# Cost-Effective Approaches to Elucidate Intergeneric Relationships of Plants: Utilizing Multiple Conserved Nuclear Genes and Whole Chloroplast Genomes

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

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

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

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

***Corresponding author:**

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

## Abstract

The chloroplast (cp) genome sequence, which is a widely used and affordable tool, is utilized for plant classification. However, its effectiveness remains unclear. This study contrasts the phylogenetics derived from whole cp genome sequences with those obtained using Read2Tree, another cost-efficient method capable of extracting multiple conserved nuclear gene sequences. To explore this topic, we focused on the Aurantioideae subfamily, which includes *Citrus* and its relatives, where phylogenetic trees based on nuclear DNA from high-throughput sequencing are available. In intergeneric relationships, phylogenetic trees generated with Read2Tree were consistent with previous high-precision nuclear DNA-based trees but differed from those generated using whole cp genome sequences. Thus, even in plants in which past polyploidization or hybridization between distantly related species had not been previously assumed, the two phylogenetic trees did not coincide. Both techniques present unique pros and cons, leaving it unclear which technique is superior for understanding a plant’s evolutionary history. However, integrating both approaches is more useful for effectively classifying plants than it is for understanding the unified evolutionary history based on phylogenetics. We propose that integrating these two cost-effective methods by utilizing the same raw sequencing data is a mainstream approach in DNA-based plant classification.

**Keywords:** Chloroplast genome sequences, Read2Tree, Aurantioideae subfamily, Phylogenetic trees

# 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. The advent of high-throughput sequencing has provided access to numerous variable sites. In recent years, chloroplast (cp) genome sequences have gained popularity in plant molecular systematics. Characterized by maternal inheritance, stable gene content, low recombination, and manageable size, the cp genome has emerged as a cost-effective analytical tool. Assembling the whole cp genome, typically spanning 120–170 kbp, from approximately 1 Gbp of raw sequencing data transforms it into a 'super-barcode' for plant species classification[1,2,3,4].

Another sequence that can provide many variable sites is nuclear DNA sequence. However, the phylogenetic analysis of many plants encounters challenges due to the limited availability of high-quality nuclear sequences. This scarcity results from the high costs and technical difficulties associated with sequencing and annotation. Consequently, investigating the phylogenetic relationships within a family becomes intricate. Moreover, comparing phylogenetic trees derived from numerous nuclear genes with those from whole cp genomes for multiple family members poses challenges.

The 'Read2Tree' tool addresses these challenges by directly converting raw sequencing reads into gene sets[5]. This innovative method bypasses traditional reconstruction steps, facilitating the integration of nuclear sequences into phylogenetic research. Additionally, the potential to employ the same short-read sequence data for both whole cp genome analysis and Read2Tree represents an unexplored yet promising approach. The untapped possibilities of Read2Tree underscore its growing significance in plant phylogenetics.

Although the efficacy of phylogenetic investigations utilizing the whole cp genome has been well-established, the validation of Read2Tree in plant research remains relatively less substantiated. This study aims to elucidate and harness the potential of Read2Tree by re-evaluating 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[6]. Notably, Read2Tree specializes in analyzing conserved nuclear gene sequences, diverging from the focus of RAD-Seq on nuclear DNA proximal to specific restriction sites. This contrast provides valuable insights through comparative analysis. Second, the genus *Citrus*, along with other members of Aurantioideae, holds global appeal due to its enticing flavor, nutritional advantages, and adaptability in culinary applications. These fruits, rich in vitamin C and other essential nutrients, play pivotal roles in maintaining health. Moreover, their extensive cultivation and widespread distribution significantly contribute to their economic value. Furthermore, beyond *Citrus*, the members of the Aurantioideae subfamily are important in various regions for their edible and medicinal attributes. Hence, a detailed phylogenetic analysis of Aurantioideae is not only scientifically pertinent but also holds substantial economic and cultural significance.

Inconsistencies between nuclear and cpDNA phylogenies are well-documented, primarily in interspecies relationships, as highlighted in some studies[7,8,9,10,11]. Although intergeneric relationships have not been extensively studied, we observed a difference in the phylogenetic tree derived from nuclear DNA using RAD-Seq compared to that based on short cpDNA sequences when analyzing the Aurantioideae subfamily[6,12]. Many studies reporting such mismatches have focused on short sequences from either nuclear or cpDNA, or both. However, few studies have compared the phylogenetic trees constructed from the whole cp genome with those derived from numerous nuclear gene sequences. One exception is a study that utilized transcriptome and whole cp genome sequences to analyze plants in which past hybridization between distantly related species was previously assumed[13]. Therefore, focusing on intergeneric relationships, in which past hybridization was not assumed, a comparative analysis of these two types of phylogenies, utilizing extensive datasets, promises to be a valuable contribution to the field of plant classification.

In this study, we attempted for the first time to compare phylogenetic trees obtained from numerous nuclear genes using the Read2Tree method and those obtained from whole cp genomes, involving numerous members within a subfamily. We constructed phylogenetic trees for approximately 40 plants employing the Read2Tree method. To assess the reliability of this approach, we compared our results with those obtained from previous RAD-Seq analyses. Subsequently, we constructed phylogenetic trees for around 100 plants based on whole cp genome sequences. We then performed a comprehensive comparative analysis between the phylogenetic insights derived from whole cp genomes and those obtained through the Read2Tree method. These investigations have elucidated the constraints associated with utilizing whole cp genome sequences for intergeneric classification. Simultaneously, we explored pragmatic approaches to integrate the whole cp genome sequences with numerous nuclear genes to delineate intergeneric relationships.

## Materials and **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[6] 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[14].

### Phylogenetic analysis based on nuclear gene sequences

For the phylogenetic analysis, we used Read2Tree (version 0.1.5)[5]. Marker genes for *Citrus clementina* were obtained from the OMA browser (https://omabrowser.org/oma/export_markers). Read2Tree was used to produce alignment files using these marker genes. Aligned data were trimmed using trimAl (v1.4. rev15) with the − automated1 option[15]. ModelTest-NG (version 0.1.7) was used to identify the optimal evolutionary model[16]. Subsequently, Maximum Likelihood (ML) trees were generated using RAxML (version 8.2.12) with 100 bootstrap replicates[17.] The visualization of the resulting trees was achieved using Dendroscope (version 3.8.8)[18].

### *De novo* assembly and annotation of cp genomes

The *de novo* assembly of cp genomes was performed using GetOrganelle (version 1.7.5)[19]. 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[20]. The sorting of the mapped data and the generation of bam files were conducted using Samtools (version 1.14)[21]. Assembly quality was confirmed, and heteroplasmic sites were identified using the Integrated Genome Viewer[22]. Despite detecting heteroplasmic sites, the assemblies obtained from GetOrganelle were considered suitable for subsequent analysis.

For annotation of the assembled cp genomes, we used GeSeq[23]. Within the GeSeq framework, protein-coding sequences (CDS) were predicted using Chloë (version 0.1.0)[24], tRNAs were identified using tRNAscan-SE (version 2.0.7)[25], and rRNAs were predicted using HMMER[26]. Thereafter, the annotation results were curated manually. Circular genome maps of the annotated cp genomes were generated using OGDRAW (version 1.3.1)[27].

### 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[28]. The optimal evolutionary model for the alignments was identified using ModelTest-NG (version 0.1.7)[16]. Subsequently, ML trees were constructed using RAxML (version 8.2.12), employing 1000 bootstrap replicates[17]. The visualization of the resultant ML trees was accomplished using Dendroscope (version 3.8.8)[18].

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[29]. 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/).

### 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[30]. To visualize these identified structural variations, Easyfig (version 2.2.5) was employed[31].

### 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[32]. For accurate sequence alignment, MAFFT (version 7.515) was used to align the amino acid sequences[33]. The aligned results were visualized using MEGA11 (version 11.0.13) for comparative analysis[34], which facilitated the identification of pseudogenes.

### Divergence Time Estimation

The ML phylogenetic tree for the Ruatceae family was constructed based on the aligned sequences of cp genomes using MEGA11（version 11.0.13)[34]. 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[35]. 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.

### 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.

## Results

### Read2Tree tool exhibits notable accuracy in generating phylogenetic trees from nuclear gene sequences

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 for further phylogenetic analyses. ML phylogenetic trees were constructed based on DNA sequences using these data (Figure 1), with rooting determined from the Rutaceae analysis. Within the genus *Citrus*, an in-depth analysis was deemed unnecessary due to recombination events during hybridization, and it was consequently excluded. The ML phylogenetic trees largely aligned with findings from a prior RAD-Seq based study[6], with noteworthy exceptions observed for *Micromelum minutum* and *Murraya koenigii*. Their phylogenetic positions, previously uncertain in a prior study[6], remained unresolved in our current analysis, as evident from the comparison between Figure 1 and Supplementary Fig. S1.

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

Initially, a phylogenetic tree encompassing 59 members of the Rutaceae family was constructed, serving as the foundation for subsequent analyses (Supplementary Fig. S2). Subsequently, phylogenetic trees for 45 members of the Aurantioideae subfamily were developed based on DNA sequences, utilizing ML and Bayesian Inference (BI) methods (Figure 2 and Supplementary Fig. S3, respectively). The topologies of these phylogenetic trees remained consistent across the analyses.

Given the absence of recombination in hybridization processes affecting cpDNA, 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 (Figure 3 and Supplementary Fig. S4). We identified eight distinct cpDNA types 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.

Significant disparities were observed when comparing the topologies of phylogenetic trees derived from whole cp genome sequences with those obtained from nuclear DNA analyses. Inconsistencies within the *Citrus*, *Atalantia*, and *Clausena* genera were noted, likely resulting from hybridization effects. Moreover, marked differences were observed in the intergeneric relationships. For instance, the cpDNA-based analysis positioned the *Limonia*/*Feroniella* group as one of the closest outgroups to *Citrus*, contrary to the nuclear DNA analysis, which identified this group 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* group. 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 again.

## Variability in structural variations and pseudogenization across phylogenetic tree branches

Our investigation aimed to assess the correspondence of identical structural variations or pseudogenizations in cp genomes with the respective branches of the phylogenetic tree, utilizing whole cp genome sequences. We identified an inversion event occurring within the *psbM* to *trnG-UCC* region in the cp genomes of *Atalantia ceylanica, Atalantia roxburghiana*, and *Naringi crenulata* (Figure 4 and Supplementary Figs. S5). This event consistently manifested at the same nucleotide sequence position in all three species, indicating that identical cpDNA structural variations can occur at different phylogenetic nodes. Notably, 53 bp inverted repeat sequences were observed at both ends of this region, suggesting their involvement in these inversion events. These repeats are present in most Aurantioideae, excluding *Citrus*, *Feroniella*, *Triphasia*, and *Aegle*.

Expanding our analysis to pseudogenization events, we examined 103 species. The *infA* gene was identified as a pseudogene in all surveyed species. Interestingly, the *rpl22* gene or its pseudogenized form is located at two distinct positions across the border between the Large Single Copy (LSC) region and Inverted Repeat (IR) regions. One location is at the LSC-Inverted Repeat B (IRB) boundary, maintaining the usual length with identifiable start and stop codons. The other is at the LSC-Inverted Repeat A (IRA) boundary, where it does not retain its typical length. Analysis of the *rpl22* region with the usual length revealed pseudogenization events, observed across various branches of the phylogenetic tree (Supplementary Figures S6 and S7). Notably, the *rpl22* region was occasionally absent at the boundary. For instance, the *rps19* gene at the LSC-IRA boundary in *Afraegle paniculata* undergoes pseudogenization. Additionally, in *Glycosmis citrifolia,* both *rpl22* and *rps19* were absent around the LSC-IRB boundary. Pseudogenization of *ycf68* and *orf56* occurred across different phylogenetic branches (Supplementary Figure S6). Pseudogenization also occurred in *cemA* and *ycf15* in *Afraegle paniculata* and *Aegle marmelos*, respectively. Furthermore, for *Atalantia kwangtungenensis*, we confirmed the previously reported pseudogenization of *accD* gene[36] and newly identified a pseudogenization event in *ycf1* gene (Supplementary Figure S8).

## Integrating cp and nuclear DNA analyses for group identification

The integration of cp and nuclear DNA analyses to unveil a cohesive evolutionary history presents challenges; nevertheless, our research demonstrates its efficacy in identifying distinct groups. The methodology involved the identification of clusters with a bootstrap value of 100. Subsequently, groups were defined based on the congruence observed between cp and nuclear DNA analysis results. These groups are consistently referenced across all Aurantioideae phylogenetic tree figures (Figures 1 and 2 and Supplementary Figures S1, S2, and S3) and detailed in Supplementary Table S2, which also includes historical nomenclature for comparative purposes.

In line with historical classifications[37], we divided all species within the Aurantioideae subfamily into the Citreae and Clauseneae tribes. Within Citreae, we identified five distinct groups. The "True citrus fruit trees" group comprises *Citrus* species, consistent with traditional taxonomy[37]. The "Yaki Naran" group encompasses all *Atalantia* species, inspired by Ayurvedic medicine in Sri Lanka including *Atalantia ceylanica*. The "Thanaka" group, named after the traditional cosmetic from Myanmar, comprises *Citropsis* and *Naringi* species. The "Wood apple" group includes *Limonia* and *Feroniella* species, while the "Bael fruit" group comprises *Aegle* and *Afraegle*, reflecting their historical names[37]. Within the Clauseneae tribe, two groups emerged: "Clauseneae 1," featuring *Murraya caloxylon* and *Murraya paniculata*, and "Clauseneae 2," encompassing the remaining species in the tribe. The division between Clauseneae 1 and Clauseneae 2 is attributed to the early branching within the Aurantioideae subfamily. However, three species –, *Swinglea glutinosa*, *Paramignya lobata*, and *Triphasia trifolia*, did not fit neatly into any defined group.

## Simultaneous plant divergence events between Asia and Africa

Within the Thanaka group, *Citropsis* are native to Africa, while *Naringi* originates from Asia. Similarly, *Aegle* is indigenous to Asia within the Bael fruit group, and *Afraegle* is indigenous to Africa. Phylogenetic trees constructed using both conserved nuclear genes and whole cp genome sequences strongly indicate simultaneous plant divergence between Asia and Africa in the past. The comprehensive data from whole cp genome sequences provide a more precise estimation of divergence times due to the absence of missing values. Based on this data, it is estimated that the two dispersion events in question occurred approximately five million years ago (Figure 5). However, this study did not explore the unique biogeographical distribution of *Clausena anisata*, the sole species within the *Clausena* genus found in both Asia and Africa, due to inconsistencies in the genus's interrelationships across various phylogenetic trees.

## Discussion

In this study, we utilized an affordable dataset to analyze numerous nuclear gene sequences and whole cp genome sequences. Unlike previous investigations relying on transcriptome data[13], our method stands out for its cost efficiency. This report represents the second study of its kind, following prior research[13], utilizing a multitude of variable sites from both nuclear and cpDNA to investigate intergeneric relationships. Notably, our study diverges from earlier research in that, even in plants where past polyploidization or hybridization between distantly related species was not previously assumed, phylogenetic trees based on whole cp genome sequences do not align with those constructed from multiple nuclear gene sequences. This discrepancy is evident not only in interspecies relationships but also in intergeneric relationships.

The conventional dependence on morphological differences for plant classification often introduces ambiguities in determining the presence or absence of reproductive isolation within and between genera. Consequently, the potential extent of hybridization in many species remains unclear. Nevertheless, certain plants, such as members of the genus *Citrus* known for intra-genus hybridization[38], have undergone extensive study. Recognizing the significant impact of hybridization on phylogenetic tree construction, a comparative analysis of the two types of phylogenetic trees was conducted, utilizing whole cp genome sequences and multiple nuclear gene sequences. The following sections, we will delve into the discussion of the advantages and disadvantages of each approach.

The strengths of phylogenetic trees derived from whole cp genome sequences are noteworthy for several reasons. First, they offer valuable insights into maternal inheritance processes. Second, due to their reduced susceptibility to DNA recombination, these trees facilitate construction in groups characterized by frequent hybridization, both within species and genera. Third, the substantial size of the cp genome (approximately 150 kbp) containing numerous variable sites allows for high-resolution phylogenetic analysis.

Nevertheless, this method presents several drawbacks. A significant limitation arises from the potential transfer of cpDNA between individuals through hybridization, a phenomenon known as 'chloroplast capture' resulting from such DNA movement[39]. Additionally, cpDNA often exhibits behavior similar to that of a single gene. Consequently, phylogenetic trees based on cpDNA may more accurately reflect gene trees than species trees[40]. Discrepancies between gene trees and species trees can be attributed to factors such as incomplete lineage sorting or hybridization[41,42,43]. These discrepancies underscore the critical challenge of relying solely on cp genome sequences for phylogenetic analysis.

The construction of phylogenetic trees from multiple nuclear gene sequences is advantageous, primarily because of the utilization of many sequences. This abundance reduces the likelihood of discrepancies between the gene and species trees. However, this method has several limitations. A notable limitation is the occurrence of recombination during hybridization, which can complicate tree construction within species or, in some cases, within genera. Furthermore, the potential for a lack of reproductive isolation between genera poses additional challenges in accurately depicting evolutionary relationships, as it may introduce complexities that are not accounted for in phylogenetic analysis. Another drawback is that the evolutionary rate of nuclear DNA is higher than that of cpDNA. This faster rate of evolution, which leads to greater variance, could potentially affect the topology of phylogenetic trees[44]. Another drawback is related to the presence of missing values in multiple sequence alignments, a problem shared by the RAD-Seq method, as highlighted in our previous paper[6]. In contrast, data from the whole cp genome generally exhibit almost identical gene content, which significantly reduces the occurrence of missing values in the alignments.

In summary, our investigation reveals that both methods of constructing phylogenetic trees, using whole cp genome sequences and multiple nuclear gene sequences, have their own set of strengths and weaknesses. In other words, it is unclear which sequence is useful for understanding a plant’s evolutionary history. The ambiguity in determining the presence or absence of reproductive isolation among plants adds to the complexity of establishing the superiority of one method over another.

Moreover, the observed discrepancies between the phylogenetic trees derived from these two methods in our study suggest that traditional phylogenetic analyses, which often combine short gene sequences from cp and nuclear DNA (such as *rbcL*/*matK* and ITS pairings), may not always provide reliable results.

The identification of structural variations or pseudogenization events has limited utility for the intergeneric classification of plants. This limitation arises because genetic changes detected in genes such as *rpl22*, *ycf68*, and *orf56* can occur independently of their positions in the phylogenetic trees (Supplementary Figure S6). The observed discrepancies between structural variations, pseudogenization patterns, and phylogenetic trees could be due to several factors. One potential cause is the transfer of cpDNA among plants through hybridization[45]or the phenomenon of incomplete lineage sorting[46].

Alternatively, these structural changes or pseudogenization events may have occurred independently within the different lineages. For instance, the inversion observed in the *psbM* to *trnG-UCC* region across different species may be the result of a similar mechanism operating independently in each lineage. These complexities highlight the challenges of relying solely on the detection of structural variation and pseudogenization for accurate classification and underscore the need for a multifaceted approach in plant phylogenetics.

The investigation of pseudogenization in our study yielded interesting findings. We observed simultaneous pseudogenization in the *accD* and *ycf1* genes, where *accD* encodes a subunit of acetyl-CoA carboxylase[47,48], and *ycf1* is hypothesized to encode a subunit of the protein translocon[49]. The co-evolution of *accD* and *ycf1* has been previously suggested[49], and our current findings provide additional evidence supporting this hypothesis. Another intriguing case in our study was the pseudogenization of *cemA* gene[50]. We speculate that the nuclear-encoded DLDG1 may have compensated for the loss of *cemA* function or that *cemA* was transferred to nuclear DNA and subsequently acquired a transit peptide sequence[51].

This study highlights a discrepancy between phylogenetic trees derived from conserved nuclear genes and those based on whole cp genome sequences. Despite this divergence, the ability to conduct analyses on both nuclear genes and cp genome sequences within the same dataset emphasizes the need to integrate these methods into plant classification, particularly at the family or subfamily levels. Informed by this approach, we embarked on a reclassification within the Aurantioideae subfamily.

Adhering to traditional morphological taxonomy[37], Aurantioideae is divided into two tribes: Clauseneae and Citreae. Our analysis substantiates and aligns with this division; Traditionally, Clauseneae was further divided into the subtribes Micromelinae, Clauseninae, and Merrillinae, whereas Citreae was categorized as Triphasiinae, Citrinae, and Balsamocitrinae[37]. However, our research identified several inconsistencies within these subtribal divisions, which led us to propose their discontinuation. Although previous morphological classifications recognized groups beneath the subtribal level, we suggest retaining only those groups associated with Bael fruit, Wood apple, and True citrus fruit trees. Additionally, we recommend introducing new groups into the Aurantioideae taxonomy: Yaki Naran, Thanaka, Clauseneae 1, and Clauseneae 2 based on our abovementioned criteria for subgrouping. These proposed groups represent a significant refinement of the taxonomy of the subfamily, offering a more accurate reflection of its genetic and evolutionary diversities. This reclassification, driven by combined analysis of nuclear and cp genome sequences, provides a more comprehensive understanding of the structure and relationships of the Aurantioideae subfamily.

The concurrent divergence events observed between Asian and African continents warrant investigation into their possible causes. These events were estimated to have occurred approximately five million years ago, a period coinciding with the closure of the Panama Isthmus, which precipitated significant global climatic changes[52]. While such abrupt phenomena played a role, it is also plausible that more gradual processes, such as the slow rise of the Himalayan-Tibetan Plateau[53], contributed to climate alterations over time, alongside single, large-scale events. These climatic shifts could have rendered the transfer of plants between India and Africa more challenging for animal or bird vectors.

Furthermore, although fossil evidence suggests a potential African origin for the Aurantioideae subfamily[54], we could not conclusively determine this aspect.

In this report, we bring to the forefront crucial considerations for plant systematists, focusing on two main areas.

1. Integrating both cpDNA-based and nuclear DNA-based phylogenetics in research; utilizing both approaches is essential for a comprehensive understanding of plant evolutionary relationships, as each provides unique and complementary insights.
2. The necessity of submitting not only the assembled whole cp genome sequences to databases but also the raw sequence data used in the assembly process. This practice is particularly important for the Read2Tree tool because it relies on the availability of raw sequence data. Complete data submission is critical for ensuring the robustness, transparency, and reproducibility of phylogenetic studies.

In conclusion, our findings propose a new trajectory for plant molecular phylogenetics. By advocating a more inclusive approach that combines diverse genetic data and emphasizes the importance of thorough data sharing, we aim to advance the field of plant systematics and pave the way for future research.

## Author contributions

E. P. W. and Y. K. designed the study. M. Y. and N. K. supplied the experimental samples. Y. N. performed the DNA experiments. E. P. W., M. H. S, and Y. N. conducted bioinformatics analysis. N. U. J. established the framework for conducting the study with 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.

## 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.

## Conflict of interest statement

The authors declare that they have no conflicts of interest.

## Data Accessibility 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 chloroplast genome sequences, assembled and annotated from publicly available raw sequencing data, are available in GenBank format in Supplementary Dataset 1. The alignments and tree data used in this study are included as separate files in the Supplementary Dataset.

## Benefit-Sharing Statement

The ensuing benefits arising from this research materialize via the impartial distribution of our comprehensive data and research findings throughout openly accessible databases, as described above.

## ORCID

| | |
|---|---|
| Eranga Pawani Witharana | https://orcid.org/0009-0002-7028-999X |
| Myat Htoo San | https://orcid.org/0009-0004-6840-1705 |
| Nadeeka U. Jayawardana | https://orcid.org/0000-0002-8986-0706 |
| Nobuhiro Kotoda | https://orcid.org/0000-0001-9107-1960 |
| Yukio Nagano | https://orcid.org/0000-0001-5030-676X |

## References

1. Chen, X. *et al.* Identification of *Ligularia* herbs using the complete chloroplast genome as a super-barcode. *Front. Pharmacol.* **9**, 695; 10.3389/fphar.2018.00695 (2018).

2. Zhang, Z., Zhang, Y., Song, M., Guan, Y. & Ma, X. Species identification of *Dracaena* using the complete chloroplast genome as a super-barcode. *Front. Pharmacol.* **10**, 1441; 10.3389/fphar.2019.01441 (2019).

3. Zhang, W. *et al.* DNA barcoding of *Oryza*: conventional, specific, and super barcodes. *Plant Mol. Biol.* **105**, 215–228 (2021).

4. Wu, L. *et al.* Plant super-barcode: a case study on genome-based identification for closely related species of *Fritillaria*. *Chin. Med.* **16**, 1–11 (2021).

5. 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).

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

7. 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).

8. Zhang, G. Q. *et al.* Phylogenetic incongruence in *Cymbidium* orchids. *Plant Divers.* **43**, 452–461 (2021).

9. 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).

10. Lee-Yaw, J. A. *et al.* An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (*Helianthus*). *New Phytol.* **221**, 515–526 (2019).

11. Ananda, G. *et al.* Phylogenetic relationships in the Sorghum genus based on sequencing of the chloroplast and nuclear genes. *Plant Genome* **14**, e20123; 10.1002/tpg2.20123 (2021).

12. 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).

13. 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).

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

15. 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).

16. 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).

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

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

19. 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).

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

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

22. 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).

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

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

25. 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).

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

27. 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).

28. 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).

29. 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).

30. 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).

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

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

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

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

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

36. 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).

37. 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).

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

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

41. Simpson, M. G. *Plant systemics*, second edition (Elsevier, 2010)

42. Kuritzin, A., Kischka, T., Schmitz, J. & Churakov, G. (2016). Incomplete lineage sorting and hybridization statistics for large-scale retroposon insertion data. *PLOS Comput. Biol.* **12**, e1004812; 10.1371/journal.pcbi.1004812 (2016).

43. Suh, A., Smeds, L. & Ellegren, H. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLOS Biol* **13**, e1002224; 10.1371/journal.pbio.1002224 (2015).

44. Rousseau-Gueutin, M. *et al.* The intertwined chloroplast and nuclear genome coevolution in plants in *Plant growth and regulation - Alterations to Sustain Unfavorable Condition* (eds Diah Ratnadewi and Hamim) 61–84 (IntechOpen, 2018)

45. 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).

46. Wang, Y. *et al.* Incomplete lineage sorting and local extinction shaped the complex evolutionary history of the Paleogene relict conifer genus, *Chamaecyparis* (Cupressaceae). *Mol Phylogenet. Evol.* **172**, 107485; 10.1016/j.ympev.2022.107485 (2022).

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

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

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

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

51. 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).

52. Thomson, J. R. *et al*. Tectonic and climatic drivers of Asian monsoon evolution. *Nat. Commun.* **12**, 4022; 10.1038/s41467-021-24244-z (2021).

53. Wager, L. R. The rise of the Himalaya. *Nature* **132**, 28 (1933)

54. Pan, A. D. Rutaceae leaf fossils from the Late Oligocene (27.23 Ma) Guang River flora of northwestern Ethiopia. *Rev. Palaeobot. Palynol.* **159**, 188–194 (2010)

## Figure Legends

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

The alignment used for the 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 2). 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 3.

**Figure 2.** ML phylogenetic tree constructed using whole cp genome sequences of 45 species in the Aurantioideae subfamily.

The alignment used for the phylogenetic tree construction included a total of 80,986 nucleotide positions with 4,599 parsimony-informative sites, as calculated using MEGA11. In the analysis, the GTRGAMMAIX model was chosen as the optimal evolutionary model and applied using RAxML. The root position of this tree aligns with that found in the phylogenetic tree of the Rutaceae family (Supplementary Fig. S2). Bootstrap values, presented as percentages over 1,000 replicates, are shown at the nodes. Bootstrap values falling below 50% are not included. The scale bar denotes genetic divergence in terms of substitutions per site. Asterisks highlight sequences assembled *de novo* in this study. Different colors denote the subgroups identified in this study. Because the DNA sequences of *Atalantia monophyla* and *Atalantia spinosa*, along with those of *Citropsis schweinfurthii* and *Citropsis gilletiana*, were completely identical, we did not include the data for *Atalantia spinosa* and *Citropsis gilletiana* in this study. The alignment data in FASTA format and the tree data in Newick format are available in Supplementary Dataset 4.

**Figure 3.** ML phylogenetic tree, created using maximum likelihood analysis, illustrating the evolutionary relationships among 56 *Citrus* species based on whole cp genome sequences.

The sequence of *Atalantia buxifolia* served as the root. For the analysis, the GTRGAMMAIX model was selected as the most suitable evolutionary model and applied with RAxML. This analysis included a total of 88,414 nucleotide positions, encompassing 1,672 parsimony-informative sites, as calculated by MEGA11. Bootstrap values, expressed as percentages over 1,000 replicates, are indicated at the nodes, and the scale bar represents the number of substitutions per site. Bootstrap values below 50% are not displayed. Asterisks mark the sequences that were assembled *de novo* in this study. Different colors are used to distinguish the clusters identified in this study. Since the same species name exists, superscript numbers were used to distinguish individuals. The alignment data in FASTA format and the tree data in Newick format are available in Supplementary Dataset 5.

**Figure 4.** Comparative analysis of the inversion event in the *psbM* to *trnG-UCC* region of cp genomes. Asterisks indicate sequences that were assembled *de novo* in this study.

(A) Visualization of the inversion event was facilitated using Easyfig software. Red and light blue arrows denote the breakpoints. The gene arrangement of this region in *Atalantia monophylla* and *Citropsis gabunensis* is the same as that of the other Aurantioideae subfamily. (B) The breakpoint for the inversion event was identified within the nucleotide sequence, illustrated using a 116-bp segment. DNA sequences in dark blue represent 53 bp inverted repeat sequences located at both ends of this region. Notably, this event consistently occurred at the same nucleotide sequence position in *Atalantia roxburghiana* as well (Supplementary Fig. S5).

**Figure 5.** Time tree for the Rutaceae family.

It was constructed using MEGA11 software, calibrated with the 7.6 million years ago (MYA) divergence time between *Murraya paniculata* and *Murraya caloxylon*. *Ailanthus altissima* serves as an outgroup, depicted by a grey branch. The GTR+G+I model was selected as the most appropriate evolutionary model for the analysis. Horizontal blue bars represent error bars, while calibration constraints are marked on the corresponding nodes. The time scale is presented in MYA. Asterisks indicate sequences that were assembled *de novo* for this study. Different colors represent the subgroups identified in this study. The alignment data in FASTA format and the tree data in Newick format are available in Supplementary Dataset 6.

Citrus trifoliata
Citrus mangshanensis
100
Citrus cavaleriei
100
Citrus japonica
Citrus australasica
100
100
Citrus australis
100
Citrus glauca
Clymenia polyandra
100
True citrus fruit trees
Citrus maxima
100
Citrus reticulata
100
100
Citrus sinensis
100
Citrus micrantha
100
Citrus medica
Citrus aurantifolia
Atalantia racemosa
100
Atalantia bilocularis
100
Yaki Naran
Atalantia buxifolia
100
100
Atalantia ceylanica
Citreae
Atalantia roxburghiana
100
Atalantia monophylla
100
Citropsis gabunensis
100
Thanaka
100
Citropsis schweinfurthii
Naringi crenulata
100
Feroniella oblata
100
Wood apple
Limonia acidissima
100
Swinglea glutinosa
100
Afraegle paniculata
Bael fruit
100
Aegle marmelos
100
Paramignya lobata
100
Triphasia trifolia
Murraya caloxylon
Clauseneae 1
100
Murraya paniculata
Clausena lansium
100
Clausena anisata
100
Clausena harmandiana
Clauseneae 2
60
Clauseneae
Glycosmis citrifolia
100
100
Glycosmis pentaphylla
100
Murraya koenigii
Micromelum minutum
0.001

True citrus fruit trees
Yaki Naran
Wood apple
Thanaka
Bael fruit
Clauseneae 1
Clauseneae 2
Citreae
Clauseneae
Citrus japonica
Citrus mangshanensis
Citrus reticulata
Citrus cavaleriei*
Citrus aurantiifolia
Citrus micrantha
Citrus maxima
Citrus sinensis
Citrus trifoliata
Citrus medica
Clymenia polyandra*
Citrus australasica
Citrus australis*
Citrus glauca*
Atalantia ceylanica*
Atalantia ceylanica
Atalantia roxburghiana*
Atalantia roxburghiana
Atalantia monophylla
Atalantia monophylla*
Atalantia racemosa
Atalantia bilocularis*
Atalantia kwangtungensis
Atalantia buxifolia
Atalantia buxifolia*
Limonia acidissima*
Feroniella oblata*
Swinglea glutinosa*
Naringi crenulata*
Citropsis schweinfurthii*
Citropsis gabunensis*
Paramignya lobata*
Triphasia trifolia*
Aegle marmelos*
Afraegle paniculata*
Murraya caloxylon*
Murraya paniculata*
Murraya koenigii*
Clausena lansium*
Clausena excavata
Clausena anisata*
Clausena harmandiana*
Glycosmis pentaphylla*
Glycosmis citrifolia*
Micromelum minutum*
78
100
99
99
100
100
86
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
88
100
100
100
100
100
100
99
79
100
100
100
60
100
100
100
100
100
100
0.003

Citrus hongheensis
100
Citrus latipes*
Citrus aurantium1
100
100
Citrus limon2
Citrus aurantium2
Citrus sinensis
100
Citrus keraji
80
Citrus tarogayo
65
100
Citrus oto
Citrus limon1
100
Citrus platymamma
Citrus erythrosa
80
Citrus x paradisi1
Citrus maxima cv Liuyuezao
100
Citrus maxima
Citrus x paradisi2
60
Citrus maxima cv Guanxi miyou
Pummelo Cluster
100
96
Citrus micrantha
100
Citrus aurantifolia*
100
Citrus aurantiifolia
Papeda Cluster
Citrus cavaleriei*
100
Citrus cavaleriei
100
Citrus junos
Ichang Papeda /
Yuzu Cluster
100
Citrus mangshanensis
Citrus depressa Okitsu strain*
100
Citrus tachibana
99
Citrus depressa Isolate:oku
100
75
Citrus reshni*
100
Citrus limonia*
100
Citrus tangerina*
100
Citrus sunki
Citrus unshiu*
Citrus unshiu isolate Satsuma-L3
100
85
Citrus reticulata cv Succosa
Citrus reticulata
100
Citrus clementina
Mandarin Cluster
100
Citrus trifoliata2
100
Citrus trifoliata1
100
Citrus polytrifolia
Poncirus Cluster
Citrus japonica2
100
Citrus hindsii2
100
Citrus hindsii1
100
Citrus japonica1
100
Citrus japonica vari.margaritia*
89
Citrus madurensis
100
Citrus madurensis*
Fortunella Cluster
Citrus medica cv mac veu mountain citron*
100
Citrus medica cv buddha's hand citron*
100
Citrus medica cv humpang citron*
100
Citrus medica
100
Citrus medica cv corsica citron*
Citron Cluster
100
Clymenia polyandra*
Citrus glauca*
100
100
Citrus australis*
100
Citrus australasica*
100
Citrus australasica
Australian Cluster
Atalantia buxifolia*
1.0E-4

***Atalantia ceylanica*/ Naringi crenulata****
psbI
psbM
petN
trnC-GCA
rpoB
rpoC1
rpoC2
rps2
atpI
atpH
atpF
atpA
trnR-UCU
trnG-UCC
trnD-GUC
trnY-GUA
***Atalantia monophylla*/ Citropsis gabunesis****
psbK
psbI
trnG-UCC
trnR-UCU
atpA
atpF
atpH
atpI
rps2
trnG-UCC
rpoC2
rpoC1
rpoB
trnC-GCA
petN
psbM
trnD-GUC
trnY-GUA
5 Kbp
100%
82%
(A)
***Atalantia ceylanica* / Naringi crenulata****
GACTATTTCCATTCTTTTGATTGAATCAGGAATCGAAGTTTGTATTTGGTGTGTGGATAAAAG AGCTGCCTATGTTATACTATAGCTGCCTATGTTATACTATTGAATTCTCGACG
***Atalantia monophylla* / Citropsis gabunesis****
GACTATTTCCATTCTTTTGATTGAATCAGGAATCGAAGTTTGTATTTGGTGTGTGGATAAAAG TTATTTTGTCGAGCCATATCTGTCAAAATTCGTCCAACAAAAGAAATTGTTTT
***Atalantia ceylanica* / Naringi crenulata****
AAAACAATTTCGTTTGTTGGACGAATTTTGACAGATATGGCTCGACAAAATAA CTTTTATCCACACACCAAATACAAACTTTGATTCCTGATTCAATCAAAAGAATTCCTTTACTT
***Atalantia monophylla* / Citropsis gabunesis****
TGACTGTCAAGTCGATTCATCGTCGAGAATTCAATAGTATAACATAGGCAGCT CTTTTATCCACACACCAAATACAAACTTTGATTCCTGATTCAATCAAAAGAATTCCTTTACTT
(B)

True citrus fruit trees
Citrus aurantiifolia
0.08
Citrus micrantha
2.07
Citrus maxima
0.20
Citrus sinensis
2.92
Citrus reticulata
1.68
Citrus cavaleriei*
3.11
Citrus mangshanensis
3.26
Citrus trifoliata
Citrus japonica
3.70
Citrus medica
2.98
Clymenia polyandra*
2.64
Citrus australasica
1.59
Citrus australis*
1.37
Citrus glauca*
9.36
Wood apple
Limonia acidissima*
3.84
Feroniella oblata*
Yaki Naran
9.08
Atalantia ceylanica*
6.13
Atalantia roxburghiana*
7.75
Atalantia monophylla*
9.67
Atalantia racemosa
7.33
Atalantia bilocularis*
6.04
Atalantia buxifolia*
1.54
Atalantia buxifolia
0.03
Atalantia kwangtungensis
Citreae
Thanaka
10.59
Naringi crenulata*
5.73
Citropsis gabunensis*
4.51
Citropsis schweinfurthii*
8.91
Swinglea glutinosa*
12.30
Paramignya lobata*
10.30
Triphasia trifolia*
Bael fruit
9.99
Aegle marmelos*
5.14
Afraegle paniculata*
Clauseneae 1
14.25
Murraya caloxylon*
8.01
Aurantioideae
Murraya paniculata*
Clauseneae 2
7.17
Glycosmis citrifolia*
13.37
Glycosmis pentaphylla*
Micromelum minutum*
14.02
Clauseneae
31.70
Murraya koenigii*
13.05
Clausena harmandiana*
9.22
Clausena lansium*
8.54
32.50
Clausena anisata*
8.32
Clausena excavata
38.17
Amyridoideae
Cneoridium dumosum
Haplophylloideae
Haplophyllum dauricum
Rutaceae
41.23
Rutoideae
Boenninghausenia albiflora*
Zanthoxyloideae
Zanthoxylum piperitum
Ailanthus altissima
40
30
20
10
0
Divergence Time
(MYA)
EOCENE
OLIGOCENE
MIOCENE
PLIOCENE
PLEISTOCENE
HOLOCENE
CENOZOIC