# An Improved Assembly Strategy Helps Parsing the Cryptic Mitochondrial Genome Evolution in Plants

**Yanlei Feng** ( fengyanlei@westlake.edu.cn )
Westlake Institute for Advanced Study https://orcid.org/0000-0002-3015-4970
**Xiaoguo Xiang**
Nanchang University
**Delara Akhter**
Sylhet Agricultural University
**Zhixi Fu**
Sichuan Normal University
**Ronghui Pan**
Zhejiang University
**Xiaohua Jin**
Institute of Botany Chinese Academy of Sciences

**Research Article**

**Keywords:** Plant mitochondrial genome, Fagales, Horizontal transfer, Evolution, Genome expansion

**Posted Date:** June 3rd, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-512282/v1

# Abstract

Although plant mitochondrial genomes (mitogenomes) are small, they exhibit considerable complexity not seen in other eukaryotic mitogenomes. Assembly and analysis of plant mitogenomes is hampered by their large variations in structure and size, and the mitogenome remains the last genome to be deciphered in many plant species. As a result, very few plant mitogenomes have been assembled and little is known regarding their evolution. In this study, a strategy was devised for assembly of mitogenomes from existing short reads from whole-genome sequencing projects. The strategy combined current tools and manual steps to resolve the two main challenges to mitogenome assembly: repeat and plastid insertion sequences. High-quality complete mitogenomes were assembled for 23 species from five families of the Fagales. Mitogenomes varied 2.4 times in size. The largest, *Carpinus cordata*, did not contain large amounts of unique sequences, but instead contained a high proportion of sequences homologous to other Fagales. Further analysis of the Fagales mitogenomes revealed highly mosaic characteristics, with horizonal transfer (HGT)-like sequences identified from almost all seed plant taxa. Independent and unequal transfers of third-party DNA may partially account for the HGT-like fragments and unbalanced size expansions observed in Fagales mitogenomes. Supporting this, a mitochondrial plasmid of nuclear origin was found in *Carpinus*, and this may represent an intermediate stage prior to incorporation into the mitogenome. The approaches used in this study are widely applicable and provide new insights into the mechanisms of mitogenome evolution in plants.


# Introduction

Organelle genomes such as mitochondrial genomes (mitogenomes) and plastid genomes (plastomes) are generally small in size and have conserved structures. However, despite the ubiquity of mitochondria across most eukaryotes, the mitogenome in angiosperms (seed plants) exhibits several unique features. The angiosperm mitogenome is highly expanded compared with other organisms, and size also varies significantly among species, from 200 Kb up to 11 Mb (Sloan et al., 2012; exception see Skippington et al., 2015). Expansions, duplications, plastid-derived insertions (referred to as mitochondrial plastid insertions, MTPTs), and horizontal gene/DNA transfers (HGTs) are common. DNA double-strand breaks are rampant and the mitogenome structure is extremely dynamic, with chromosomes in several species exhibiting linear, multipartite, branched, or a combination of several structure types (Alverson et al., 2011a; Davila et al., 2011; Cheng et al., 2017; Christensen, 2018; Kozik et al., 2019). Variability is extremely common, with even close relatives or individuals of the same species exhibiting mitogenome differences.

The unique characteristics of plant mitogenomes hinder the production of complete and high-quality assemblies. For plastomes, published relatives can be used as reference genomes during assembly. However, the poor conservation and uncertain chromosome structures of plant mitogenomes preclude this strategic approach. Similarly, *de novo* assembly is challenging, in part due to the presence of dispersed repeats that result in fragmented contigs during assembly. Furthermore, the usual DNA source for plant sequences is foliage and, as a consequence, plastome coverage is often higher than mitogenome coverage. This further complicates attempts to understand the precise MTPT content in mitogenomes.

Several software packages have been used for the analysis of plastomes and animal mitogenomes, such as GetOrganelle, Novoplasty, and mitoZ, and bacterial genome tools such as Unicycle can also be used (Dierckxsens et al., 2017; Wick et al., 2017; Meng et al., 2019; Jin et al., 2020). However, these packages have proved inadequate for the analysis of plant mitogenomes, and no specialized tools are available for assembly of plant mitogenomes.

The difficulties in assembling and analyzing plant mitogenomes has hampered progress in this area of research. In angiosperms, far fewer sequenced mitogenomes are available compared with plastomes and nuclear genomes. At the species level, no more than 300 mitogenomes are available (NCBI), compared with more than 4000 plastomes (NCBI) and more than 500 nuclear genomes (https://www.plabipd.de), as of January 2021. In many plant species, the mitogenome is the last genome remaining to be deciphered. Mitogenome publications to date have usually focused on one or a few species, and large-scale mitogenome comparisons remain rare, leaving questions of mitogenome evolution unanswered. Recent research has adopted PacBio or Nanopore long-read sequencing technologies, which offer some advantages over short-read Next-Generation Sequencing (NGS) systems. However, despite the longer read lengths, long repeats and long MTPTs remain challenging to resolve even with the newer sequencing technologies, and the high costs limit sampling capacity. Ongoing research with plastomes and nuclear plant genomes has accumulated abundant NGS data. If an appropriate analysis pipeline can be developed, these existing NGS datasets could be used to obtain high-quality mitogenomes with reduced cost and time inputs.

The Fagales order of flowering plants belongs to the Rosid clade of the Eudicotidae. Fagales contains more than 1,000 species in seven families and 33 genera, according to the Angiosperm Phylogeny Group (APG) system (Sennikov et al., 2016). Many Fagales species are important in food supply and ecosystem support. Several well-known tree species belong to this order, including beeches, oaks, and birches, and many provide nuts and fruits, such as walnuts, chestnuts, hazels, and bayberries. The Fagales group also includes nitrogen-fixing species among its many diverse and ecologically dominant species. At least 24 Fagales genomes have been sequenced (https://www.plabipd.de), and 150 Fagales plastomes have been released. Despite this, only three mitogenomes have been produced to date, for *Betula pendula*, *Quercus variabilis*, and *Fagus sylvatica*. The *B. pendula* mitogenome was derived from a whole-genome sequencing (WGS) study, but only minimal information regarding the mitogenome was included (Salojärvi et al., 2017). The *Q. variabilis* mitogenome was similarly sparsely described (Bi et al., 2019). A detailed *F. sylvatica* mitogenome was published recently and is discussed in the context of our research (Mader et al., 2020). For many plant genomes, mitogenomes are the only remaining unexplored genetic material, and are crucial for understanding the adaptive capacity and genetic and genomic resources of these species.

In this study, the difficulties of mitogenome assembly with short reads are discussed and a strategy to address these problems is explored. Raw short reads from the NCBI Sequence Read Archive (SRA) were used to assemble mitogenomes of 23 Fagales species, including 16 genera from five families, covering almost half of the total Fagales genera and 71% of the total families, respectively. Differences among the resulting high-quality complete mitogenomes were assessed, particularly with regard to size variation. This

is the most comprehensive study of mitogenomes to date, comprising the largest number of new and complete angiosperm mitogenomes yet produced.

# Materials And Methods

*Sequenced data acquisition.*

Fagales DNA sequences were retrieved from the NCBI SRA database. Data for 23 species from 16 genera and five families were selected for assembly (Table S1). All the data were from whole-genome sequencing (WGS) datasets, meaning that reads included sequences from the nuclear, mitochondrial, and plastid (including chloroplast) genomes. Organelle genome sequences are usually smaller than nuclear sequences but are present at much higher copy numbers. Therefore, relatively small amounts of data are needed to obtain mitogenomes and plastomes (Table S1). *Genome assembly*

Raw reads of each species were filtered for low-quality bases using TRIMMOMATIC v0.36 (Bolger et al., 2014). Clean reads of approximately 2–4 Gb were used for do novo assembly with SPADES v3.13 (Bankevich et al., 2012; Table S1). Plastid contigs of *Casuarina equisetifolia*, *Lithocarpus fenestratus*, and *Quercus suber* were obtained by BLASTN v2.9.0 (Camacho et al., 2009) searches of all assembled contigs against the B. pendula plastome (GenBank ID: NC_044852). Clean reads were then mapped to plastid contigs using GENEIOUS R10 (Biomatters, Inc.), and contigs were extended and connected manually until joined. Inverted repeat (IR) boundaries were identified by searching repeats using the GENEIOUS “Repeat Finder” plugin. Mitogenomes are often more variable than nuclear genomes in terms of their content and structure. Preliminary mitogenome contigs were identified from total contigs by BLASTN with the *B. pendula* mitogenome (GenBank ID: LT855379) as a reference (word size: 16, evalue: 1e-20). All hit sequences longer than 500 bp were extracted. Two subsequent strategies were used to improve completeness and sequence content. For completeness, contigs were annotated using the GENEIOUS “Annotate from Database” function, where the “Database” comprised all known mitochondrial genes. If known mitochondrial genes were absent, reads were mapped to the gene to check coverage and confirm presence or absence. Mitochondrial genes that were missing from preliminary contigs were used to search all contigs and identified contigs were added to the preliminary mitogenome contig set. This strategy ensured that gene sets for mitogenome assemblies were complete. For DNA content, clean reads were mapped back to the selected contigs in GENEIOUS. Plastid (higher coverage) and other contigs (unbalanced coverage) were removed from the set to provide approximate mitogenome coverage, which was then used to bait other potential mitochondrial contigs from all contigs. New selected contigs were mapped back to reads and non-mitochondrial contigs were removed as before. This strategy reduced the amount of missing sequences and ensured that mitogenome assemblies were as complete as possible.

Next, the comprehensive mitochondrial contig sets were joined together. Contigs normally ended with repeat and/or MTPT sequences. Repeats longer than 50 bp in contigs were found using GENEIOUS “Repeat Finder”, and paired reads were mapped to contigs. Repeat regions were identified and resolved using sequencing coverage. Connections of long repeats may introduce artificial rearrangements (Fig. S1). MTPTs are very similar to plastome sequences and it is not usually possible to assemble MTPTs directly

into contigs. MTPTs can be identified on plastomes (or plastid contigs); however, unlike repeat sequences, MTPT regions cannot be easily resolved by coverage as plastome coverage is usually much higher than mitogenome coverage. Repeats were filled and contigs were connected at both ends, after which MTPT ends were mapped to the plastome. After plastome mapping, the closest ends in the same orientation were most likely derived from the same MTPT. Rearrangements or recombination can occur within MTPTs, resulting in extended distances between sequences or opposite orientations with respect to plastome mapping (Fig. S2a). In these circumstances, paired reads could be used to identify the correct connections (Fig. S2b-d). MTPTs and their plastid counterparts may not be 100% identical, and additional steps were needed to correct MTPTs identified in the previous step (Fig. S3). Reads were either re-mapped and divergent bases manually checked and corrected (Fig. S3c), or reads that were 100% identical to the plastome were filtered, with the remaining “Used reads” re-mapped to mitogenomes to enhance the identification of divergent bases.

Several iterations of the map-check-connect strategies outlined above were usually sufficient to resolve all the repetitive and MTPT ends and retrieve one or more circular chromosomes. As a last step, paired-end reads were re-mapped a final time to check and correct any misassemblies and ensure that all bases were correct.

*Annotation*

Putative mitochondrial protein-coding and rRNA genes were annotated by similarity to known mitochondrial genes using GENEIOUS, followed by manual corrections. The *B. pendula* mitogenome was also annotated, and tRNA genes were predicted using tRNAscan-SE v2.0 (Chan and Lowe, 2019). Coding genes with disrupted reading frames, premature stop codons, or non-triplet frameshifts were annotated as pseudogenes.

MTPTs were determined by BLASTN comparison to a collection of plastomes. Hits smaller than 100 bp were masked. Dispersed repeats within the genome were searched by BLASTN against itself. Hits with identity less than 95% were filtered. Repeat lengths were determined using a custom Perl script. Only one part of each repeat pair was calculated and overlapping bases were counted only once.

Mitogenome annotations and syntenies were plotted using CIRCOS v0.69 (Krzywinski et al., 2009). Links were searched by BLASTN with default parameters, and hits shorter than 500 bp were excluded. Homologs within species were masked and only those between species were used.

*Phylogeny*

Although the substitution rate of mitochondrial genes is very low (Palmer and Herbon, 1988), mitogenomes contain hundreds of RNA editing sites (Small et al., 2020). These sites can affect phylogenetic analysis and are often removed during analysis to reduce their impact. There are relatively few plastid RNA editing sites, and these can be ignored during analysis due to their low impact. Mitochondrial protein-coding sequences (CDS) were extracted and aligned using MAFFT software with “auto” mode. RNA editing sites were predicted using the PREP website (Mower, 2009). Identification of an editing site within a codon

prompted the removal of corresponding codons in all species. After editing site removal, aligned genes were concatenated into one matrix and used to build a maximum likelihood tree on CIPRES v3.3 (Miller et al., 2010) with RAxML-HPC2 on XSEDE (Stamatakis, 2014), with the GTRGAMMA model and 1000 bootstrap iterations. Plastid CDS were aligned and used to construct a similar tree, without removal of RNA editing sites. The mitochondrial and plastid matrices were 31,551 and 69,243 bp in length with 750 and 6,495 parsimony-informative characters, respectively. Both the mitochondrial and plastid trees (saved in Fig. S7) in this analysis showed a polyphyletic relationship for *Quercus*, which was consistent with previous studies (Manos et al., 2008; Simeone et al., 2016). The mitochondrial tree was poorly supported, and the plastid tree was used for discussion in this study.

*Genus-specific DNA Analysis*

BLAST was used to compare mitogenomes to a database comprising all Fagales mitogenomes, with e-value 1e-5 and word size 16. Genus-specific sequences (*i.e.*, sequences present only within the specific Fagales genus) longer than 300 bp were isolated using a custom Python script. *Quercus* species exhibited non-monophyletic relationships (Fig. S7), and *Q. suber* and *Q. variabilis* were considered as a single genus, with *Q. robur* excluded from the analysis. BLAST was used to compare the genus-specific sequences to the NCBI nr database, with parameters as before, and the first 100 hits were saved. The best hit or hits for each genus-specific sequence were examined (more than one best hit was possible if sequences matched different targets) using a custom Python script. Only best hits longer than 100 bp were used, and MTPTs were removed from the results. Subsequently, the best hits were grouped into orders and a face-to-face tree plotted in R using the ape package cophyloplot function (Paradis et al., 2004). Connections were colored using RColorBrewer (https://colorbrewer2.org/) and orders were positioned with reference to the Angiosperm Phylogeny Group website (Stevens, 2001 onwards).

# Results

*The necessity of an appropriate assembly strategy for mitogenomes*

Genomic researchers have used different strategies for mitogenome analysis, including *de novo* assembly, mapping to reference genomes, and seed-mapping reads, and these have produced mitogenome sequences of variable quality. A survey of NCBI mitogenome data showed that, in some assemblies, MTPTs and repeat sequences were unfeasibly long compared with other individuals of the same species or close relatives of the same genus (Fig. 1). Although some differences among individuals of the same species have been observed, these differences are minimal for most species. The unusually long repeat sequences in some assemblies are also unlikely given the error-prone nature of MTPT and repeat assembly, which increases the chance of misassemblies. However, unfortunately, in most studies, mtDNA content was not conserved and misassemblies were undetectable, including for ribosomal protein genes and succinate dehydrogenase subunit genes *sdh3* and *sdh4*. Some poorly assembled mitogenomes included artificial structures, such as inappropriate circularization, or had missing sequences, such as absence of ribosomal RNA genes *rrnS* and *rrnL*. In addition, assembly difficulties prompted some researchers to focus on analysis of fragmentary sequences without producing complete assemblies. The

full scope of mitogenome evolution remains obscure in these cases, and it is challenging to reuse data when mitogenomes have been assembled and analyzed with a variety of methods. Improved assembly methods are needed to address the issues of existing assemblies.

*Improvement of assembly*

In most genome sequencing projects, total DNA is directly sequenced without separation of organellar and nuclear material. At present, mitochondrial sequence reads cannot be easily distinguished in whole-genome datasets. In our novel strategy, initial *de novo* assembly was followed by identification of mitochondrial contigs through gene identification and sequence coverage. Experience from our previous work and published mitogenomes suggests that angiosperm mitogenomes usually lack AT-rich regions so the read coverage is generally balanced. Mitochondrial contig ends occur when repeat sequences or MTPTs are encountered during assembly, *i.e.*, contigs usually end either with repeats (repetitive ends) or MTPTs (MTPT ends). Our strategic workflow is outlined in Fig. 2. First, sequencing coverage was used to resolve repeat ends. Next, all MTPT ends were mapped to plastome and the highest numbers of potential connections were identified using their positions and directions. Circular mitogenomes were produced once all repetitive and MTPT ends were resolved. Finally, clean reads were mapped to the correct MTPTs (Fig. S3) and the final assembly.

*Assembly results and completeness assessment*

Our mitogenome assembly approach focused on solving issues caused by repeat and MTPT sequences in 23 Fagales species. Sequences of 2–3 Gb in size were used and coverage depth was 33–174. Of the 23 species, 13 yielded one or more circular mitogenomes, and the remaining 10 species contained one or more linear chromosomes (Table 1). Mitogenomes housed in multiple circular chromosomes did not share long repeats between the structures. In principle, circular chromosome assemblies can be achieved if all the repetitive and plastid sequence ends are connected. Sometimes, however, repetitive ends appeared unstable, with decreasing coverage towards the ends, or the paired end for an MTPT end could not be found. In these cases, ends could not then be connected properly (Fig. S4; Table S2). The sequencing datasets were derived from several different studies, and assemblies might therefore have been affected by the sequencing and library type, and read and insert lengths. For example, coverage of *B. platyphylla* and *Lithocarpus*, both produced by an Illumina Genome Analyzer, was much lower in some regions than others.

During this study, the mitogenome of *Fagus sylvatica* was published. Long and short reads were used to produce an assembly with a single circular chromosome of 504,715 bp in length (Mader et al., 2020). The sequence content of the published assembly was almost identical to that of the *F. sylvatica* assembly produced in this study, differing only in two bases (differences between individuals rather than differences in assembly). The only disparity between the two assemblies was an inversion of a sequence located between 900 bp repeats (Fig. S1). The repeat region was much longer than the insert length (450 bp; Table S1), and this inversion was therefore not unexpected. The consistency between our assembly and that of the previous study provided support for the practicability and reliability of our assembly methods.

Furthermore, the mitogenomes in the two independent *F. sylvatica* projects were almost identical, indicating preservation of mitogenomes among individuals in at least some plant species.

*Mitogenome size and content*

Characteristics of the mitogenome assemblies produced in this study, as well as previously published *B. pendula* and *Q. variabilis* assemblies, are provided in Table 1. Mitogenome sizes in Casuarinaceae, Fagaceae, and Myricaceae resembled those of distant relatives from Rosales or Fabales (400 Kb and 480 Kb on average, respectively, NCBI data). By contrast, mitogenome sizes were substantially expanded in Betulaceae and Juglandaceae. The largest mitogenome (922 Kb) was found in *Carpinus* (Betulaceae), and was much larger than those of confamiliar species. DNA content of mitogenomes did not differ substantially within families, but structures were often highly rearranged (Fig. S5). Mitogenome sequences were less similar between families, with some sequences having no homologs in other families (Fig. 3). The proportion of repeats in Fagales mitogenomes was small, normally less than 3% and no more than 6.2% of the total mitogenome length (Table 1). In Betulaceae, short repeats of less than 200 bp were more apparent, especially in *Alnus* (Table S3). MTPT percentages were also low, with only two species having more than 6% (*Casu. equisetifolia*, 13.5%; and *Corylus*, approximately 9.5%).

Conservation of some ribosomal protein genes was poor (Fig. S6), as in many plant species. Five of the seven Betulaceae species had *rps11* sequences with approximate identities of 100%. Comparison of Betulaceae *rps11* sequences with those in the NCBI nr database indicated similarities with *rps11* in monocots or basal core angiosperms such as *Triantha glutinosa* (KX808303, Alismatales) and *Liriodendron tulipifera* (NC_021152, Magnoliales), consistent with previous research (Bergthorsson et al., 2003). These similarities suggested that HGT of *rps11* may have occurred in a common Betulaceae ancestor, followed by differential losses in some species. Exon 4 of *nad1* (*nad1*e4), *matR*, and *nad1*e5 form a colinear block in most angiosperms. This block was disrupted between *matR* and *nad1*e5 at least twice in Fagales species but, surprisingly, was recovered in *Juglans sigillata* and *J. regia* (Fig. S6).

*Identification of a mitochondrial plasmid*

A small circular mitochondrial plasmid, 2,888 bp in length, was found in *Carpinus*. Sequencing coverage was similar to that of the mitogenome. Plasmid GC content was 37.6%, which was much lower than normal mtDNA (Table 1) but was similar to *Carpinus* nuclear genomes (*Car. fangiana*: 37.6%, Yang et al., 2020). With the exception of a small 240 bp plastid-like region, the plasmid had no sequence similarities with angiosperm mitogenomes. The plasmid was fully encompassed by *Car. avellana* or *Car. fangiana* nuclear sequences from different chromosomes. Two large open reading frames (ORFs), ORF244 (732 bp) and ORF162 (486 bp), were found on the plasmid. BLASTP comparison against the nr database identified homologs of ORF244 in several angiosperm species, including a nearly full-length match in Arabidopsis (AT1G74875, identical 34%). ORF244 homologs were annotated as putative F-box proteins and homologs of ORF162 were annotated as DNA methylation 4 factors in several Rosids. It was unclear whether the two plasmid ORFs were expressed, but there was sufficient evidence to conclude that the plasmid was of nuclear origin.

*Genus-specific mitogenome sequences and mosaic origins*

Repeat and MTPT sequences were not solely sufficient to explain the substantial size variation observed among mitogenomes from different species (Table 1). Genus-specific sequences were identified (*i.e.*, sequences with no homologs in other Fagales genera) and used to explore the causes of metagenome size divergence. *Quercus* species were found to have non-monophyletic relationships (Fig. S7), and *Q. robur* was not included with other *Quercus* species when identifying *Quercus*-specific sequences. Surprisingly, the species with the largest mitogenome, *Carpinus*, did not contain correspondingly long genus-specific sequences. By contrast, *Casuarina* species, which had relatively small mitogenomes, had the most unique sequences (Table S5). Plant mitogenomes are prone to absorbing foreign DNA, which might therefore be the source of additional sequences in large mitogenomes. Genus-specific sequences were used to search the NCBI nt database, and best-hits were assessed by order and compartment (Fig. 4; Table S4-S5). Overall, the genus-specific sequences were related to a range of seed plant lineages and were mainly of mitogenomic origin (Fig. 4).

Mitovirus-like sequences were found in several of the 23 Fagales species, including nearly full-length sequences in two *Betula* species and an approximately 1500 bp sequence in *Castanea* (Fig. 5). Mitoviruses, which belong to the Narnaviridae family, are positive single-stranded RNA viruses that replicate in host mitochondria. Mitovirus genomes are small, approximately 2.1–4.4 Kb in length, and contain a single ORF encoding a viral RNA-dependent RNA polymerase (RdRP) required for replication (Nibert, 2017). The phylogeny of Fagales mitovirus-like sequences is incongruent with the species tree (Fig. 5), indicating that these sequences were not introduced into Fagales via a single event.

Fagales belong to the nitrogen-fixing lineage of angiosperms, and at least three genera in this study have nitrogen-fixing capacity: *Casuarina*, *Morella*, and *Alnus* (Yelenik and D'Antonio, 2013; Huisman and Geurts, 2020). However, there was no indication that these genera contained more sequences similar to bacteria than other Fagales species.

# Discussion

*A strategy helps precise assembly*

Plant mitogenomes can be used to explore several aspects of molecular evolution, including structure variation, genome expansion, intron splicing, and DNA double-strand breaks and repairs. However, assembly of plant mitogenomes is challenging due to their considerable size variability and their often unusual structural architectures. In this study, we developed a strategy to address the assembly difficulties and used the process to obtain complete high-quality mitogenomes. Visual processes in the powerful GENEIOUS software allowed full verification of every base. Our assemblies showed that, despite the complex mitogenome structures observed under microscopy (Backert and Börner, 2000; Manchekar et al., 2006; Cheng et al., 2017), the majority of mitogenomes were composed of one or more circular chromosomes. The assemblies in this study were compiled from existing NGS short-read datasets produced using different sequencing technologies and with different read and insert lengths. This

demonstrated the utility of our assembly strategy and showed that the abundant existing short-read datasets can be mined to produce new assemblies without the need for additional expensive sequencing.

One disadvantage of short reads is their inability to process long repeats. In our mitogenome assemblies, connections proximal to long repeats are likely to be pseudoconnections or are only representative of one potential type (Fig. S1). Plant mitogenomes experience frequent rearrangements through their long repeats (Kozik et al., 2019), and it is thus unclear whether these mitogenomes can be considered to have a standard structure. One way to address potential misassemblies is by checking completeness of gene colinear blocks (Chaw et al., 2008), such as the *nad1*e4-*matR*-*nad1*e5 block in Fagales, and these strategies were used in our assemblies.

Intracellular gene transfer (IGT) between genome compartments is a common phenomenon. Interactions between nuclear and mitochondrial genomes may occur frequently, producing structures like the small nuclear-derived plasmid found in *Carpinus*. These types of interaction may explain mitogenome expansion and size variability, as well as explaining nuclear gene innovation (O'Conner and Li, 2020). Unfortunately, despite recent increases in genomic studies and new developments in sequencing technologies, precise assembly of IGTs remains unexplored in the majority of projects. Nuclear genomes almost always contain mitochondrial contigs (*e.g.*, Alverson et al., 2011b), and our MTPT assembly method may provide a baseline to develop strategies for assembling other IGT sequences.

*Evolution of plant mitogenomes*

Size variation between close species is a common feature of plant mitogenomes and has been observed in a range of taxa, such as, *Viscumalbum* and *V. scurruloideum* (565 Kb *vs.* 66 Kb; Petersen et al., 2015; Skippington et al., 2015), *Sliene conica* and *S. noctiflora* (11.1 Mb vs. 6.7 Mb; Wu et al., 2015; Wu and Sloan, 2018), *Cucumis melo* and *C. sativus* (2.7 Mb vs. 1.7 Mb; Alverson et al., 2011a; Rodríguez-Moreno et al., 2011). The reasons for this size variability may be complex. Duplications, IGT events, and introduction of foreign DNA all contribute to mitogenome size expansion (Alverson et al., 2011a; Rice et al., 2013). In Fagales, the mitogenome of *Carpinus* is notably larger than those of close relatives. However, lengths of repeats, MTPTs, and genus-specific sequences were insufficient to fully explain the size divergence. Another possibility is that the *Carpinus* mitogenome has an unusually high number of homologs with other Fagales, indicating increased acquisition or decreased loss of mitogenome sequences compared with other Fagales. Homolog searches between *Carpinus* and other lineages confirmed a high number of homologs in *Carpinus*, raising questions regarding the ancestral mitogenome in Fagales. One possibility is that the ancestral mitogenome was similarly large to that of *Carpinus*, and sequences were then lost independently in different lineages during evolution. This model was used to explain mitogenome size variation in kiwifruits (Wang et al., 2019). However, it appears unlikely that all Fagales genera other than *Carpinus* would experience such large and variable sequence losses, suggesting that sequence transfer may be a more likely scenario for Fagales.

Mitochondrial plasmids are small autonomous circular or linear extrachromosomal DNA molecules in mitochondria, and these have been found in several species, including maize, rice, and carrot (McDermott

et al., 2008). The origins and functions of mitochondrial plasmids remain unclear. The small mitochondrial plasmid found in *Carpinus* was of nuclear origin, and may represent an intermediary stage prior to incorporation into the chromosomal mitogenome.

Genus-specific sequence analysis showed that Fagales mitogenomes exhibited mosaic characteristics. All species had sequences that were most similar to those of distant seed plant taxa, including those of gymnosperms. Some other plant orders that are distantly related to Fagales also have mitogenome sequences that are similar to those in Fagales mitogenomes. *Amborella* contain HGT sequences from many species, including Fagales (Rice et al., 2013), and we found that these HGTs were mainly shared with Casuarinaceae. As genus-specific sequences were used in the analysis, the direction of HGT events remain unclear. Parasitic plants are notable for their acquisition of large HGT sequences from their hosts (Bellot et al., 2016; Sanchez-Puerta et al., 2017; Kovar et al., 2018; Sanchez-Puerta et al., 2018). However, in our analysis, parasitic plants had similar amounts of Fagales HGT-like sequences than non-parasitic species (Table S4). Previous research proposed a “wounding-HGT model” to explain HGT events between non-parasitic plants and “mitochondrial fusion occurs in a fundamentally similar manner” (Rice et al., 2013). These hypotheses may also be relevant for Fagales as no sequences appeared to be derived from other cellular organisms besides seed plants, even though some species were symbionts with nitrogen-fixing bacteria. Alternatively, the HGT-like sequences in angiosperms may be similar to the expanded set of homologs observed in *Carpinus*, suggesting alternative means of sequence origin and dispersal.

Active mitoviruses have only been identified in fungi to date. However, mitovirus sequences, particularly those corresponding to the RdRP region, are widespread in plant nuclear and mitochondrial genomes (Alverson et al., 2011a; Bruenn et al., 2015; Nibert, 2017; Silva et al., 2017; Chu et al., 2018; Nibert et al., 2018; Charon et al., 2020). Plant mitovirus-like sequences are thought to be derived from plant pathogenic fungal interactions and HGT events (Bruenn et al., 2015). However, direct HGT from fungal to plant mitogenomes is unlikely, as incompatibility hampers fusion between mitochondria in fungi and plants (Rice et al., 2013). An alternative pathway is transfer from fungi to the plant nuclear genome, and then from the nucleus to the plant mitogenome. To assess this, the full-length mitovirus sequence from the *Betula* mitogenome was used to search the *B. nana* and *B. pendula* nuclear genomes (Wang et al., 2013; Salojarvi et al., 2017), but only short sections were found, and only in *B. nana*. Furthermore, phylogenetic analysis of mitovirus-like sequences in Fagales do not support common origins (Fig. 5). It is therefore possible that mitoviruses can infect plants directly and frequently.

In conclusion, the “third-party” DNA, including mitoviruses and nuclear insertions, may account partially for the mosaic composition of plant mitogenomes. If two species get DNA from the same source, sometimes can make an illusion that similar sequences are shared with far-away lineages; if different content was transferred in independent events, some species may share more homologs with others, like the situation in *Carpinus* (Fig. 6). Since the transfers between the third-party and mitogenomes could happen independently and not limited to time, and mitogenomes themselves also encountered continuous rearrangements and deletion, from time to time it would finally create extremely mosaic mitogenomes.

# Declarations

*Funding*

This work was supported by the Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (grant number 2019R01002) and Westlake Postdoc project (grant number 101196582003).

*Conflicts of interest*

The authors declare that they have no conflict of interest.

*Availability of data and material*

The assembled sequences have been deposited to CNGB Sequence Archive of China National GeneBank DataBase (CNGBdb, https://db.cngb.org/) under Project CNP0001491 (mitogenomes: accessions N_000011064 - N_000011115; plastomes: accessions N_000011061 - N_000011063; *Carpinus* mitochondrial plasmid: accession N_000011116).

*Contributions*

YF and XJ designed the project. YF assembled and annotated the genomes. XX finished the phylogeny, YF did all rest analyses. YF wrote the manuscript, and XJ supervised the writing. DA, ZF and RP polished the language and provided suggestions about the content.

*Code availability*

The used scripts can be found in Github (https://github.com/fengyanlei33/Fagales_mitogenome).

*Ethics approval*

Not applicable

*Consent to participate*

Not applicable

*Consent for publication*

All authors approved the final version of the manuscript.

**Acknowledgements**

We gratefully acknowledged Xingxing Shen (Zhejiang University) for his valuable comments and suggestions.

# References

1. Alverson, A.J., Rice, D.W., Dickinson, S., Barry, K., and Palmer, J.D. (2011a). Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. Plant Cell *23*, 2499-2513.https://doi.org/10.1105/tpc.111.087189
2. Alverson, A.J., Zhuo, S., Rice, D.W., Sloan, D.B., and Palmer, J.D. (2011b). The mitochondrial genome of the legume Vigna radiata and the analysis of recombination across short mitochondrial repeats. Plos One *6*, e16404.https://doi.org/10.1371/journal.pone.0016404
3. Backert, S., and Börner, T. (2000). Phage T4-like intermediates of DNA replication and recombination in the mitochondria of the higher plant Chenopodium album (L.). Curr Genet *37*, 304-314.https://doi.org/10.1007/s002940050532
4. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D*., et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol *19*, 455-477.https://doi.org/10.1089/cmb.2012.0021
5. Bellot, S., Cusimano, N., Luo, S., Sun, G., Zarre, S., Groger, A., Temsch, E., and Renner, S.S. (2016). Assembled Plastid and Mitochondrial Genomes, as well as Nuclear Genes, Place the Parasite Family Cynomoriaceae in the Saxifragales. Genome Biol Evol *8*, 2214-2230.https://doi.org/10.1093/gbe/evw147
6. Bergthorsson, U., Adams, K.L., Thomason, B., and Palmer, J.D. (2003). Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature *424*, 197.https://doi.org/10.1038/nature01743
7. Bi, Q., Li, D., Zhao, Y., Wang, M., Li, Y., Liu, X., Wang, L., and Yu, H. (2019). Complete mitochondrial genome of Quercus variabilis (Fagales, Fagaceae). Mitochondrial DNA Part B *4*, 3927-3928.https://doi.org/10.1080/23802359.2019.1687027
8. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.https://doi.org/10.1093/bioinformatics/btu170
9. Bruenn, J.A., Warner, B.E., and Yerramsetty, P. (2015). Widespread mitovirus sequences in plant genomes. PeerJ *3*, e876.https://doi.org/10.7717/peerj.876
10. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST plus : architecture and applications. BMC Bioinformatics *10*.https://doi.org/10.1186/1471-2105-10-421
11. Chan, P.P., and Lowe, T.M. (2019). tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. Methods Mol Biol *1962*, 1-14.https://doi.org/10.1007/978-1-4939-9173-0_1
12. Charon, J., Marcelino, V.R., Wetherbee, R., Verbruggen, H., and Holmes, E.C. (2020). Meta-transcriptomic detection 1 of diverse and divergent RNA viruses in green and chlorarachniophyte algae.https://doi.org/10.1101/2020.06.08.141184
13. Chaw, S.M., Shih, A.C., Wang, D., Wu, Y.W., Liu, S.M., and Chou, T.Y. (2008). The mitochondrial genome of the gymnosperm Cycas taitungensis contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol Biol Evol *25*, 603-615.https://doi.org/10.1093/molbev/msn009

14. Cheng, N., Lo, Y.S., Ansari, M.I., Ho, K.C., Jeng, S.T., Lin, N.S., and Dai, H. (2017). Correlation between mtDNA complexity and mtDNA replication mode in developing cotyledon mitochondria during mung bean seed germination. New Phytol *213*, 751-763.https://doi.org/10.1111/nph.14158
15. Christensen, A.C. (2018). Mitochondrial DNA Repair and Genome Evolution. 11-32.https://doi.org/10.1002/9781119312994.apr0544
16. Chu, H., Jo, Y., Choi, H., Lee, B.C., and Cho, W.K. (2018). Identification of viral domains integrated into Arabidopsis proteome. Mol Phylogenet Evol *128*, 246-257.https://doi.org/10.1016/j.ympev.2018.08.009
17. Davila, J.I., Arrieta-Montiel, M.P., Wamboldt, Y., Cao, J., Hagmann, J., Shedge, V., Xu, Y.Z., Weigel, D., and Mackenzie, S.A. (2011). Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. BMC Biol *9*, 64.https://doi.org/10.1186/1741-7007-9-64
18. Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res *45*, e18.https://doi.org/10.1093/nar/gkw955
19. Huisman, R., and Geurts, R. (2020). A Roadmap toward Engineered Nitrogen-Fixing Nodule Symbiosis. Plant Commun *1*, 100019.https://doi.org/10.1016/j.xplc.2019.100019
20. Jin, J.J., Yu, W.B., Yang, J.B., Song, Y., dePamphilis, C.W., Yi, T.S., and Li, D.Z. (2020). GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol *21*, 241.https://doi.org/10.1186/s13059-020-02154-5
21. Kovar, L., Nageswara-Rao, M., Ortega-Rodriguez, S., Dugas, D.V., Straub, S., Cronn, R., Strickler, S.R., Hughes, C.E., Hanley, K.A., Rodriguez, D.N.*, et al.* (2018). PacBio-Based Mitochondrial Genome Assembly of Leucaena trichandra (Leguminosae) and an Intrageneric Assessment of Mitochondrial RNA Editing. Genome Biol Evol *10*, 2501-2517.https://doi.org/10.1093/gbe/evy179
22. Kozik, A., Rowan, B.A., Lavelle, D., Berke, L., Schranz, M.E., Michelmore, R.W., and Christensen, A.C. (2019). The alternative reality of plant mitochondrial DNA: One ring does not rule them all. PLoS Genet *15*, e1008373.https://doi.org/10.1371/journal.pgen.1008373
23. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res *19*, 1639-1645.https://doi.org/10.1101/gr.092759.109
24. Mader, M., Schroeder, H., Schott, T., Schoning-Stierand, K., Leite Montalvao, A.P., Liesebach, H., Liesebach, M., Fussi, B., and Kersten, B. (2020). Mitochondrial Genome of Fagus sylvatica L. as a Source for Taxonomic Marker Development in the Fagales. Plants (Basel) *9*.https://doi.org/10.3390/plants9101274
25. Manchekar, M., Scissum-Gunn, K., Song, D., Khazi, F., McLean, S.L., and Nielsen, B.L. (2006). DNA recombination activity in soybean mitochondria. J Mol Biol *356*, 288-299.https://doi.org/10.1016/j.jmb.2005.11.070
26. Manos, P.S., Cannon, C.H., and Oh, S.-H. (2008). Phylogenetic Relationships and Taxonomic Status Of the Paleoendemic Fagaceae Of Western North America: Recognition Of A New Genus, Notholithocarpus. Madrono *55*, 181-190.https://doi.org/10.3120/0024-9637-55.3.181

27. McDermott, P., Connolly, V., and Kavanagh, T.A. (2008). The mitochondrial genome of a cytoplasmic male sterile line of perennial ryegrass (Lolium perenne L.) contains an integrated linear plasmid-like element. Theor Appl Genet *117*, 459-470.https://doi.org/10.1007/s00122-008-0790-7
28. Meng, G., Li, Y., Yang, C., and Liu, S. (2019). MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Res *47*, e63.https://doi.org/10.1093/nar/gkz173
29. Miller, M.A., Pfeiffer, W., and Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Paper presented at: 2010 Gateway Computing Environments Workshop (GCE).
30. Mower, J.P. (2009). The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res *37*, W253-259.https://doi.org/10.1093/nar/gkp337
31. Nibert, M.L. (2017). Mitovirus UGA(Trp) codon usage parallels that of host mitochondria. Virology *507*, 96-100.https://doi.org/10.1016/j.virol.2017.04.010
32. Nibert, M.L., Vong, M., Fugate, K.K., and Debat, H.J. (2018). Evidence for contemporary plant mitoviruses. Virology *518*, 14-24.https://doi.org/10.1016/j.virol.2018.02.005
33. O'Conner, S., and Li, L. (2020). Mitochondrial Fostering: The Mitochondrial Genome May Play a Role in Plant Orphan Gene Evolution. Front Plant Sci *11*, 600117.https://doi.org/10.3389/fpls.2020.600117
34. Palmer, J.D., and Herbon, L.A. (1988). Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J Mol Evol *28*, 87-97
35. Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289-290.https://doi.org/10.1093/bioinformatics/btg412
36. Petersen, G., Cuenca, A., Moller, I.M., and Seberg, O. (2015). Massive gene loss in mistletoe (Viscum, Viscaceae) mitochondria. Sci Rep *5*, 17588.https://doi.org/10.1038/srep17588
37. Rice, D.W., Alverson, A.J., Richardson, A.O., Young, G.J., Sanchez-Puerta, M.V., Munzinger, J., Barry, K., Boore, J.L., Zhang, Y., dePamphilis, C.W.*, et al.* (2013). Horizontal Transfer of Entire Genomes via Mitochondrial Fusion in the Angiosperm Amborella. Science *342*, 1468
38. Rodríguez-Moreno, L., González, V.M., Benjak, A., Martí, M.C., Puigdomènech, P., Aranda, M.A., and Garcia-Mas, J. (2011). Determination of the melon chloroplast and mitochondrial genome sequences reveals that the largest reported mitochondrial genome in plants contains a significant amount of DNA having a nuclear origin. BMC Genomics *12*, 424-424.https://doi.org/10.1186/1471-2164-12-424
39. Salojärvi, J., Smolander, O.-P., Nieminen, K., Rajaraman, S., Safronov, O., Safdari, P., Lamminmäki, A., Immanen, J., Lan, T., Tanskanen, J.*, et al.* (2017). Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. Nat Genet *49*, 904-912.https://doi.org/10.1038/ng.3862
40. Salojarvi, J., Smolander, O.P., Nieminen, K., Rajaraman, S., Safronov, O., Safdari, P., Lamminmaki, A., Immanen, J., Lan, T., Tanskanen, J.*, et al.* (2017). Genome sequencing and population genomic

analyses provide insights into the adaptive landscape of silver birch. Nat Genet *49*, 904-912.https://doi.org/10.1038/ng.3862

41. Sanchez-Puerta, M.V., Edera, A., Gandini, C.L., Williams, A.V., Howell, K.A., Nevill, P.G., and Small, I. (2018). Genome-scale transfer of mitochondrial DNA from legume hosts to the holoparasite Lophophytum mirabile (Balanophoraceae). Mol Phylogenet Evol *132*, 243-250.https://doi.org/10.1016/j.ympev.2018.12.006
42. Sanchez-Puerta, M.V., Garcia, L.E., Wohlfeiler, J., and Ceriotti, L.F. (2017). Unparalleled replacement of native mitochondrial genes by foreign homologs in a holoparasitic plant. New Phytol *214*, 376-387.https://doi.org/10.1111/nph.14361
43. Sennikov, A.N., Soltis, D.E., Mabberley, D.J., Byng, J.W., Fay, M.F., Christenhusz, M.J.M., Chase, M.W., Stevens, P.F., Soltis, P.S., Judd, W.S.*, et al.* (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc *181*, 1-20.https://doi.org/10.1111/boj.12385
44. Silva, S.R., Alvarenga, D.O., Aranguren, Y., Penha, H.A., Fernandes, C.C., Pinheiro, D.G., Oliveira, M.T., Michael, T.P., Miranda, V.F.O., and Varani, A.M. (2017). The mitochondrial genome of the terrestrial carnivorous plant Utricularia reniformis (Lentibulariaceae): Structure, comparative analysis and evolutionary landmarks. Plos One *12*, e0180484.https://doi.org/10.1371/journal.pone.0180484
45. Simeone, M.C., Grimm, G.W., Papini, A., Vessella, F., Cardoni, S., Tordoni, E., Piredda, R., Franc, A., and Denk, T. (2016). Plastome data reveal multiple geographic origins of Quercus Group Ilex. PeerJ *4*, e1897.https://doi.org/10.7717/peerj.1897
46. Skippington, E., Barkman, T.J., Rice, D.W., and Palmer, J.D. (2015). Miniaturized mitogenome of the parasitic plant Viscum scurruloideum is extremely divergent and dynamic and has lost all nad genes. P Natl Acad Sci USA *112*, E3515-E3524.https://doi.org/10.1073/pnas.1504491112
47. Sloan, D.B., Alverson, A.J., Chuckalovcak, J.P., Wu, M., McCauley, D.E., Palmer, J.D., and Taylor, D.R. (2012). Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol *10*, e1001241.https://doi.org/10.1371/journal.pbio.1001241
48. Small, I.D., Schallenberg-Rudinger, M., Takenaka, M., Mireau, H., and Ostersetzer-Biran, O. (2020). Plant organellar RNA editing: what 30 years of research has revealed. Plant J *101*, 1040-1056.https://doi.org/10.1111/tpj.14578
49. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312-1313.https://doi.org/10.1093/bioinformatics/btu033
50. Stevens, P.F. (2001 onwards). Angiosperm Phylogeny Website. Version 14, July 2017.
51. Wang, N., Thomson, M., Bodles, W.J., Crawford, R.M., Hunt, H.V., Featherstone, A.W., Pellicer, J., and Buggs, R.J. (2013). Genome sequence of dwarf birch (Betula nana) and cross-species RAD markers. Mol Ecol *22*, 3098-3111.https://doi.org/10.1111/mec.12131
52. Wang, S., Li, D., Yao, X., Song, Q., Wang, Z., Zhang, Q., Zhong, C., Liu, Y., and Huang, H. (2019). Evolution and Diversification of Kiwifruit Mitogenomes through Extensive Whole-Genome

Rearrangement and Mosaic Loss of Intergenic Sequences in a Highly Variable Region. Genome Biol Evol *11*, 1192-1206.https://doi.org/10.1093/gbe/evz063

53. Wick, R.R., Judd, L.M., Gorrie, C.L., and Holt, K.E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol *13*, e1005595.https://doi.org/10.1371/journal.pcbi.1005595
54. Wu, Z., Cuthbert, J.M., Taylor, D.R., and Sloan, D.B. (2015). The massive mitochondrial genome of the angiosperm Silene noctiflora is evolving by gain or loss of entire chromosomes. Proc Natl Acad Sci U S A *112*, 10185-10191.https://doi.org/10.1073/pnas.1421397112
55. Wu, Z., and Sloan, D.B. (2018). Recombination and intraspecific polymorphism for the presence and absence of entire chromosomes in mitochondrial genomes. Heredity (Edinb).https://doi.org/10.1038/s41437-018-0153-3
56. Yang, X., Wang, Z., Zhang, L., Hao, G., Liu, J., and Yang, Y. (2020). A chromosome-level reference genome of the hornbeam, Carpinus fangiana. Sci Data *7*, 24.https://doi.org/10.1038/s41597-020-0370-5
57. Yelenik, S.G., and D'Antonio, C.M. (2013). Self-reinforcing impacts of plant invasions change over time. Nature *503*, 517-520.https://doi.org/10.1038/nature12798

# Tables

**Table 1.** Basic information of Fagales mitogenomes. In column “Chr”, the number means chromosome number, while “C” and “L” behind represent “circular” and “linear”, respectively. The two species with asterisks mean published by others.

| Species | Family | length (bp) | Chr | GC (%) | CDS | tRNA | rRNA | Repeat (bp) | MTPT (bp) |
|---|---|---|---|---|---|---|---|---|---|
| *Alnus glutinosa* | Betulaceae | 629389 | 16L | 45.44 | 35 | 18 | 3 | 8581 | 29856 |
| *Betula pendula** | Betulaceae | 581505 | 1L | 45.52 | 36 | 19 | 3 | 3703 | 31908 |
| *Betula platyphylla* | Betulaceae | 581519 | 2L | 45.53 | 36 | 19 | 3 | 3724 | 32032 |
| *Carpinus cordata* | Betulaceae | 922154 | 3C | 44.97 | 34 | 20 | 3 | 16557 | 46637 |
| *Corylus avellana* | Betulaceae | 635030 | 2L | 44.58 | 35 | 22 | 3 | 39128 | 60416 |
| *Ostrya chinensis* | Betulaceae | 688786 | 1L | 45.23 | 34 | 18 | 3 | 2601 | 31075 |
| *Ostryopsis nobilis* | Betulaceae | 669332 | 1C | 45.21 | 35 | 18 | 3 | 4810 | 24710 |
| *Casuarina equisetifolia* | Casuarinaceae | 492230 | 2C | 44.15 | 35 | 23 | 3 | 2431 | 66400 |
| *Casuarina glauca* | Casuarinaceae | 445851 | 2C | 44.92 | 34 | 19 | 3 | 1758 | 21776 |
| *Fagus sylvatica* | Fagaceae | 504715 | 1C | 45.85 | 34 | 17 | 3 | 2702 | 4615 |
| *Castanea mollissima* | Fagaceae | 388038 | 1C | 45.67 | 36 | 20 | 3 | 10888 | 10169 |
| *Lithocarpus fenestratus* | Fagaceae | 485396 | 6L | 45.76 | 35 | 17 | 3 | 13254 | 7122 |
| *Quercus robur* | Fagaceae | 390878 | 1C | 45.92 | 35 | 17 | 3 | 22006 | 8140 |
| *Quercus suber* | Fagaceae | 478989 | 1L | 45.85 | 36 | 19 | 3 | 26635 | 4967 |
| *Quercus variabilis** | Fagaceae | 412886 | 1C | 45.76 | 36 | 17 | 3 | 20747 | 4537 |
| *Cyclocarya paliurus* | Juglandaceae | 628759 | 1C | 44.89 | 37 | 19 | 3 | 2422 | 34143 |
| *Juglans cathayensis* | Juglandaceae | 740307 | 1C | 45.16 | 37 | 20 | 3 | 25019 | 22632 |
| *Juglans hindsii* | Juglandaceae | 716397 | 1C | 45.25 | 36 | 17 | 3 | 2118 | 12514 |
| *Juglans microcarpa* | Juglandaceae | 623287 | 1L | 45.2 | 35 | 18 | 3 | 6984 | 12448 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Juglans nigra* | Juglandaceae | 716680 | 1C | 45.26 | 37 | 17 | 3 | 2022 | 12667 |
| *Juglans regia* | Juglandaceae | 775914 | 3L | 45.19 | 35 | 17 | 3 | 16232 | 15813 |
| *Juglans sigillata* | Juglandaceae | 778034 | 1L | 44.87 | 36 | 17 | 3 | 19731 | 34425 |
| *Platycarya strobilacea* | Juglandaceae | 502903 | 1C | 45.26 | 37 | 19 | 3 | 3253 | 28548 |
| *Pterocarya stenoptera* | Juglandaceae | 603233 | 1L | 45.35 | 36 | 17 | 3 | 5269 | 3524 |
| *Morella rubra* | Myricaceae | 523452 | 2C | 45.33 | 38 | 18 | 3 | 5668 | 9312 |

# Figures

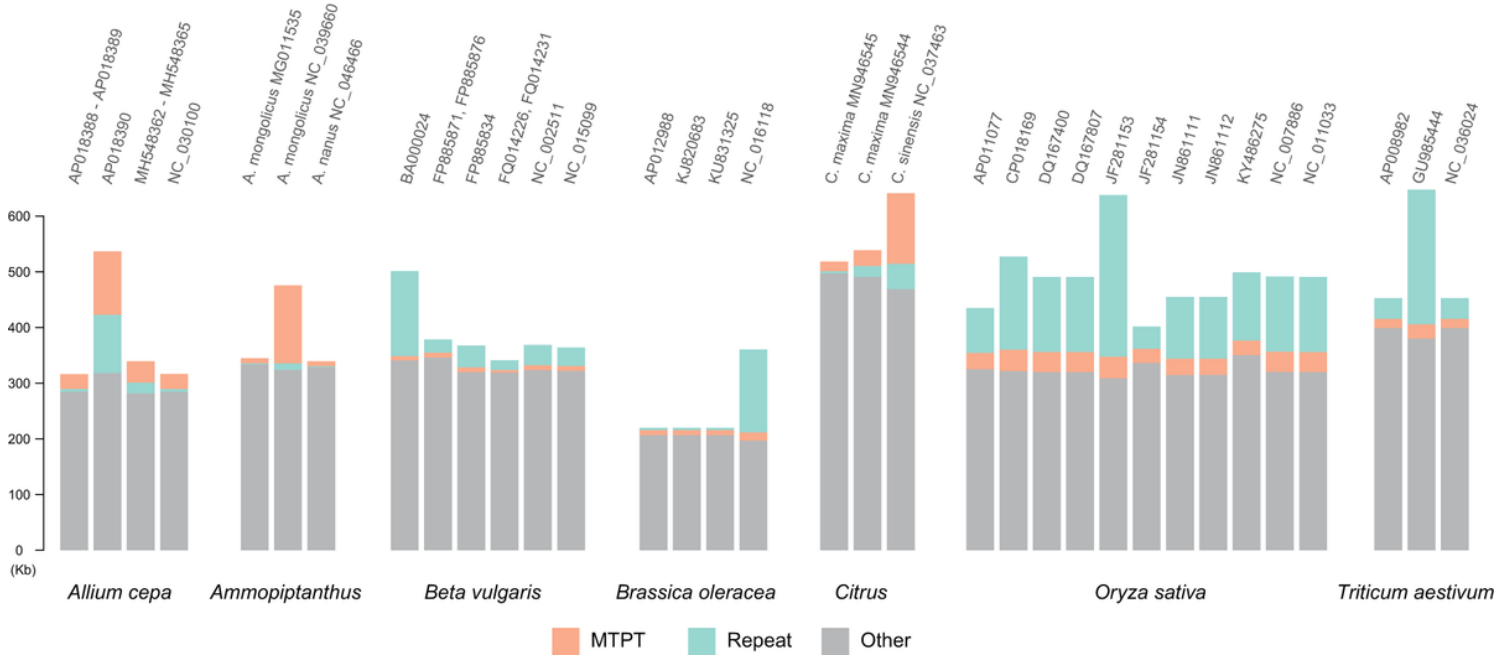
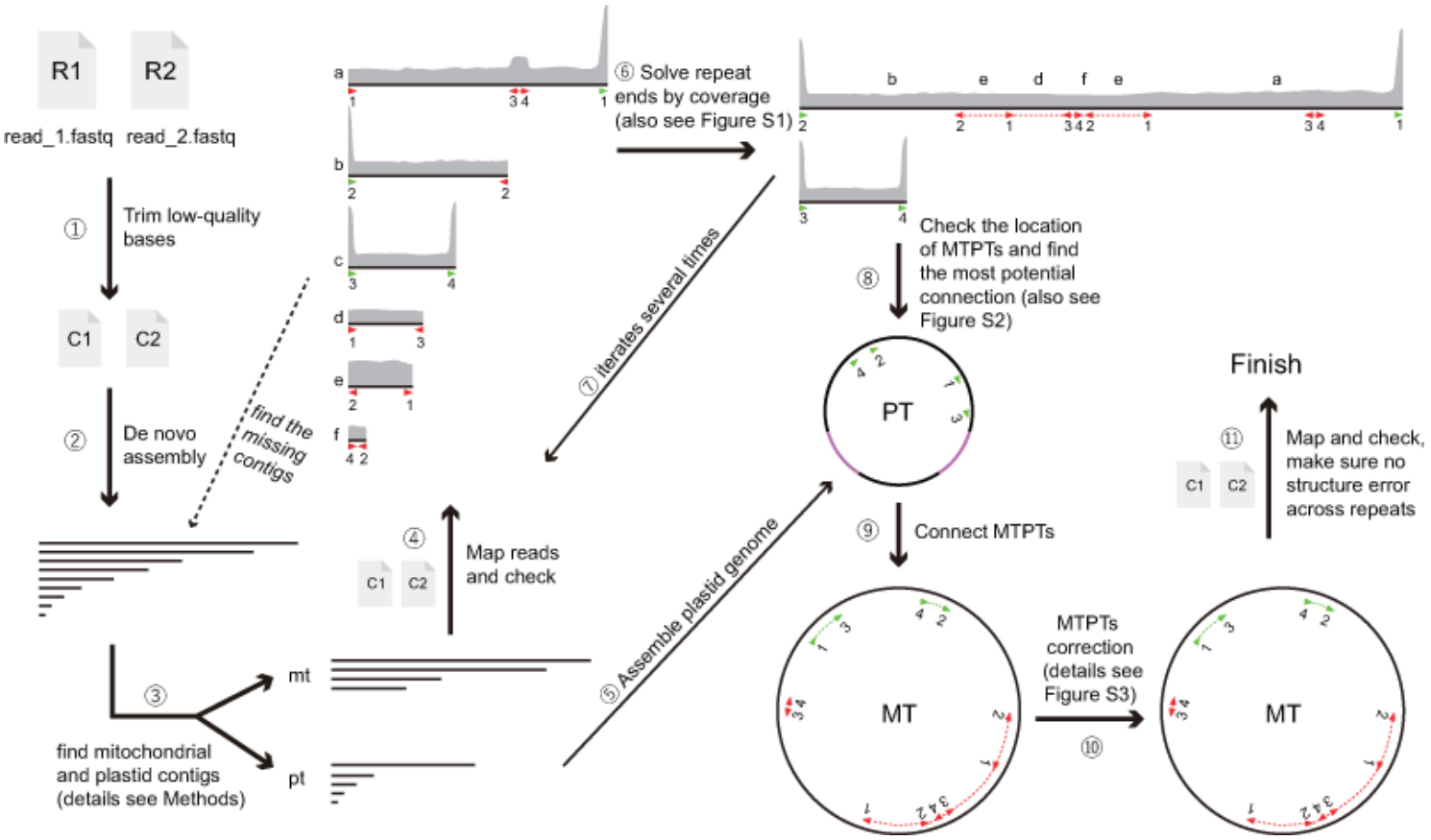


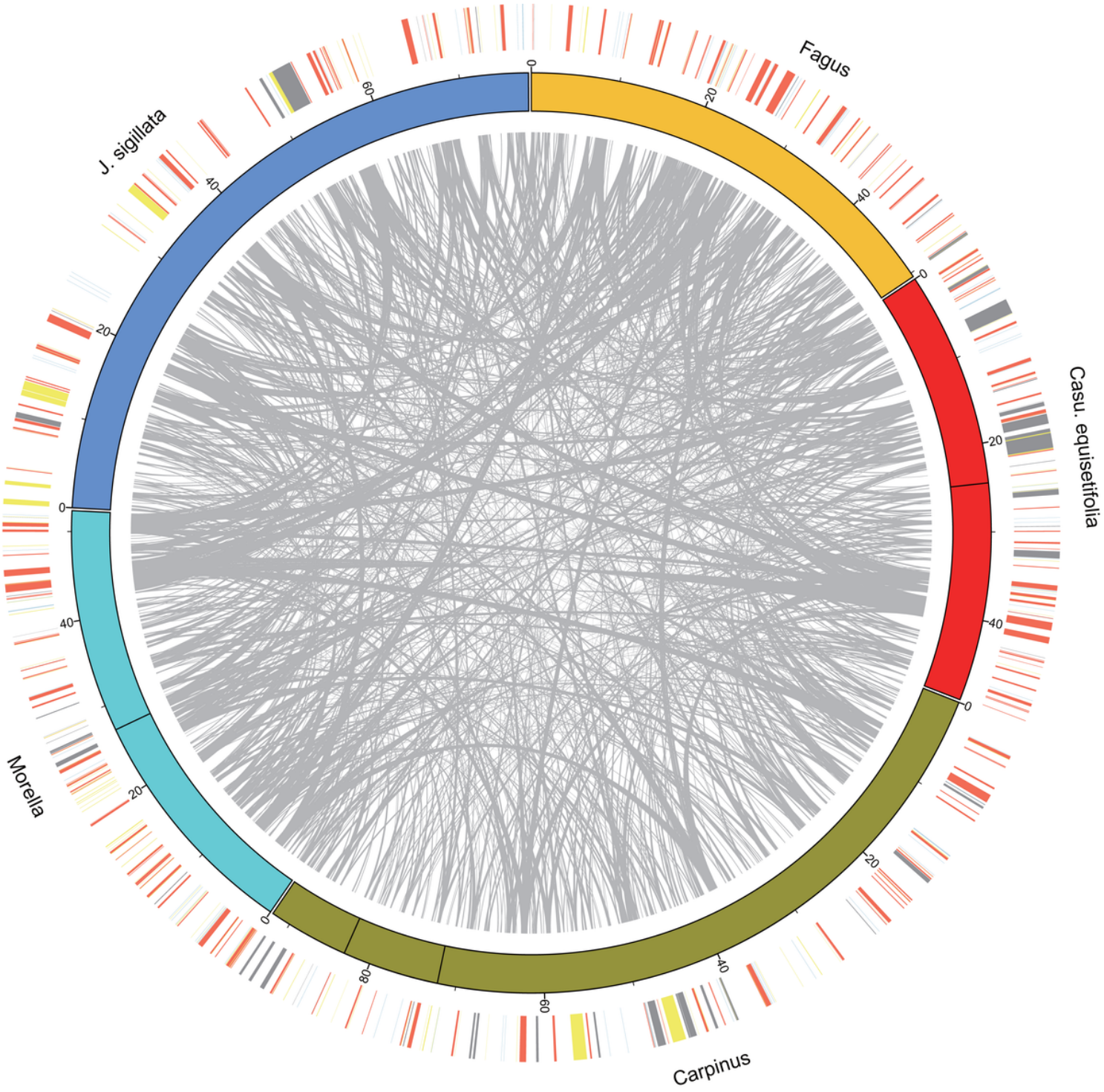

**Figure 1**

Potential misassemblies of repeat or/and MTPT sequences in individuals of a species or species of a genus. Total bar heights indicate mitogenome size. Orange, cyan, and gray indicate the proportions of MTPT, repeat, and other sequences, respectively, ordered according to their proportions. Some assemblies have extremely long MTPTs and repeats compared with related individuals or species. Accessions and species names within genera are indicated.



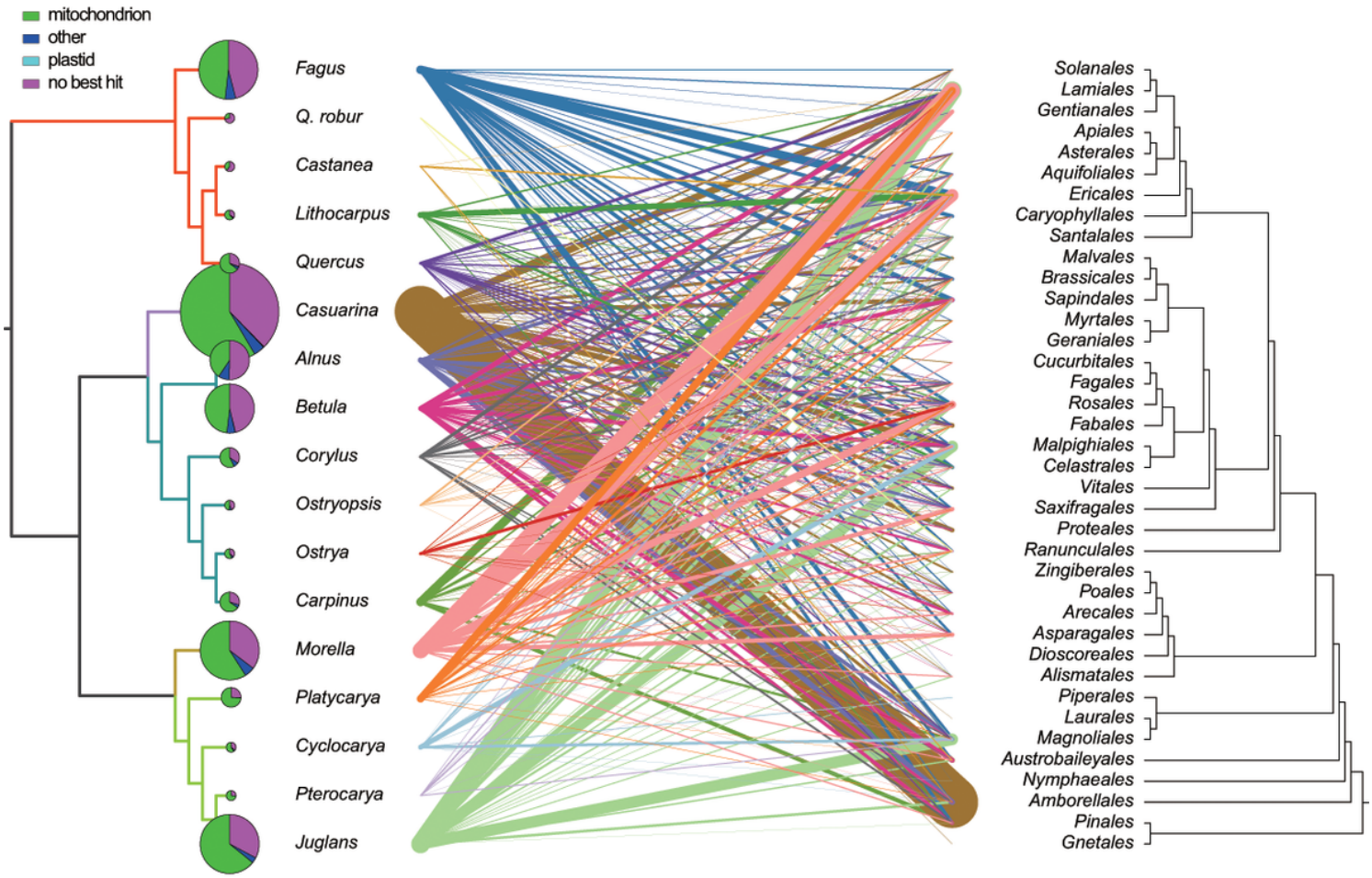

## Figure 2

Assembly strategy workflow. Red and green arrows indicate repetitive and MTPT ends, respectively. The arrow direction indicates the direction in which the end should extend. The same number under red or green arrows indicates copies of the same repeat. R1, R2: raw reads; C1, C2: clean reads.



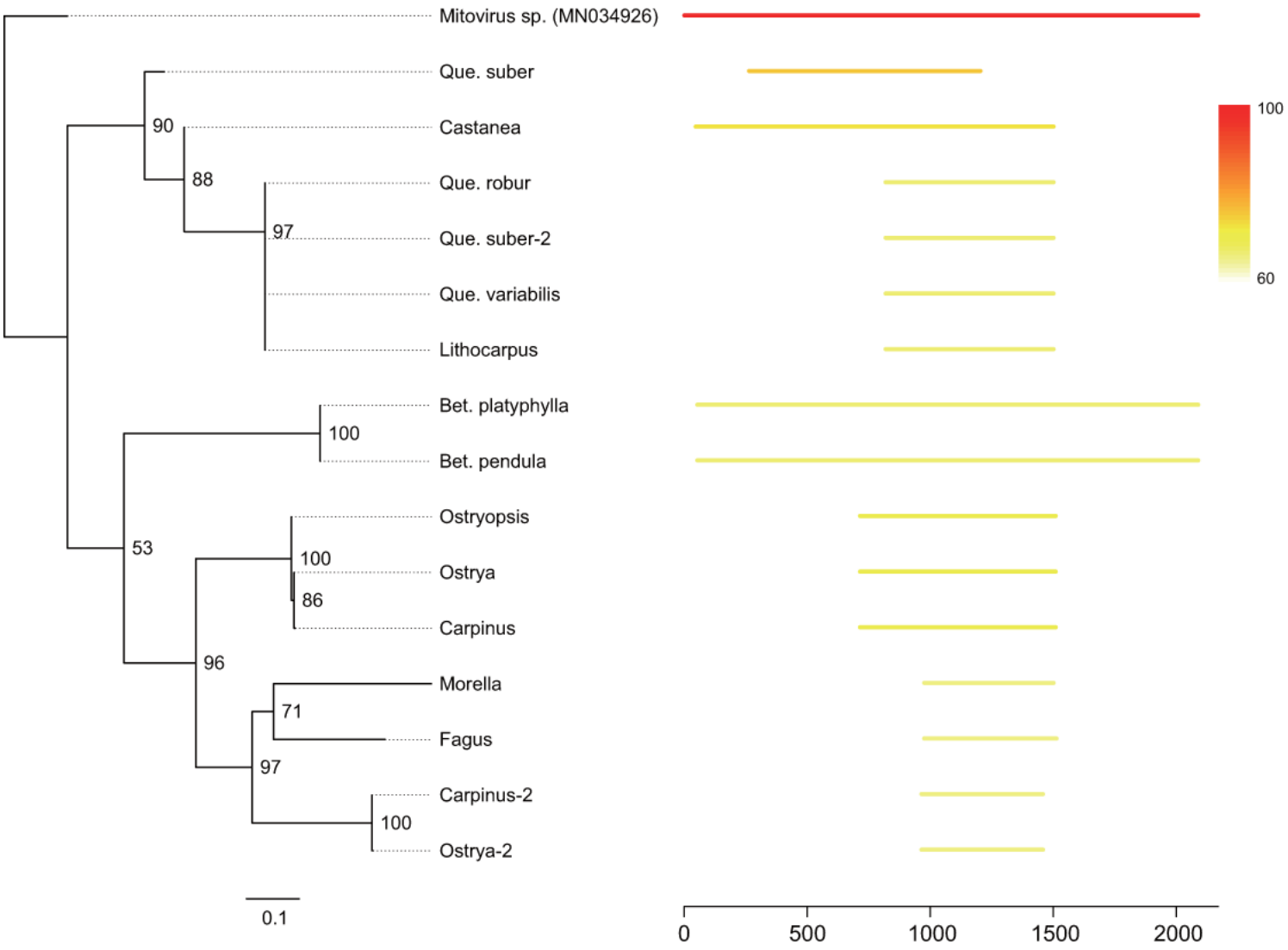

## Figure 3

CIRCOS plot of species from five families. The longest mitogenomes of each family were used. The outer ring shows the position of protein-coding genes and rRNA (red), tRNA (blue), repeat (yellow), and MTPT (gray) sequences.




## Figure 4

Analysis of genus-specific mitogenome sequences. Best-hits of genus-specific sequences between Fagales species and their hit taxa (combined into orders) are connected by lines. Each species is represented with a single color, and line thickness indicates the total sequence hit length. Pie charts indicate the proportions of mitochondrial, plastid, and other sequence hits. Pie size represents the total genus-specific DNA length. Full details are in Table S4 and S5.




## Figure 5

Mitovirus-like sequences in Fagales. Lines at the right indicate the position and similarity of sequence hits against Mitovirus (MN034926). Only matches longer than 400 bp are shown. A tree (left) of hit sequences was constructed using the maximum likelihood method.

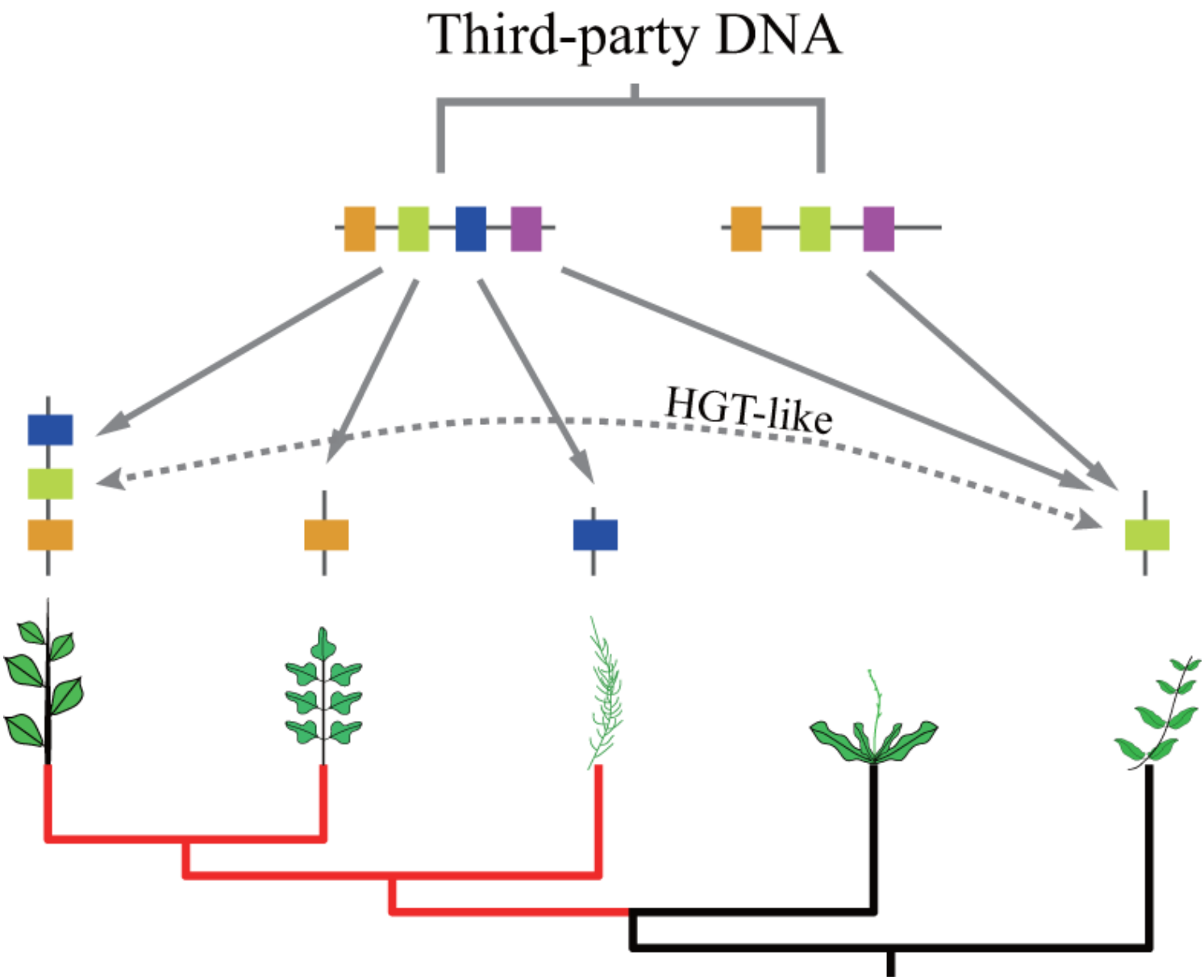



**Figure 6**

HGT-like sequences and unequal transfers from extraneous sequences. Red branches show three close species. Colored blocks indicate different genes or sequence fragments. Solid lines indicate different transfer events. In the red lineage, unequal transfers result in some species acquiring additional homologs. The dotted line indicates the creation of an HGT-like sequence upon the transfer of a single gene on two independent occasions in distant lineages.

## Supplementary Files

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

- TableS1.xls
- TableS2.xls

- TableS3.xls
- TableS4.xls
- TableS5.xls
- Fig.S1.png
- Fig.S2.png
- Fig.S3.png
- Fig.S4.png
- Fig.S5.png
- Fig.S6.png
- Fig.S7.png
- Fig.S8.png