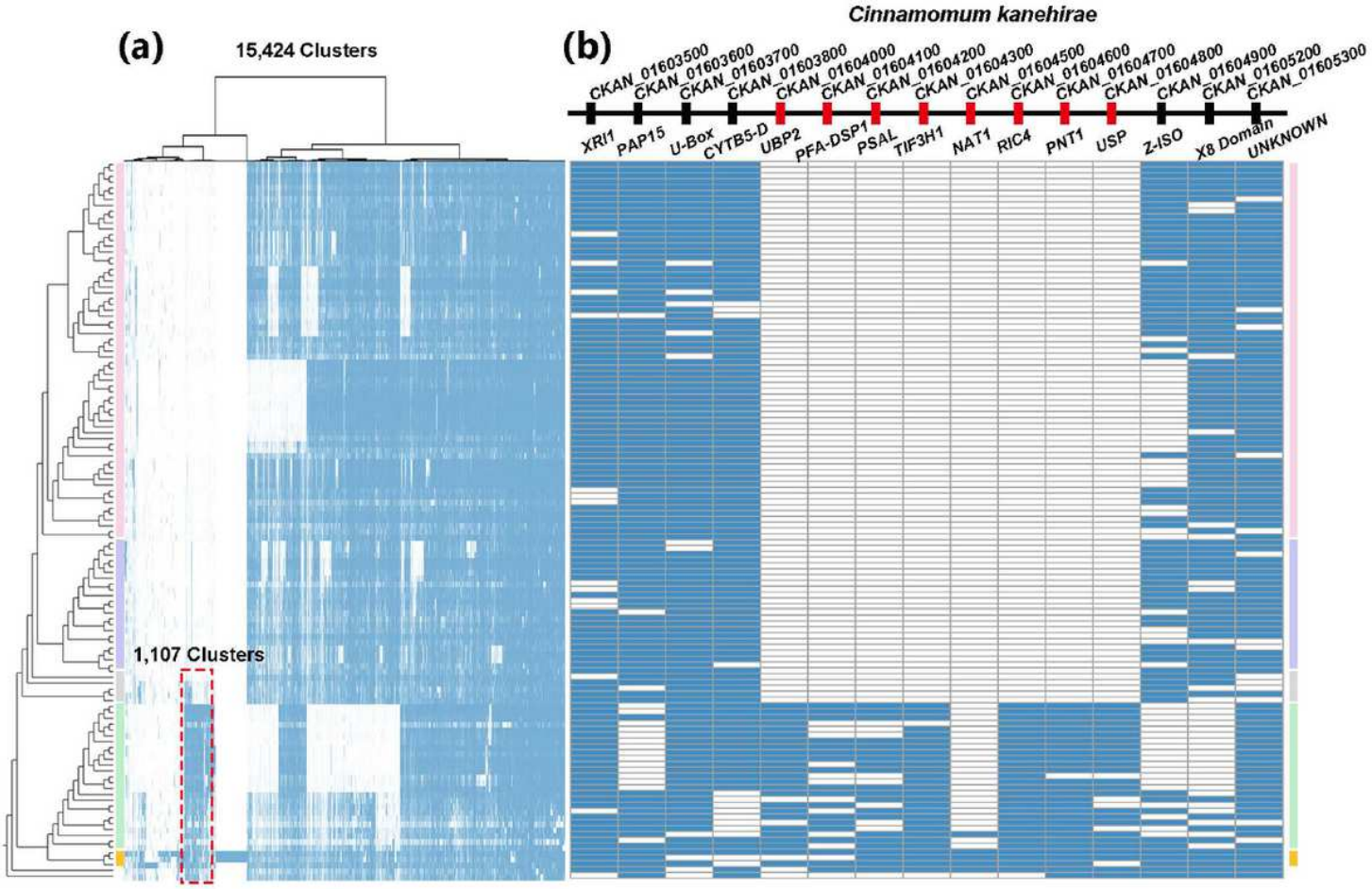


Figure 4

Magnoliids-associated signals and a representative example of phylogenetically informative microsynteny. (a) Hierarchical clustering (ward.D) of 15,424 magnoliids-associate cluster profiles based on jaccard distance. On the far left, the synteny-based species tree is displayed (same as Fig. 1). Superrosids, superasterids, early-diverging eudicots, monocots, and magnoliids are shaded in light-red, light-purple, light-grey, light-green, and light-yellow, respectively. 1,107 clusters supporting a grouping of magnoliids and monocots (Supplemental Table S2). (b) One example from all supporting signals. A fifteen-gene context in the genome of Cinnamomum kanehirae (a magnoliid) shows eight neighboring genes (highlighted in orange) only present in magnoliids and monocot genomes, while the flanking genes (colored blue) are generally conserved angiosperm-wide.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf

- TableS1.xlsx
- TableS2.xlsx