Comprehensive study of subfamily evolution through sequences of conserved nuclear genes and whole chloroplast genomes extracted from the same raw read data

Eranga Pawani Witharana[1,2,3, *], Takaya Iwasaki[4], Myat Htoo San[2,5], Nadeeka U. Jayawardana[1,6], Nobuhiro Kotoda[3,5,7], Masashi Yamamoto[5,8], Yukio Nagano[2,3,5, *]

[1]*Faculty of Agriculture, University of Peradeniya, Peradeniya, Sri Lanka*

[2]*Analytical Research Center for Experimental Sciences, Saga University, Saga, Japan*

[3]*Graduate School of Advanced Health Science, Saga University, Saga, Japan*

[4]*Faculty of Science, Ochanomizu University, Tokyo, Japan*

[5]*The United Graduate School of Agricultural Sciences, Kagoshima University, Kagoshima, Japan*

[6]*Applied BioSciences,205B, Culloden Road, Macquarie University, Sydney, NSW, Australia*

[7]*Faculty of Agriculture, Saga University, Saga, Japan*

[8]*Faculty of Agriculture, Kagoshima University, Kagoshima, Japan*

***Corresponding author:**

E-mail: erangav8@gmail.com

E-mail: nagano@cc.saga-u.ac.jp

Abstract

The chloroplast (cp) genome is a widely used tool for exploring plant evolutionary relationships, yet its effectiveness in fully resolving these relationships remains uncertain. Integrating cp genome data with nuclear DNA information offers a more comprehensive view but often requires separate datasets. Nuclear DNA analyses are hindered by the need for high-quality sequences and high costs. In response, we employed the same raw read sequencing data to construct cp genome-based trees and nuclear DNA phylogenetic trees using Read2Tree, a cost-efficient method for extracting conserved nuclear gene sequences from raw read data, focusing on the Aurantioideae subfamily, which includes *Citrus* and its relatives. The resulting nuclear DNA trees were consistent with existing nuclear evolutionary relationships derived from high-throughput sequencing, but diverged from cp genome-based trees. To elucidate the underlying complex evolutionary processes causing these discordances, we implemented an integrative workflow that utilized multiple alignments of each gene generated by Read2Tree, in conjunction with other phylogenomic methods. Our analysis revealed that incomplete lineage sorting predominantly drives these discordances, while introgression and ancient introgression also contribute to topological discrepancies within certain clades. This study underscores the cost-effectiveness of using the same raw sequencing data for both cp and nuclear DNA analyses in understanding plant evolutionary relationships.

Keywords: Chloroplast genome sequences, Read2Tree, Aurantioideae subfamily, Phylogenetic discordance, Incomplete lineage sorting, Introgression

Introduction

Phylogenetic analysis plays a crucial role in comprehending organism evolution and taxonomy. Maximizing the utilization of variable sites during the construction of a phylogenetic tree enhances accuracy[1,2]. The advent of high-throughput sequencing has vastly increased the number of available variable sites. In recent years, chloroplast (cp) genome sequences have gained popularity in plant molecular systematics[3,4,5]. Characterized by typical maternal inheritance, stable gene content, low recombination, and manageable size, the cp genome has emerged as a cost-effective tool for phylogenetic studies[6,7,8]. Assembling the complete cp genome, typically ranging from 120 to 170 kbp, from as little as 1 Gbp of raw sequencing data enables detailed analysis of evolutionary relationships among plant species[9,10]. However, sole reliance on cp genome sequences in plant phylogenetics has its limitations. These include potential instances of biparental or paternal inheritance of cpDNA in some species[6,11], and the phenomenon of 'chloroplast capture,' where cpDNA is transferred between different lineages through hybridization[12]. Additionally, because cpDNA represents a single genetic lineage inherited separately from the nuclear genome, it is subject to independent genetic drift and chloroplast capture that do not affect the nuclear genome in the same way. These factors can lead to differences in the phylogenetic signals from cpDNA and nuclear DNA, resulting in cytonuclear discordance[13,14,15]. Given these limitations, integrating cpDNA data with other molecular markers, particularly nuclear DNA, can provide a more comprehensive approach to phylogenetic reconstruction.

Nuclear DNA markers, owing to their biparental inheritance and greater genetic diversity, offer a broader and more accurate understanding of evolutionary relationships compared to cpDNA[16,17]. However, the phylogenetic analysis of many plants encounters challenges due to the limited availability of high-quality nuclear sequences[18,19]. This scarcity results from the high costs and technical difficulties associated with sequencing and annotation[20]. To address these limitations, transcriptomic data are increasingly being employed in phylogenetic reconstructions[21,22]. However, variations in gene expression among species, dependent on tissue type, can affect the accuracy of these reconstructions, potentially leading to skewed interpretations of phylogenetic signals[23,24].

The Read2Tree tool may offer a promising alternative for overcoming the limitations associated with nuclear molecular markers in reconstructing phylogenies[25]. This method allows for the direct extraction of conserved nuclear gene sequences from raw sequencing reads, effectively bypassing the conventional steps such as genome assembly and annotation. Additionally, the potential to employ the same short-read sequence data for both whole cp genome analysis and Read2Tree offers a promising approach for phylogenetic reconstructions. Read2Tree produces two primary outputs: concatenated alignment data for tree inference and multiple alignments for each gene across the analyzed species. The concatenated alignment data is a primary focus of Read2Tree's publication, which has been employed in reconstructing the phylogeny of yeast and various coronaviruses[25]. Its effectiveness was confirmed by validating the reconstructed yeast phylogeny through comparison with results from a previous study based on whole genome assemblies[25]. Additionally, the versatility of Read2Tree was demonstrated by the alignment of the reconstructed coronavirus phylogeny with lineage classifications from UniProt reference proteomes[25]. However, only one study has applied Read2Tree to plant phylogenetics[26], and it did not address the validation of Read2Tree results as comprehensively as the original publication[25]. The application of Read2Tree in plant phylogenetics is particularly significant because plants are more prone to interspecies hybridization compared to animals[27,28]. This frequent hybridization can lead to complex evolutionary histories, including phenomena like chloroplast capture, which are challenging to resolve using chloroplast DNA alone. Therefore, employing phylogenetic analyses based on many nuclear genes becomes crucial for accurately reconstructing plant evolutionary relationships. Given this context, the effectiveness of Read2Tree in plant phylogenetics remains relatively underexplored and requires further validation through comprehensive comparisons with existing nuclear-based phylogenetic findings.

The second output type, multiple alignments for each gene, may hold significant potential for analyzing detailed genetic differentiation processes as a downstream application of Read2Tree. However, attempts to use Read2Tree for this purpose have not been comprehensively explored in previous studies [25,26]. The genetic differentiation processes such as incomplete lineage sorting (ILS), introgression, and ancient introgression often generate discordances between gene trees and species

trees, driving significant interest in understanding the underlying sources of these discrepancies[29,30,31,32,33,34]. Consequently, constructing gene trees from multiple gene alignments can facilitate species tree construction while also identifying conflicting phylogenetic signals. Integrating Read2Tree with other phylogenomic methods can address these processes more effectively. Additionally, this integration can pinpoint their contributions to cytonuclear discordance.

This study aims to elucidate and harness the potential of Read2Tree in advancing phylogenetic reconstructions and discordance analysis, focusing on the phylogenetic relationships within Aurantioideae, a subfamily of Rutaceae that encompasses *Citrus* and its relatives. We selected Aurantioideae as the ideal model for this investigation for two primary reasons. First, our previous research successfully utilized nuclear DNA sequences for this subfamily based on RAD-Seq data[35]. While RAD-Seq focuses on nuclear DNA near specific restriction sites, Read2Tree specializes in analyzing a broader set of conserved nuclear gene sequences. This allows for a more comprehensive evaluation of nuclear gene sequences. By comparing the nuclear gene trees generated by Read2Tree with RAD-Seq results, we can more effectively evaluate the phylogenetic relationships and identify potential sources of discordance within the subfamily. Second, the genus *Citrus*, along with other members of Aurantioideae, holds global importance due to its nutritional value, high vitamin C content, and widespread culinary use. Understanding the phylogenetic relationships within this group has implications not only for evolutionary biology but also for agriculture and conservation efforts. Many Aurantioideae species are valued for their edible and medicinal properties across various regions, making accurate phylogenetic reconstructions crucial for biodiversity preservation and resource management[36,37].

In this study, we employed the Read2Tree method to conduct nuclear-based phylogenetic reconstructions of approximately 40 plant species within the subfamily Aurantioideae. We integrated the nuclear phylogenies with whole cp genomes analyses to explore evolutionary relationships more precisely and to investigate the potential causes of cytonuclear discordance. By comparing the nuclear phylogenetic trees generated by Read2Tree with those obtained from previous RAD-Seq analyses, we aimed to validate the robustness and utility in plant phylogenetics. Additionally, we

developed a comprehensive workflow that integrates Read2Tree with other phylogenomic methods to analyze both species-gene tree discordance and cytonuclear discordance. This workflow enabled us to assess whether the observed discordances were due to factors such as ILS, introgression, or ancient introgression followed by ILS. By identifying the sources of these discordances, we aimed to gain insights into the complex evolutionary history of Aurantioideae and the potential mechanisms driving cytonuclear discordance in plant phylogenetics.

Results

The phylogenetic tree constructed using Read2Tree demonstrates strong concordance with the tree derived from RAD-Seq data

The samples utilized in this investigation are detailed in Supplementary Table S1. Initially, Read2Tree facilitated the retrieval of sequences from various conserved nuclear genes. To establish the root of the phylogenetic tree within the subfamily Aurantioideae, a Maximum Likelihood (ML) tree was generated for 43 species of the Rutaceae family (Supplementary Fig. S1). Subsequently, we focused on 39 Aurantioideae species to generate a ML phylogenetic tree using the conserved nuclear gene sequences (Fig.1), with rooting determined from the Rutaceae analysis. The ML phylogenetic trees largely aligned with findings from a prior RAD-Seq based study[35], with noteworthy exceptions observed for *Micromelum minutum* and *Murraya koenigii*. Their phylogenetic positions, previously uncertain in the prior study[35], remained unresolved in our current analysis, as evident from the comparison between Fig.1 and Supplementary Fig. S1. An in-depth analysis within the genus *Citrus* was not conducted using Read2Tree due to the presence of recombination events occurring during hybridization[38].

Phylogenetic analysis employed whole cp genome sequences revealing disparities when compared with nuclear gene-based trees

Initially, a cp phylogenetic tree encompassing 59 members of the Rutaceae family was constructed, to determine the root position for subsequent analyses (Supplementary Fig. S2). Subsequently, cp phylogenetic trees were developed based on DNA sequences for the same 39 members of the Aurantioideae subfamily used in Read2Tree, utilizing ML and Bayesian Inference (BI) methods (Supplementary Fig. S3 and Supplementary Fig. S4, respectively). The inclusion of identical taxa in both the whole cp genome and nuclear analyses was a deliberate approach to minimize discrepancies arising from taxon sampling, thereby reducing variations associated with

lineage-specific evolutionary histories. The topologies of these cp phylogenetic trees remained consistent across the two analyses.

Comparison of the topologies of the two phylogenetic trees, one from whole cp genome sequences and the other from the concatenated nuclear gene tree analysis (Fig. 2A), revealed two primary groups within the Aurantioideae subfamily, referred to as group A and group B. However, this comparison detected cytonuclear discordances within each group (Fig. 2A). Inconsistencies within the *Citrus*, *Atalantia*, and *Clausena* genera were noted due to interspecific hybridization within the genus. Moreover, marked differences were observed in the intergeneric relationships. For instance, the cpDNA-based analysis positioned the *Limonia*/*Feroniella* subgroup as one of the closest outgroups to *Citrus*, contrary to the nuclear DNA analysis, which identified this subgroup as the nearest outgroup to a clade comprising *Citrus*, *Atalantia*, *Citropsis*, and *Naringi*. Furthermore, the cpDNA analysis grouped *Swinglea* with *Citropsis* and *Naringi*, while nuclear DNA analysis aligned *Swinglea* with the *Limonia*/*Feroniella* subgroup. Discrepancies between the two analyses were also evident in the phylogenetic structures formed by *Paramignya*, *Triphasia*, and *Aegle*/*Afraegle*. The phylogenetic positions of *Micromelum minutum* and *Murraya koenigii* remained ambiguous.

Integrating Read2Tree for species tree estimation and revealing phylogenetic discordances

The detection of cytonuclear discordances strongly suggests the presence of broader phylogenetic discordances. By integrating Read2Tree with other phylogenomic methods, these discordances were determined. Orthologous groups (OGs) consisting of genes or DNA sequences in different species that diverged from a common ancestral gene through speciation, typically maintain similar functions across species, making them critical for studying evolutionary relationships[39]. In this analysis, Read2Tree generated multiple sequence alignments for each OG, which were used for constructing each gene tree. Specifically, alignments from 579 OGs with a ratio of N characters to total nucleotides below 10%, as identified through Read2Tree analyses, were selected for gene trees

constructions. The gene trees were subsequently used to infer the ASTRAL species tree, as shown in Fig. 2B (left panel), Fig. 3A (right panel), and detailed in Supplementary Table S3. The resulting species tree has a normalized quartet score (QT) of 0.7, indicating that 70% of the quartets in the gene trees are concordant with the species tree. This score reflects a high level of overall concordance. Additionally, the grouping into groups A and B in the ASTRAL species tree aligns with those in the cp phylogeny and concatenated nuclear phylogeny as demonstrated in the Fig.2B and Fig.3A.

When comparing the species tree to the concatenated nuclear tree constructed using Read2Tree, notable discrepancies are observed within the *Citrus* genus and in the intergeneric relationships involving *Swinglea* and *Murraya koenigii* (Fig. 3A). In the species tree, *Swinglea* is placed as an outgroup to *Citrus*, *Atalantia*, and *Citropsis/Naringi*, whereas in the concatenated tree, *Swinglea* clusters with *Limonia* and *Feroniella*. Additionally, *Murraya koenigii* acts as a near outgroup exclusively to *Clausena* species in the species tree, while in the concatenated tree, it serves as an outgroup to both *Clausena* and *Glycosmis* species. This ambiguous positioning of *Murraya koenigii* mirrors the unresolved status observed in previous RAD-Seq studies as well as in our Read2Tree and cp genome trees (Fig.2A).

PhyParts conflict analysis was performed to investigate the complex discordances between gene trees and the species tree, with *Boenninghausenia albiflora* used as the outgroup (Fig. 3B). The analysis revealed that 511 out of 579 gene trees were concordant with the topology supporting the presence of two primary groups, groups A and B, within the subfamily Aurantioideae. However, significant conflict was observed among informative gene trees at the node encompassing the *Citrus*, *Atalantia*, *Citropsis/Naringi*, *Swinglea*, *Paramignya*/*Triphasia*, *Murraya koenigii*, and *Micromelum minutum*.

As mentioned above, cytonuclear discordance was observed between the concatenated nuclear and cp trees (Figure 2A). Additionally, significant cytonuclear discordances were evident between the species tree and cp tree, particularly within the *Citrus*, *Atalantia*, and *Clausena* genera, as well as

in intergeneric relationships (Fig.2B). The species tree positions the *Limonia/Feroniella* subgroup as an outgroup to a clade containing *Citrus*, *Atalantia*, *Citropsis/Naringi*, and *Swinglea*, while the cp tree places this subgroup as the closest outgroup to *Citrus* and *Atalantia*. Furthermore, in the species tree, *Swinglea* is positioned as an outgroup to *Citrus*, *Atalantia*, *Citropsis*, and *Naringi*, but the cp tree groups *Swinglea* with *Citropsis/Naringi*. Discrepancies between the species tree and cp tree are also apparent in the phylogenetic structures of *Paramignya*, *Triphasia*, and *Aegle/Afraegle*. Additionally, in the ASTRAL species tree, *Micromelum minutum* is positioned as an outgroup to both *Clausena* and *Glycosmis*, whereas in the cp tree, it appears as a near outgroup exclusively to *Glycosmis*.

Consistent subgrouping patterns unveiled by nuclear and cp trees

Both nuclear and cp analyses demonstrated similar grouping and subgrouping patterns, as depicted in Figures 2 and 3. The phylogenies derived from both nuclear and cp data consistently revealed two primary groups, designated as group A and group B, within the subfamily Aurantioideae (Figs. 1, 2, and 3). Historically, Aurantioideae has been classified into the Citreae and Clauseneae tribes[40]. However, our analysis led to a reassignment of species within these tribes. Specifically, *Murraya caloxylon* and *Murraya paniculata*, historically classified within the Clauseneae tribe, have been reassigned to the Citreae tribe, now defined as group A. The remaining species historically classified under the Clauseneae tribe have been consolidated into group B, which corresponds to the newly defined Clauseneae tribe in this study (Fig.2 and Fig.3).

Within group A, six distinct subgroups were identified. The "True citrus fruit trees" subgroup comprises *Citrus* species, consistent with traditional taxonomy[40]. The "Yaki Naran" subgroup encompasses all *Atalantia* species, inspired by Ayurvedic medicine in Sri Lanka including *Atalantia ceylanica*. The "Thanaka" subgroup, named after the traditional cosmetic from Myanmar, comprises *Citropsis* and *Naringi* species. The "Wood apple" subgroup includes *Limonia* and *Feroniella* species, while the "Bael fruit" subgroup comprises *Aegle* and *Afraegle*, reflecting their historical names[40].

The final subgroup, the *Murraya* subgroup, consists of *Murraya caloxylon* and *Murraya paniculata*. However, three species *Swinglea glutinosa*, *Paramignya lobata*, and *Triphasia trifolia*, did not fit neatly into any defined subgroup.

Signatures of ILS, introgression, and ancient introgression

Discordances were detected among three main subgroups: (1) True citrus fruit trees and its near subgroups, (2) the Bael fruit subgroup and its associated species, and (3) Group B (Fig.4). These discordances underscore the significant influence of genetic differentiation processes, such as ILS, introgression, and ancient introgression. To assess the impact of these processes on the discordance patterns observed in these subgroups, one representative species from each subgroup was selected for further analysis. To elucidate whether ILS or introgression was responsible for the observed phylogenetic discordances, Quantifying Introgression via Branch Lengths (QuIBL) was initially employed[41]. This method evaluates triplet topologies from gene trees by comparing branch length distributions under two models: one accounting for ILS alone and the other incorporating both ILS and introgression. Model selection was guided by Bayesian Information Criterion (BIC) values, with ΔBIC used to differentiate between ILS and introgression. Specifically, a ΔBIC greater than 10 indicates ILS, a value lower than -10 suggests introgression, and values within this range imply that ILS and introgression cannot be clearly distinguished. To further assess introgression, Patterson's D statistic[42] was applied ($P < 0.05$), while D_{FOIL} was employed to detect ancient introgression events ($P < 0.05$)[43].

In the True citrus fruit trees and its near subgroups, we assessed 860 gene trees, focusing on 30 triplets. Among these, 63.33% (19 of 30) indicated ILS (ΔBIC > 10), 10% (3 of 30) supported introgression (ΔBIC < -10), and 26.67% were inconclusive (ΔBIC between -10 and 10). This suggests that ILS was the predominant cause of phylogenetic conflict (Fig.5A and Supplementary Table S4). Specifically, the discordant topology involving wood apple, *Swinglea glutinosa*, and *Citropsis gabunensis* (used as an outgroup) was primarily attributed to introgression (ΔBIC < -10).

Patterson's D statistic and D_{FOIL} analyses indicated that ancient introgression was the main factor, followed by ILS (Supplementary Table S5). Additionally, QuIBL and Patterson's D statistic identified introgression events between True citrus fruit trees and Yaki Naran, as well as between Yaki Naran and wood apple (Supplementary Table S4 and Supplementary Table S6). Furthermore, PhyloNet analyses provided evidence of introgression events within True citrus fruit trees and its near subgroups (Fig.5B and Supplementary Fig.S5).

In the analysis of the Bael fruit subgroup and its associates, based on 1964 gene trees, ILS was identified as the primary cause of discordance (66.67%). However, triplet analysis involving *Paramignya lobata*, *Triphasia trifolia*, and *Aegle marmelos* (used as an outgroup) yielded ΔBIC values between -10 and 10, suggesting the potential presence of either ILS or introgression (Fig.5A, and Supplementary Table S7). Further clarification from Patterson's D statistic and PhyloNet analyses indicated that the observed inconsistencies were due to introgression between *Paramignya lobata* and *Triphasia trifolia* (Fig.4C, Supplementary Fig.S6 and Supplementary Table S8).

Within the Group B, based on 1135 gene trees, 58.33% of the triplets provided evidence for ILS (7 of 12 triplets, ΔBIC > 10), 16.67% indicated introgression (2 of 12 triplets, ΔBIC < -10), and 25% were ambiguous (ΔBIC between -10 and 10). This suggests that ILS was the predominant cause of phylogenetic discordance. Discordant topologies involving *Glycosmis citrifolia*, *Murraya koenigii*, and *Clausena anisata* were mainly attributed to ILS (Supplementary Table S11, Supplementary Table S12 and Supplementary Table S13). Additionally, D_{FOIL} analysis suggested ancient introgression involving *Glycosmis citrifolia* and *Micromelum minutum* with other taxa (Fig.4A). Consequently, the inconsistencies within this group were primarily attributed to ancient introgression, followed by ILS (Supplementary Table S11 and Supplementary Table S13).

The source of cytonuclear discordance was inferred using QuIBL and D-statistics results. The QuIBL analysis suggested that ILS is the primary driver of the observed cytonuclear discordance within True citrus fruit trees and its near subgroups, as well as the Bael fruit subgroup and its associates (Supplementary Table S4, Supplementary Table S7, and Supplementary Table S9).

Additionally, both QuIBL and D_{FOIL} analyses revealed that the discordant topologies involving *Micromelum minutum* and *Murraya koenigii* are predominantly due to ancient introgression, followed by ILS (Supplementary Table S11 and Supplementary Table S13).

Introgression events between Asian and African Plants

The generated phylogenetic trees revealed two instances of close relationships between Asian and African plants in the Thanaka and Bael fruit subgroups (Fig.2 and Fig.3). Within the Thanaka subgroup, *Citropsis* is native to Africa, while *Naringi* originates from Asia. Similarly, within the Bael fruit subgroup, *Aegle* is indigenous to Asia, and *Afraegle* is native to Africa. To investigate the discordance in topologies involving *Citropsis* and *Naringi*, we analyzed a total of 1,084 gene trees. Similarly, we evaluated 1,080 gene trees to explore the discordance in topologies involving *Afraegle* and *Aegle*. Our QuIBL and Patterson's D-statistic analyses revealed compelling evidence of introgression events between these Asian and African plant lineages (Fig.5A, Supplementary Table S12 and Supplementary Table S13). Specifically, the introgression event within the Bael fruit subgroup occurred over 1 to 2 coalescent time scales, while the introgression event in the Thanaka subgroup occurred over 0.5 to 2 coalescent time scales (Supplementary Table S12).

Variability in structural variations and pseudogenization across phylogenetic tree branches using publicly available whole cp genome sequences

Due to the lack of publicly available raw sequencing data, an in-depth phylogenetic analysis using Read2Tree was not feasible. Instead, we focused on analyzing publicly accessible whole cp genome sequences to assess phylogenetic relationships within the Aurantioideae subfamily. We constructed a phylogenetic tree for 45 Aurantioideae members using both ML and BI methods (Supplementary Figs. S7 and S8), with congruent topologies across analyses. These cp phylogenetic trees reinforced the close relationships between Asian and African plants within the Thanaka and Bael subgroups and allowed us to estimate divergence times, suggesting that these dispersal events occurred around five million years ago (Supplementary Fig. S9). However, our study did not resolve the unique biogeographical distribution of *Clausena anisata*, the sole *Clausena* species found in both Asia and Africa, due to inconsistencies in its interrelationships across different phylogenetic trees.

Given the little or absence of recombination in hybridization processes affecting cpDNA[11], the construction of phylogenetic trees for the *Citrus* genus using whole cp genome sequences was considered valuable. Consequently, phylogenetic trees for 56 *Citrus* species were generated (Supplementary Fig. S10 and Supplementary Fig. S11). We identified eight clusters within the *Citrus* genus based on whole cp genome sequences. The cpDNAs of the two wild species, *Citrus latipes*, and *Citrus hongheensis*, were closely related to those of the pummelo cluster.

We identified a ~20,000 bp inversion in the *psbM* to *trnG-UCC* region of *Atalantia ceylanica*, *Atalantia roxburghiana*, and *Naringi crenulata*, consistently occurring at the same nucleotide position (Supplementary Fig. S12 and Supplementary Fig. S13). This suggests that similar cpDNA structural variations can arise at different phylogenetic nodes. Notably, 53 bp inverted repeats at both ends of this region suggest their involvement in these inversions. These repeats were found in most Aurantioideae, except in *Citrus*, *Feroniella*, *Triphasia*, and *Aegle*. Additionally, pseudogenes were identified through frameshifts, premature stop codons, or missing start codons. The pseudogenization of *infA* occurred in all species, with *rpl22* and *rps19* present in various configurations, along with *ycf68*, *orf56*, *cemA*, and *ycf15* (Supplementary Fig. S14, and Supplementary Fig. S15). Furthermore, in *Atalantia kwangtungenensis*, the pseudogenization of the *accD* gene was confirmed[44], and a new pseudogenization event was identified in the *ycf1* gene (Supplementary Fig. S16).

Discussion

In this study, we leveraged an accessible dataset to analyze a large number of nuclear gene sequences alongside whole cp genome sequences for phylogenetic reconstructions. Rather than relying on a single method, the integration of both nuclear and cp genome sequences offers a more comprehensive and informative perspective in plant phylogenetics[14,15,45,46,47]. Unlike previous studies that required separate datasets for nuclear and cp genome sequences[45,46,47], our method is cost-efficient as it uses the same short-read sequence data for both analyses. This not only reduces costs but also enhances research efficiency, demonstrating its validity as a robust tool for plant phylogenetic reconstructions.

Our analysis revealed discrepancies between phylogenetic trees derived from nuclear gene sequences and those based on whole cp genome sequences. Despite these differences, all trees consistently identified two primary groups with similar subgrouping patterns. These results underscore the need for reclassification within the Aurantioideae subfamily to better align with the evolutionary relationships revealed by both nuclear and cp genome data. In response, we propose a reclassification that refines the taxonomy of Aurantioideae, providing a more accurate reflection of the subfamily's genetic and evolutionary relationships.

The Read2Tree method extends its utility beyond core phylogenetic reconstruction by addressing phylogenetic discordance. It facilitates the direct acquisition of multiple alignments for each OG, crucial for constructing individual gene trees. Integrating Read2Tree with other phylogenomic tools allows for the effective identification of genetic differentiation processes such as ILS, introgression, and ancient introgression that contribute to discordances. These evolutionary processes pose challenges to phylogenomics by complicating the inference of accurate evolutionary relationships[13,31,32,33,34]. ILS, often identified as a primary cause of gene tree discordance, arises when ancestral polymorphisms persist through successive speciation events, especially during rapid radiations with short intervals between these events[31,32,33,34,48,49]. This persistence can lead to discordant gene tree topologies. Introgression, which involves the transfer of genetic material

between species through hybridization, can blur species boundaries and distort gene trees[29,32,33,50,51]. Ancient introgression, which occurs between species that have diverged significantly, can incorporate historical gene flow into the phylogenetic signal[29,30,31,32]. Our analysis of nuclear gene-derived trees indicates that ILS is the primary driver of discordance, rather than introgression. The presence of ILS suggests that the Aurantioideae subfamily has experienced rapid diversification, which has obscured the true phylogenetic relationships and complicating the reconstruction of its evolutionary history.

Our discordance analysis concentrated on three key groups: True citrus fruit trees and its near subgroups, the Bael fruit subgroup and its associates, and group B. We found that some of the discordant topologies in these areas were primarily attributed to introgression and ancient introgression, followed by ILS. These introgressive events likely facilitated gene flow between distinct lineages, resulting in discrepancies between gene trees and the species tree. Additionally, ILS preserves ancestral genetic polymorphisms, further complicating the reconstruction of accurate phylogenetic histories[32,33,48,51,52]. In the context of True citrus fruit trees and its near subgroups, our results align with previous studies that reported hybridization between *Citrus* and intergeneric species, emphasizing the substantial impact of introgression on shaping phylogenetic relationships within this group[53,54]. Prior research also highlights how hybridization frequently leads to incongruence among gene trees, underscoring its importance in accurately inferring the phylogeny of *Citrus* species[55].

Topological conflicts between nuclear and organellar phylogenies, known as cytonuclear discordance, are common in phylogenetic studies[14,15,45,46,47,56,57]. In our investigation of the Aurantioideae subfamily, we found that ILS is often a more likely cause of cytonuclear discordance than introgression. Rapid speciation events within Aurantioideae may have occurred with insufficient time for gene lineages to coalesce, resulting in discordant gene trees between nuclear and organellar genomes[58]. Additionally, large effective population sizes can also extend the persistence of multiple gene lineages, increasing the likelihood of discordant phylogenies[58,59]. While historical hybridization and introgression introduce additional genetic material and create a mosaic

of gene trees[60,61,62], our findings suggest that ILS plays a primary role in the observed cytonuclear discordance, underscoring the complex evolutionary history of the Aurantioideae subfamily.

Structural variations and pseudogenization events provide valuable insights into the evolutionary history and genetic diversity within a genus but have limited utility for intergeneric classification. Variations in genes such as *rpl22*, *ycf68*, and *orf56* have appeared independently of their phylogenetic positions, likely due to hybridization[63] or ILS[15]. Our analysis highlighted significant patterns of pseudogenization, including the co-pseudogenization of *accD* and *ycf1*, which supports the co-evolution hypothesis[64,65,66]. Additionally, the pseudogenization of the *cemA* gene suggests functional compensation by the nuclear-encoded *DLDG1* or a transfer to the nuclear genome with a transit peptide sequence[67,68]. Independent structural changes, such as inversions in the *psbM* to *trnG-UCC* region, may arise from similar mechanisms across different lineages, complicating classification efforts. These findings emphasize the necessity for a multifaceted approach in plant phylogenetics, rather than relying solely on structural variations or pseudogenization patterns.

In conclusion, our study highlights the effectiveness of the Read2Tree method in plant phylogenetics, showcasing its utility not only in reconstructing phylogenies but also in addressing phylogenetic discordance. By integrating Read2Tree with other phylogenomic approaches, we were able to identify key genetic differentiation processes, including ILS, introgression, and ancient introgression, which contribute to phylogenetic conflicts. This sheds light on the method's robustness and versatility, advancing our understanding of complex evolutionary relationships and offering a valuable tool for future research. Additionally, our findings emphasize the critical importance of complete data submission for ensuring the transparency, reproducibility, and robustness of phylogenetic analyses. While many studies have submitted only assembled whole cp genome sequences to public databases, the inclusion of the raw sequence data used in the assembly process is essential, especially for methods like Read2Tree that rely on these data. Ensuring the availability of raw data allows for more comprehensive analyses, facilitating validation, replication, and deeper insights into evolutionary relationships.

Methods

DNA sequence data collection and processing

The DNA used in this study was either sourced from the same batch used in a previous experiment[35] or purified using the method described therein. DNA quality was assessed by electrophoresis on a 1% agarose gel, and DNA concentration was measured using a DNA BR Assay Kit (Thermo Fisher Scientific, USA). Sequencing libraries were prepared from total DNA samples by Novogene (Singapore), using the NEBNext Ultra DNA Library Prep Kit for Illumina (NEB, USA). These libraries were then sequenced to produce 150-bp paired-end reads using a NovaSeq 6000 instrument (Illumina, USA) by Novogene. Additional DNA sequence data were acquired from public databases, with accession numbers detailed in Supplementary Table S1. For preprocessing all DNA sequence data, FastP (version 0.23.2) was employed with its default settings[69].

Concatenated phylogenetic analysis based on nuclear gene sequences

For the concatenated phylogenetic analysis, we used Read2Tree (version 0.1.5)[25]. Marker genes for *Citrus clementina*, *Eucalyptus grandis*, all five species of Malvaceae, and all seven species of Brassicaceae were obtained from the OMA browser (https://omabrowser.org/oma/export_markers). The OMA browser parameters were set with a minimum fraction of covered species at 0.8 and a maximum number of markers at –1. In total, 16,258 OGs were identified among the selected 14 species in the OMA browser. Read2Tree was utilized to generate alignment files using these marker genes. To ensure the specificity of the phylogenetic trees, the 14 species obtained from the OMA browser were excluded from the resulting alignment files. Aligned data were trimmed using trimAl (v1.4. rev15) with the − automated1 option[70]. ModelTest-NG (version 0.1.7) was used to identify the optimal evolutionary model[71]. Subsequently, Maximum Likelihood (ML) trees were generated using RAxML (version 8.2.12) with 100 bootstrap replicates[72]. The visualization of the resulting trees was achieved using Dendroscope (version 3.8.8)[73].

De novo assembly and annotation of cp genomes

The *de novo* assembly of cp genomes was performed using GetOrganelle (version 1.7.5)[74]. The parameters for this process were set to − R 15, − k 21,45,65,85,105, and − F embplant_pt. To validate the assembly, we utilized Bowtie2 (version 2.4.4) for mapping[75]. The sorting of the mapped data and the generation of bam files were conducted using Samtools (version 1.14)[76]. Assembly quality was confirmed, and heteroplasmic sites were identified using the Integrated Genome Viewer[77]. Despite detecting heteroplasmic sites, the assemblies obtained from GetOrganelle were used for subsequent analysis.

For annotation of the assembled cp genomes, we used GeSeq[78]. Within the GeSeq framework, protein-coding sequences (CDS) were predicted using Chloë (version 0.1.0)[79], tRNAs were identified using tRNAscan-SE (version 2.0.7)[80], and rRNAs were predicted using HMMER[81]. Thereafter, the annotation results were curated manually. Circular genome maps of the annotated cp genomes were generated using OGDRAW (version 1.3.1)[82].

Phylogenetic analysis based on cp genome sequences

For phylogenetic analysis of cp genome sequences, DNA sequence alignments were performed using HomBlocks (version 1.0) with its default settings[83]. The optimal evolutionary model for the alignments was identified using ModelTest-NG (version 0.1.7)[71]. Subsequently, ML trees were constructed using RAxML (version 8.2.12), employing 1000 bootstrap replicates[72]. The visualization of the resultant ML trees was accomplished using Dendroscope (version 3.8.8)[73].

For the construction of Bayesian Inference (BI) trees, BEAST (version 1.10.4) was used, implementing the Markov Chain Monte Carlo (MCMC) method over a total of 10,000,000 generations[84]. Tree construction samples were collected every 10,000 generations, with the initial 10% of generations discarded as 'burn-in' to reduce potential bias from the early stages of the MCMC run. The completed BI trees were visualized using FigTree (version 1.4.4) (http://tree.bio.ed.ac.uk/software/figtree/).

Species tree inference from Read2Tree data

To construct the species tree, alignments for 16,284 OGs across all species were utilized from the Read2Tree output file "06_align_merge_dna". The OG alignments were processed using trimAl (version 1.4, revision 15)[70] with the -automated1 option and eliminate species present in the OMA browser. The alignments were then assessed based on the ratio of missing data (N characters) to total nucleotide characters, with a threshold set at 10%. From this, 579 OGs were selected for further analysis (Supplementary Dataset 16). Model testing[71] was performed for each OG, and the ML gene tree of each OG was constructed using RAxML (version 8.2.12) with 100 bootstrap replicates[72]. The species tree was subsequently inferred using ASTRAL-III (version 5.7.8)[85].

Phylogenetic discordance analysis

PhyParts was utilized to evaluate the concordance and discordance between gene trees and the species tree[86]. Gene trees were processed by collapsing branches with bootstrap support below 33% and subsequently rerooted using *Boenninghausenia albiflora* as the outgroup. Similarly, the ASTRAL species tree was rerooted with *B. albiflora*, and this rerooted tree served as the reference for the conflict analysis. The PhyParts output was visualized on the ASTRAL topology using the phypartspiecharts.py script available at github.com/mossmatters/phyloscripts. The analysis provided metrics on the number of gene trees that were concordant or discordant with the species tree topology at each branch, as well as those that were uninformative (i.e., gene trees with support values below 33% at the given branch).

The discordance analysis was focused on three specific issues: (1) *Citrus* and its near subgroups using 860 gene trees (Supplementary Dataset 17), (2) Bael fruit subgroup and its associated species using 1,964 gene trees (Supplementary Dataset 18), and (3) Group B using 1,135 gene trees (Supplementary Dataset 19), with one representative from each subgroup included. Gene trees for these analyses were generated using similar criteria and procedures as those applied in species tree estimation. Additionally, 1,084 gene trees were generated to assess discordance in topologies involving *Citropsis* and *Naringi* (Supplementary Dataset 20), while 1,080 gene trees were used to examine discordance involving *Afraegle* and *Aegle* (Supplementary Dataset 21). QuIBL was

employed to distinguish between ILS and introgression as potential causes of the observed discordances[41]. Model selection was guided by BIC values, with ΔBIC values interpreted as follows: ΔBIC > 10 indicates ILS only, ΔBIC < -10 suggests introgression, and -10 < ΔBIC < 10 indicates either ILS or introgression. Patterson's D-statistic analysis was performed using D-suite (version 0.4)[42] with the Dtrios option to identify introgression events ($P < 0.05$). Additionally, D_{FOIL} statistics[43] were employed to detect ancient introgression events, with significance determined by $P < 0.05$.

To further investigate reticulation events, Maximum Pseudolikelihood (MPL) analysis in PhyloNet (version 3.8.2)[87] was conducted. We performed 10 runs per search, each returning five optimal networks. These networks were then optimized for branch lengths and inheritance probabilities under full likelihood, using default settings and the -po option. The best-fitting network was selected based on the Akaike Information Criterion (AIC)[88,89], with the number of parameters (k) being the sum of branch lengths and reticulations, and L representing the likelihood value.

Detection of Structural Variations in Chloroplast Genomes

The process of identifying structural variations within the cp genomes involved a thorough visual inspection of the circular cp genome map combined with an analysis of BLAST results[90]. To visualize these identified structural variations, Easyfig (version 2.2.5) was employed[91].

Divergence Time Estimation

The ML phylogenetic tree for the Rutaceae family was constructed based on the aligned sequences of cp genomes using MEGA11（version 11.0.13)[92]. The optimal evolutionary model was determined using a model test conducted within MEGA11. To generate the ML tree used in divergence time estimation, the General Time Reversible (GTR) model incorporating gamma distribution (G), and proportion of invariable sites (I) was employed. Additionally, complete deletion parameters were used. The RelTime-ML option was utilized to produce the time tree by employing the GTR+G+I model and complete deletion parameters. Calibration of the time tree was established

based on the divergence time between *Murraya paniculata* and *Murraya caloxylon* retrieved from the TimeTree database[93]. The calibration parameters were defined using a Normal Distribution, with a mean of 10, a standard deviation of 2, and a confidence interval ranging from 6.08 to 13.92 million years ago.

Identification of Pseudogenes

Protein sequences were extracted from each annotation file using genbank_to (version 0.42) (https://github.com/linsalrob/genbank_to). OrthoFinder (version 2.5.4) was used to extract all combinations of orthologous proteins[94]. For accurate sequence alignment, MAFFT (version 7.515) was used to align the amino acid sequences[95]. The aligned results were visualized using MEGA11 (version 11.0.13) for comparative analysis[92], which facilitated the identification of pseudogenes. Pseudogenes were characterized by the presence of frameshift mutations, premature stop codons, and the absence of start codons. These features were systematically compared to the intact gene sequences, with additional support provided by translation mapping using the tool available at https://www.bioinformatics.org/sms2/trans_map.html.

Use of generative AI

ChatGPT (https://chat.openai.com/) was used to rewrite the manuscript into more appropriate English and proofread the English text. Subsequently, the authors verified the correctness of the revisions.

References

1. Kaya, G., Ezekannagha, C., Heider, D. & Hattab, G., Context-aware phylogenetic trees for phylogeny-based taxonomy visualization. *Front. Plant Sci*. **13**, 891240; 10.3389/fgene.2022.891240 (2022).

2. Steenwyk, J.L. & King, N., The promise and pitfalls of synteny in phylogenomics. *Plos Biology* **22(5),** e3002632; 10.1371/journal.pbio.3002632 (2024).

3. Tian, Y. *et al.* Research progress in plant molecular systematics of lauraceae. *Biology* **10(5)**, 391 (2021).

4. Jiang, D. *et al*. Complete chloroplast genomes provide insights into evolution and phylogeny of *Zingiber* (Zingiberaceae). *BMC genomics* **24(1)**, 30 (2023).

5. Li, C.C. *et al.* Insights into chloroplast genome evolution in Rutaceae through population genomics. *Horticulture Advances* **2(1)**, 1-14 (2024).

6. Park, H.S. *et al.* Inheritance of chloroplast and mitochondrial genomes in cucumber revealed by four reciprocal F1 hybrid combinations. *Sci. Rep*. **11(1)**, 2506; 10.1038/s41598-021-81988-w (2021).

7. Sun, J. *et al.* Evolutionary and phylogenetic aspects of the chloroplast genome of *Chaenomeles* species. *Sci. Rep*. **10(1),** 11466; 10.1038/s41598-020-67943-1 (2020).

8. Feng, J. et al. Analysis of complete chloroplast genome: Structure, phylogenetic relationships of *Galega orientalis* and evolutionary inference of Galegeae. *Genes* **14(1)**,176 (2023).

9. Kahraman, K. & Lucas, S.J., Comparison of different annotation tools for characterization of the complete chloroplast genome of *Corylus avellana* cv Tombul. *Bmc Genomics* **20**,1-12 (2019).

10. Wang, Y., Wang, S., Liu, Y., Yuan, Q., Sun, J. & Guo, L., Chloroplast genome variation and phylogenetic relationships of *Atractylodes* species. *BMC genomics* **22**,1-12 (2021).

11. Greiner, S., Sobanski, J. & Bock, R., Why are most organelle genomes transmitted maternally?. *Bioessays* **37(1)**, 80-94 (2015).

12. Rieseberg, L. H. & Soltis, D. E. Phylogenetic consequences of cytoplasmic gene flow in plants. *Evol. Trend. Plant.* **5**, 65–84 (1991).

13. Pamilo, P. & Nei, M. Relationships between gene trees and species trees. *Mol. Biol. Evol.* **5**, 568–583 (1988).

14. Soltis, D.E. and Kuzoff, R.K., Discordance between nuclear and chloroplast phylogenies in the Heuchera group (Saxifragaceae). *Evolution* **49**, 727-742 (1995).

15. Thureborn, O., Wikström, N., Razafimandimbison, S.G. & Rydin, C., Plastid phylogenomics and cytonuclear discordance in Rubioideae Rubiaceae. *Plos one* **19**, e0302365; 10.1371/journal.pone.0302365 (2024).

16. Krawczyk, K., Nobis, M., Nowak, A., Szczecińska, M. & Sawicki, J., Phylogenetic implications of nuclear rRNA IGS variation in *Stipa* L.(Poaceae). *Sci. Rep*. **7(1)**,11506; 10.1038/s41598-017-11804-x (2017).

17. Wang, W., Li, H. & Chen, Z., Analysis of plastid and nuclear DNA data in plant phylogenetics—evaluation and improvement. *Science China Life Sciences* **57**, 280-286 (2014).

18. Lutz, K.A., Wang, W., Zdepski, A. & Michael, T.P., Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. *BMC biotechnology* **11**,1-9 (2011).

19. Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. & Gouil, Q., Opportunities and challenges in long-read sequencing data analysis. *Genome biology* **21(1)**, 30; 10.1186/s13059-020-1935-5 (2020).

20. Mani, I., Current status and challenges of DNA sequencing. *Advances in synthetic biology*, 71-80 (2020).

21. Liu, X., Griffiths, J.I., Bishara, I., Liu, J., Bild, A.H. & Chang, J.T., Phylogenetic inference from single-cell RNA-seq data. *Sci. Rep.* **13(1)**, 12854; 10.1038/s41598-023-39995-6 (2023).

22. Washburn, J.D. *et al.* Genome-guided phylo-transcriptomic methods and the nuclear phylogenetic tree of the Paniceae grasses. *Sci. rep.* **7(1)**, 13528; 10.1038/s41598-017-13236-z (2017).

23. Spillane, J.L., LaPolice, T.M., MacManes, M.D. & Plachetzki, D.C., Signal, bias, and the role of transcriptome assembly quality in phylogenomic inference. *BMC ecology and evolution* **21**, 1-17(2021).

24. Quek, Z.B.R. and Huang, D., Effects of missing data and data type on phylotranscriptomic analysis of stony corals (Cnidaria: Anthozoa: Scleractinia). *Molecular phylogenetics and evolution* **134**, 12-23 (2019).

25. Dylus, D., Altenhoff, A., Majidian, S., Sedlazeck, F. J. & Dessimoz, C. Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. *Nat. Biotechnol.* **42**, 139–147 (2024).

26. Park, S., Kwak, M. & Park, S., Complete organelle genomes of Korean fir, *Abies koreana* and phylogenomics of the gymnosperm genus *Abies* using nuclear and cytoplasmic DNA sequence data. *Sci. Rep*. **14(1)**, 7636; 10.1038/s41598-024-58253-x (2024).

27. Soltis, P. S., & Soltis, D. E., The Role of Hybridization in Plant Speciation, *Annual Review of Plant Biology* **60**, 561–588 (2009).

28. Rieseberg, L. H., & Willis, J. H., Plant Speciation. *science* **317(5840)**, 910-914 (2007).

29. Nelson, T.C. *et al.* Ancient and recent introgression shape the evolutionary history of pollinator adaptation and speciation in a model monkeyflower radiation (*Mimulus* section *Erythranthe*). *PLoS Genetics* **17(2)**, e1009095; 10.1371/journal.pgen.1009095 (2021).

30. Suvorov, A. *et al.* Deep ancestral introgression shapes evolutionary history of dragonflies and damselflies. *Syst. Biol.* **71(3)**, 526-546 (2022).

31. Maddison, W. P. Gene trees in species trees. *Syst. Biol.* **46**, 523–536 (1997).

32. Meleshko, O. et al. Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated *bryophyte* genus. *Mol Biol Evol*. **38**, 2750-2766 (2021).

33. Kandziora, M., Sklenář, P., Kolář, F. & Schmickl, R., How to tackle phylogenetic discordance in recent and rapidly radiating groups? Developing a workflow using *Loricaria* (Asteraceae) as an example. *Front. Plant Sci*. **12**, 765719; 10.3389/fpls.2021.765719 (2022).

34. Wanke, S. and Wicke, S., Phylogenomic discordance in plant systematics. *Front. Plant Sci*. **14**, 1308126; 10.3389/fpls.2023.1308126 (2023).

35. Nagano, Y. *et al.* Phylogenetic relationships of Aurantioideae (Rutaceae) based on RAD-Seq. *Tree Genet. Genomes* **14**, 6; 10.1038/nrg.2015.28 (2018).

36. Herrero, R., Asins, M.J., Carbonell, E.A. & Navarro, L., Genetic diversity in the orange subfamily Aurantioideae. I. Intraspecies and intragenus genetic variability. *Theoretical and Applied Genetics* **92**, 599-609 (1996).

37. Penjor, T. *et al.* Phylogenetic relationships of *Citrus* and its relatives based on *matK* gene sequences. *PLOS ONE* **8**, e62574; 10.1371/journal.pone.0062574 (2013).

38. Iwamasa, M., Nito, N. & Ling, J. T. Intra- and intergeneric hybridization in the orange subfamily, Aurantioideae *in Proceedings of the Sixth International Citrus Congress* (eds. Goren, R. & Mendel, K.) 123–130 (Balaban Publishers, Philadelphia/Rehovot Margraf Scientific Books, Weikersheim, 1988).

39. Gabaldón, T. & Koonin, E.V., Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics* **14(5)**, 360-366 (2013).

40. Swingle, W. T., Reece, P. C. The botany of Citrus and its wild relatives in the orange subfamily in *The citrus industry, vol 1* (eds. Reuther, W., Webber, H. J. & Bachelor, L. D.) 190–430 (University of California, Berkeley, 1967).

41. Edelman, N.B. *et al*. Genomic architecture and introgression shape a butterfly radiation. *Science* **366(6465)**,594-599 (2019).

42. Malinsky, M., Matschiner, M. & Svardal, H., Dsuite-Fast D-statistics and related admixture evidence from VCF files. *Molecular ecology resources* **21(2)**, 584-595 (2021).

43. Pease, J.B. & Hahn, M.W., Detection and polarization of introgression in a five-taxon phylogeny. *Syst. Biol.* **64(4)**, 651-662 (2015).

44. Shi, W., Song, W., Zhao, Y., Shi, C., & Wang, S. Complete chloroplast genomes of four *Atalantia* (Rutaceae) species: insights into comparative analysis, phylogenetic relationships, and divergence time estimation. *Plant Syst. Evol.* **309**, 31; 10.1007/s00606-023-01868-w (2023).

45. Hodel, R. G. J., Zimmer, E. A., Liu, B. B., & Wen, J. Synthesis of nuclear and chloroplast data combined with network analyses supports the polyploid origin of the apple tribe and the hybrid origin of the Maleae-Gillenieae Clade. *Front. Plant Sci.* **12**, 820997; 10.3389/fpls.2021.820997 (2022).

46. Yu, W. B., Huang, P. H., Li, D. Z. & Wang, H. Incongruence between nuclear and chloroplast DNA phylogenies in *Pedicularis* section *Cyathophora* (Orobanchaceae). *PLOS ONE* **8**, e74828; 10.1371/journal.pone.0074828 (2013).

47. Bruun-Lund, S., Clement, W. L., Kjellberg, F., & Rønsted, N. First plastid phylogenomic study reveals potential cyto-nuclear discordance in the evolutionary history of *Ficus* L. (Moraceae). *Mol. Phylogenet. Evol.* **109**, 93–104 (2017).

48. Morales-Briones, D.F. et al. Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae sl. *Syst. Biol.* **70**, 219-235 (2021).

49. Hibbins, M.S., Breithaupt, L.C. & Hahn, M.W., Phylogenomic comparative methods: Accurate evolutionary inferences in the presence of gene tree discordance. *PNAS* **120**, e2220389120; 10.1073/pnas.2220389120 (2023).

50. Wen, D., Yu, Y., Zhu, J. and Nakhleh, L., Inferring phylogenetic networks using PhyloNet. *Syst. Biol.* **67**, 735-740 (2018).

51. Liu, X. *et al*. Genomic insights into zokors' phylogeny and speciation in China. *PNAS* **119**, e2121819119; 10.1073/pnas.2121819119 (2022).

52. Hernández-Gutiérrez, R. *et al.* Localized Phylogenetic Discordance Among Nuclear Loci Due to Incomplete Lineage Sorting and Introgression in the Family of Cotton and Cacao (Malvaceae). *Front. Plant Sci.* **13**, 850521; 10.3389/fpls.2022.850521 (2022).

53. Yasuda, K., Yahata, M., Shigyo, M., Matsumoto, R., Yabuya, T. & Kunitake, H., Identification of parental chromosomes in sexual intergeneric hybrid progenies between *Citrus* cultivar 'Nanpu'tangor and *Citropsis schweinfurthii* in the subfamily Aurantioideae. *Journal of the Japanese Society for Horticultural Science* **79(2)**, 129-134 (2010).

54. Smith, M.W., Gultzow, D.L. & Newman, T.K., First fruiting intergeneric hybrids between *Citrus* and *Citropsis. Journal of the American Society for Horticultural Science* **138(1)**, 57-63 (2013).

55. Ramadugu, C., Pfeil, B.E., Keremane, M.L., Lee, R.F., Maureira-Butler, I.J. & Roose, M.L., A six nuclear gene phylogeny of *Citrus* (Rutaceae) taking into account hybridization and lineage sorting. *PLoS One* **8(7)**, e68410; 10.1371/journal.pone.0068410 (2013).

56. Yu, W. B., Huang, P. H., Li, D. Z. & Wang, H. Incongruence between nuclear and chloroplast DNA phylogenies in *Pedicularis* section *Cyathophora* (Orobanchaceae). *PLOS ONE* **8**, e74828; 10.1371/journal.pone.0074828 (2013).

57. Bruun-Lund, S., Clement, W. L., Kjellberg, F., & Rønsted, N. First plastid phylogenomic study reveals potential cyto-nuclear discordance in the evolutionary history of *Ficus* L. (Moraceae). *Mol. Phylogenet. Evol.* **109**, 93–104 (2017).

58. Duan, L., Fu, L. & Chen, H.F., Phylogenomic cytonuclear discordance and evolutionary histories of plants and animals. *Science China Life Sciences* **66(12)**, 2946-2948 (2023).

59. Phuong, M.A., Bi, K. & Moritz, C., Range instability leads to cytonuclear discordance in a morphologically cryptic ground squirrel species complex. *Molecular Ecology* **26(18)**, 4743-4755 (2017).

60. Hibbins, M.S. & Hahn, M.W., The timing and direction of introgression under the multispecies network coalescent. *Genetics* **211(3)**,1059-1073 (2019).

61. Seixas, F.A., Boursot, P. & Melo-Ferreira, J., The genomic impact of historical hybridization with massive mitochondrial DNA introgression. *Genome Biology* **19**, 1-20 (2018).

62. Nge, F.J., Biffin, E., Thiele, K.R. & Waycott, M., Reticulate evolution, ancient chloroplast haplotypes, and rapid radiation of the Australian plant genus *Adenanthos* (Proteaceae). *Frontiers in Ecology and Evolution* **8**, 616741; 10.3389/fevo.2020.616741 (2021).

63. Stull, G. W., Pham, K. K., Soltis, P. S., & Soltis, D. E. Deep reticulation: the long legacy of hybridization in vascular plant evolution. *Plant J.* **114**, 743–766 (2023).

64. Sasaki, Y. *et al.* Chloroplast-encoded protein as a subunit of acetyl-CoA carboxylase in pea plant. *J. Biol. Chem*. **268**, 25118–25123 (1993).

65. Sasaki, Y. & Nagano, Y. Plant acetyl-CoA carboxylase: structure, biosynthesis, regulation, and gene manipulation for plant breeding. *Biosci. Biotechnol. Biochem*. **68**, 1175–1184 (2004).

66. de Vries, J., Sousa, F. L., Bölter, B., Soll, J. & Gould, S. B. YCF1: a green TIC?. *Plant Cell* **27**, 1827–1833 (2015).

67. Sasaki, Y., Sekiguchi, K., Nagano, Y. & Matsuno, R. Chloroplast envelope protein encoded by chloroplast genome. *FEBS Lett* **316**, 93–98 (1993).

68. Trinh, M. D. L. *et al.* Lack of plastid-encoded Ycf10, a homolog of the nuclear-encoded DLDG1 and the cyanobacterial PxcA, enhances the induction of non-photochemical quenching in tobacco. *Plant Direct* **5**, e368; 10.1002/pld3.368 (2021).

69. Chen, S., Zhou, Y., Chen, Y., & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinform.* **34**, i884–i890 (2018).

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

71. Darriba, D. *et al.* ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* **37**, 291–294 (2020).

72. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinform.* **30**, 1312–1313 (2014).

73. Huson, D. H. & Scornavacca, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61**, 1061–1067 (2012).

74. Jin, J. J. *et al.* GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 241; 10.1186/s13059-020-02154-5 (2020).

75. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

76. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinform.* **25**, 2078–2079 (2009).

77. Robinson, J. T., Thorvaldsdóttir, H., Turner, D. & Mesirov, J. P. igv.js: An embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinform.* **39**, btac830; 10.1093/bioinformatics/btac830 (2023).

78. Tillich, M. *et al.* GeSeq: versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11 (2017).

79. Zhong, X. Assembly, annotation and analysis of chloroplast genomes. Doctoral thesis in *the University of Western Australia*, *Australia* (2020).

80. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).

81. Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204 (2018).

82. Greiner, S., Lehwark, P. & Bock, R. *OrganellarGenomeDRAW (OGDRAW)* version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **47**, 59–64 (2019).

83. Bi, G., Mao, Y., Xing, Q. & Cao, M. HomBlocks: A multiple-alignment construction pipeline for organelle phylogenomics based on locally collinear block searching. *Genomics* **110**, 18–22 (2018).

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

85. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S., ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC bioinformatics* **19,**15-30 (2018).

86. Smith, S.A., Moore, M.J., Brown, J.W. & Yang, Y., Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC evolutionary biology* **15**,1-15 (2015).

87. Wen, D., Yu, Y., Zhu, J., & Nakhleh, L. Inferring phylogenetic networks using PhyloNet. *Syst. Biol.* **67(4)**, 735-740 (2018).

88. Sullivan, J., & Joyce, P. Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* **36(1)**, 445-466 (2005).

89. Akaike, H. In *Selected Papers of* Hirotugu Akaike (eds Emanuel Parzen, Kunio Tanabe, & Genshiro Kitagawa) 199–213 (Springer New York, 1998).

90. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

91. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinform.* **27**, 1009–1010 (2011).

92. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **38**, 3022–3027 (2021).

93. Kumar, S. *et al.* TimeTree 5: An expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174; 10.1093/molbev/msac174 (2022).

94. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).

95. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment software version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

Acknowledgements

In the initial phase of our study's data analysis conducted in Sri Lanka; we extend our profound gratitude to Dr. Dhammika Elkaduwe from the Faculty of Engineering at the University of Peradeniya for his exceptional support in granting access to high-performance computing facilities. We would like to thank Ms. Akiko Sakai for technical assistance, and Mr. Takashi Arita for providing the experimental samples. This work was partially supported by a Grant-in-Aid for Scientific Research (18K05623) from the Japan Society for the Promotion of Science to Yukio Nagano. E.P.W. was supported by the Japanese Government (MEXT) scholarship. We would also like to thank Editage (https://www.editage.com) for the English language editing. This study was part of a dissertation submitted by the first author toward the partial fulfillment of her Ph.D. All authors provided consent for submission and publication.

Author contributions

E.P.W., Y.N., and T.I. designed the study. M.Y. and N.K. supplied the experimental samples. Y.N. performed the DNA experiments. E.P.W., T.I., M.H.S., and Y.N. conducted the bioinformatics analysis. N.U.J. established the framework for conducting the study, including an initial phase of data analysis conducted in Sri Lanka. As experts on Aurantioideae, M.Y. and N.K. reviewed the descriptions of Aurantioideae. All authors were involved in the data analysis. The manuscript was written by E.P.W. and Y.N. and reviewed by all authors.

Data Availability Statement

Raw sequencing data have been deposited in the DDBJ Sequence Read Archive (https://www.ddbj.nig.ac.jp/dra/index-e.html; accession no. DRA017638). The cp genome sequences assembled and annotated from these raw sequencing data have been deposited in DDBJ/EMBL/GenBank under accession numbers LC794878 to LC794908. The cp genome sequences, assembled and annotated from publicly available raw sequencing data, are available in GenBank format in Supplementary Dataset 1. The alignments, tree data and Nexus files used in this study are included in the Supplementary Datasets 2 through 21.

Competing interests

The authors declare that they have no conflicts of interest.

Figure Legends

Figure 1. ML phylogenetic tree based on conserved nuclear gene sequences of 39 species in the Aurantioideae subfamily.

The alignment used for phylogenetic tree construction included a total of 6,674,283 nucleotide positions with 605,597 parsimony-informative sites, as calculated using custom Python scripts (Supplementary Dataset 15). For the analysis, the optimal evolutionary model, GTRGAMMAIX, was employed in RAxML. The root position corresponded to that identified in the phylogenetic tree of the Rutaceae family (Supplementary Fig. S1). Bootstrap values (percentages over 100 replicates) are displayed at the nodes, with the scale bar indicating genetic divergence in substitutions per site. Bootstrap values below 50% are omitted. The subgroups identified in this study are represented by different colors. The alignment data in FASTA format and tree data in Newick format are available in Supplementary Dataset 2.

Figure 2. Discordance between nuclear gene-derived trees and cp tree

(A) Comparison of the concatenated nuclear tree and the cp tree. Left panel: ML phylogenetic tree generated using Read2Tree, based on conserved nuclear gene sequences from 39 species within the Aurantioideae subfamily. Bootstrap support values (percentages from 100 replicates) are displayed at each node, with values below 50% omitted. Subgroups identified in this study are represented by different colors.

(B) Comparison of the ASTRAL species tree and the cp tree. Left panel: Species tree inferred using ASTRAL-III, with local posterior probability (LPP) values displayed as percentages for each corresponding node. Subgroups identified in this study are represented by different colors. Tree data in Newick format are available in Supplementary Dataset 4.

Right panel in both A and B: ML phylogenetic tree constructed using the whole cp genome sequences of 39 species in the Aurantioideae subfamily. Bootstrap support values (percentages from 1,000 replicates) are displayed at the nodes, with values below 50% omitted. Asterisks mark

sequences that were assembled *de novo* in this study. Subgroups identified are represented by distinct colors. Dotted lines and shapes highlight conflicting relationships between the nuclear and cp phylogenies, revealing areas of topological incongruence between the two trees. Alignment data (FASTA format) and tree data (Newick format) are available in Supplementary Dataset 3.

Figure 3. Species tree and concatenated nuclear gene tree

(A) Comparison of the concatenated nuclear gene tree and ASTRAL species tree. Left panel: ML phylogenetic tree generated using Read2Tree, based on conserved nuclear gene sequences from 39 species within the Aurantioideae subfamily. Bootstrap support values (percentages from 100 replicates) are displayed at each node, with values below 50% omitted. Subgroups identified in this study are highlighted in distinct colors. Right panel: Species tree inferred using ASTRAL-III, with local posterior probability (LPP) values displayed as percentages for each corresponding node. Subgroups identified in this study are denoted by distinct colors. Dotted lines indicate areas of incongruence between the two trees.

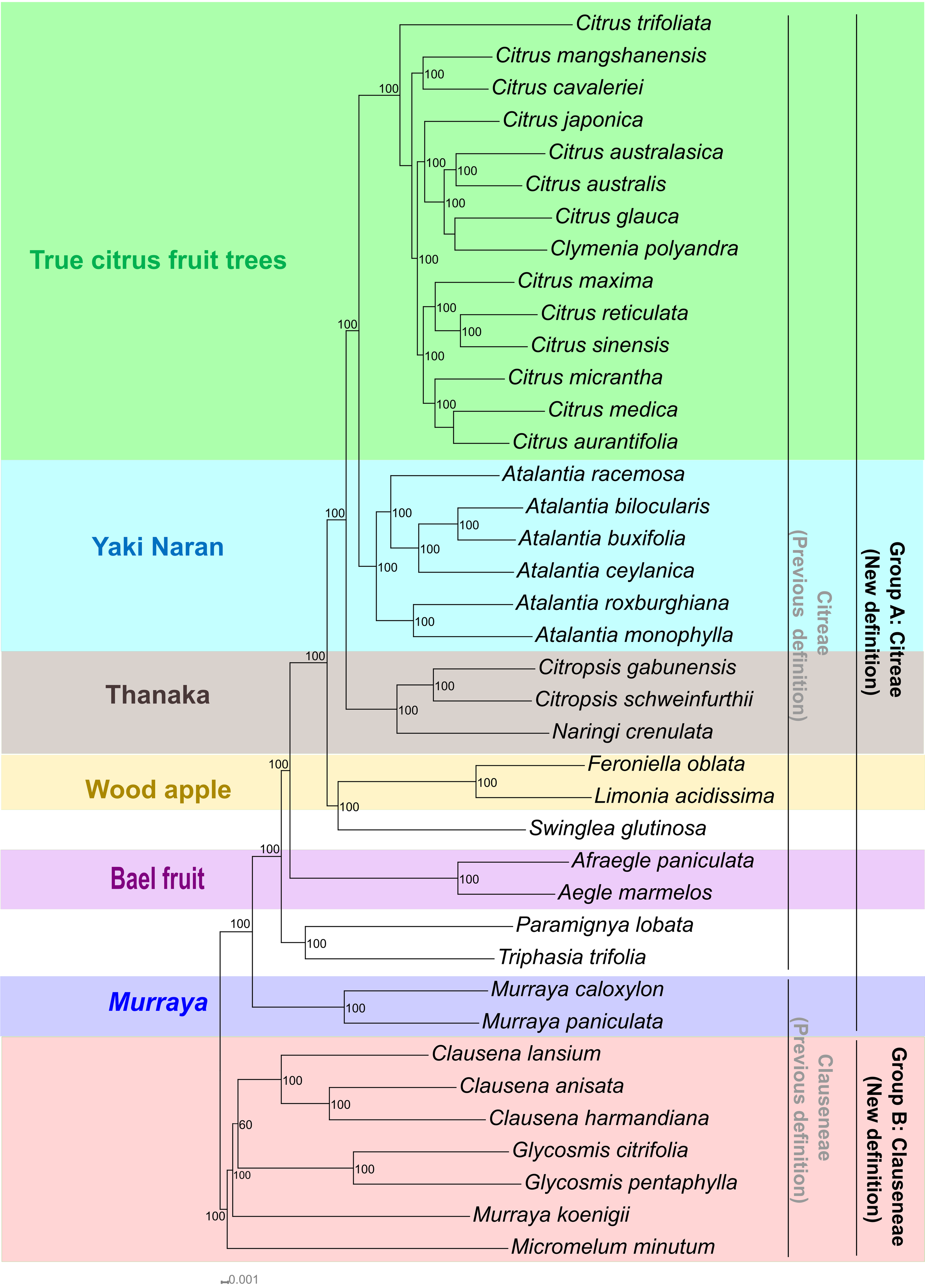
(B) Gene tree discordance displayed on the ASTRAL-inferred species tree. Pie charts at each node illustrate the proportion of gene trees falling into one of four categories: concordant with the species tree (blue), discordant but representing the most common alternative topology (green), discordant with other alternative topologies (red), and uninformative due to low branch support (gray, <33% bootstrap support at the node). The number above each branch indicates the count of gene trees concordant with the species tree topology at that node (blue portion of the pie). The number below each branch represents the sum of informative discordant topologies at that node (green + red portions of the pie). Different colors denote the subgroups identified in this study.

Figure 4. Discordance analysis among three subgroups: (1) True citrus fruit trees and its near subgroups, (2) Bael fruit subgroup and its associates, and (3) the Group B. One representative from each subgroup/genus is used.

Figure 5. Summary of Phylogenetic Discordance Sources

(A) ASTRAL-inferred species tree depicting the relationships among members of the Aurantioideae subfamily. Arrows indicate various evolutionary events: the dashed blue arrow represents ILS; green arrows represent introgression; and pink arrows indicate ancient hybridization followed by ILS. Detailed results are provided in Supplementary Tables 4 to 13. One representative from each genus/subgroup is used.

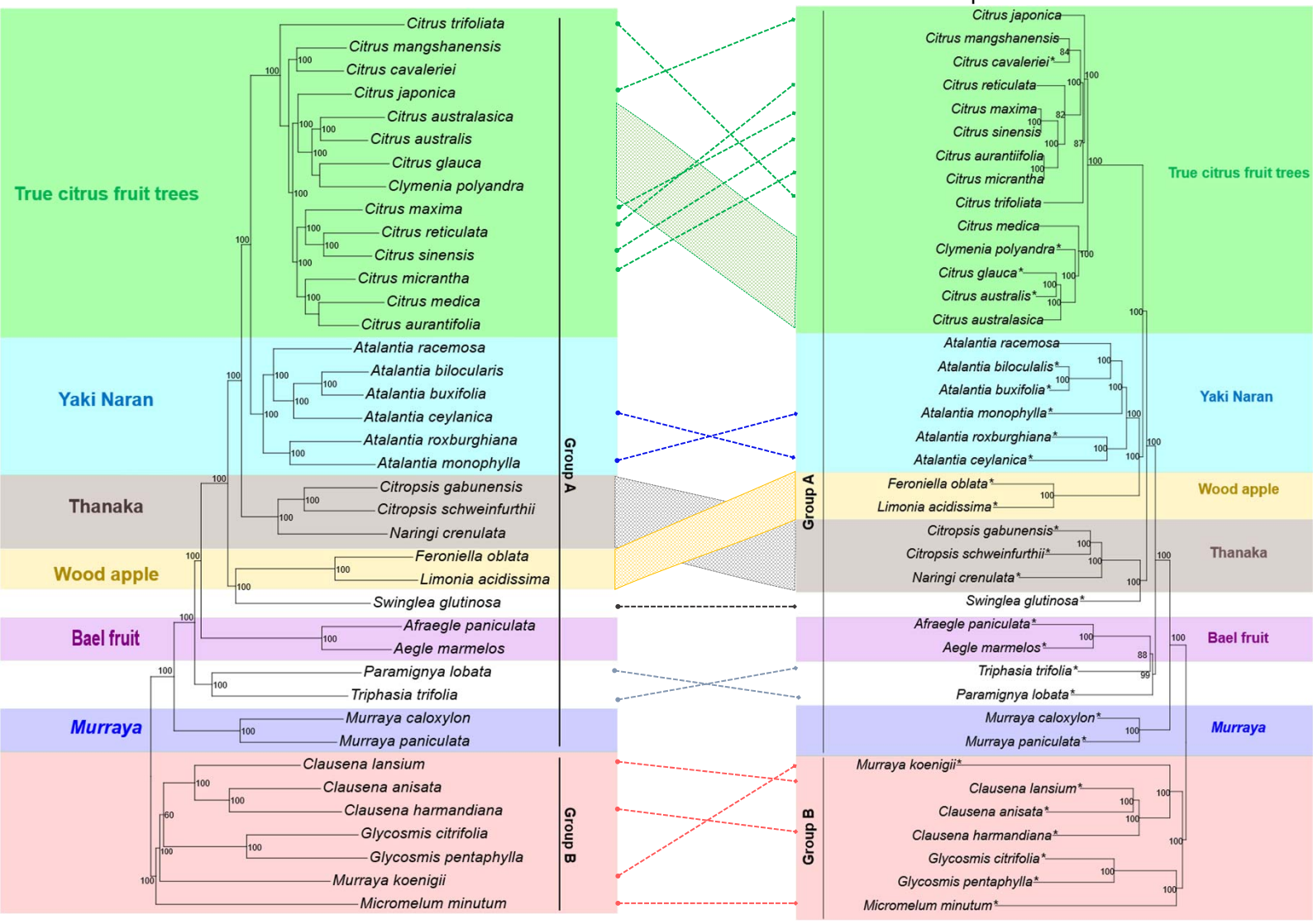
Phylogenetic networks for True citrus fruit trees and its near subgroups (B), as well as the Bael fruit subgroup and associated species (C), were inferred using PhyloNet with Maximum Pseudolikelihood under models allowing for one, two, and three reticulations. The Akaike Information Criterion (AIC) indicates that the three-reticulation model is the best fit for both groups. In Figure (B), green branches highlight lineages involved in reticulate evolutionary histories, with numerical values representing the inheritance probabilities for each reticulation event. In Figure (C), purple branches similarly indicate lineages associated with reticulation, along with their corresponding inheritance probabilities. The Nexus files for each reticulation model can be found in Supplementary Datasets 5 and 6. Supplementary Figs. S5 and S6 present all inferred phylogenetic networks generated by PhyloNet for True citrus fruit trees and its near subgroups and the Bael fruit subgroup with associated species, respectively.


True citrus fruit trees
Yaki Naran
Thanaka
Wood apple
Bael fruit
Murraya
Citrus trifoliata
Citrus mangshanensis
Citrus cavaleriei
Citrus japonica
Citrus australasica
Citrus australis
Citrus glauca
Clymenia polyandra
Citrus maxima
Citrus reticulata
Citrus sinensis
Citrus micrantha
Citrus medica
Citrus aurantifolia
Atalantia racemosa
Atalantia bilocularis
Atalantia buxifolia
Atalantia ceylanica
Atalantia roxburghiana
Atalantia monophylla
Citropsis gabunensis
Citropsis schweinfurthii
Naringi crenulata
Feroniella oblata
Limonia acidissima
Swinglea glutinosa
Afraegle paniculata
Aegle marmelos
Paramignya lobata
Triphasia trifolia
Murraya caloxylon
Murraya paniculata
Clausena lansium
Clausena anisata
Clausena harmandiana
Glycosmis citrifolia
Glycosmis pentaphylla
Murraya koenigii
Micromelum minutum
Citreae (Previous definition)
Group A: Citreae (New definition)
Clauseneae (Previous definition)
Group B: Clauseneae (New definition)
0.001

A)

Concatenated nuclear gene tree from Read2Tree

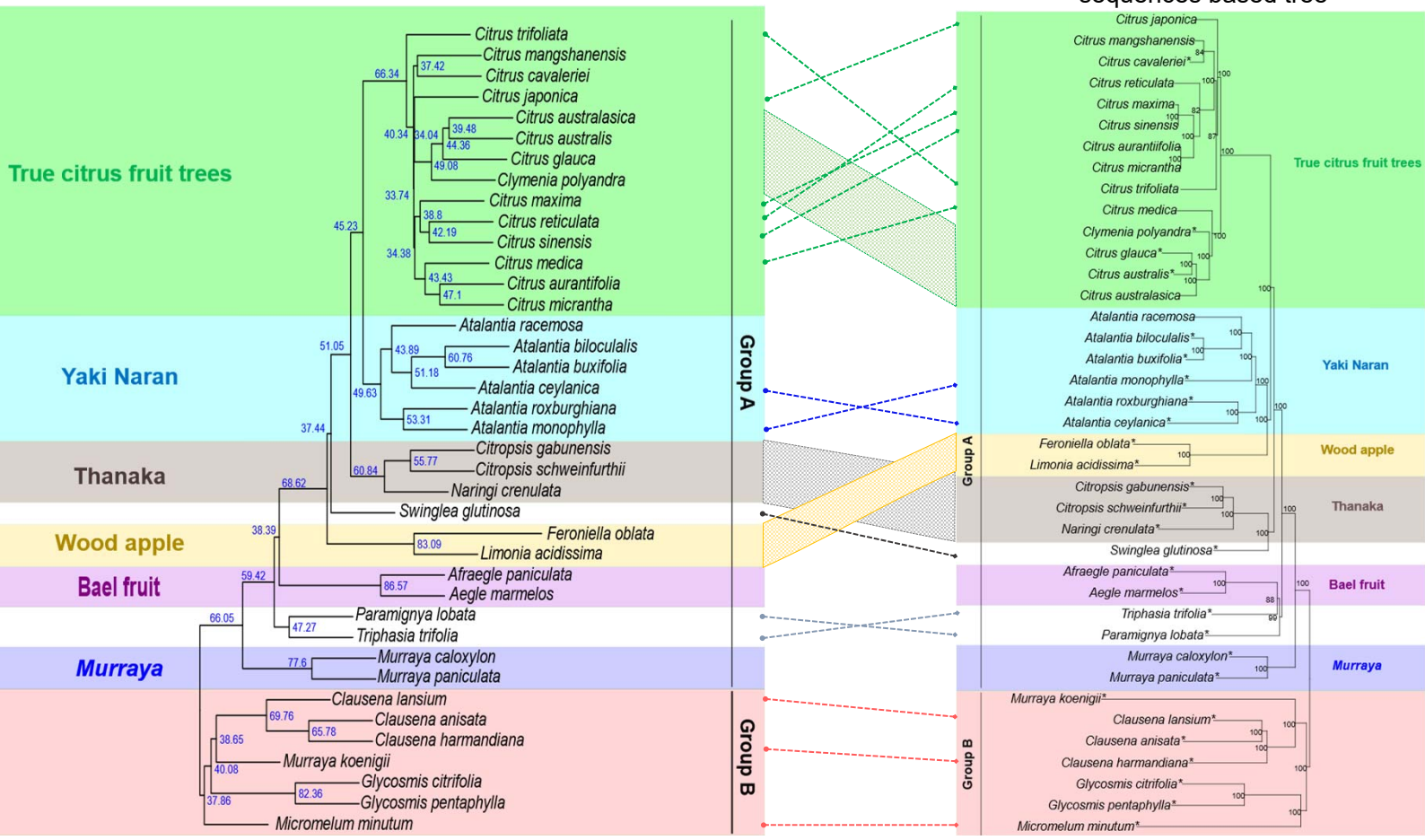
Whole cp genome sequences based tree



B)

ASTRAL species tree

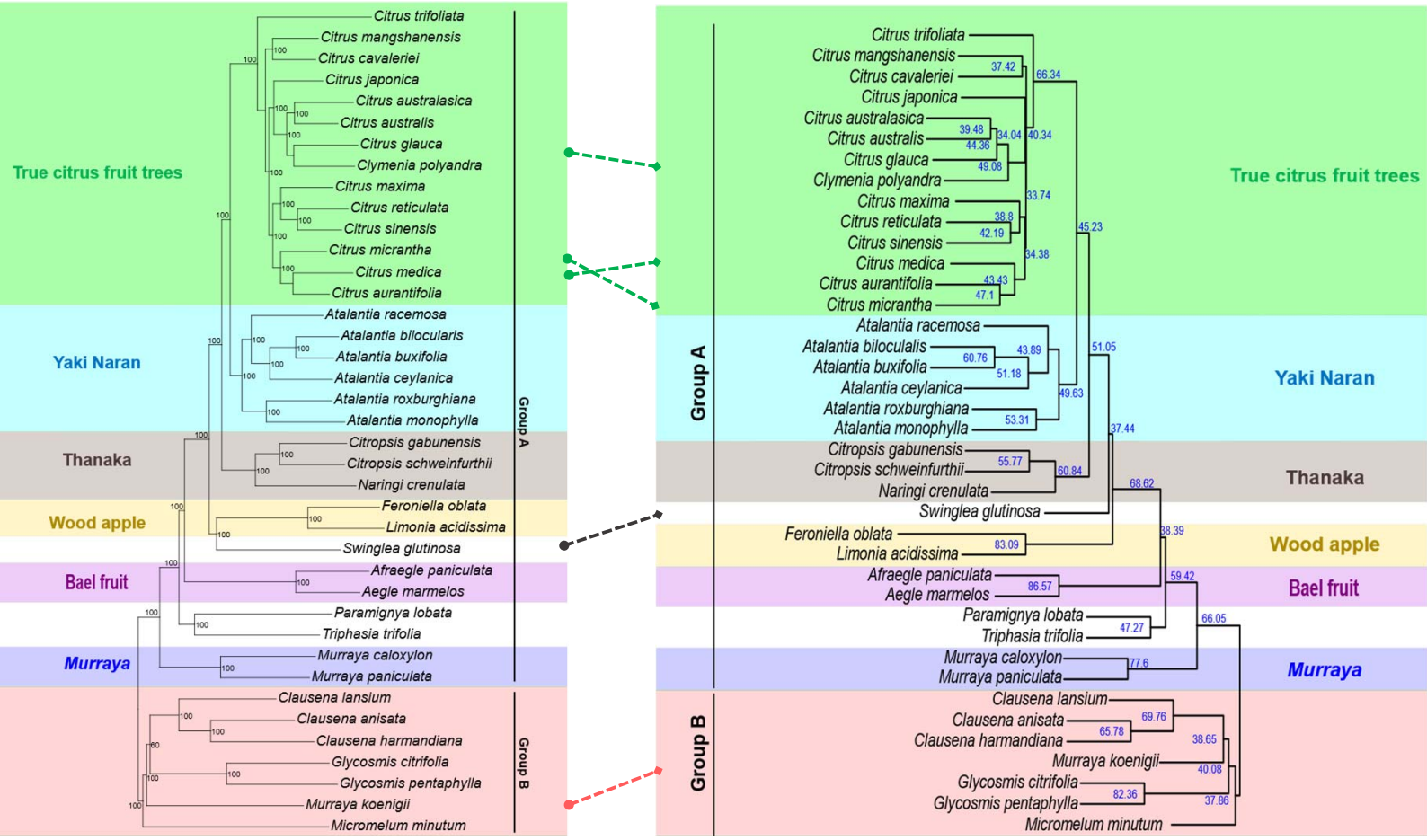
Whole cp genome sequences based tree



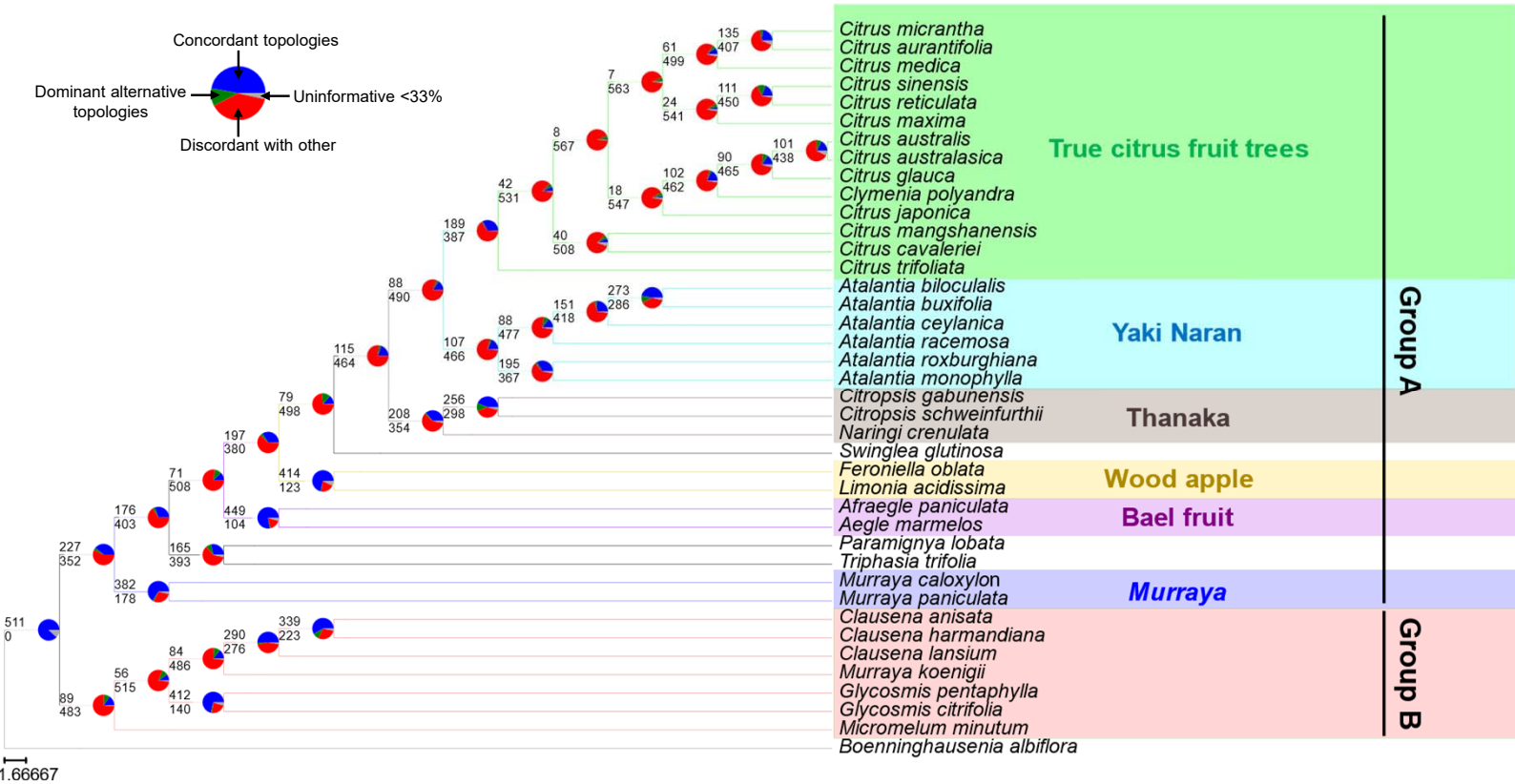
A)

Concatenated nuclear gene tree from Read2Tree

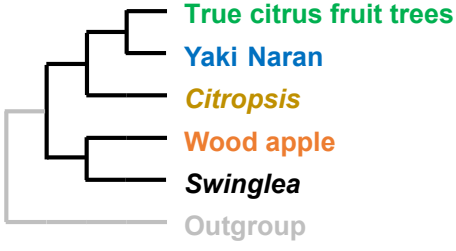
ASTRAL species tree



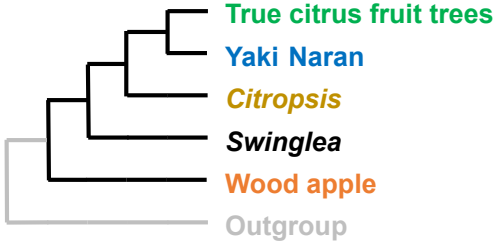
B)



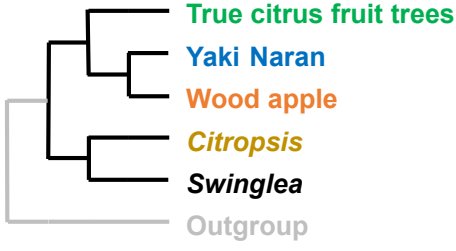
1. Conflicts in True citrus fruit trees and its near subgroups



Concatenated gene tree

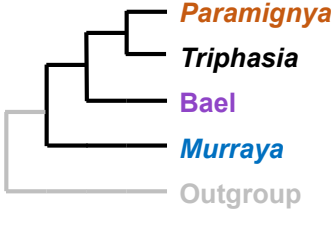


ASTRAL-III Species tree

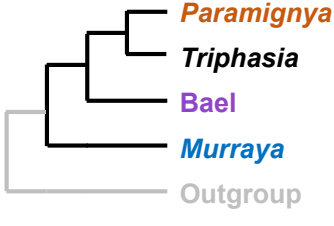


Whole Chloroplast tree

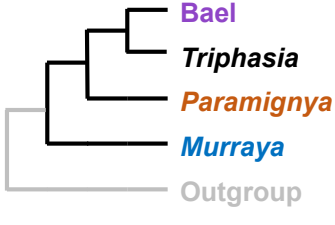
2. Conflicts in Bael fruit subgroup and its associates



Concatenated gene tree

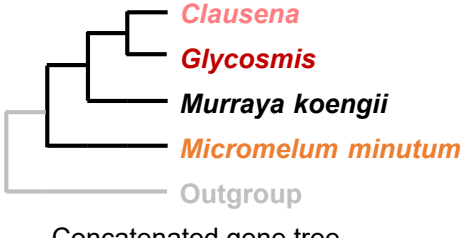


ASTRAL-III Species tree

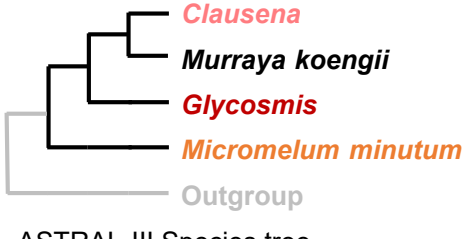


Whole Chloroplast tree

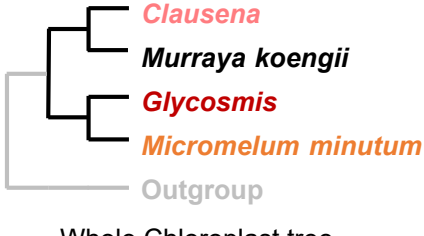
3. Conflicts in Group B



Concatenated gene tree

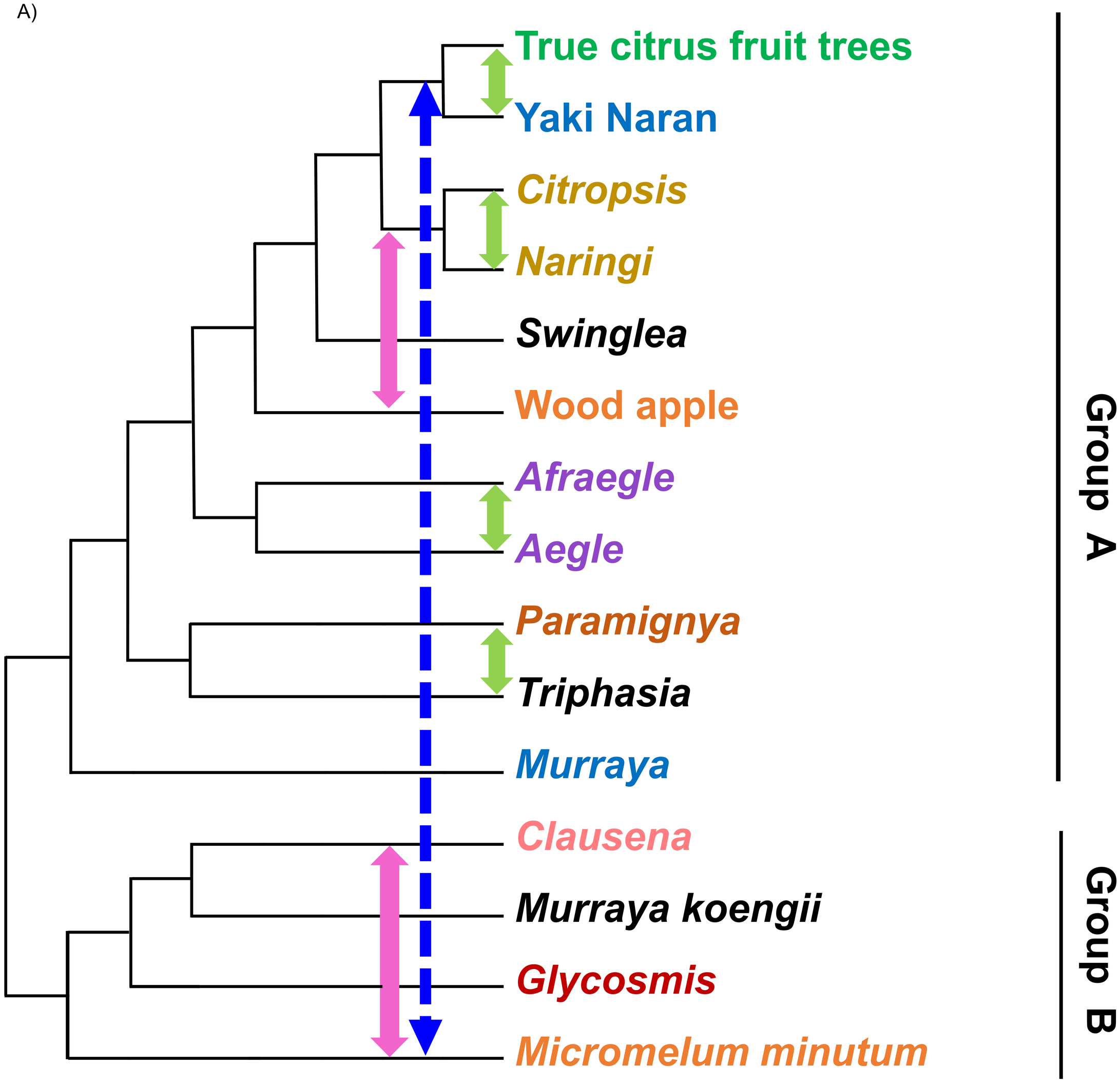


ASTRAL-III Species tree

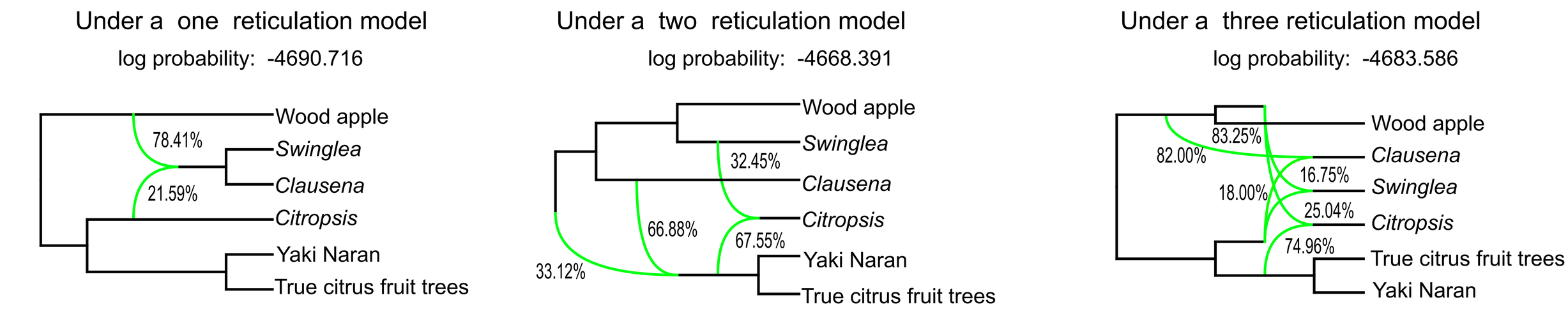


Whole Chloroplast tree

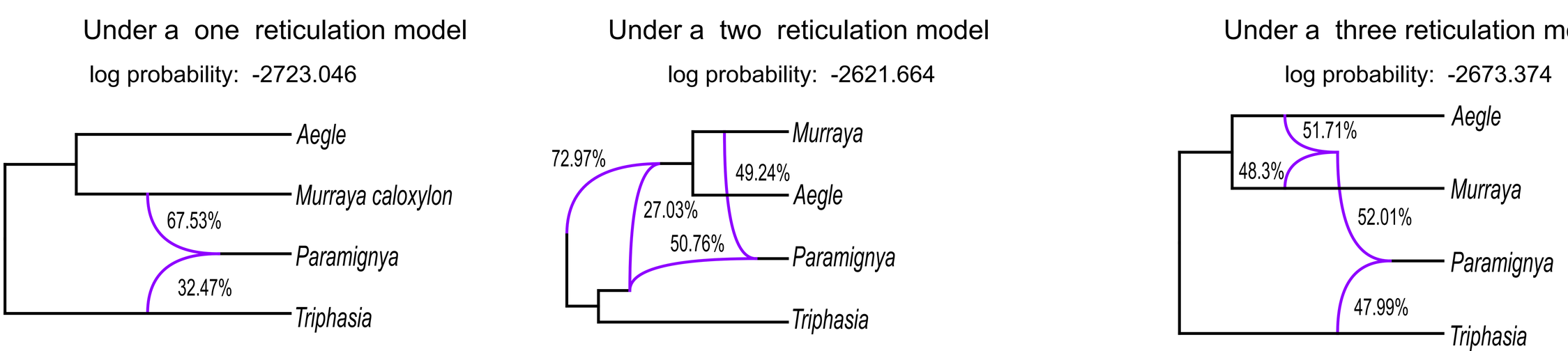
A)


True citrus fruit trees
Yaki Naran
Citropsis
Naringi
Swinglea
Wood apple
Afraegle
Aegle
Paramignya
Triphasia
Murraya
Clausena
Murraya koengii
Glycosmis
Micromelum minutum
Group A
Group B

B)


Under a one reticulation model
log probability: -4690.716
78.41%
21.59%
Wood apple
Swinglea
Clausena
Citropsis
Yaki Naran
True citrus fruit trees
Under a two reticulation model
log probability: -4668.391
32.45%
66.88%
67.55%
33.12%
Wood apple
Swinglea
Clausena
Citropsis
Yaki Naran
True citrus fruit trees
Under a three reticulation model
log probability: -4683.586
83.25%
82.00%
16.75%
18.00%
25.04%
74.96%
Wood apple
Clausena
Swinglea
Citropsis
True citrus fruit trees
Yaki Naran

C)


Under a one reticulation model
log probability: -2723.046
67.53%
32.47%
Aegle
Murraya caloxylon
Paramignya
Triphasia
Under a two reticulation model
log probability: -2621.664
72.97%
49.24%
27.03%
50.76%
Murraya
Aegle
Paramignya
Triphasia
Under a three reticulation model
log probability: -2673.374
51.71%
48.3%
52.01%
47.99%
Aegle
Murraya
Paramignya
Triphasia