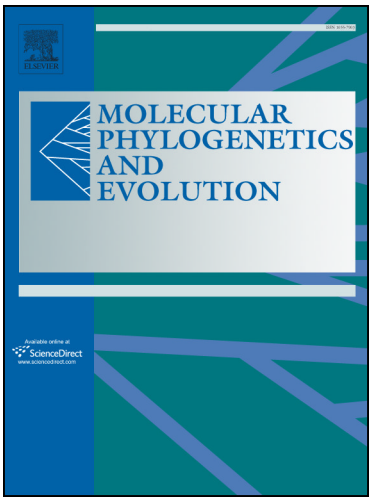


The Complete Organelle Genomes of *Physochlaina orientalis:* Insights into Short Sequence Repeats across Seed Plant Mitochondrial Genomes

Carolina L. Gandini, Laura E. Garcia, Cinthia C. Abbona, M. Virginia Sanchez-Puerta

PII: S1055-7903(19)30195-2
DOI: https://doi.org/10.1016/j.ympev.2019.05.012
Reference: YMPEV 6498

To appear in: *Molecular Phylogenetics and Evolution*

Received Date: 29 March 2019
Revised Date: 14 May 2019
Accepted Date: 17 May 2019

Please cite this article as: Gandini, C.L., Garcia, L.E., Abbona, C.C., Virginia Sanchez-Puerta, M., The Complete Organelle Genomes of *Physochlaina orientalis:* Insights into Short Sequence Repeats across Seed Plant Mitochondrial Genomes, *Molecular Phylogenetics and Evolution* (2019), doi: https://doi.org/10.1016/j.ympev.2019.05.012

The Complete Organelle Genomes of *Physochlaina orientalis:* Insights into Short Sequence Repeats across Seed Plant Mitochondrial Genomes

Carolina L. Gandini [1]

Laura E. Garcia[1,2]

Cinthia C. Abbona[1,3]

M. Virginia Sanchez-Puerta[1,2]

[1] IBAM, Universidad Nacional de Cuyo, CONICET, Facultad de Ciencias Agrarias, Almirante Brown 500, M5528AHB, Chacras de Coria, Argentina.

[2] Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Cuyo, 5500, Mendoza, Argentina.

[3] IDEVEA, Universidad Tecnologica Nacional, CONICET, 5600, San Rafael, Argentina.

*Author for correspondence:

M. Virginia Sanchez-Puerta

Email: mvsanchezpuerta@fca.uncu.edu.ar

IBAM, Universidad Nacional de Cuyo, CONICET, Facultad de Ciencias Agrarias, Almirante Brown 500, M5528AHB, Chacras de Coria, Argentina.

Phone +54 9 (261) 4135000 ext. 1307.

Abstract

Short repeats (SR) play an important role in shaping seed plant mitochondrial genomes (mtDNAs). However, their origin, distribution, and relationships across the different plant lineages remain unresolved. We focus on the angiosperm family Solanaceae that shows a wide diversity in repeat content and extend the study to a wide diversity of seed plants. We determined the complete nucleotide sequences of the organellar genomes of the medicinal plant *Physochlaina orientalis* (Solanaceae), member of the tribe Hyoscyameae. To understand the evolution of the *P. orientalis* mtDNA we made comparisons with those of five other Solanaceae. *P. orientalis* mtDNA presents the largest mitogenome (~685 kb in size) among the Solanaceae and has an unprecedented 8-copy repeat family of ~8.2 kb in length and a great number of SR arranged in tandem-like structures. We found that the SR in the Solanaceae share a common origin, but these only expanded in members of the tribe Hyoscyameae. We discuss a mechanism that could explain SR formation and expansion in *P. orientalis* and *Hyoscyamus niger*. Finally, the great increase in plant mitochondrial data allowed us to systematically extend our repeat analysis to a total of 136 seed plants to characterize and analyze for the first time families of SR among seed plant mtDNAs.

Keywords: *Physochlaina orientalis*, Solanaceae, plastid genome, mitochondrial genome, short repeats, seed plants

Abbreviations: ptDNA, plastid genome; **mtDNA**, mitochondrial genome; **PE**, paired-end; **MTPT**, mitochondrial sequence of plastid origin; **SSC**, small single copy region; **LSC**, large single copy region; **IR**, inverted repeat; **TE**, transposable element; **LR**, large repeat; **IntR**, intermediate repeat; **SR**, short repeat; **TR**, tandem repeat; **HGT**, horizontal gene transfer.

1. Introduction

Among eukaryotes, flowering plant (*i.e.* angiosperm) mitochondria exhibit unique genomes with intra- and inter-species contrasting features that makes them one of the most interesting genomes to study. They are the largest mitochondrial genomes reported so far and can vary from 66 kb (Skippington et al., 2015) to more than 11.3 Mb (Sloan et al., 2012a), with cases of multi-chromosomal mitogenomes (Alverson et al., 2011a, Sanchez-Puerta et al., 2016; Sloan et al., 2012a). The increase in genome length is attributed to the incredible variation in non-coding content, which can include duplicated regions (Alverson et al., 2010; Dong et al., 2018), sequences from the nuclear or plastid genomes acquired through intracellular gene transfer, or foreign mitochondrial sequences acquired through horizontal gene transfer (HGT) (Bergthorsson et al., 2004; Rice et al., 2013; Sanchez-Puerta et al., 2019). In addition, gene content can also fluctuate considerably between species as functional gene transfer to the nucleus is still an ongoing and frequent process (Adams et al., 2002; Covello and Gray, 1992). Angiosperm mitochondrial genomes (mtDNAs) are generally characterized by very low rates of sequence substitution and high rates of genome rearrangements (Cole et al., 2018; Palmer and Herbon, 1988; Sloan et al., 2012c). As a result, gene order is highly scrambled between mitogenomes of even closely related species (Alverson et al., 2010; Ogihara et al., 2005). Nonetheless, collinear gene blocks are maintained within lineages and a few persist since the endosymbiotic event that gave rise to mitochondria (Lelandais et al., 1996; Richardson et al., 2013; Sugiyama et al., 2005; Takemura et al., 1992).

Seed plant mtDNAs frequently contain a large fraction of repeated sequences that can be classified as large (LR), intermediate (IntR), and short (SR) repeats (Maréchal and Brisson, 2010; Mower et al., 2012). Their distribution among seed plants is not uniform (Alverson et al., 2011b; Wynn and Christensen, 2019). For example, *Cucurbita pepo* and *Nymphaea colorata* present a great extent of their mtDNAs covered by SR (Alverson et al., 2010; Dong et al., 2018), *Brassica oleracea* and *Oryza sativa* contain several LR (Chang et al., 2011; Notsu et al., 2002), and *Viscum scurruloideum* harbors a large amount of both kinds of repeats (Skippington et al., 2015). To date, we know that LR are present in most but not all angiosperms (Alverson et al., 2011b). They generally exhibit two or three copies in nearly equal stoichiometry, implying that they undergo frequent and reversible homologous recombination that is responsible for the typical multipartite structure of seed plant mtDNAs (Lonsdale et al., 1988; Oldenburg and Bendich, 1996; Sugiyama et al., 2005). In contrast, recombination across SR is infrequent and irreversible (Arrieta-Montiel et al., 2001; Small et

al., 1987). Rare recombination across SR gives rise to DNA molecules in very low number, so-called sublimons, and different mitochondrial genotypes coexist within an organism (Arrieta-Montiel et al., 2001; Small et al., 1987). Comparative analyses of closely related genomes showed that the endpoints of large genome rearrangements are usually associated with SR (Allen et al., 2007; Fauron et al., 1990; Nishizawa et al., 2007). Short interspersed repeats are widespread among many plant species and they certainly have an important role in shaping the mtDNA structure (André et al., 1992; Kanazawa et al., 1998; Small et al., 1989; Woloszynska, 2010). However, little is known about their origin, expansion, and distribution across the different seed plant lineages. It has been suggested that reverse transcription of non-functional mitochondrial transcripts and subsequent insertion into the genome might be responsible for the generation of SR (André et al., 1992; Kanazawa et al., 1998). Nonetheless, the origin and expansion of SR in plant mitochondria remain a mystery.

The present study focuses on the mtDNAs of the family Solanaceae, where we identified great variation in repeat content. We provide the complete organelle sequences of *Physochlaina orientalis*, a Solanaceae of the tribe Hyoscyameae. Comparisons of the *P. orientalis* mtDNA with those available from five Solanaceae species revealed extensive diversity in repeat content. Our data suggest that SR share a common origin, although they only expanded in the tribe Hyoscyameae. We discuss a mechanism that could explain SR generation and expansion in *P. orientalis*. Finally, given the limited information on short repeated sequences in plant mitochondria, we take advantage of the rapid increase in publicly available mtDNAs to study for the first time the short repeat content and their relationships in 136 seed plants providing valuable information of their origin, characteristics, and distribution.

2. Materials and Methods

2.1. Plant material and mitochondrial DNA extraction

Seeds of *Physochlaina orientalis* (NBG944750045) were obtained from the Nijmegen Botanical Garden (The Netherlands). Plants were cultivated *in vitro* and total DNA was extracted from young leaves using the DNeasy Plant Mini kit (Qiagen).

2.2. Genome sequencing and assembly

Whole-genome shotgun sequencing was performed using the Illumina HiSeq 2500 platform at the Beijing Genomics Institute. A total of 15.88 Gb data containing 26.5 M clean 2x125 bp paired-end (PE) reads with an average insert size of ~800 bp were generated. Clean sequence data are available from the NCBI Bioproject ID PRJNA542896. The plastid genome (ptDNA) assembly was performed *de novo* using NOVOplasty v.2.6.2 (Dierckxsens et al., 2017) with *Hyoscyamus niger* ptDNA as a reference (GenBank accession KF248009), followed by visualization and manual curation using Consed v.29.0 (Gordon et al., 1998).

The mitochondrial genome (mtDNA) was assembled as described below (Figure S1):

1. The PE Illumina reads were assembled *de novo* using Velvet v.1.2.03 (Zerbino and Birney, 2008) with multiple *k-mer* values. The assembly with the largest N50 was selected (N50=2,413 bp; 660 contigs larger than 1 kb; maximum contig length 86,061 bp, *k-mer* 111).
2. Given that several mtDNAs from the family Solanaceae (15 in total) and from other angiosperm families (>200) were available, we identified putative mitochondrial contigs through BLAST searches. A BLAST v.2.7.1+ (Camacho et al., 2009) search against a custom database containing angiosperm mitochondrial sequences available at GenBank identified a total of 131 putative mitochondrial contigs, with e-value < $1x10^{-4}$.
3. Following Silva *et al.* (2017), the 131 putative mitochondrial contigs were extended with SSAKE (Warren et al., 2007) to detect flanking repeats and connect contigs easily. Because of computational limitations, a subset of the Illumina reads was generated for the extension step using the programs BWA, SAMtools and seqtk (https://github.com/lh3/seqtk):
    a. The average read depth was calculated for each of the 660 contigs using the software BWA (Li and Durbin, 2010), SAMtools (Li et al., 2009) and BEDtools (Quinlan and Hall, 2010). Most of the putative mitochondrial contigs showed an average read depth that ranged between 74 and 652 reads.

A total of 244 additional contigs fell into the estimated mitochondrial read-depth but had no BLAST hits to the mitochondrial database.

b. To generate a subset of reads for contig extension, the Illumina reads were mapped to the 131 putative mitochondrial contigs plus the 244 contigs within the putative mitochondrial range using BWA mem with the following presets: -B 5 -O 10 -E 15. The aligned reads were extracted from the original data set to produce a subset (6,9M PE reads).
c. To avoid extensions of plastid sequences that can reside at contig ends, all plastid reads were subtracted from the subset producing a smaller subset of 6M PE reads.
d. Finally, each of the 131 putative mitochondrial contigs was extended individually using SSAKE v3.8.5 –i option. The extension process increased the total length of the putative mitochondrial contigs from 837,007 to 965,005 bp.

4. Visualization, manual curation, and genome finishing were achieved with Consed v.29.0 and GapFiller v.1.10 (Boetzer and Pirovano, 2012). Contigs with no links to the main mitochondrial assembly and with an aberrant read-depth were not considered. Most of those contigs had BLASTN hits to nuclear chromosomes or belonged to the plastid genome. Finally, the mitochondrial genome assembled into a single contig of 684,857 bp.

The depth of sequencing of the ptDNA (average 1,453.6 reads) and mtDNA (average 197.4 reads) are shown in Figure S2. The read depth was calculated using Bowtie2 (Langmead and Salzberg, 2012) with the following presets: --end-to-end --very-fast --no-discordant --no-mixed --no-contain --rdg 20,5 --rfg 20,5 --score-min C,-40,0. Read depth plots were generated using the 'Sushi.R' package (Phanstiel et al., 2014).

2.3. Annotation of the Organelle Genomes

The plastid and mitochondrial genomes of *P. orientalis* were annotated using Geneious R11 (Kearse et al., 2012). Annotations of the plastid genome were automatically transferred from the *Solanum dulcamara* plastome (GenBank accession KY863443) (Amiryousefi et al., 2018). For the mitogenome, annotations were transferred from the publicly available Solanaceae mitochondrial genomes. Each mitochondrial and plastidial annotation was manually curated (start and stop codons and exon/intron boundaries) by

inspection of gene alignments. The tRNA genes were identified using tRNAscan-SE (Lowe and Eddy, 1996). Genome maps were drawn with OGDRAW (Lohse et al., 2007) and edited in Adobe Illustrator CC 2015. The annotated organelle genomes of *P. orientalis* were deposited in GenBank (accession numbers MK490961 and MK492324).

Plastidial and mitochondrial intergenic regions were analyzed. For this, intergenic sequences were extracted and searched against the nt and nr Genbank databases using BLASTN and BLASTX. The total amount of regions with no BLAST hits were calculated.

2.4. Analyses of Solanaceae mitochondrial genomes

The *P. orientalis* mitochondrial genome was compared to five other mtDNAs from the family Solanaceae available in Genbank: *Capsicum annuum* (NC_024624), *Hyoscyamus niger* (NC_026515) (Sanchez-Puerta et al., 2015), *Nicotiana tabacum* (NC_006581) (Sugiyama et al., 2005), *Solanum lycopersicum* (MF034193), and *Solanum pennellii* (MF034194) (Kim and Lee, 2018). The proportion of the genome with similarity to other available complete angiosperm mitochondrial genomes at GenBank database was calculated using discontiguous megaBLAST. The amount of homologous regions in *P. orientalis* and different species or lineages (i.e. Angiosperms, Solanaceae, *H. niger, N. tabacum, C. annuum, S. lycopersicum* and *S. pennellii*) was calculated using BEDtools and BEDops (Neph et al., 2012). Plots were generated using the 'Sushi.R' package. In order to get an approximation of the origin of *P. orientalis* regions that do not match any sequenced Solanaceae mtDNA, we performed a discontiguous megaBLAST search of the 'non-Solanaceae' regions against the NCBI nucleotide databases using the option -culling_limit 1. We then use the ETE 3 toolkit to get the NCBI taxonomic information for each hit (Huerta-Cepas et al., 2016).

RNA editing sites were predicted using PREP-mt (Mower, 2005) with a cutoff value of 0.2. To identify putative transposable elements (TEs), mitochondrial sequences were searched against the CENSOR database with default settings and 'green plants' as a reference source. All hits were considered to compute the total length of the genome covered by TEs. Mitochondrial sequences of plastid origin (MTPTs) were identified by searching each Solanaceae mtDNA against a custom plastid database, using NCBI- BLASTN v.2.7.1+ with default parameters. BLAST hits longer than 200 bp and with an e-value < $1x10^{-4}$ were considered, except for ancient plastid homologs that were excluded (Hao and Palmer, 2009; Sloan and Wu, 2014; Wang et al., 2007). Putative foreign MTPTs were identified by searching for those MTPTs with hits showing higher similarity to plastid sequences from a

lineage unrelated to the one containing the MTPT. This search was achieved with a custom R script that makes use of the R package 'taxonomizr'. To confirm the identity of the donor lineage, Maximum Likelihood analyses (1,000 rapid bootstrapping replicates) under a GTR+G substitution model were performed with RAxML v.8.0.0 (Stamatakis, 2014). Flanking regions to foreign MTPTs (5 kb to each side) were analyzed with NCBI-BLASTN v.2.7.1+ against the nucleotide collection database to identify their origin.

To detect collinear gene blocks among the studied Solanaceae, whole mitochondrial genome alignments were conducted in MAUVE (Darling et al., 2004) using the progressive Mauve algorithm with default parameters.

The fraction of the genome covered by genes, MTPTs, repeats, and transposable elements was calculated combining overlapping intervals using the BEDtools merge option so that overlapping features (e.g. repeats) were not over-represented.

2.5 Repeat analyses across 136 seed plants

Mitochondrial genomes from a wide diversity of seed plants were downloaded from NCBI. Repetitive sequences were identified by searching each genome against itself using ungapped BLASTN with a word_size of 7. All BLAST hits with an e-value < $1x10^{-4}$ and a minimum sequence identity of 80% were considered. To analyze the similarity among repeats, cluster analysis of short (<100 bp) repeated sequences was done with VSEARCH (Rognes et al., 2016). To reduce the redundancy given by BLAST, repeats with the same start and end positions, which were part of different repeat pairs, were discarded. Sequences were first clustered with the option --cluster_fast and 100% of identity. Then, centroids from the first step were clustered with the option --cluster_size and 80% of identity. To elucidate relationships among clusters within and among species, all clusters with more than 50 sequences were re-clustered with VSEARCH using the option --cluster_fast and 80% of identity. Tandem repeats were identified with Tandem Repeat Finder v.4.09 (Benson, 1999). Finally, repeat maps were generated using ClicO FS (Cheong et al., 2015) and edited in Adobe Illustrator CC 2015.

The scripts used to extend contigs, to generate a subset of reads, and to analyze the repeat content can be found in https://github.com/cgandini/Physochlaina_orientalis.

3. Results

3.1. The *Physochlaina orientalis* plastid genome assembly and organization

The *P. orientalis* plastid genome is 156,321 bp long and exhibits a typical quadripartite structure like most land plant plastomes. It is composed of a pair of inverted repeats (IRa and IRb) of 25,867 bp and a small and large single-copy region (SSC and LSC) of 17,989 and 86,598 bp respectively (Figure S3). The genome encodes 80 protein, 30 tRNA, and 4 rRNA coding genes, totaling 114 unique genes (Table S1). The global GC content is 37.70% (LSC: 35.8%, SSC: 32.0%, and IR: 42.9%) comparable to that of other Solanaceae. Twelve protein-coding and 6 tRNA genes contain at least one intron; the genes *clpP*, *rps12*, and *ycf3* have two introns (Table S1). Except for intron 1 in *rps12* and the *trnL-UAA* intron that are *trans*-spliced, the rest are *cis*-spliced introns. Unusual start codons are present in *ndhD* (TTG), *rps19* (GTG), *ycf15* (GTG), and *psbL* (ACG). The latter is probably modified to AUG by RNA-editing as observed in several Solanaceae ptDNAs (Amiryousefi et al., 2018; Kahlau et al., 2006; Sasaki et al., 2003). Foreign sequences were not detected in the *P. orientalis* plastome by BLAST search analyses.

3.2. Features of the mitochondrial genome of *Physochlaina orientalis*

The *P. orientalis* mitochondrial genome assembled into a single molecule of 684,857 bp in length with 44.8% of GC content (Figure 1). The mitogenome encodes 37 protein, 21 tRNA, and 3 rRNA coding genes, totaling 61 unique genes (65 including repeats) (Table S2). The total number of *cis*-spliced introns is 18 (17 group II introns and the *cox1* group I intron) while 6 are *trans*-spliced. As in other sequenced angiosperms, the tRNA set is incomplete, the genes *rps2*, *rps7* and *rps11* are missing, and the gene *rps14* is present as a pseudogene (Mower et al., 2012). The gene content covers 9.94% of the genome. PREP-Mt predicted a total of 479 non-synonymous RNA editing sites in 35 protein-coding genes of *P. orientalis* (Table S2). Uncharacterized sequences with no match against the NCBI databases account for ~16% of the genome. Given that most of the putative nuclear-derived regions have BLAST hits to *Solanum spp*. nuclear genomes, recently known for harboring multiple copies of mitochondrial and plastidial DNA (Kim and Lee, 2018), accurate calculation of the nuclear contribution to the *P. orientalis* mtDNA was not feasible.

A total of 29 sequences of plastid origin (MTPTs) ranging from 200 to 6,593 bp are found in the *P. orientalis* mtDNA (Table S3). The total amount of MTPTs is 25,992 bp representing 3.80% of the mitochondrial genome and 25% of the plastid genome. Out of the 29 MTPTs, four show a foreign origin in the phylogenetic analyses (Figure S4). One MTPT grouped within members of the family Apocynaceae (Figure S4a), two within the

Orobanchaceae (Figure S4b), and the last one with *Cannabis* and an MTPT of *Hyoscyamus niger* mtDNA (Figure S4c) (Gandini and Sanchez-Puerta, 2017). The presence of a similar MTPT in the mtDNA of *P. orientalis* and *H. niger* indicates that this foreign sequence was acquired by the ancestor of these two Solanaceae. BLAST searches of the flanking sequences (5 kb at each side) of each MTPT found mitochondrial fragments (615 to 6,287 bp in length) with high similarity (>97%) to members of the donor lineage in all four cases (Figure S4, ii). This supports the hypothesis that MTPTs were acquired inside foreign mitochondrial DNA via mitochondrion-to-mitochondrion horizontal gene transfer, rather from plastid-to-mitochondria transfers (Gandini and Sanchez-Puerta, 2017).

Analysis of repeated sequences in the *P. orientalis* mtDNA revealed that they cover 219,427 bp, representing 32.04% of the mitogenome (Table S3). Large (LR, >1,000 bp), intermediate (IntR, 100-1,000 bp), short (SR, <100 bp), and tandem (TR) repeats account for 22.17%, 5.28%, 16.35%, and 1.69% of the genome, respectively. The *P. orientalis* mtDNA presents 29 LR pairs ranging from 8.2 to 35.65 kb in length with more than 99.82% of pairwise identity among them (Table S3). Twenty-eight of these repeats form part of an 8-copy repeat family that share an ~8.2 kb identical 'central core'. Extensions of ~1 to 2.6 kb can be found flanking the central core of some repeats within the family. A BLAST search of the central core against the NCBI databases revealed that 75% of the 8.2 kb sequence has no BLAST hits, suggesting that it was gained and multiplied after the divergence of *P. orientalis* from the other Solanaceae. In addition, a great extent of the genome is covered by SR (16.35%), represented by 12,925 repeat pairs that are usually imperfect and appear overlapping larger repeated regions and themselves. For example, ~37% of LR are covered by SR. Moreover, the sum of the length of all the SR pairs is 1,035,932 bp, that is ~1.5 times the entire mitogenome and ~9 times the actual genome coverage of these repeats. These observations are the result of overlapping sequences appearing in multiple repeat pairs. CENSOR searches against the RepBase databases found only a small fraction (7.73%) of short and intermediate repeats with partial hits to transposable elements (Table S3).

3.3. Comparison of the *Physochlaina orientalis* mtDNA with other five Solanaceae mitochondrial genomes

We compared the mtDNA of *P. orientalis* with five other Solanaceae mtDNAs (Figure 2, Table S3). The mitochondrial size varies from 423,596 bp in *Solanum pennellii* to 684,857 bp in *P. orientalis* (Figure 2a). As expected due to its phylogenetic affinity, *P. orientalis* shares more sequences with *Hyoscyamus niger* (55.81%) than with the other

Solanaceae (*Nicotiana tabacum* – 42.90%, *Capsicum annuum* – 36.88%, *Solanum pennellii* – 38.46% and *Solanum lycopersicum* – 38.13%) (Figure 2b). In total, all the sequenced Solanaceae cover 70.93% of *P. orientalis* mitogenome and from this, 17.07% is only shared with *H. niger* (Figure 2b). Overall, 83.90% of the genome presents detectable homology with angiosperm mitochondrial sequences available at Genbank. Not considering the regions with homology to Solanaceae, the *P. orientalis* mtDNA shares 8,886 bp, 7,862 bp, and 5,951 bp with the angiosperm families Apocynaceae, Cannabaceae, and Orobanchaceae, respectively. Interestingly, these are the same lineages identified as donors of the foreign MTPTs in *P. orientalis*.

The GC and protein-coding gene content are similar in all species (GC: 44.8 – 45.2% and 37 protein genes), except for *C. annuum* that shows the lowest GC content (42.7%) and the pseudogenization of the genes *mttB*, *rpl16*, and *rps1* (Table S3 and Table S4). The exclusion of a large fraction of MTPTs (10.66% of the genome) from the *C. annuum* mtDNA results in a GC content of 45.1%, indicating that the MTPTs are responsible for the low GC content calculated for the whole mtDNA. The *cox1* group I intron is only present in *P. orientalis* and *H. niger*, as previously reported (Sanchez-Puerta et al., 2011). The proportion of the genome covered by transposable elements (TE) is similar in all Solanaceae, ranging from 5.07 to 6.60% of the mtDNAs. The majority of the TE are LTR-retrotransposons *copia* and *gypsy*-like (Figure 2, Table S3).

We examined the presence of conserved gene clusters in the family Solanaceae, defined as two or more adjacent genes shared by all the studied Solanaceae species. The analysis revealed 16 shared blocks, representing ~62% of all *P. orientalis* genes: *atp8-cox3-sdh4*, *ccmB-trnK*, *ccmC-trnL*, *nad1.x4-matR*, *nad3-rps12*, *nad4L-atp4*, *nad5.x1.x2-trnC-trnN(cp)-trnY-nad2.x3.x4.x5*, *rpl2-rpl10*, *rpl5-rps14*, *rps13-nad1.x2.x3*, *rps19-rps3-rpl16-cox2*, *rrnS-rrn5*, *sdh3-nad2.x1.x2*, *trnD-trnS*, *trnP(cp)-trnW(cp)*, and *trnS-trnF-trnP* (Table S5). In addition, the clusters *rps4-nad6* and *rps10-cox1* were present in all Solanaceae but in *C. annuum* and the cluster *trnG-trnQ* was found in all but in *H. niger* (Table S5). All gene clusters except one (*ccmB-trnK*) included genes encoded on the same strand.

3.4. Repeat content across the Solanaceae mitochondrial genomes

The repeat content is probably the most variable feature across the Solanaceae mtDNAs. The *P. orientalis* mtDNA shows more repeated DNA in both number of repeat pairs and the extent of the genome covered by all repeat categories (LR, IntR, SR, and TR) (Figure 3 and Table S3). In addition, *H. niger* also exceeds the average amount of imperfect

and partially overlapping SR and TR across the Solanaceae (Figure 3). While *P. orientalis* and *H. niger* present 12,925 and 7,402 SR pairs respectively, the other Solanaceae mtDNAs range from 286 in *N. tabacum* to 555 in *S. pennellii*. SR in *P. orientalis* and *H. niger* mtDNAs are mainly found as multiple copy arrays or tandem-like structures distributed mostly over intergenic regions. Exceptionally, SR arranged in tandem are found within the *nad1.x2.x3* intron of both species and within the *cox2*, *nad4*, *rpl2*, and *rps3* introns of *H. niger*, giving origin to expansions of ~100-300 bp in those genes. Despite differences in the number of repeat pairs, all species present similar distributions of short repeats in terms of length and sequence identity (Figure S5). Most short repeat pairs are 20-39 bp in length and have 95-100% pairwise identity.

Given the great disparity of SR across the Solanaceae, we tested whether or not these sequences were similar to each other within and among species by clustering the SR (Table 1). A cluster is defined by a centroid, that is a representative sequence for which all other sequences in the cluster must have an identity above 80%. In general, a large fraction of SR in *P. orientalis* and *H. niger* mtDNAs form part of clusters. For example, clusters of 10 or more SR sequences account for 88.97% of the total number of SR in *H. niger*, 85.09% in *P. orientalis*, 51.49% in *S. pennellii*, 33.02% in *S. lycopersicum*, 30.51% *C. annuum*, and 24.59% in *N. tabacum*. However, clusters in *H. niger* are fewer and larger than those in *P. orientalis*. While *H. niger* formed 27 groups of 10 or more sequences, *P. orientalis* formed 117 groups. Furthermore, 40.54% and 11.05% of the SR in *H. niger* and *P. orientalis* group in a single cluster, respectively (Table 1). These results could be indicating that repeats in *H. niger* are more conserved than those in *P. orientalis* or that repeats in *P. orientalis* have diverse origins.

A drawback in the clustering analysis is that a given repeated sequence may match two different centroids with the same identity and in that case, sequence assignment into the 'best cluster' is arbitrary. To overcome this weakness and to elucidate relationships among species, we re-clustered all sequences contained in clusters with 20 or more sequences. As expected, we found that some clusters within a species were related to each other (Figure S6). Moreover, clusters found in different species show connections between all the analyzed Solanaceae.

A detailed analysis of the *nad1.x2.x3* intron in the six Solanaceae revealed the presence of a 39 bp insertion in all except *N. tabacum* (shaded sequence in Figure 4). This insertion, which likely took place after the divergence of *N. tabacum*, resulted in the duplication of an 11 bp sequence present immediately upstream the insertion site creating a

pair of direct repeats. An in-depth analysis of the inserted sequence within the *nad1.x2.x3* intron revealed that it is present in the largest SR clusters of all Solanaceae, including *N. tabacum*. In addition, this sequence is located in the upstream region of the gene *rrnL*, as previously reported for *N. tabacum* and other angiosperms (reviewed in Andre *et al*., 1992). Additionally, several short tandem duplications occurred at the 5' end of the inserted sequence producing intron expansions of 281 and 303 bp in *P. orientalis* and *H. niger*, respectively (Figure 4). Some of the duplicated motifs in *H. niger* and *P. orientalis* are similar to the 39 bp initial insertion. However, the remaining tandem duplications differ between the two species suggesting that these elements have expanded independently in *P. orientalis* and *H. niger*.

3.5. Short repeats across seed plant mitochondrial genomes

We wish to understand the evolution and dynamics of perfect and imperfect repeats in plant mitochondria, with emphasis on short repeats. A recent study on plant mitochondrial repeats focused on non-tandem larger repeats with high similarity (Wynn and Christensen, 2019). Here, the repeat content in the plant mtDNAs was analyzed by using a more sensitive BLAST search strategy that can detect repeats as short as 20 bp. To understand the contribution of repeated sequences to mtDNA size variation and whether the SR were related across plant lineages, we expanded our repeat content analyses to screen 130 additional publicly available seed plants mtDNAs (Table S6).

Of the 136 plants compared, *Nymphaea colorata* ranked first with almost half of the genome covered by SR (46.59%), followed by *Silene conica* (35.63%), *Cucumis melo* (33.32%), *Cucumis sativus* (28.20%), *Cucurbita pepo* (24.62%), *Cycas taitungensis* (22.71%), *Pinus taeda* (18.64%), *Stratiotes aloides* (17.12%), *Corchorus capsularis* (16.52%) and *P. orientalis* (16.34%). The median fraction of the genome covered by SR (3.94%) is extremely low among plants, suggesting that the proliferation of these elements is not a universal feature of seed plant mitogenomes (Table S6). Moreover, the expansion of short repeated sequences is not conserved within plant orders. While some angiosperm orders (e.g. Poales, Fabales, Rosales, Brassicales, Lamiales) present a similar extent of the genome covered by SR, others exhibit great diversity (e.g. Cucurbitales, Malvales, Caryophyllales, Solanales). In agreement with our observations in the family Solanaceae, SR pairs in other plant mitochondria range mainly between 20-39 bp and, in general, present a high sequence identity (95-100%) (Figure 5 and Table S6). This high level of identity between repeats is probably due to the typical low mutation rates of seed plant mitogenomes. For example,

species with high mutation rates as *Silene noctiflora* (Sloan et al., 2012b) present the lowest proportion of SR within the 95-100% category (21.12%) (Table S6).

We wish to evaluate the origin of the SR that proliferated in different land plant lineages and assess whether they were related within and among species. The clustering analysis showed that, in general, species with a high proportion of the genome covered by SR have a great amount of those sequences grouped in clusters (Figure 5, Figure S7 and Table S6). However, these repeats were not necessarily grouped in a single cluster. For example, *Nymphaea colorata* has 99.24% of the SR in clusters of 50 or more sequences, while the largest cluster comprises only 28.35% of its SR. On the other hand, *Ginkgo biloba* has 79.90% in clusters of 50 or more sequences and 72% belong to the largest cluster, indicating that these sequences are highly conserved and probably have a common origin. Other species with large amount of SR clustered in one group include: *Nepenthes ventricosa x Nepenthes alata* (69.90%), *C. taitungensis* (60.39%), *Amborella trichopoda* (41.28%), *H. niger* (40.54%), *Betula pendula* (35.77%), *Viscum scurruloideum* (32.92%), *Lagerstroemia indica* (30.44%), *Spinacia oleracea* (29.26%) and *Citrullus lanatus* (28.59%). The median of the fraction of SR found in the largest cluster in the 136 species is 7.93% (Table S6). Interestingly, re-clustering the clusters containing 50 or more sequences of all the species show that connections between SR are more common among related species; e.g. within Poales, Cucurbitales, Brassicales, or Solanales (Figures 5 and S7). However, relationships between clusters of distant species were also detected. A few of these could be explained by HGT events (violet lines in Figure 5). For example, clustered sequences were related between *P. orientalis* and *Cannabis sativa*, for which HGT events were described (this study) and between the holoparasite *Lophophytum mirabile* and members of its host lineage, the mimosoid species *Acacia ligulata* and *Leucaena trichandra*, also known for having a parasitic interaction and extensive HGT (Sanchez-Puerta et al., 2019).

4. Discussion

4.1. Assembly, a challenging task

We successfully sequenced and assembled the organellar genomes of *Physochlaina orientalis* using Illumina technology. The high degree of conservation, the relatively small size, and the high read depth of plastid genomes allow *de novo* assembly algorithms to use known plastid genomes (or even short sequences or genes) as seeds from which extension by overlapping reads can easily form a circular genome (Dierckxsens et al., 2017). On the contrary, plant mitochondrial genomes usually exhibit repetitive sequences that make the assembly of short-reads a challenging task. The iterative individual extension of each contig of the initial assembly followed in this work and others (Silva et al., 2017) helped us: (i) to detect repeated sequences with multiple copies as they appeared flanking contigs during the extension phase and (ii) to connect contigs easily, avoiding misassemblies due to the high frequency of repeats. We failed to assemble the mtDNA into a master circle and obtained a linear assembly that lacked paired-end connections to other parts of the genome at one end, suggesting that some sequences may be missing from our assembly. Notwithstanding, the master circle configuration may not exist *in vivo* at all. Instead, a mix of branched and linear molecules permutable by recombination to circular structures are thought to coexist within angiosperm tissues (Oldenburg and Bendich, 1996; Sloan, 2013).

4.2. The *Physochlaina orientalis* mitogenome is rich in foreign and repeated sequences

The *P. orientalis* mitochondrial genome ranked 28th in size among 136 seed plant mitochondrial genomes studied here, which had a median mtDNA size of 497,367 bp (Table S6). Flowering plant mtDNA expansions and contractions are the result of a number of processes. While some genomes increased in size through the proliferation of repetitive sequences or the incorporation of plastid DNA, as in *Cucurbita* (Alverson et al., 2010), others expanded by the acquisition of foreign sequences. The horizontal incorporation of mitochondrial sequences has been widely described among angiosperms (Bergthorsson et al., 2004; Mower et al., 2012), including the transfer of entire mitogenomes (Rice et al., 2013; Sanchez-Puerta et al., 2019; 2017). The *P. orientalis* mtDNA exhibits several foreign regions as a result of independent HGT events: (1) the presence of foreign MTPTs surrounded by extensive mitochondrial sequences from the putative donor lineages (e.g. Apocynaceae, Cannabaceae); and (2) the acquisition of the *cox1* intron (Sanchez-Puerta et al., 2011; 2008).

The size of the *P. orientalis* mtDNA is likely the result of horizontally-transferred sequences and the proliferation of both large and small repeats.

With rare exceptions, large repeated sequences are common among angiosperm mitochondrial genomes (Alverson et al., 2011b). They are mainly found in 2-3 copies and they frequently undergo homologous recombination contributing to the typical multipartite structure of angiosperm mitochondrial genomes (Cole et al., 2018; Palmer and Herbon, 1988; Sloan et al., 2012a). In general recombination across LR is frequent and reciprocal generating isomeric subgenomic molecules (Maréchal and Brisson, 2010; Mower et al., 2012). In a much lesser extent, SR were also found to recombine resulting in great impact to the mitogenome structure (André et al., 1992; Kanazawa et al., 1998; Small et al., 1989; Woloszynska, 2010). The *P. orientalis* mtDNA presents an elevated number of both types of repeats. In fact, *P. orientalis* ranked 18th and 10th in terms of the extent of the genome covered by LR and SR, respectively, among 136 seed plant mitochondria. Probably the most distinctive feature of *P. orientalis* is the presence of a rare 8-copy repeat family of large repeats (>8.2 kb). Similar results were only described in *Silene latifolia* in which a 6-copy repeat family of ~1.3 kb was found to be in recombinational equilibrium (Sloan et al., 2010). The large number of possible recombination events through large and short repeats in *P. orientalis* suggests that the genome structure shown here represents only one of thousands of alternatives that could arise via intramolecular recombination.

4.3. Similar and distinct evolutionary features across the Solanaceae mtDNAs

The mitochondrial genome of six species of the family Solanaceae are now available for comparison to assess their evolution along >24 million years since their divergence (Tu et al., 2010). The Solanaceae mtDNAs present 37 protein-coding genes, except for *C. annuum* in which three genes became pseudogenes. Of those three genes, the ribosomal protein-coding genes *rpl16* and *rps1* had been frequently lost from the mtDNA during angiosperm evolution (Adams et al., 2002; Covello and Gray, 1992). Less common is the loss of the translocase gene *mttB*, although it is a pseudogene in *Viscum* spp. and it is missing in *Vitis vinifera*, *Malus x domestica*, and *Boea hygrometrica* (Petersen et al., 2016). We found 16 collinear gene blocks shared among the studied Solanaceae, and three more shared by all but one of them. That is, ~70% of the entire *P. orientalis* gene set is found in conserved gene clusters. Two clusters date back to the original bacterial ancestor of mitochondria (Takemura et al., 1992), seven were unique to angiosperms, and six were unique to eudicots (Richardson et al., 2013). Interestingly, 9 and 13 gene clusters were shared by the Solanaceae and

Liriodendron tulipifera (Richardson et al., 2013) or *Nelumbo nucifera* (Gui et al., 2016), respectively. As previously suggested, gene blocks could be potential co-transcription units (Lelandais et al., 1996; Sugiyama et al., 2005) and gene cluster fragmentation could affect mitochondrial functions, explaining why almost all of the clusters contain genes encoded on the same strand (Richardson et al., 2013).

The repeat content is the most contrasting feature across the Solanaceae. Repeat pairs vary ~40-fold between *N. tabacum* and *P. orientalis*. LR covered similar proportions in all genomes, but the overlapping arrangement of LR is unique to *P. orientalis*. In addition, LR are not homologous among species, except for the genus *Solanum* in which entire LR were shared between *S. pennellii* and *S. lycopersicum* (Kim and Lee, 2018). A comparative study of *Brassica* species showed that LR can remain static for long periods but rapidly diverge through genome recombination (Wynn and Christensen, 2019). In contrast to the other studied Solanaceae species, both Hyoscyameae species shared an explosion of short repeats mainly disposed in a tandem structure with imperfect repeat units. Clustering analysis showed that groups of repeats were connected within and between Solanaceae species. This observation suggests a common origin for at least some of the Solanaceae SR followed by independent expansion in both Hyoscyameae species. Independent multiplication of a repeat family was previously reported for putatively mobile elements, so-called Bpu sequences, in *C. taitungensis* and *G. biloba* mitogenomes (Chaw et al., 2008; Guo et al., 2016). The *P. orientalis* and *H. niger* SR share several features with the Bpu elements: they are short, many are flanked by direct repeats, and most are found as multiple copy arrays. However, the SR identified in the Hyoscyameae are more variable and are not found associated with retrotransposon-like sequences.

4.4. The enigmatic expansion of short repeats

Despite the impact of SR in the size and structure of plant mitogenomes (André et al., 1992; Kanazawa et al., 1998; Small et al., 1989; Woloszynska, 2010), little is known about the molecular mechanism leading to their formation. It has been proposed that short repeats can arise from ‘nonfunctional’ regions of mitochondrial transcripts relocated into the genome through reverse transcription (André et al., 1992; Gualberto et al., 1988; Kanazawa et al., 1998). Even though there is sparse direct evidence of reverse transcription in plant mitochondria (Moenne et al., 1996; Schuster and Brennicke, 1989), indirect evidence has accumulated from the loss of introns and editing sites through the retro-transcription of mature transcripts (Edera et al., 2018; Sloan et al., 2010). Therefore, it is not difficult to

imagine that the mechanism of SR generation proposed by André et al. (1992) could play a role in plant genome evolution.

A sequence repeated several times and flanked by direct repeats of 7-10 bp was found upstream of the gene *rrnL* of the mitochondrial genomes of *Zea*, *Oryza*, *Oenothera*, *Petunia,* and *Nicotiana* (reviewed in André *et al*., 1992). We have identified the former sequence not only in the 5' end of the gene *rrnL* in all Solanaceae species but repeated several times across these mitogenomes. Interestingly, one of the many insertions was found in the *nad1.x2.x3* intron of all Solanaceae but *N. tabacum* allowing us to compare an expansion event along the evolution of this lineage. While the first insertion event in the *nad1.x2.x3* intron seems to have taken place in the ancestor of all but *N. tabacum*, subsequent tandem duplications were restricted to the tribe Hyoscyameae and seem to have evolved independently in *P. orientalis* and *H. niger*. Even though the origin of these short repeats could be through reverse transcription of non-processed mitochondrial transcripts, their subsequent tandem duplication appears to be the result of an entirely different process.

Slipped-strand mispairing (SSM), also known as replication slippage, can account for the formation, expansion, and contraction of short contiguous repeats. In this model, simple tandem repeats originated by chance are expanded or contracted by a process that involves the mispairing of complementary bases at the site of an existing short repeat during DNA repair or replication (Levinson and Gutman, 1987; Levinson et al., 1985). In a similar manner, SSM at noncontiguous short repeats could generate longer tandem duplications flanked by one unit of the original noncontiguous repeat (Levinson and Gutman, 1987; Taylor and Breden, 2000). We propose that a combination of both processes, retrotranscription and SSM at noncontiguous repeats, could be playing a role in the expansion of short repeats in the Hyoscyameae. This would explain the presence of duplicated motifs flanked by direct repeats. The presence of the initial structure (i.e. a motif flanked by noncontiguous direct repeats) in all the Solanaceae and the arousal of tandem-like structures only in *P. orientalis* and *H. niger* mtDNAs suggest that differences in the replication, transcription, or recombination machinery impact the SR content of the mtDNAs across the Solanaceae.

4.5. Short repeat clusters are shared by closely related species

The origin, distribution, and expansion of SR across seed plants has not been previously investigated. We present here a detailed analysis comparing families of short repeats among seed plant mitogenomes in terms of SR amount, genome coverage, size, and,

similarity. All the studied seed plant species exhibit SR, although the expansion of these elements is highly variable. Indeed, SR expansion seems to be restricted to some lineages or even, to individual species. The clustering analysis grouped the known Bpu elements of *C. taitungensis* and *G. biloba* (Chaw et al., 2008; Guo et al., 2016) in a well-conserved single cluster per species. These elements were not identified in other plants, not even in the gymnosperm *Welwitschia mirabilis* (Guo et al., 2016). Consistently, the SR of *C. taitungensis* were only shared with those of *G. biloba* in the re-clustering analysis. Even though it is highly probable that relationships of SR across species were underrepresented as we only compared clusters with 50 or more sequences, several connections among SR from different species were detected. We observed that the expansion of SR appears to have taken place in ancestral repeats as connections between clusters of the same taxonomic group were quite frequent. This is clearly exemplified in the orders Poales, Brassicales, and Solanales. In some cases, repeats were connected between species in which HGT events had been reported (this study and Sanchez-Puerta et al., 2019; 2017). Whether all the shared sequences were transferred as a result of the HGT event or were expanded after the HGT event we do not know. Finally, highly clustered repeats were also common between species without an obvious biological or physical association, e.g. *Nelumbo* and *Sorghum*. However, connections between SR families from distant species does not necessary imply a common origin; instead, they could be the result of convergent evolution. That is, it is possible that mitochondrial insertions of retro-transcribed, highly expressed transcripts have taken place independently in distant lineages. This may explain why *P. orientalis* shares highly repetitive sequences with clusters of several distant species, such as those of the order Poales. Subsequent spreading by SSM at noncontiguous repeats could account for tandem-like structures and therefore high repeat clustering in species, such as *N. colorata* (Dong et al., 2018), *C. taitungensis* (Chaw et al., 2008) and *G. biloba* (Guo et al., 2016). The availability of more plant mitogenomes will likely help to elucidate the expansion of SR during plant evolution.

5. Reference

Adams, K.L., Qiu, Y.-L., Stoutemyer, M., Palmer, J.D., 2002. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci 99, 9905–9912. doi:10.1073/pnas.042694899

Allen, J.O., Fauron, C.M., Minx, P., Roark, L., Oddiraju, S., Lin, G.N., Meyer, L., Sun, H., Kim, K., Wang, C., Du, F., Xu, D., Gibson, M., Cifrese, J., Clifton, S.W., Newton, K.J., 2007. Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. Genetics 177, 1173–1192. doi:10.1534/genetics.107.073312

Alverson, A.J., Rice, D.W., Dickinson, S., Barry, K., Palmer, J.D., 2011a. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. Plant Cell 23, 2499–2513. doi:10.1105/tpc.111.087189

Alverson, A.J., Wei, X., Rice, D.W., Stern, D.B., Barry, K., Palmer, J.D., 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol Biol Evol 27, 1436–1448. doi:10.1093/molbev/msq029

Alverson, A.J., Zhuo, S., Rice, D.W., Sloan, D.B., 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. doi:10.1371/journal.pone.0016404

Amiryousefi, A., Hyvönen, J., Poczai, P., 2018. The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in Solanaceae. Plos One 13, e0196069. doi:10.1371/journal.pone.0196069

André, C., Levy, A., Walbot, V., 1992. Small repeated sequences and the structure of plant mitochondrial genomes. Trends Genet. 8, 128–132. doi:10.1016/0168-9525(92)90370-J

Arrieta-Montiel, M., Lyznik, A., Woloszynska, M., Janska, H., Tohme, J., Mackenzie, S., 2001. Tracing evolutionary and developmental implications of mitochondrial stoichiometric shifting in the common bean. Genetics 158, 851–864.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573–580.

Bergthorsson, U., Richardson, A.O., Young, G.J., Goertzen, L.R., Palmer, J.D., 2004. Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm *Amborella*. Proc. Natl. Acad. Sci. U.S.A. 101, 17747–17752. doi:10.1073/pnas.0408336102

Boetzer, M., Pirovano, W., 2012. Toward almost closed genomes with GapFiller. Genome Biol. 13, R56. doi:10.1186/gb-2012-13-6-r56

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10, 421. doi:10.1186/1471-2105-10-421

Chang, S., Yang, T., Du, T., Huang, Y., Chen, J., Yan, J., He, J., Guan, R., 2011. Mitochondrial genome sequencing helps show the evolutionary mechanism of mitochondrial genome formation in *Brassica*. BMC Genomics 12, 497. doi:10.1186/1471-2164-12-497

Chaw, S.-M., Chun-Chieh Shih, A., Wang, D., Wu, Y.-W., Liu, S.M., 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. doi:10.1093/molbev/msn009

Cheong, W.-H., Tan, Y.-C., Yap, S.-J., Ng, K.-P., 2015. ClicO FS: An interactive web-based service of Circos. Bioinformatics 31, 3685–3687. doi:10.1093/bioinformatics/btv433

Cole, L.W., Guo, W., Mower, J.P., Palmer, J.D., 2018. High and variable rates of repeat-mediated mitochondrial genome rearrangement in a genus of plants. Mol Biol Evol 35, 2773–2785. doi:10.1093/molbev/msy176

Covello, P.S., Gray, M.W., 1992. Silent mitochondrial and active nuclear genes for subunit 2 of cytochrome c oxidase (cox2) in soybean: Evidence for RNA-mediated gene transfer. EMBO Journal 11, 3815–3820. doi:10.1002/j.1460-2075.1992.tb05473.x

Darling, A.C.E., Mau, B., Blattner, F.R., Perna, N.T., 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Research 14, 1394–1403. doi:10.1101/gr.2289704

Dierckxsens, N., Mardulyn, P., Smits, G., 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res 45, e18. doi:10.1093/nar/gkw955

Dong, S., Zhao, C., Chen, F., Liu, Y., Zhang, S., Wu, H., Zhang, L., Liu, Y., 2018. The complete mitochondrial genome of the early flowering plant *Nymphaea colorata* is highly repetitive with low recombination. BMC Genomics 19, 614. doi:10.1186/s12864-018-4991-4

Edera, A.A., Gandini, C.L., Sanchez-Puerta, M.V., 2018. Towards a comprehensive picture of C-to-U RNA editing sites in angiosperm mitochondria. Plant Mol Biol 97, 215–231. doi:10.1007/s11103-018-0734-9

Fauron, C.M.R., Havlik, M., Brettell, R.I.S., 1990. The mitochondrial genome organization of a maize fertile cmsT revertant line is generated through recombination between two sets of repeats. Genetics 124, 423–428.

Gandini, C.L., Sanchez-Puerta, M.V., 2017. Foreign plastid sequences in plant mitochondria are frequently acquired via mitochondrion-to-mitochondrion horizontal transfer. Sci Rep 7, 123. doi:10.1038/srep43402

Gordon, D., Abajian, C., Green, P., 1998. Consed: A graphical tool for sequence finishing. Genome Research 8, 195–202. doi:10.1101/gr.8.3.195

Gualberto, J.M., Wintz, H., Weil, J.-H., Grienenberger, J.M., 1988. The genes coding for subunit 3 of NADH dehydrogenase and for ribosomal protein S12 are present in the wheat and maize mitochondrial genomes and are co-transcribed. Mol. Gen. Genet. 215, 118–127. doi:10.1007/BF00331312

Gui, S., Wu, Z., Zhang, H., Zheng, Y., Zhu, Z., Liang, D., Ding, Y., 2016. The mitochondrial genome map of *Nelumbo nucifera* reveals ancient evolutionary features. Sci. Rep. 6, 1–11. doi:10.1038/srep30158

Guo, W., Grewe, F., Fan, W., Young, G.J., Knoop, V., Palmer, J.D., Mower, J.P., 2016. *Ginkgo* and *Welwitschia* mitogenomes reveal extreme contrasts in gymnosperm mitochondrial evolution. Mol Biol Evol 33, 1448–1460. doi:10.1093/molbev/msw024

Hao, W., Palmer, J.D., 2009. Fine-scale mergers of chloroplast and mitochondrial genes create functional, transcompartmentally chimeric mitochondrial genes. Proc. Natl. Acad. Sci. U.S.A. 106, 16728–16733. doi:10.1073/pnas.0908766106

Huerta-Cepas, J., Serra, F., Bork, P., 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol 33, 1635–1638. doi:10.1093/molbev/msw046

Kahlau, S., Aspinall, S., Gray, J.C., Bock, R., 2006. Sequence of the tomato chloroplast DNA and evolutionary comparison of solanaceous plastid genomes. J Mol Evol 63, 194–207. doi:10.1007/s00239-005-0254-5

Kanazawa, A., Tozuka, A., Kato, S., Mikami, T., Abe, J., Shimamoto, Y., 1998. Small interspersed sequences that serve as recombination sites at the *cox2* and *atp6* loci in the mitochondrial genome of soybean are widely distributed in higher plants. Curr Genet 33, 188–198. doi:10.1007/s002940050326

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: An integrated and extendable desktop software platform for

the organization and analysis of sequence data. Bioinformatics 28, 1647–1649. doi:10.1093/bioinformatics/bts199

Kim, H.T., Lee, J.M., 2018. Organellar genome analysis reveals endosymbiotic gene transfers in tomato. Plos One 13, e0202279. doi:10.1371/journal.pone.0202279

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357–359. doi:10.1038/nmeth.1923

Lelandais, C., Gutierres, S., Mathieu, C., Vedel, F., Remacie, C., Maréchal-Drouard, L., Brennicke, A., Binder, S., Chétrit, P., 1996. A promoter element active in run-off transcription controls the expression of two cistrons of *nad* and *rps* genes in *Nicotiana sylvestris* mitochondria. Nucleic Acids Res 24, 4798–4804.

Levinson, G., Gutman, G.A., 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evo 4, 203–221. doi:10.1093/oxfordjournals.molbev.a040442

Levinson, G., Marsh, J.L., Epplen, J.T., Gutman, G.A., 1985. Cross-hybridizing snake satellite, Drosophila, and mouse DNA sequences may have arisen independently. Mol Biol Evo 2, 494–504. doi:10.1093/oxfordjournals.molbev.a040374

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595. doi:10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Lohse, M., Drechsel, O., Bock, R., 2007. Organellar Genome DRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet 52, 267–274. doi:10.1007/s00294-007-0161-y

Lonsdale, D.M., Brears, T., Hodge, T.P., Melville, S.E., Rottmann, W.H., 1988. The plant mitochondrial genome: homologous recombination as a mechanism for generating heterogeneity. Philosophical Transactions of the Royal Society B: Biological Sciences 319, 149–163. doi:10.1098/rstb.1988.0039

Lowe, T.M., Eddy, S.R., 1996. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25, 955–964. doi:10.1093/nar/25.5.0955

Maréchal, A., Brisson, N., 2010. Recombination and the maintenance of plant organelle genome stability. New Phytol 186, 299–317. doi:10.1111/j.1469-8137.2010.03195.x

Moenne, A., Bégu, D., Jordana, X., 1996. A reverse transcriptase activity in potato mitochondria. Plant Mol Biol 31, 365–372. doi:10.1007/BF00021796

Mower, J.P., 2005. PREP-Mt: Predictive RNA editor for plant mitochondrial genes. BMC Bioinformatics 6, 96. doi:10.1186/1471-2105-6-96

Mower, J.P., Sloan, D.B., Alverson, A.J., 2012. Plant mitochondrial genome diversity: The genomics revolution, in: Wendel, J.F., Greilhuber, J., Doležel, J., Leitch, I.J. (Eds.), Plant Genome Diversity Volume 1: Plant Genomes, Their Residents, and Their Evolutionary Dynamics. pp. 123–144. doi:10.1007/978-3-7091-1130-7_9

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., Stamatoyannopoulos, J.A., 2012. BEDOPS: High-performance genomic feature operations. Bioinformatics 28, 1919–1920. doi:10.1093/bioinformatics/bts277

Nishizawa, S., Mikami, T., Kubo, T., 2007. Mitochondrial DNA phylogeny of cultivated and wild beets: Relationships among cytoplasmic male-sterility-inducing and nonsterilizing cytoplasms. Genetics 177, 1703–1712. doi:10.1534/genetics.107.076380

Notsu, Y., Masood, S., Nishikawa, T., Kubo, N., Akiduki, G., Nakazono, M., Hirai, A., Kadowaki, K., 2002. The complete sequence of the rice (*Oryza sativa L.*) mitochondrial genome: Frequent DNA sequence acquisition and loss during the evolution of flowering plants. Molecular Genetics and Genomics 268, 434–445. doi:10.1007/s00438-002-0767-1

Ogihara, Y., Yamazaki, Y., Murai, K., Kanno, A., Terachi, T., Shiina, T., Miyashita, N., Nasuda, S., Nakamura, C., Mori, N., Takumi, S., Murata, M., Futo, S., Tsunewaki, K., 2005. Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. Nucleic Acids Res 33, 6235–6250. doi:10.1093/nar/gki925

Oldenburg, D.J., Bendich, A.J., 1996. Size and structure of replicating mitochondrial DNA in cultured tobacco cells. Plant Cell 8, 447–461. doi:10.1105/tpc.8.3.447

Palmer, J.D., Herbon, L.A., 1988. Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. J Mol Evol 28, 87–97. doi:10.1007/BF02143500

Petersen, G., Cuenca, A., Møller, I.M., Seberg, O., 2016. Massive gene loss in mistletoe (Viscum, Viscaceae) mitochondria. Sci. Rep. 5, 1–7. doi:10.1038/srep17588

Phanstiel, D.H., Boyle, A.P., Araya, C.L., Snyder, M.P., 2014. Sushi.R: Flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. Bioinformatics 30, 2808–2810. doi:10.1093/bioinformatics/btu379

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. doi:10.1093/bioinformatics/btq033

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., Knox, E.B., Palmer, J.D., 2013. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm Amborella. Science 342, 1468–1473. doi:10.1126/science.1246275

Richardson, A.O., Rice, D.W., Young, G.J., Alverson, A.J., Palmer, J.D., 2013. The "fossilized" mitochondrial genome of *Liriodendron tulipifera*: Ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. BMC Biol 11, 29. doi:10.1186/1741-7007-11-29

Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F., 2016. VSEARCH: A versatile open source tool for metagenomics. PeerJ 2016, e2584. doi:10.7717/peerj.2584

Sanchez-Puerta, M.V., Abbona, C.C., Zhuo, S., Tepe, E.J., Bohs, L., Olmstead, R.G., Palmer, J.D., 2011. Multiple recent horizontal transfers of the *cox1* intron in Solanaceae and extended co-conversion of flanking exons. BMC Evol. Biol. 11, 277. doi:10.1186/1471-2148-11-277

Sanchez-Puerta, M.V., Cho, Y., Mower, J.P., Alverson, A.J., Palmer, J.D., 2008. Frequent, phylogenetically local horizontal transfer of the *cox1* group I Intron in flowering plant mitochondria. Mol Biol Evol 25, 1762–1777. doi:10.1093/molbev/msn129

Sanchez-Puerta, M.V., Edera, A., Gandini, C.L., Williams, A.V., Howell, K.A., Nevill, P.G., Small, I., 2019. Genome-scale transfer of mitochondrial DNA from legume hosts to the holoparasite *Lophophytum mirabile* (Balanophoraceae). Mol Phylogenet Evol 132, 243–250. doi:10.1016/j.ympev.2018.12.006

Sanchez-Puerta, M.V., García, L.E., Wohlfeiler, J., Ceriotti, L.F., 2017. Unparalleled replacement of native mitochondrial genes by foreign homologs in a holoparasitic plant. New Phytol 1–12. doi:10.1111/nph.14361

Sanchez-Puerta, M.V., Zubko, M.K., Palmer, J.D., 2015. Homologous recombination and retention of a single form of most genes shape the highly chimeric mitochondrial genome of a cybrid plant. New Phytol 206, 381–396. doi:10.1111/nph.13188

Sasaki, T., Yukawa, Y., Miyamoto, T., Obokata, J., Sugiura, M., 2003. Identification of RNA editing sites in chloroplast transcripts from the maternal and paternal progenitors of tobacco (*Nicotiana tabacum*): Comparative analysis shows the involvement of distinct trans-factors for *ndhB* editing. Mol Biol Evol 20, 1028–1035. doi:10.1093/molbev/msg098

Schuster, W., Brennicke, A., 1989. Conserved sequence elements at putative processing sites in plant mitochondria. Curr Genet 15, 187–192. doi:10.1007/BF00435505

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., 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. doi:10.1371/journal.pone.0180484

Skippington, E., Barkman, T.J., Rice, D.W., Palmer, J.D., 2015. Miniaturized mitogenome of the parasitic plant *Viscum scurruloideum* is extremely divergent and dynamic and has lost all nad genes. Proc. Natl. Acad. Sci. U.S.A. 112, E3515–24. doi:10.1073/pnas.1504491112

Sloan, D.B., 2013. One ring to rule them all? Genome sequencing provides new insights into the "master circle" model of plant mitochondrial DNA structure. New Phytol 200, 978–985. doi:10.1111/nph.12395

Sloan, D.B., Alverson, A.J., Chuckalovcak, J.P., Wu, M., McCauley, D.E., Palmer, J.D., Taylor, D.R., 2012a. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. 10, e1001241. doi:10.1371/journal.pbio.1001241

Sloan, D.B., Alverson, A.J., Štorchová, H., Palmer, J.D., Taylor, D.R., 2010. Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm *Silene latifolia*. BMC Evol. Biol. 10, 274. doi:10.1186/1471-2148-10-274

Sloan, D.B., Alverson, A.J., Wu, M., Palmer, J.D., Taylor, D.R., 2012b. Recent acceleration of plastid sequence and structural evolution coincides with extreme mitochondrial divergence in the angiosperm genus *Silene*. Genome Biol Evol 4, 294–306. doi:10.1093/gbe/evs006

Sloan, D.B., Müller, K., McCauley, D.E., Taylor, D.R., Štorchová, H., 2012c. Intraspecific variation in mitochondrial genome sequence, structure, and gene content in *Silene vulgaris*, an angiosperm with pervasive cytoplasmic male sterility. New Phytol 196, 1228–1239. doi:10.1111/j.1469-8137.2012.04340.x

Sloan, D.B., Wu, Z., 2014. History of plastid DNA insertions reveals weak deletion and AT mutation biases in angiosperm mitochondrial genomes. Genome Biol Evol 6, 3210–3221. doi:10.1093/gbe/evu253

Small, I., Suffolk, R., Leaver, C.J., 1989. Evolution of plant mitochondrial genomes via substoichiometric intermediates. Cell 58, 6976–76. doi:10.1016/0092-8674(89)90403-0

Small, I.D., Isaac, P.G., Leaver, C.J., 1987. Stoichiometric differences in DNA molecules containing the *atpA* gene suggest mechanisms for the generation of mitochondrial genome diversity in maize. EMBO J. 6, 865–869. doi:10.1002/j.1460-2075.1987.tb04832.x

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. doi:10.1093/bioinformatics/btu033

Stern, D.B., Lonsdale, D.M., 1982. Mitochondrial and chloroplast genomes of maize have a 12-kilobase DNA sequence in common. Nature 299, 698–702. doi:10.1038/299698a0

Sugiyama, Y., Watase, Y., Nagase, M., Makita, N., Yagura, S., Hirai, A., Sugiura, M., 2005. The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: Comparative analysis of mitochondrial genomes in higher plants. Molecular Genetics and Genomics 272, 603–615. doi:10.1007/s00438-004-1075-8

Takemura, M., Oda, K., Yamato, K., Ohta, E., Nakamura, Y., Nozato, N., Akashi, K., Ohyama, K., 1992. Gene clusters for ribosomal proteins in the mitochondrial genome of a liverwort, *Marchantia polymorpha*. Nucleic Acids Res 20, 3199–3205. doi:10.1093/nar/20.12.3199

Taylor, J.S., Breden, F., 2000. Slipped-strand mispairing at noncontiguous repeats in *Poecilia reticulata:* A model for minisatellite birth. Genetics 155, 1313–1320.

Tu, T., Volis, S., Dillon, M.O., Sun, H., Wen, J., 2010. Dispersals of Hyoscyameae and Mandragoreae (Solanaceae) from the New World to Eurasia in the early Miocene and their biogeographic diversification within Eurasia. Mol Phylogenet Evol 57, 1226–1237. doi:10.1016/j.ympev.2010.09.007

Wang, D., Wu, Y.-W., Shih, A.C.-C., Wu, C.-S., Wang, Y.-N., Chaw, S.-M., 2007. Transfer of chloroplast genomic DNA to mitochondrial genome occurred at least 300 MYA. Mol Biol Evol 24, 2040–2048. doi:10.1093/molbev/msm133

Warren, R.L., Sutton, G.G., Jones, S.J.M., Holt, R.A., 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23, 500–501. doi:10.1093/bioinformatics/btl629

Woloszynska, M., 2010. Heteroplasmy and stoichiometric complexity of plant mitochondrial genomes-though this be madness, yet there's method in't. J Exp Bot 61, 657–671. doi:10.1093/jxb/erp361

Wynn, E.L., Christensen, A.C., 2019. Repeats of unusual size in plant mitochondrial genomes: Identification, incidence and evolution. G3: Genes, Genomes, Genetics 9, 549–559. doi:10.1534/g3.118.200948

Zerbino, D.R., Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18, 821–829. doi:10.1101/gr.074492.107

6. Tables

Table 1. Clustering analysis of short repeats (SR) across Solanaceae mitochondrial genomes using VSEARCH.

	Capsicum annuum	*Hyoscyamus niger*	*Nicotiana tabacum*	*Physochlaina orientalis*	*Solanum lycopersicum*	*Solanum pennellii*
Proportion of SR in the single largest cluster	8.61%	40.54%	16.12%	11.05%	21.74%	28.13%
Proportion of SR in clusters with ≥ 50 sequences (# of clusters)	8.61% (1)	82.19% (8)	16.12% (1)	68.00% (40)	30.57% (2)	28.13% (1)
Proportion of SR in clusters with ≥ 100 sequences (# of clusters)	0.00% (0)	81.09% (6)	0.00% (0)	47.97% (15)	21.74% (1)	28.13% (1)
Proportion of SR in clusters with ≥ 500 sequences (# of clusters)	0.00% (0)	71.78% (3)	0.00% (0)	18.85% (2)	0.00% (0)	0.00% (0)
Proportion of SR in clusters with ≥ 1000 sequences (# of clusters)	0.00% (0)	62.24% (2)	0.00% (0)	0.00% (0)	0.00% (0)	0.00% (0)

7. Figure Legends

Figure 1: Map of the mitochondrial genome of *Physochlaina orientalis.* Full-length genes, repeats greater than 1 kb in length, and plastid-derived regions greater than 200 bp in length are shown. Large repeat pairs are connected with light gray lines.

Figure 2: Genomic comparisons among Solanaceae mitochondria. (a) Genome size and content of *Physochlaina orientalis* and five other Solanaceae mitogenomes. The genome size and the fraction of each genome covered by genes, sequences of plastid origin (MTPTs), tandem repeats (TR), other repeat sequences, and transposable elements (TE) are shown. Also, the overlap between the repeat fraction and TE or MTPTs is indicated. A schematic tree on the left shows the phylogenetic relationships among the species; **(b) Similarity between *Physochlaina orientalis* mtDNA and other angiosperm mitochondria.** Each line depicts the BLAST hits of the *P. orientalis* mitogenome against a custom mitochondrial database. The fraction of the *P. orientalis* mtDNA with similarity to each species or lineage is detailed on the right.

Figure 3: Distribution of repetitive DNA in the mitochondrial genomes of six Solanaceae. Short and intermediate repeat pairs are connected with light gray lines within each circular mitochondrial map. Large repeat pairs are connected with dark gray lines. Black lines on the outer circle indicate tandem repeats. The number of repeat pairs and the fraction of the genome covered by all (R), large (LR), intermediate (IntR), short (SR), and tandem (TR) repeats are shown inside each circle.

Figure 4: Alignment of a region of the *nad1.x2.x3* intron. *Physochlaina orientalis* is the reference sequence. Dots and dashes indicate equal bases and gaps, respectively. Direct repeats (DR) are underlined. Tandem repeat expansions in *P. orientalis* and *H. niger* are shown as rectangles and are similar to the shaded sequence in *P. orientalis*. Colors depict sequence similarity between tandem repeats. The total length of the expanded region is shown on the right.

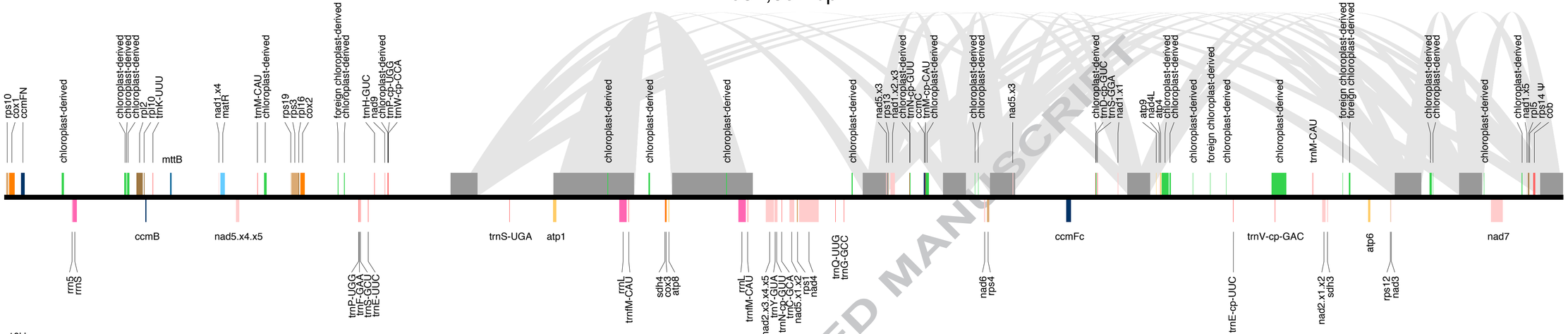
Figure 5: Distribution of short repeated sequences among land plants. Plant species are arranged in phylogenetic order. On the right, the fraction of the mtDNA covered by short repeats (SR) is depicted. Within each bar, the proportion of SR of 20-39 bp (green), 40-59 bp (dark blue), 60-79 bp (dark red), and 80-99 bp (pink) in length are shown. For each species, the proportion of SR clustered in groups of 50 or more sequences is painted with white lines. On the left, inter-species connections between SR found in clusters of 50 or more sequences are depicted in different colors. Those linking species of the same angiosperm order are shown with dark blue lines, between species of different orders with gray lines, and

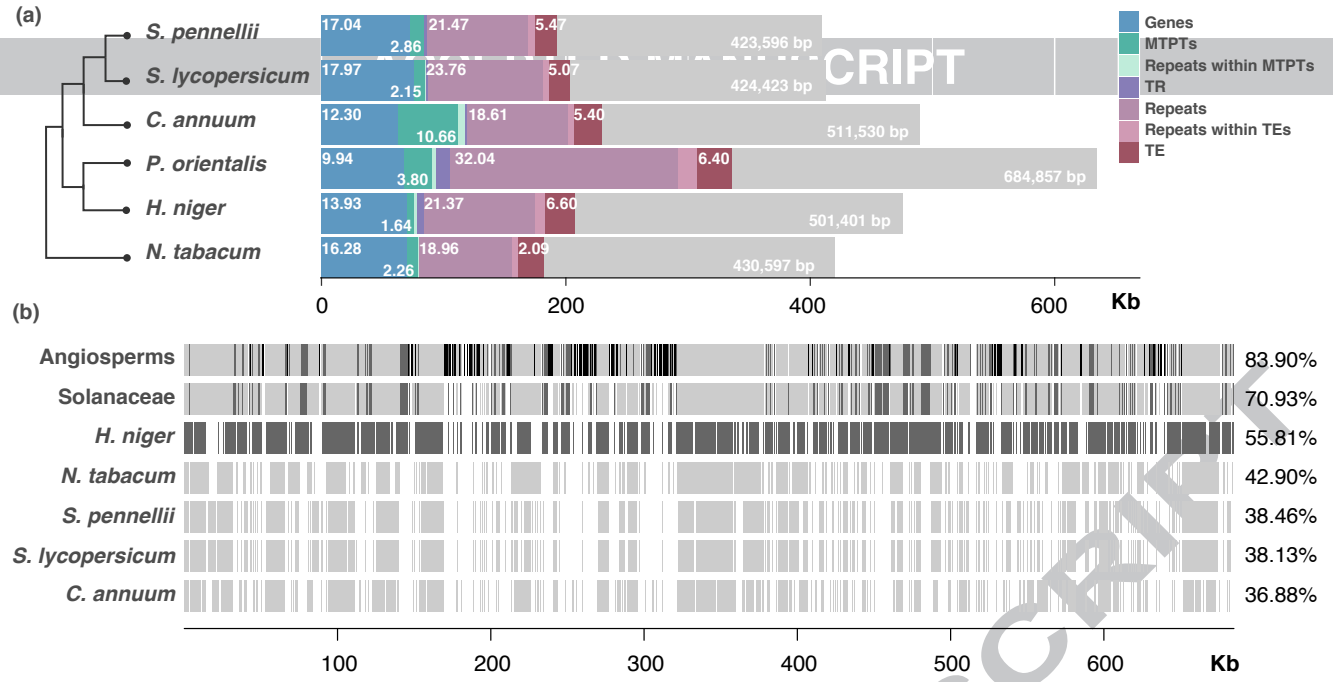
those that connect species involved in known horizontal transfer events with violet lines. Species in normal and boldface indicate that they lack or contain SR in clusters of 50 or more sequences. Connections between two species are shown only once.

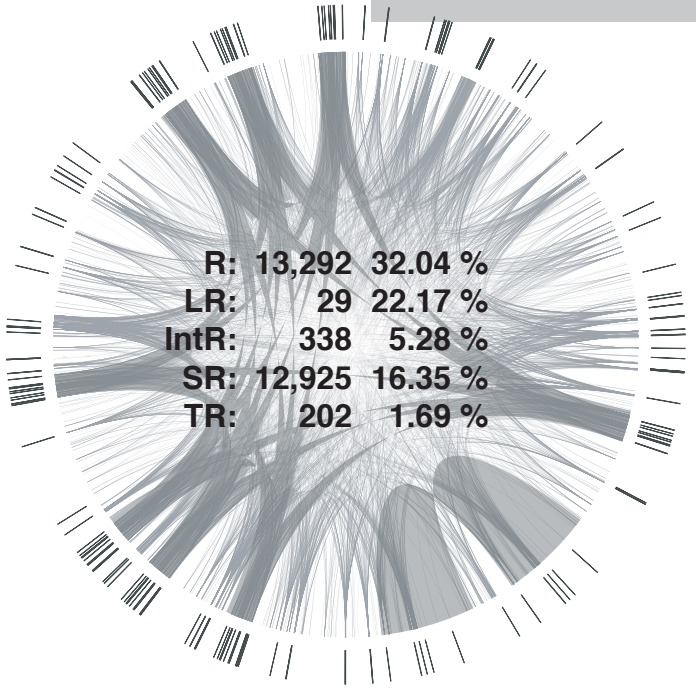
Physochlaina orientalis

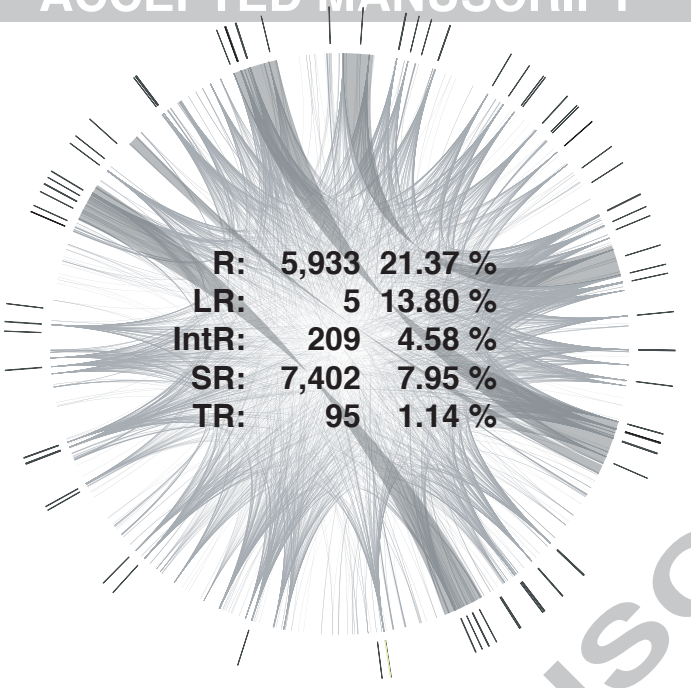
mitochondrial genome

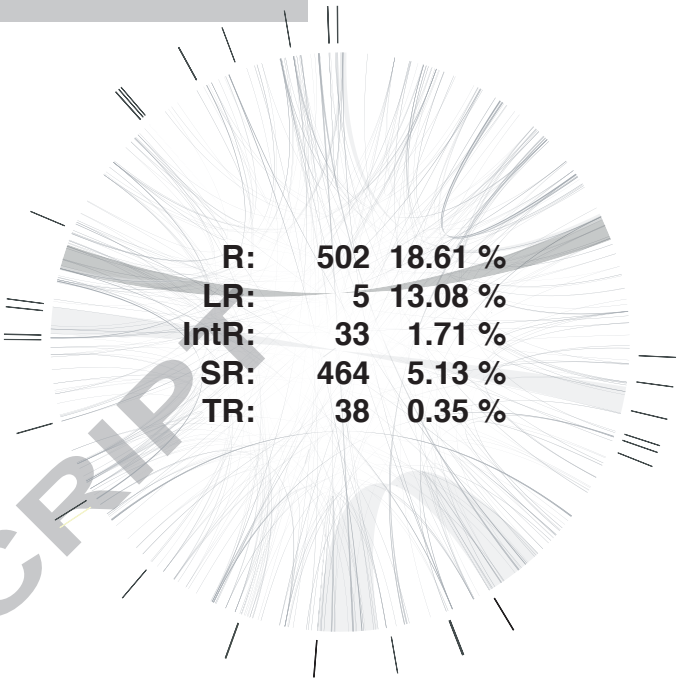
684,857 bp

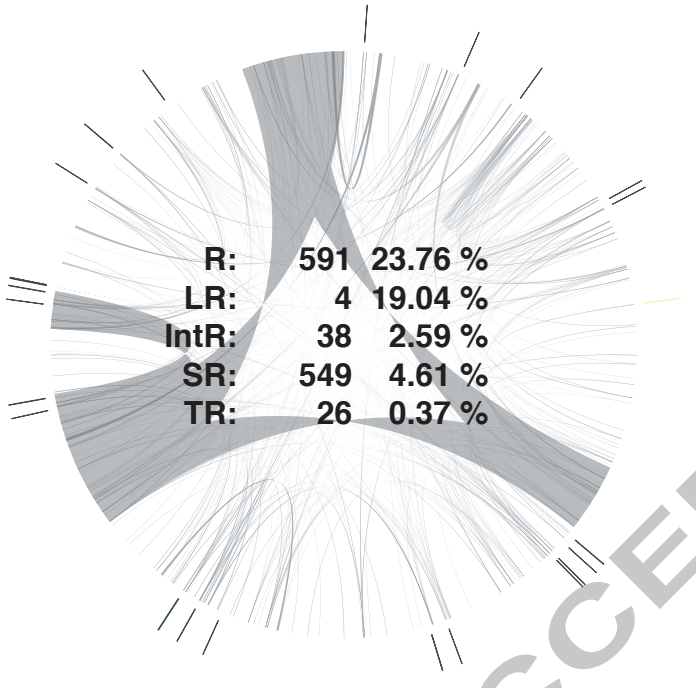


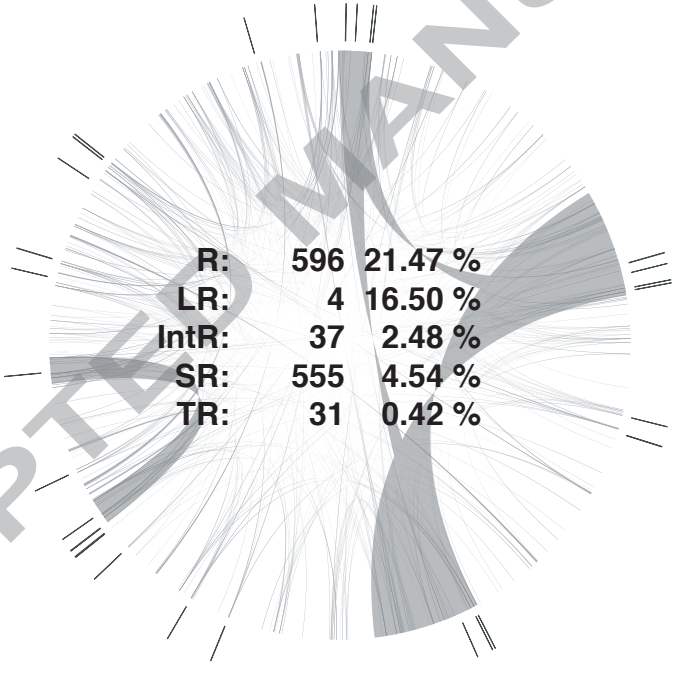

(a)
S. pennellii
S. lycopersicum
C. annuum
P. orientalis
H. niger
N. tabacum
17.04
2.86
21.47
5.47
423,596 bp
17.97
2.15
23.76
5.07
424,423 bp
12.30
10.66
18.61
5.40
511,530 bp
9.94
3.80
32.04
6.40
684,857 bp
13.93
1.64
21.37
6.60
501,401 bp
16.28
2.26
18.96
2.09
430,597 bp
0
200
400
600
Kb
Genes
MTPTs
Repeats within MTPTs
TR
Repeats
Repeats within TEs
TE
(b)
Angiosperms
Solanaceae
H. niger
N. tabacum
S. pennellii
S. lycopersicum
C. annuum
83.90%
70.93%
55.81%
42.90%
38.46%
38.13%
36.88%
100
200
300
400
500
600
Kb

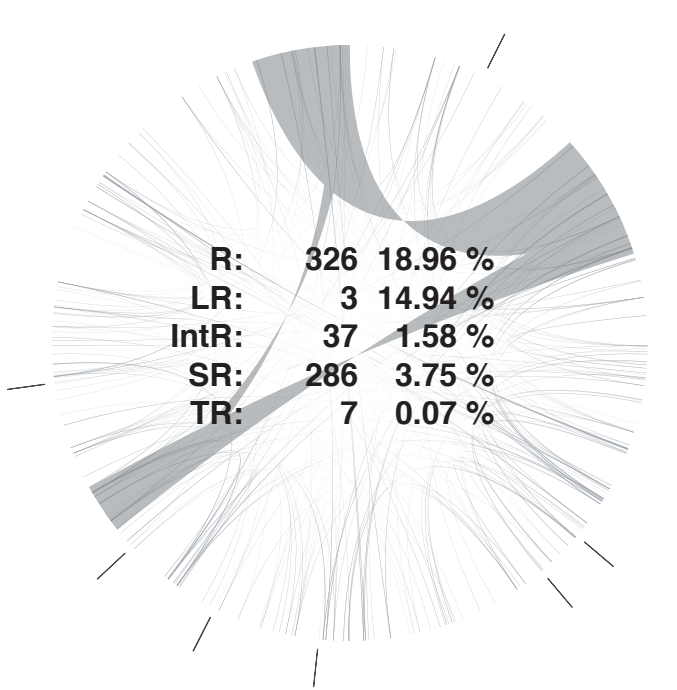


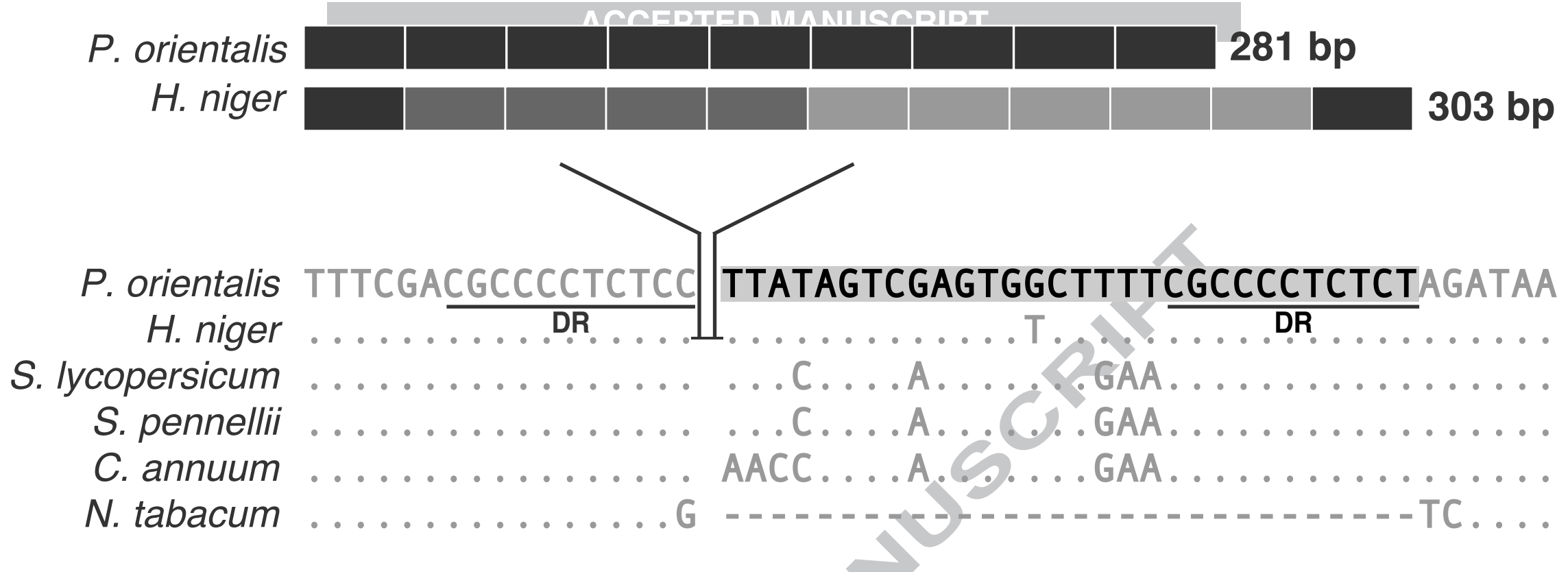












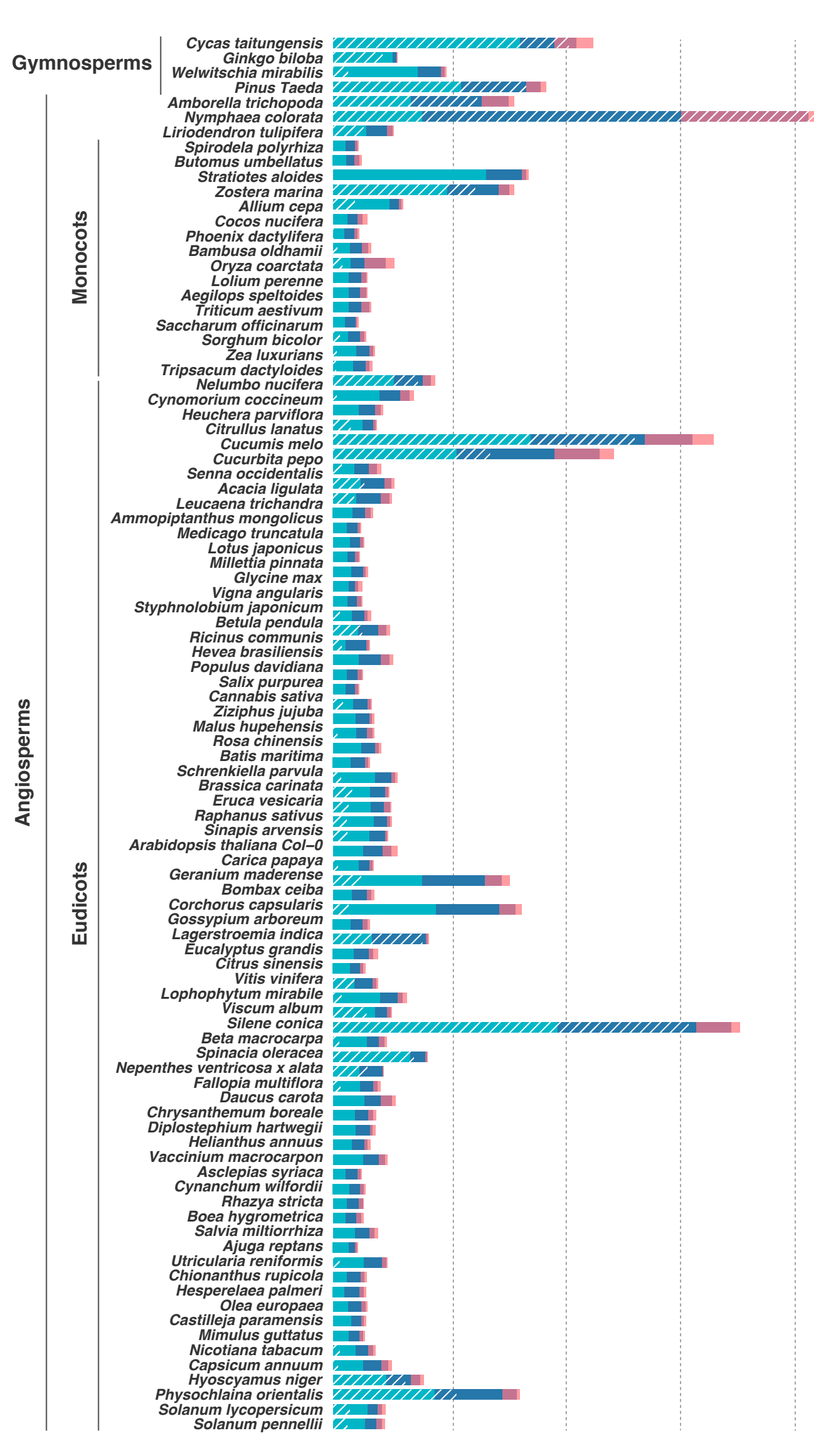
Gymnosperms
Angiosperms
Monocots
Eudicots

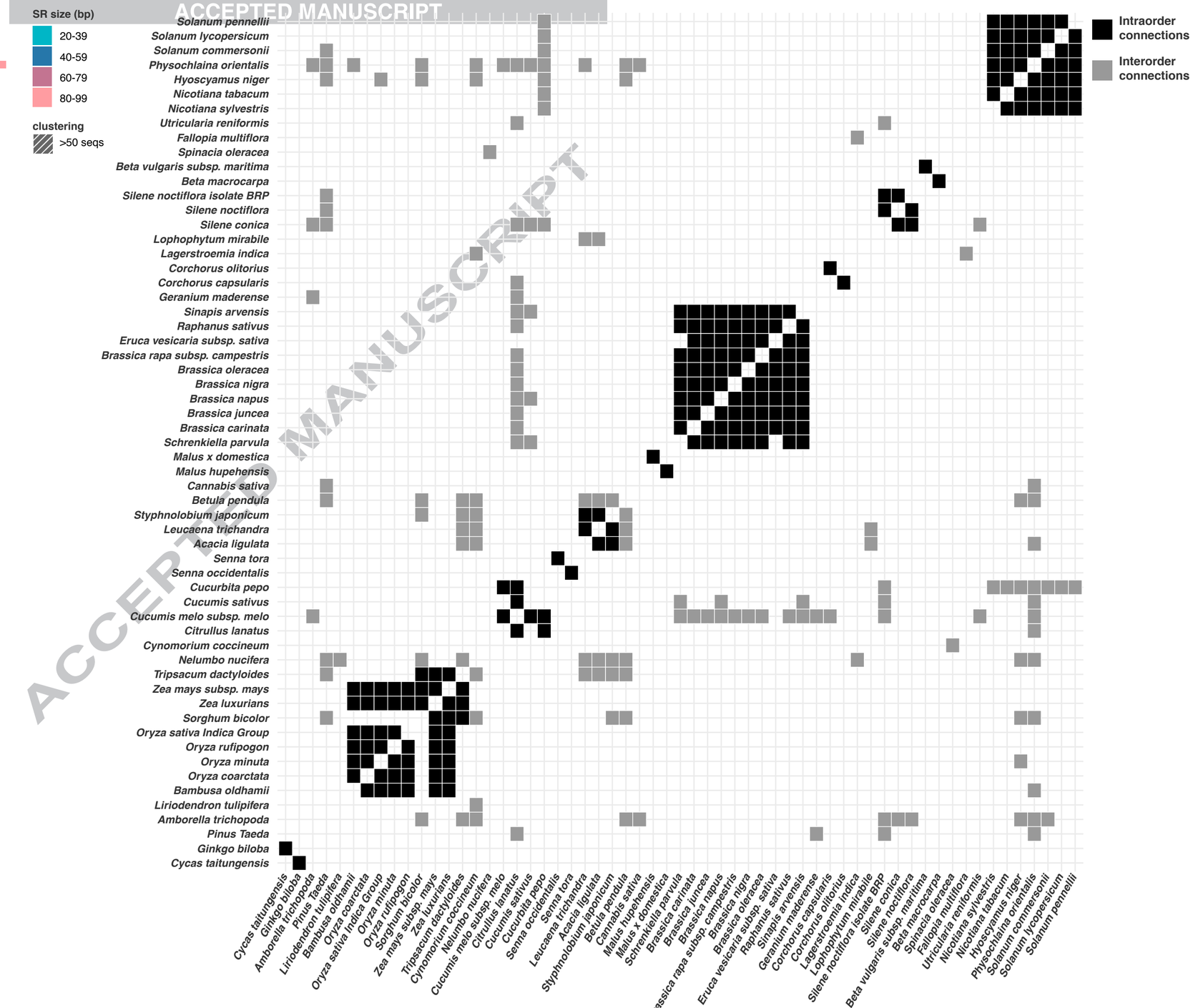
Cycas taitungensis
Ginkgo biloba
Welwitschia mirabilis
Pinus Taeda
Amborella trichopoda
Nymphaea colorata
Liriodendron tulipifera
Spirodela polyrhiza
Butomus umbellatus
Stratiotes aloides
Zostera marina
Allium cepa
Cocos nucifera
Phoenix dactylifera
Bambusa oldhamii
Oryza coarctata
Oryza minuta
Oryza rufipogon
Oryza sativa Indica
Lolium perenne
Aegilops speltoides
Triticum aestivum
Triticum timopheevii
Saccharum officinarum
Sorghum bicolor
Zea luxurians
Zea mays mays
Zea perennis
Tripsacum dactyloides
Nelumbo nucifera
Cynomorium coccineum
Heuchera parviflora
Citrullus lanatus
Cucumis melo
Cucumis sativus
Cucurbita pepo
Senna occidentalis
Senna tora
Acacia ligulata
Leucaena trichandra
Ammopiptanthus mongolicus
Vicia faba
Medicago truncatula
Lotus japonicus
Millettia pinnata
Glycine max
Glycine soja
Vigna angularis
Vigna radiata
Styphnolobium japonicum
Betula pendula
Ricinus communis
Hevea brasiliensis
Populus davidiana
Populus tremula
Populus tremula x alba
Salix purpurea
Salix suchowensis
Cannabis sativa
Ziziphus jujuba
Malus hupehensis
Malus x domestica
Rosa chinensis
Batis maritima
Schrenkiella parvula
Brassica carinata
Brassica juncea
Brassica napus
Brassica nigra
Brassica oleracea
Brassica rapa
Eruca vesicaria
Raphanus sativus
Sinapis arvensis
Arabidopsis thaliana Col–0
Carica papaya
Geranium maderense
Bombax ceiba
Corchorus capsularis
Corchorus olitorius
Gossypium arboreum
Gossypium barbadense
Gossypium davidsonii
Gossypium harknessii
Gossypium hirsutum
Gossypium raimondii
Gossypium thurberi
Gossypium trilobum
Hibiscus cannabinus
Lagerstroemia indica
Eucalyptus grandis
Citrus sinensis
Vitis vinifera
Lophophytum mirabile
Viscum album
Viscum scurruloideum
Silene conica
Silene latifolia
Silene noctiflora
Silene noctiflora BRP
Silene vulgaris
Silene vulgaris KOV
Silene vulgaris KRA
Silene vulgaris MTV
Silene vulgaris S9L
Beta macrocarpa
Beta vulgaris
Spinacia oleracea
Nepenthes ventricosa x alata
Fallopia multiflora
Daucus carota
Chrysanthemum boreale
Diplostephium hartwegii
Helianthus annuus
Vaccinium macrocarpon
Asclepias syriaca
Cynanchum wilfordii
Rhazya stricta
Boea hygrometrica
Salvia miltiorrhiza
Ajuga reptans
Utricularia reniformis
Chionanthus rupicola
Hesperelaea palmeri
Olea europaea
Castilleja paramensis
Mimulus guttatus
Nicotiana attenuata
Nicotiana sylvestris
Nicotiana tabacum
Capsicum annuum
Hyoscyamus niger
Physochlaina orientalis
Solanum commersonii
Solanum lycopersicum
Solanum pennellii

Magnoliales
Alismatales
Arecales
Poales
Proteales
Saxifragales
Cucurbitales
Fabales
Fagales
Malpighiales
Rosales
Brassicales
Geraniales
Malvales
Myrtales
Sapindales
Vitales
Santalales
Caryophyllales
Apiales
Asterales
Ericales
Gentianales
Lamiales
Solanales

SR size (bp)
20-39
40-59
60-79
80-99
clustering
>50 seqs

0
10
20
30
40
Proportion of the genome covered by short repeats (%)





Highlights

- The mitochondrial genome (mtDNA) of *Physochlaina orientalis* is the largest among the sequenced Solanaceae
- The mtDNA of *P. orientalis* presents a rare 8-copy repeat family of 8.2 kb in length and a great number of short repeats (SR) arranged in tandem-like structures
- SR share a common origin in the Solanaceae, but only expanded in the tribe Hyoscyameae
- We propose a mechanism that could explain SR generation and expansion in *P. orientalis* and *Hyoscyamus niger* mtDNAs.
- We study for the first time the short repeat content and their relationships in 136 seed plants