# The complete mitochondrial genome assembly of Capsicum pubescens reveals key evolutionary characteristics of mitochondrial genes of two Capsicum subspecies

**Lin Li**
College of Horticulture, Hainan University
**Xu Lu**
College of Horticulture, Hainan University
**Huizhen Fu**
College of Horticulture, Hainan University
**Zhiwei Wang** ( wangzhiwei@hainanu.edu.cn )
College of Horticulture, Hainan University

**Research Article**

**Keywords:** Pepper, Capsicum pubescens, Mitochondrial genome, Organelle genome, Phylogenetic, Homologous sequences, Positive selection

**Posted Date:** July 19th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-3167937/v1

# Abstract

Pepper (*Capsicum pubescens*), one of the five domesticated pepper species, boasts unique characteristics such as numerous hairs on the epidermis of its leaves and stems, black seeds, and vibrant purple flowers. Previously, no studies have reported on the complete assembly of the mitochondrial genome (mitogenome) of *C. pubescens*. Understanding the mitogenome is crucial for further research on *C. pubescens*. In our study, we successfully assembled the first mitogenome of *C. pubescens*, which has been assigned the GenBank accession number OP957066. This mitogenome has a length of 454,165 bp and exhibits the typical circular structure observed in most mitogenomes. We have annotated a total of 70 genes, including 35 protein-coding genes (PCGs), 30 tRNA genes, 3 rRNA genes, and 2 pseudogenes. Compared to the other three pepper mitogenomes (KJ865409, KJ865410, and MN196478), *C. pubescens* OP957066 exhibits 4 unique PCGs (*atp4*, *atp8*, *mttB*, and *rps1*), while 2 PCGs (*rpl10* and *rps3*) are absent. It's worth mentioning that each of the three pepper mitogenomes (KJ865409, KJ865410, and MN196478) experienced the loss of 4 PCGs (*atp4*, *atp8*, *mttB*, and *rps1*). To further explore the evolutionary relationships, we reconstructed a phylogenetic tree using the mitogenomes of C. pubescens and fourteen other species. The structural comparison and synteny analysis of the above four pepper mitogenomes showed that *C. pubescens* has a higher sequence similarity with KJ865409, and *C. pubescens* experienced rearrangements with the other three pepper mitogenomes. Interestingly, we observed 72 homologous sequences between the mitochondrial and chloroplast genomes, which accounted for 12.60% of the mitogenome, with a total length of 57,207 bp. These sequences encompassed 12 tRNA genes and the rRNA gene (*rrn18*). Remarkably, selective pressure analysis suggests that the *nad5* gene underwent obvious positive selection. Furthermore, the single base mutation in three genes (*nad1*, *nad2*, and *nad4*) resulted in amino acid change. This study has provided a high-quality mitogenome of pepper, offering valuable molecular data for future investigations into the genetic information exchange between organelle genomes in pepper.


# Introduction

The *Capsicum* genus includes five domesticated cultivars, which are *C. pubescens*, *Capsicum annuum*, *Capsicum baccatum*, *Capsicum chinense*, and *Capsicum frutescens* (Ďúranová et al. 2021). Among them, *C. annuum*, *C. chinense*, and *C. frutescens* have been cultivated in many countries and regions worldwide (Jarret et al. 2019, (Palombo et al. 2022). *C. pubescens* has an abundance of hairs on the leaves and stems' epidermis and produces purple flowers. Additionally, it is distinctive from other pepper cultivars in the presence of black seeds (Ibiza et al. 2012). *C. pubescens* holds considerable economic value. It is widely planted in the Andes (Palombo et al. 2022) and is often used as a condiment and as a fruit for consumption (Rodríguez-Burruezo et al. 2009). Moreover, it provides support and shade in orchards and possesses medicinal properties, including anti-inflammatory, antibacterial, and digestive benefits (Barboza et al. 2022). Interestingly, Of the five domesticated pepper cultivars, *C. pubescens* is the hardy cultivar (Gao et al. 2022).

Mitochondria play a crucial role in converting the proton concentration gradient in living cells into ATP, serving as the "energy factories" of cells. However, it's essential to note that mitochondria have diverse functions beyond their primary role as the powerhouse of the cell. They are involved in maintaining calcium homeostasis, facilitating biosynthesis of heme and ubiquinone, assembling ironsulfur clusters, regulating fatty acid metabolism, and performing various other vital functions. In terms of inheritance, while most plants inherit nuclear genetic information from both parents, chloroplast and mitochondrial DNA are typically inherited maternally (Ye et al. 2017). Mitochondrial DNA differs from chloroplast DNA and nuclear DNA, with a generally lower substitution rate in plant mitochondria compared to chloroplast and nuclear DNA (Fan et al. 2022). Mitochondrial DNA is commonly used to study evolutionary relationships, hybridization events, and nuclear interactions among species (Zhang et al. 2019, (Wang et al. 2021, (Wang et al. 2023).

With the rapid development of second-generation sequencing and genome assembly technology, completing genome sequencing has become efficient and affordable. As a result, there has been a substantial increase in the publication of information regarding organelle genomes (Mahapatra et al. 2021, (Odahara et al. 2021). As of July 2022, the GenBank database contains 7,408 published complete organelle genomes of land plants, including 326 mitogenomes and 7,182 chloroplast (cp) genomes. The number of published cp genomes surpasses that of mitogenomes because assembling mitogenomes is more complex compared to cp genomes (Ye et al. 2017, (Bi et al. 2020). Obtaining and assembling complete mitogenome sequences from plants pose challenges due to the presence of large repetitive sequences (Xiong et al. 2021). Additionally, mitogenomes often exhibit extensive recombinations and rearrangements, predominantly caused by the abundance of repetitive sequences (Woloszynska et al. 2009). Some repeats maintained a certain degree of specificity between species, so these specific repeats played an essential role as detailed genetic markers in studying the evolutionary connection between species (Wu et al. 2014).

Typically, mitogenomes of plants are circular, except for *Oryza sativa*, which has linear mitogenomes (Notsu et al. 2002). The size of plant mitogenomes ranges from 66 kb in *Viscum scurruloideum* (Skippington et al. 2015) to 11.3 Mb in *Silene conica (*Sloan et al. 2012), with most ranging from 200 to 800 kb. The size variation is primarily due to the absorption of foreign DNA from other organisms by the mitogenome and the complex nature of repetitive sequences, leading to extensive changes in size (Bergthorsson et al. 2003, (Wynn et al. 2019). In other words, besides variations in the size of repeat regions, it's essential to consider another significant contributor to mitogenome diversity—plastid DNA (cpDNA)—which enters mitochondria through the process of endosymbiotic gene transfer (EGT). In addition to size, plant mitogenomes also exhibit significant variation in gene length and content (Richardson et al. 2013, (Ye et al. 2017).

Although various plant mitogenomes have been published, there has been no relevant report on the sequencing of the *C. pubescens* mitogenome. In our study, we have sequenced and assembled the *C. pubescens* mitogenome and obtained a circular structure with a size of 454,165 bp. We are comprehensively describing the *C. pubescens* mitogenome, including its genomic features, codon usage

bias, repetitive sequences, and RNA editing sites. Additionally, we are combining other sequenced and published mitogenomes of higher plants to comprehensively analyze the structure and evolution of the mitogenome of *C. pubescens*.

## Materials and methods

# Plant material, DNA extraction, and sequencing

The pepper (HNUCP0006) seeds used in this study are obtained from the Pepper Germplasm Bank of Hainan University (E: 110°33'52.29", N: 20°06'22.48"). Wrap the seeds with gauze and place them in water at 50°C for treatment, allowing the water to cool naturally. After 6 hours, remove the seeds from the water bath and keep them in a dark environment for 5–7 days while ensuring they remain moist. Sow the treated seeds in 50-cell plug trays and plant them in the plant growth room at Hainan University, where the temperature is maintained at 26°C and there is a light-dark cycle of 16/8 hours. After a period of two months, transfer the healthy pepper seedlings from the plug trays into a dark environment and allow them to grow for 7 days. During this time, the temperature is maintained at room temperature, and watering is done once on the 0th day, the third day, and the sixth day. On the eighth day, discontinue the dark treatment of the pepper seedlings, resulting in pepper seedlings with etiolated leaves. Collect 10 g of etiolated leaves, reserving 5 g as backup, and conduct the isolation process at Nanjing Jisihuiyuan Biotechnology Co., Ltd. Extract mitochondrial genomic DNA from the samples and use the Illumina Novaseq6000 platform for next-generation sequencing. Filter the second-generation raw data to obtain high-quality reads using fastp software (v0.20.0). Perform third-generation sequencing using an Oxford Nanopore PromethION sequencer and filter the third-generation sequencing data using filtlong software (v0.2.1).

# Assembly and annotation of the mitogenome

Acquire all three-generation sequencing data for the mitogenome using minimap2 software (v2.1), and correct the obtained three-generation data using canu software. Align the second-generation data to the corrected sequences using bowtie2 software (v2.3.5.1). Then use Unicycle (v0.4.8, default parameter) to splice the second-generation data on the comparison and the third-generation data after correction to obtain the assembly result.

Annotate the tRNA of the *C. pubescens* mitogenome using tRNAscanSE software. Use Open Reading Frame Finder to annotate the ORFs (Open Reading Frame), with the minimum sequence length set to 102 bp. Remove redundant sequences and align those over 300 bp to the NR database for annotation. Finally, compare the encoded protein and rRNA with published plant mitogenome sequences through a blast search, and make manual adjustments based on closely related species for further refinement. Use the OGDRAW online tool for mapping the mitogenome.

# Prediction of RNA editing using software and validation with RNA-seq data

Utilize the online website (http://prep.unl.edu/) to predict RNA editing sites. Align the RNA-seq data (project number PRJNA822667) with the mitogenome data to identify potential RNA editing sites. First, employ bowtie2 software (version 2.3.5.1) to align the RNA sequencing data with the coding sequence (CDS) of the PCGs in the mitogenome. Next, use Samtools software (version 1.9) to filter out sequences with alignment quality scores greater than or equal to 40. Identify sites with single nucleotide polymorphisms (SNPs) between the sequencing data and the reference genome.

For a site to be considered a potential RNA editing site, it must meet the following criteria: SNP depth exceeding 3x, and the SNP depth being at least 20% of the total depth at that site.

# Codon composition and RSCU analysis

Codons exhibit degeneracy, with each amino acid having 1–6 corresponding codons. The uneven usage of synonymous codons is commonly referred to as codon bias or "Relative Synonymous Codon Usage (RSCU)". Use the self-written Perl script to filter the unique CDS and calculate RSCU.

# Repeat sequences analysis

We utilize the vmatch software (v2.3.0) to identify interspersed repeats. The criteria we set are a minimum length of 30 base pairs and a Hamming distance of 3. To investigate simple sequence repeats (SSRs), we use the MISA online software. We set a minimum distance of 100 base pairs between two SSRs. Through this analysis, we identified 8, 4, 4, 3, 3, and 3 repeats, corresponding to 1, 2, 3, 4, 5, and 6 bases, respectively. For the detection of tandem repeats, we employ the Tandem Repeats Finder v4.09 software, which identifies matches more significant than 95%.

# Phylogenetic analysis, comparison of genome size and GC content

We download 18 green plant mitogenomes from NCBI and perform phylogenetic analysis. This analysis includes the mitogenomes of *C. pubescens* and 14 other species, including 9 eudicots, 3 monocots, and 2 gymnosperms. Multiple sequence alignments of 23 common PCGs in 15 species are performed using MAFFT v7.490 software (using default parameters). For the construction of the phylogenetic tree, we utilize MEGA 11 and set the Bootstrap value to 1,000, using the Maximum-likelihood method. To enhance the visual presentation of the evolutionary tree, we utilize Evolview v2 software. Furthermore, we compare the size and GC content of the *C. pubescens* mitogenome with 18 other green plants' mitogenomes and visualize the results using Origin 2023 software.

# Analysis of non-synonymous (Ka) and synonymous (Ks) mutation rates

The CDS sequences of 30 common PCGs in *C. pubescens* and three pepper species are calculated using Tbtools (v1.0987663) software. The pepper species include *C. annuum* cultivar Jeju (KJ865410), *C. annuum* cultivar CMS line FS4401 (KJ865409), and *C. annuum* var. *glabriusculum* (MN196478).

# Mitogenome structure comparison and collinearity analysis

The online tool proksee (https://proksee.ca/) is used for conducting a comparative analysis of the mitogenome structure between *C. pubescens* and other species. The species included in the analysis are *A. thaliana*, *S. melongena*, *S. lycopersicum*, *C. annuum* cultivar Jeju, *C. annuum* var. *glabriusculum*, and *C. annuum* cultivar CMS line FS4401. Mauve (https://darlinglab.org/mauve/mauve.html) software performed sequence homology and collinearity analysis between the *C. pubescens* mitogenomes and three other peppers, including *C. annuum* cultivar CMS line FS4401, *C. annuum* var. *glabriusculum*, and *C. annuum* cultivar Jeju.

# Chloroplast and mitochondrial homologous sequence analysis

The BLAST (v2.10.1) software is detected homologous sequence fragments (identity% > 70%, E-value = 10E-5) between the mitogenome and cp genome of *C. pubescens*. The cp genome of *C. pubescens* is assembled using the next-generation sequencing results obtained from this study. The circular map of the cp genome is created using the web application OGDRAW. The nuclear genomes of *C. baccatum* and *C. chinense* are obtained from the NCBI database, while the nuclear genome of *C. annuum* is retrieved from the Ensembl plants database.

# Results

### Genomic characterization of the mitogenome of C. pubescens

The *C. pubescens* mitogenome is 454,165 bp in length and has a typical circular structure (Fig. 1). The genome nucleotide composition is 28.04% A, 27.68% T, 22.29% G, and 22.00% C. Moreover, the GC content is 44.29% (Table 1). We have annotated a total of 70 genes in the *C. pubescens* mitogenome, which includes 35 PCGs, 30 tRNA genes, 3 rRNA genes, and 2 pseudogenes (Table S1). The coding genes have a combined length of 37,303 bp, accounting for 8.21% of the entire mitogenome. Among them, PCGs make up 6.50% of the mitogenome, while rRNA and tRNA genes account for 1.22% and 0.50%, respectively (Table 1). The mitogenome of *C. pubescens* encodes 35 proteins, which can be categorized into different groups (Table S2): ribosomal small subunit (SSU), NADH dehydrogenase, ATP synthase, cytochrome c biogenesis, cytochrome c oxidase, ribosomal large subunit (LSU), transport membrane protein, maturases, succinate dehydrogenase, ubiquinol cytochrome c reductase, NADH dehydrogenase subunits, ATP synthase subunits, cytochrome c oxidase subunits, ubiquinol cytochrome c reductase subunits, and succinate dehydrogenase subunits. There are 10 intronic genes (*ccmFC*, *cox2*, *nad1*, *nad2*, *nad4*, *nad5*, *nad7*, *rpl2*, *rps1*, and *trnY-GTA*) in the mitogenome of *C. pubescens*, with a total of 24 introns. The *nad1*, *nad2*, *nad5*, and *nad7* genes each contain 4 introns, while the *nad4* gene contains 3. Additionally, the *ccmFC*, *cox2*, *rpl2*, *rps1*, and *trnY-GTA* genes each have only one intron (Table S2).

Table 1
Genomic features of the *C. pubescens* mitogenome.

| Feature | Size (bp) | A (%) | T (%) | G (%) | C (%) | GC (%) | Proportion in Genome (%) |
|---|---|---|---|---|---|---|---|
| Mitogenome | 454,165 | 28.04 | 27.68 | 22.29 | 22 | 44.29 | 100 |
| PCGs | 29,514 | 26.24 | 31.02 | 21.63 | 21.11 | 42.74 | 6.5 |
| tRNAs | 2,262 | 22.55 | 26.22 | 28.43 | 22.81 | 51.24 | 0.5 |
| rRNAs | 5,527 | 26.13 | 22.24 | 28.86 | 22.78 | 51.64 | 1.22 |

# Protein coding genes (PCGs) and codon usage analysis

The length of the 35 PCGs in the *C. pubescens* mitogenome is 29,514 bp, accounting for 79.12% of the total length of all coding genes (37,303 bp). All PCGs start with the ATG codon as the start codon. The stop codons are used with the following frequencies: TAA (51.43%), TGA (34.29%), and TAG (14.29%) (Table S3). Analyzing the frequency of amino acid usage in the PCGs, it can be observed that Leu is the most commonly used amino acid, followed by Ser, Ile, Gly, Arg, Phe, Ala, and Val (Table S4). We investigate the relative synonymous codon usage (RSCU) of the PCGs in the *C. pubescens* mitogenome and find 30 optimal codons (RSCU > 1): GCU, UAU, CCU, UAA, CAA, GGA, CAU, ACU, UUA, AGA, GAU, UCU, AAU, GAA, AUU, CGA, GGU, CUU, GUU, CGU, AAA, UUG, GUA, UGU, UCA, UUU, CCA, AGU, ACA, and UGA. The most commonly used codons are UUU (Phe), AUU (Ile), and UUC (Phe). From Fig. 2, we can observe that the two amino acids, Met (ATG) and Trp (TGG), present no preference due to having only one codon, while every other amino acid has its preferred codon (Table S4).

### Analysis of repeat sequences in the C. pubescens mitogenome

We have discovered 473 interspersed repeats in the *C. pubescens* mitogenome (> 30 kb). Among these repeats, there are 246 forward repeats and 227 palindromic repeats, but no complement or reverse repeats (Fig.s 3a and 3b). The total length of these interspersed repeats is 38,393 bp, accounting for 8.45% of the mitogenome's length (454,165 bp). Most of the repeats we find range from 30 to 39 bp in length, with a total of 285 repeats, making up approximately 60.25% of the total repeats. Additionally, we identify 5 repeats that exceed 1,000 bp in length, ranging from 1,347 bp to 4,425 bp (Table S5). Genes that appear in large repeats may generate multiple copies (Ye et al. 2017). In the *C. pubescens* mitogenome, we find that the gene with the most extensive repeat length of 4,425 bp is *trnL-CAA*, which currently has three copies (Table S2).

Microsatellites or simple sequence repeats (SSRs) are DNA fragments containing 1–6 base pairs, and they are widely used in genotyping (Liu et al. 2014, (Grover et al. 2016). We have discovered 359 SSRs in the *C. pubescens* mitogenome (Table S6). When we examine the different proportions of SSR sites, we find that dinucleotide repeats have the highest number with 154, accounting for 42.90% of the total SSR

count. Monomer repeats followed closely with 143 repeats, accounting for 39.83%. The number of hexanucleotide repeats is the least, with only 2, accounting for a mere 0.56% (Fig. 4). Among monomer SSRs, 86.71% are monomer repeats composed of A/T bases. Among dinucleotide SSRs, 64.29% are dinucleotide repeats composed of AG/CT bases.

Tandem repeats can be found in eukaryotic and some prokaryotic genomes (Wang 2005). They usually refer to repetitive sequences formed by taking 1–200 bases as the repeating unit and then connecting them in series, also known as satellite DNA. By setting the matching degree to be greater than 95%, we have identified 17 tandem repeats with a length of 7–39 bp in the *C. pubescens* mitogenome (Table S7).

# Prediction of RNA editing sites

The online website (http://prep.unl.edu/) is used to make predictions of RNA editing sites. Prediction of RNA editing sites for 35 PCGs in the mitogenome of *C. pubescens* reveals 448 RNA editing sites (Table S8). Among these RNA editing sites, 89 (19.87%) appear in the first base position of the codon, 359 (80.13%) in the second base position, and there are no RNA editing occurrences in the third base position. These PCGs have varying numbers of RNA editing sites, ranging from 2 to 35. The *rps10* gene and the *sdh3* gene have the least 2 RNA editing sites, while the *ccmB* gene has the most 35 RNA editing sites (Fig. 5). Table S9 shows that 11.83% (53 sites) of amino acids do not change their hydrophilicity after RNA editing. On the other hand, 47.32% (212 sites) are predicted to change from hydrophilic to hydrophobic. 32.81% (147 sites) of amino acids remain unchanged in hydrophobicity, while 7.37% (33 sites) are predicted to change from hydrophobic to hydrophilic. Additionally, 0.67% (3 sites) of amino acids change from hydrophilic to terminated. Furthermore, most amino acids tend to be converted from Ser to Leu (23.88%, 107 sites), Pro to Leu (21.88%, 98 sites), and Ser to Phe (14.51%, 65 sites). The remaining 178 RNA editing sites are distributed among other RNA editing types, including His to Tyr, Arg to Cys, Thr to Ile, Thr to Met, Arg to Trp, Ser to Leu, Ser to Phe, Pro to Ser, Pro to Leu, Pro to Phe, Leu to Phe, Ala to Val, and Gln to X (X = stop codon).

We align the RNA-seq data (NCBI project number PRJNA822667) with the mitogenome data and search for potential RNA editing sites (Table S10). The results reveal the prediction of 454 RNA editing sites in 35 PCGs, with a frequency of approximately 15.6% (71 out of 454) at the third codon position. It is evident that this result is not solely achieved through software prediction, highlighting the necessity of verifying it using RNA-seq data.

# Phylogenetic analysis of mitogenomes in higher plants

A phylogenetic tree is created for the *C. pubescens* mitogenome and the mitogenomes of 14 other species, including 9 eudicots, 3 monocots, and 2 gymnosperms, to explain the evolution of the *C. pubescens* mitogenome. Twenty-three PCGs from 15 species are used for analysis, including *matR*, *cox2*, *cox3*, *cox1*, *atp9*, *nad4L*, *atp1*, *rpl16*, *rps4*, *rps12*, *nad4*, *nad3*, *nad2*, *nad1*, *ccmFN*, *ccmFC*, *nad7*, *nad6*, *nad5*, *cob*, *ccmC*, *ccmB*, and *nad9* genes. The gene sequences of each species are concatenated head-to-tail, and then multiple sequence alignments are performed using MAFFT v7.490

(https://mafft.cbrc.jp/alignment/software/) software (default parameters). The aligned sequences are imported into MEGA 11 to predict the most appropriate amino acid substitution model (GTR + G). A maximum-likelihood phylogenetic tree is constructed using MEGA 11 (Bootstrap value is set to 1,000). As shown in Fig. 6, among the 12 nodes of the phylogenetic tree, only one node's bootstrap support value is less than or equal to 90, while the other 11 nodes' bootstrap support value is above 90. Phylogenetic trees strongly support the division of eudicots and monocots into two clades and the division of angiosperms and gymnosperms into two clades. The clustering results of this phylogenetic tree conform to the hierarchical relationship among species at different family and genus levels. It is displayed that our clustering results of common PCGs based on the mitogenome are reliable. The scientific names and GenBank accession numbers are shown in Table S11.

### Comparison of mitogenome size, GC content, and PCGs between C. pubescens and other species

Organelle genomes have two main parameters, size and GC content. The mitogenome size and GC content of *C. pubescens* are compared with 18 other green plants, including 9 eudicots, 3 monocots, 2 gymnosperms, 2 bryophytes, and 2 phycophyta. The species abbreviations of these plants and their mitogenome GenBank accession numbers are displayed in Table S11. Fig.s 7a and 7b show that the mitogenomes range in size from 62,477 bp (*Chlorella heliozoae*) to 680,603 bp (*Zea mays*). Compared with land plants, the mitogenomes of phycophyta and bryophytes are generally smaller, while *C. pubescens* (474,330 bp) has an average size. Likewise, the GC content of the mitogenome varies, ranging from 32.24% in *C. heliozoae* to 50.36% in *Ginkgo biloba*. Generally, the GC content of angiosperms, including eudicots and monocots, is lower than that of gymnosperms but higher than that of phycophyta and bryophytes. Remarkably, *C. heliozoae* and *Nitella hyalina*, with GC contents of 32.24% and 41.00%, respectively, indicate that the GC contents of phycophyta fluctuate wildly. In contrast, although mitogenome sizes in angiosperms vary widely, ranging from 281,132 bp (*Daucus carota*) to 680,603 bp (*Z. mays*), their GC content has been remarkably conserved over evolution, usually stable at around 44%.

PCGs in plants are often lost during evolution (Bi et al. 2020). We investigated the loss of PCGs in eudicots, monocots, gymnosperms, bryophytes, and phycophyta (Fig. 8). Compared to highly variable ribosomal proteins and genes in succinate dehydrogenase, most PCGs are conserved in different plant mitogenomes, especially in the following genes: ATP synthase, cytochrome c biogenesis, ubichinol cytochrome c reductase, cytochrome c oxidase, maturases, transport membrane protein, and NADH dehydrogenase (Chang et al. 2013). To some extent, the conservation of these genes indicates that they play a crucial role in plants' mitochondrial function. As indicated by the red dashed box, when compared with the other three pepper mitogenomes, 4 PCGs (*atp4*, *atp8*, *mttB*, and *rps1*) are unique to *C. pubescens*, while 2 PCGs (*rpl10* and *rps3*) are missing (Fig. 8).

# Non-synonymous (Ka) and synonymous (Ks) mutation rate analysis of PCGs

In this study, we calculate the Ka/Ks values of 30 PCGs shared by the *C. pubescens* mitogenome and the mitogenomes of three peppers (KJ865409, KJ865410, and MN196478) using TBtools software (Table S12). Following a previous study, we change the part of the gene with Ka/Ks ratio that does not apply (NA) to zero (Androsiuk et al. 2022). The results indicate that the PCGs shared by *C. pubescens* and the other three peppers are close homologs because most PCGs have zero Ka and Ks values. When comparing *C. pubescens* and KJ865409, only two genes (*sdh3* and *rps10*) have Ka/Ks values less than 0, signifying that they have undergone negative selection. In contrast, the Ka/Ks values of the remaining genes are all equal to 0 or -Infinity (Ks = 0). In the other two comparison combinations, the Ka/Ks values of the *sdh3* and *rps10* genes are also less than 0. Furthermore, the only gene that undergoes negative selection is *rpl16* in the combination of *C. pubescens* vs. MN196478. The genes undergoing positive selection during the evolution process, with a Ka/Ks value greater than 1 (*nad5*), exist only in the combination of *C. pubescens* vs. KJ865410 and *C. pubescens* vs. MN196478.

To further comprehend the sequence differences among the 30 PCGs in the four pepper mitogenomes, we use the MAFFT v7.490 software (default parameters) to carry out multiple sequence alignments for the 30 PCGs, respectively (Fig. 9). In general, 13 PCGs have sequence differences caused by at least one base change, although the remaining 17 PCGs have no sequence differences. PCGs with sequence differences include *matR*, *rps10*, *rpl2*, *nad1*, *nad2*, *nad4*, *nad5*, *rps4*, *cox2*, *ccmFC*, *ccmFN*, *rpl16*, and *sdh3*. Among them, *rps10*, *rpl16*, and *sdh3* have a wide range of base deletions (more than 10 bases) in the four peppers, but more deletions occur in KJ865409, KJ865410, and MN196478. There are a few basic changes for the remaining PCGs. Some PCGs, such as *matR*, *rpl2*, *rps4*, and *ccmFC*, are not affected in their amino acids as a result of base substitution. Some PCGs, including *nad1*, *nad2*, *ccmFN*, and *nad4*, have amino acid changes as a result of base substitution.

It is worth mentioning that the only gene undergoing positive selection is the *nad5* gene. With *C. pubescens* as a reference, the *nad5* gene of the other two peppers (KJ865410 and MN196478) has a substitution of the C/T base at the 1,863rd base (Fig. 9a). However, due to the base change, the amino acid remains unchanged, resulting in a synonymous mutation. Furthermore, between the 1,447th and 1,468th bases, KJ865410 and MN196478 have multiple base substitutions, leading to changes in amino acids, causing non-synonymous mutations. However, the sequence difference of the *nad5* gene between *C. pubescens* and KJ865409 is insignificant, with only a substitution of the C/T base at the 1,863rd position. Additionally, the amino acid remains unchanged due to the base change.

In summary, a significantly greater number of genes undergo negative selection compared to positive selection. Moreover, base substitutions in some PCGs may result in changes in their corresponding amino acids.

# Structural comparison and synteny analysis of mitogenomes

The online tool proksee (https://proksee.ca/) is used to perform comparative analyses of mitogenome structures across species. The first to sixth tracks, moving from the outside to the inside of the circles,

indicate the results of blast alignment of DNA sequences of *C. pubescens* with *A. thaliana* (BK010421), *S. melongena* (MT122962), *S. lycopersicum* (MF034193), *C. annuum* cultivar Jeju (KJ865410), *C. annuum* var. *glabriusculum* (MN196478), and *C. annuum* cultivar CMS line FS4401 (KJ865409), respectively (Fig. 10a). It can be observed that *C. pubescens* exhibits higher sequence similarity with KJ865409 compared to several other groups. Additionally, KJ865410 shows a high degree of sequence similarity with MN196478, which is consistent with the results demonstrated by the phylogenetic tree constructed using the conserved PCGs between them (Fig. 6). The Fig. clearly illustrates that, in comparison to BK010421 of Brassicaceae, *C. pubescens* shares higher sequence similarity with other species of Solanaceae at the family and genus level.

*C. pubescens* is being compared with three other pepper mitogenomes, namely KJ865409, MN196478, and KJ865410, using mauve software (default parameters) to analyze the sequence homology and synteny relationship between them (Fig. 10b). Using *C. pubescens* as the reference genome, the results indicate that *C. pubescens* exhibits extensive recombination with the other three pepper mitogenomes. The relative positions and order of the homology blocks of KJ865410 and MN196478 are mostly identical, indicating a high degree of homology between them. Furthermore, KJ865409, MN196478, and KJ865410 display numerous inverted homologous blocks, suggesting that rearrangements between the *C. pubescens* mitogenome may have occurred and these three pepper mitogenome inverted homologous blocks.

# Homologous sequence analysis of cp genome and mitogenome

The cp genome is highly conserved when compared with the mitogenome. Furthermore, gene transfer from the mitogenome to the cp genome, known as PTMT, rarely occurs, while gene transfer from the cp genome to the mitogenome, known as MTPT, is relatively common. Finally, we have detected 72 fragments with a combined length of 57,207 bp that have migrated from the *C. pubescens* cp genome to the mitogenome, accounting for 12.60% of the mitogenome's size (454,165 bp) (Fig. 11a, Table S13). These fragments contain 13 annotated genes, including 12 tRNA genes (*trnL-CAA*, *trnD-GTC*, *trnE-TTC*, *trnT-GGT*, *trnY-GTA*, *trnR-ACG*, *trnP-TGG*, *trnW-CCA*, *trnM-CAT*, *trnS-GGA*, *trnN-GTT*, and *trnH-GTG*), and rRNA gene (*rrn18*). Our analysis also reveals that some genes, such as *ycf3*, *atpF*, *rpl2*, and *ndhB*, have migrated from the cp genome to the mitogenome. However, during migration, most of these genes lost their integrity, and only partial fragments can now be found in the mitogenome (Table S13). Figure 11b and 11c show the circular map of the cp genome and a schematic diagram of the partial fragments of the *ycf3* gene migrating to the mitogenome, respectively. After the cp genome's *ycf3* gene (1,991 bp) migrates to the mitogenome, only a 282 bp fragment remains, while the remaining 1,709 bp fragment is lost.

# Discussion

### Genomic characteristics, PCGs, and intronic genes of C. pubescens mitogenome

Mitochondria, often referred to as the "powerhouse" of the cell, play a vital role in plant life. In recent years, advances in sequencing technology have facilitated the gradual assembly of numerous plant mitogenomes. Compared to animal mitogenomes, plant mitogenomes tend to exhibit more complexity, as reflected in their variable sizes and multiple repetitive sequences (Ye et al. 2017, (Bi et al. 2020, (Chevigny et al. 2020). In this study, we have conducted a detailed characterization of the *C. pubescens* mitogenome, providing the first comprehensive account of its features. The mitogenome of *C. pubescens* possesses a circular structure with a length of 454,165 bp and includes 35 PCGs. The GC content is significant for evaluating the application of species. The GC content of the *C. pubescens* mitogenome is 44.29%, which is basically consistent with the GC content of several peppers that have been sequenced (*C. annuum* var. *glabriusculum*, 44.50% (Magdy et al. 2020); *C. annuum* cultivar Jeju, 44.50% (Jo et al. 2014); C. *annuum* cultivar CMS line FS4401, 44.50%) (Jo et al. 2014). During the course of evolution, plant mitogenomes often experience the loss of PCGs (Bi et al. 2020, (Kan et al. 2020, (Sullivan et al. 2020). When we compare the mitogenomes of *C. pubescens* with those of three other pepper species (*C. annuum* var. *glabriusculum*, *C. annuum* cultivar Jeju, and *C. annuum* cultivar CMS line FS4401), we observe that *C. pubescens* possesses four unique PCGs (*atp4*, *atp8*, *mttB*, and *rps1*), while two genes (*rpl10* and *rps3*) are absent (Fig. 8). Interestingly, our findings align with previous research, which shows that the *rpl10* gene is lost in monocots and gymnosperms but is later regained during the evolutionary processes in eudicots (Notsu et al. 2002, (Ye et al. 2017). Additionally, we observed that the *rps7* gene is present solely in *D. carota* (JQ248574) and *A. thaliana* (BK010421) of the eudicots, suggesting its loss after the common ancestor of Solanaceae and Convolvulaceae began to diverge. Conversely, the *rps10* gene appears in all Solanaceae and Convolvulaceae, but it is lost in *D. carota* of Apiaceae and *A. thaliana* of Brassicaceae. These findings indicate that the *rps10* gene might have been lost during the separation of Umbelliferae and Brassicaceae. However, during subsequent evolutionary processes, the common ancestor of Solanaceae and Convolvulaceae began to diverge and reacquire the gene.

Introns are indeed prevalent in fully sequenced eukaryotic genomes (Gilson et al. 2006, (Gozashti et al. 2022). Despite the fact that introns need to be removed during RNA splicing in eukaryotic cells, they possess significant functions. One notable function of introns in eukaryotes is their contribution to increased protein abundance in intronic genes (Chorev et al. 2012, (Grabski et al. 2021). As the number of introns increases, so does the number of intronic genes. Moreover, certain introns can greatly enhance gene expression levels. Typically, substances containing introns tend to exhibit much higher expression levels compared to those without introns (Buchman et al. 1988, (Clark et al. 1993, (Fu et al. 2022). In our study, we have identified 10 intronic genes (*ccmFC*, *cox2*, *nad1*, *nad2*, *nad4*, *nad5*, *nad7*, *rpl2*, *rps1*, and *trnY-GTA*) in the *C. pubescens* mitogenome, totaling to 24 intronic genes. Among them, 4 genes (*nad1*, *nad2*, *nad5*, and *nad7*) contain 4 introns, while only the *nad4* gene contains 3 introns. Additionally, 5 genes (*ccmFC*, *cox2*, *rpl2*, *rps1*, and *trnY-GTA*) consist of a single intron. The presence of introns in these PCGs within the mitogenome of *C. pubescens* suggests that they warrant attention and exploration in terms of their functional significance in future studies.

# Repeat sequences

Repeated sequences are indeed a widespread feature in plant mitogenomes and play an important role in the intermolecular recombination that occurs within mitochondria (Dong et al. 2018). Among these repeated sequences, one type that is particularly notable is Microsatellites or Simple Sequence Repeats (SSRs). SSRs consist of short stretches of DNA composed of repetitive units, typically ranging from 1 to 6 nucleotides in length (Li et al. 2021). SSRs have a variety of functions and have been widely used to identify molecular markers of species, QTL mapping, marker-assisted breeding, analysis of diversity, and establishment of evolutionary relationships (Cao et al. 2014, (El-Esawi 2017, (Ma et al. 2017, (Li et al. 2020, (Shukla et al. 2021). In our analysis of the *C. pubescens* mitogenome, we found a diverse array of SSRs. Dinucleotide repeats are the most abundant, accounting for 42.90% of the total SSRs with 154 occurrences. Monomeric repeats are the second most frequent, making up 39.83% (143 occurrences), while hexanucleotide repeats are the least common, representing only 0.56% (2 occurrences). Previous studies have indicated that SSRs containing AT base repeats are more prevalent in organelle genomes due to the relative ease of breaking AT bonds compared to GC bonds (Qiao et al. 2022). By applying a matching threshold of 95%, we identified 17 tandem repeat sequences in the *C. pubescens* mitogenome, ranging in length from 7 to 39 bp. Tandem repeats, found in many plant mitogenomes, can be utilized as molecular markers for population identification (Sperisen et al. 2001). Additionally, we identified 473 interspersed repeats longer than 30 bp, with a cumulative length of 38,393 bp. These larger interspersed repeats are primarily responsible for genome rearrangements (Ogihara et al. 2005, (Alverson et al. 2011). Notably, we found five repeats exceeding 1 kb in length, ranging from 1,347 bp to 4,425 bp.

# RNA editing

RNA editing plays a crucial role in gene expression in plants (Ma et al. 2022). Specifically, the conversion of cytosine (C) to uracil (U) at specific sites in the cp genomes and mitogenomes alters genetic information (Cheng et al. 2021). In our analysis of the *C. pubescens* mitogenome, we identified 448 RNA editing sites in 35 PCGs. These editing sites encompassed 30 different codon transfer types. Among them, the TCA to TTA transition had the highest number of RNA editing sites, with 70 occurrences. Interestingly, this type of editing resulted in amino acid changes from hydrophilic to hydrophobic properties. Previous studies have shown that RNA editing at the second position within a codon is widespread, often accounting for more than half of the total editing events (Cheng et al. 2021, (Qiao et al. 2022) Similarly, we found that 359 RNA editing sites (80.13%) occurred at the second base position, consistent with these previous findings. Furthermore, we aligned RNA sequencing (RNA-seq) data from project number PRJNA822667 with the *C. pubescens* mitogenome data. This analysis revealed a high frequency of RNA editing events at the second codon position. Out of a total of 454 editing sites, 236 are observed at the second codon position, comprising over half of the total editing sites (Table S10). It is notable that all observed editing sites are of the CT type, which is the most common form of RNA editing in plant mitogenomes (Edera et al. 2021).

Curiously, we observe no RNA editing events at the third base position in our study. Some researchers suggest that this absence may be due to limitations in predicting RNA editing using the PREP-mt program rather than the actual absence of editing (Mower 2005, (Bi et al. 2020). To address this, we rigorously

align the RNA-seq data (project number PRJNA822667) with the mitogenome data and identify potential RNA editing sites (Table S10). Through this analysis, we uncover 71 RNA editing events at the third codon position out of the 454 identified editing sites. This highlights the fact that these results are not solely obtained through software prediction but also emphasizes the importance of validating them using RNA-seq data.

# Non-synonymous mutation rate (Ka) and synonymous mutation rate (Ks)

Calculating the Ka and Ks of PCGs provides valuable insights into phylogenetic reconstruction and understanding gene evolutionary dynamics (Fay et al. 2003). The Ka/Ks ratio can indicate different selective pressures acting on genes. If Ka/Ks > 1, it suggests positive selection; if Ka/Ks = 1, it implies neutral evolution; and if Ka/Ks < 1, it indicates negative selection. During our study, we find that the Ka/Ks values of the *sdh3* and *rps10* genes in the three comparison combinations (*C. pubescens* vs. KJ865409, KJ865410, and MN196478) are all less than 1. This signifies that these genes experienced negative selection. Mitochondrial genes undergoing negative selection likely play a crucial role in maintaining the normal function of mitochondria (Bi et al. 2020). Additionally, we observe that only the Ka/Ks value of the *nad5* gene is greater than 1 when comparing *C. pubescens* with KJ865410 and MN196478. This indicates that the *nad5* gene may have undergone positive selection since its divergence from the last common ancestor. These results suggest that different mitochondrial genes may experience various selection pressures throughout evolution (Bi et al. 2020). To visually observe the changes in Ka and Ks caused by sequence differences among the four pepper species, we conduct multiple sequence alignments for the 30 PCGs (Fig. 9). Among them, *rps10*, *rpl16*, and *sdh3* had a wide range of base deletions (more than 10 bases) in the four peppers. Though, more deletions occurred in KJ865409, KJ865410, and MN196478. As for the remaining PCGs, there are a few base substitutions. Some PCGs did not change their amino acids due to base substitution, such as *matR*, *rpl2*, *rps4*, and *ccmFC*. Some PCGs, such as *nad1*, *nad2*, *ccmFN*, and *nad4*, have amino acid changes due to base substitutions (Fig. 9). Typically, a change in the third base of a codon results in a synonymous mutation (Ks), while changes in the first and second bases lead to non-synonymous mutations (Ka) (Hurst 2002). According to the definition of Ka and Ks, when an amino acid change occurs due to a base change, it generally causes Ka. Conversely, when a base change does not result in an amino acid change, it generally causes Ks. Regardless of whether it is Ka or Ks, it often affects the mRNA expression level of the mutated gene (Shen et al. 2022). In the sequence alignment of the *nad5* gene, there is a C/T base substitution at the 1,863rd base between *C. pubescens* and KJ865409, which is located at the third base of the Val (V) amino acid. However, this base change does not result in an amino acid change, so the *nad5* gene between *C. pubescens* and KJ865409 only causes Ks (Ks = 0.0021, Fig. 9 and Table S12). For the *nad5* genes of KJ865410 and MN196478, Ks also exists at the 1,863rd base, and there are multiple base substitutions between the 1,447th and 1,468th bases, resulting in amino acid changes, thus causing Ka. The sequence alignment also reveals that the number of synonymous mutation sites is less than the number of non-synonymous mutation sites. Overall, when Ka/Ks > 1, it indicates that the gene is under

positive selection. Such a gene is a gene that is currently undergoing evolution, just like the *nad5* gene in this study, which has paramount significance for species evolution research in *C. pubescens*.

# Migration of homologous sequence fragments from the cp genome to the mitogenome

The transfer of DNA fragments from the organelle genome to the nuclear genome often occurs, and it is also common for DNA fragments to transfer from the cp genome to the mitogenome (Martin et al. 1998, (Bergthorsson et al. 2003). Nevertheless, the transfer of DNA fragments from the mitogenome to the cp genome is reported infrequently. Researchers believe this phenomenon may be because the plant cp genome is highly conserved, so foreign DNA rarely enters the cp genome (Richardson et al. 2007, (Smith 2011). Previous studies have shown that the migration of mitochondrial DNA to chloroplast has been found in four species (including *D. carota*, *Asclepias syriaca*, *Anacardium occidentale*, and herbaceous bamboos). In the mitogenome of *D. carota*, a 74 bp *cox1* gene sequence migrates from the mitogenome to the cp genome, and this sequence is named *Dc*MP (Goremykin et al. 2009, (Iorizzo et al. 2012, (Iorizzo et al. 2012). A ~ 2.7 kb sequence is inserted into the IR region of the herbaceous bamboos cp genome, originating from the mitogenome and confirmed (Ma et al. 2015). *A. syriaca* transfers the *rpl2* pseudogene (*ψrpl2*) from the mitogenome to the cp genome (Straub et al. 2013). In the *A. occidentale*, *ccmB* is transferred (Rabah et al. 2017). Through identification, 72 fragments totaling 57,207 bp migrated from the cp genome to the mitogenome of *C. pubescens*, accounting for 12.60% of the mitogenome (454,165 bp). Transfer of tRNA genes from the cp genome to the mitogenome is more common in angiosperms than in bryophytes and gymnosperms (Oda et al. 1992, (Notsu et al. 2002, (Ogihara et al. 2005, (Chaw et al. 2008). We find 13 annotated genes on these transferred fragments, including 12 tRNA genes (*trnL-CAA*, *trnD-GTC*, *trnE-TTC*, *trnT-GGT*, *trnY-GTA*, *trnR-ACG*, *trnP-TGG*, *trnW-CCA*, *trnM-CAT*, *trnS-GGA*, *trnN-GTT*, and *trnH-GTG*), and rRNA gene (*rrn18*). Some chloroplast PCGs, such as *ycf3*, *atpF*, *rpl2*, and *ndhB*, have also migrated from the cp genome to the mitogenome, although many of them have lost their integrity over evolution. Researchers suggest that these lost genes are likely to be found in the nuclear genome, where they are regulated as coding genes through related expression regulation (Bryant et al. 2011, (Savage et al. 2013). Let's take the *ycf3* gene of the cp genome as an example. We can search for homologous gene sequences in the nuclear genomes of three peppers (*C. annuum*, *C. baccatum*, and *C. chinense*) using the *ycf3* gene as a query sequence (Table S14). The obtained sequences can then be screened for GO enrichment function analysis. We can find that these genes are mainly involved in the Generation of precursor metabolites and energy (Biological Process, GO:0006091), NADH dehydrogenase activity (Molecular Function, GO:0003954), and Membrane protein complex (Cellular Component, GO:0098796). We show only one GO term for each functional enrichment process, and more results can be found in Table S15.

# Conclusions

In this study, we assembled the first mitogenome of *C. pubescens* (GenBank accession number OP957066), 454,165 bp in length, with a typical circular structure. The mitogenome of *C. pubescens* has a GC content of 44.29% and we annotate 70 genes in it. Compared to the other three pepper mitogenomes (KJ865409, KJ865410, and MN196478), the mitogenome of *C. pubescens* shows four unique PCGs (*atp4*, *atp8*, *mttB*, and *rps1*), while two PCGs (*rpl10* and *rps3*) are absent. Structural comparison and synteny analysis of these four pepper mitogenomes reveal that *C. pubescens* is most similar to KJ865409 but exhibits rearrangements with the other three. We identify 72 homologous sequences, accounting for 12.60% (57,207 bp) of the mitogenome, between the mitochondrial and chloroplast genomes. These sequences include 12 tRNA genes and the rRNA gene (*rrn18*). Of note, the *nad5* gene shows signs of positive selection based on selective pressure analysis. Additionally, three genes (*nad1*, *nad2*, and *nad4*) exhibit single base mutations that result in amino acid changes. Through comparative genome analysis, we can gain insight into the genetic information exchange between pepper organelle genomes. This exploration will be helpful for future studies.

# Declarations

**Acknowledgments** We acknowledge the support of the Hainan Province Science and Technology Special Fund (ZDYF2023XDNY028), the National Key Research and Development Program of China (2018YFD1000800), and the Project of the Administrative Bureau of Sanya Yazhou Bay Science and Technology City (HNF202210). Thank you to our group members for (Yuanyuan Hao, Yu Zhang, Muhammad Ali Mumtaz, Huangying Shu, Liping Zhang, and Shanhan Cheng) their excellent suggestions in research and article writing. In addition, special thanks to the anonymous reviewers for their valuable time and suggestions on this manuscript.

**Author contributions** L.L. and Z.W. designed the experiments, and supervised and participated in the entire work. L.L., X.L., and H.F. wrote the frst manuscript draft; all authors contributed by comments and approved the fnal manuscript.

**Funding** This research is funded by grants from the Hainan Province Science and Technology Special Fund (ZDYF2023XDNY028), the National Key Research and Development Program of China (2018YFD1000800), and the Project of the Administrative Bureau of Sanya Yazhou Bay Science and Technology City (HNF202210).

**Data availability** Mitochondrial DNA sequence from *C. pubescens* is stored in NCBI with the accession number OP957066.

**Ethical approval** Not applicable.

**Competing interests** The authors declare no competing interests.

# References

1. Alverson A J, D W Rice, S Dickinson, K Barry and J D Palmer (2011). Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. Plant Cell 23(7): 2499-2513. https://doi.org/10.1105/tpc.111.087189.
2. Androsiuk P, A Okorski, L Paukszto, J P Jastrzebski, S Ciesielski and A Pszczolkowska (2022). Characterization and phylogenetic analysis of the complete mitochondrial genome of the pathogenic fungus *Ilyonectria destructans*. Sci Rep 12(1): 2359. https://doi.org/10.1038/s41598-022-05428-z.
3. Barboza G E, C C García, L de Bem Bianchetti, M V Romero and M Scaldaferro (2022). Monograph of wild and cultivated chili peppers (*Capsicum* L., Solanaceae). PhytoKeys 200: 1. https://doi.org/10.3897/phytokeys.200.71667.
4. Bergthorsson U, K L Adams, B Thomason and J D Palmer (2003). Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature 424(6945): 197-201. https://doi.org/10.1038/nature01743.
5. Bi C, N Lu, Y Xu, C He and Z Lu (2020). Characterization and analysis of the mitochondrial genome of common bean (*Phaseolus vulgaris*) by comparative genomic approaches. Int J Mol Sci 21(11): 3778. https://doi.org/10.3390/ijms21113778.
6. Bryant N, J Lloyd, C Sweeney, F Myouga and D Meinke (2011). Identification of nuclear genes encoding chloroplast-localized proteins required for embryo development in Arabidopsis. Plant Physiol 155(4): 1678-1689. https://doi.org/10.1104/pp.110.168120.
7. Buchman A R and P Berg (1988). Comparison of intron-dependent and intron-independent gene expression. Mol Cell Biol 8(10): 4395-4405. https://doi.org/10.1128/mcb.8.10.4395-4405.1988.
8. Cao Z, P Wang, X Zhu, H Chen and T Zhang (2014). SSR marker-assisted improvement of fiber qualities in *Gossypium hirsutum* using *G. barbadense* introgression lines. Theor Appl Genet 127(3): 587-594. https://doi.org/10.1007/s00122-013-2241-3.
9. Chang S, Y Wang, J Lu, J Gai, J Li, P Chu, R Guan and T Zhao (2013). The mitochondrial genome of soybean reveals complex genome structures and gene evolution at intercellular and phylogenetic levels. PLoS One 8(2): e56502. https://doi.org/10.1371/journal.pone.0056502.
10. Chaw S-M, A Chun-Chieh Shih, D Wang, Y-W Wu, S-M Liu and T-Y Chou (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(3): 603-615. https://doi.org/10.1093/molbev/msn009.
11. Cheng Y, X He, S Priyadarshani, Y Wang, L Ye, C Shi, K Ye, Q Zhou, Z Luo, F Deng, L Cao, P Zheng, M Aslam and Y Qin (2021). Assembly and comparative analysis of the complete mitochondrial genome of *Suaeda glauca*. BMC Genomics 22(1): 167. https://doi.org/10.1186/s12864-021-07490-9.
12. Chevigny N, D Schatz-Daas, F Lotfi and J M Gualberto (2020). DNA repair and the stability of the plant mitochondrial genome. Int J Mol Sci 21(1): 328. https://doi.org/10.3390/ijms21010328.
13. Chorev M and L Carmel (2012). The function of introns. Front. Genet 3: 55. https://doi.org/10.3389/fgene.2012.00055.

14. Clark A J, A L Archibald, M McClenaghan, J P Simons, R Wallace and C B Whitelaw (1993). Enhancing the efficiency of transgene expression. Philos Trans R Soc Lond B Biol Sci 339(1288): 225-232. https://doi.org/10.1098/rstb.1993.0020.
15. Dong S, C Zhao, F Chen, Y Liu, S Zhang, H Wu, L Zhang and Y Liu (2018). The complete mitochondrial genome of the early flowering plant *Nymphaea colorata* is highly repetitive with low recombination. BMC Genomics 19(1): 614. https://doi.org/10.1186/s12864-018-4991-4.
16. Ďúranová H, E Ivanišová, L Galovičová, L Godočíková, P Borotová, S Kunová, K Miklášová, Ľ Lopašovský and E Mňahončáková (2021). Antioxidant and antimicrobial activities of fruit extracts from different fresh chili peppers. Acta Sci Pol Technol Aliment 20(4): 465-472. https://doi.org/10.17306/J.AFS.2021.0997.
17. Edera A A and M V Sanchez-Puerta (2021). Computational detection of plant RNA editing events. RNA Editing: Methods and Protocols 2181: 13-34. https://doi.org/10.1007/978-1-0716-0787-9_2.
18. El-Esawi M A (2017). SSR analysis of genetic diversity and structure of the germplasm of faba bean (*Vicia faba* L.). C R Biol 340(11-12): 474-480. https://doi.org/10.1016/j.crvi.2017.09.008.
19. Fan W, F Liu, Q Jia, H Du, W Chen, J Ruan, J Lei, D Z Li, J P Mower and A Zhu (2022). Fragaria mitogenomes evolve rapidly in structure but slowly in sequence and incur frequent multinucleotide mutations mediated by microinversions. New Phytol 236(2): 745-759. https://doi.org/10.1111/nph.18334.
20. Fay J C and C I Wu (2003). Sequence divergence, functional constraint, and selection in protein evolution. Annu Rev Genomics Hum Genet 4(1): 213-235. https://doi.org/10.1146/annurev.genom.4.020303.162528.
21. Fu J, Y-W Fu, J-J Zhao, Z-X Yang, S-A Li, G-H Li, Z-J Quan, F Zhang, J-P Zhang and X-B Zhang (2022). Improved and flexible HDR editing by targeting introns in iPSCs. Stem Cell Reviews and Reports 18(5): 1822-1833. https://doi.org/10.1007/s12015-022-10331-1.
22. Gao C, M A Mumtaz, Y Zhou, Z Yang, H Shu, J Zhu, W Bao, S Cheng, L Yin and J Huang (2022). Integrated transcriptomic and metabolomic analyses of cold-tolerant and cold-sensitive pepper species reveal key genes and essential metabolic pathways involved in response to cold stress. Int J Mol Sci 23(12): 6683. https://doi.org/10.3390/ijms23126683.
23. Gilson P R, V Su, C H Slamovits, M E Reith, P J Keeling and G I McFadden (2006). Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. Proc Natl Acad Sci U S A 103(25): 9566-9571. https://doi.org/10.1073/pnas.060070710.
24. Goremykin V V, F Salamini, R Velasco and R Viola (2009). Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. Mol Biol Evol 26(1): 99-110. https://doi.org/10.1093/molbev/msn226.
25. Gozashti L, S W Roy, B Thornlow, A Kramer, M Ares, Jr. and R Corbett-Detig (2022). Transposable elements drive intron gain in diverse eukaryotes. Proc Natl Acad Sci U S A 119(48): e2209766119. https://doi.org/10.1073/pnas.2209766119.

26. Grabski D F, L Broseus, B Kumari, D Rekosh, M L Hammarskjold and W Ritchie (2021). Intron retention and its impact on gene expression and protein diversity: A review and a practical guide. WIRES RNA 12(1): e1631. https://doi.org/10.1002/wrna.1631.
27. Grover A and P C Sharma (2016). Development and use of molecular markers: past and present. Crit Rev Biotechnol 36(2): 290-302. https://doi.org/10.3109/07388551.2014.959891.
28. Hurst L D (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 18(9): 486. https://doi.org/10.1016/s0168-9525(02)02722-1.
29. Ibiza V P, J Blanca, J Canizares and F Nuez (2012). Taxonomy and genetic diversity of domesticated *Capsicum* species in the Andean region. Genet Resour Crop Ev 59(6): 1077-1088. https://doi.org/10.1007/s10722-011-9744-z.
30. Iorizzo M, D Grzebelus, D Senalik, M Szklarczyk, D Spooner and P Simon (2012). Against the traffic: The first evidence for mitochondrial DNA transfer into the plastid genome. Mob Genet Elements 2(6): 261-266. https://doi.org/10.4161/mge.23088.
31. Iorizzo M, D Senalik, M Szklarczyk, D Grzebelus, D Spooner and P Simon (2012). De novo assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome. BMC Plant Biol 12: 1-17. https://doi.org/10.1186/1471-2229-12-61.
32. Jarret R L, G E Barboza, F R da Costa Batista, T Berke, Y-Y Chou, A Hulse-Kemp, N Ochoa-Alejo, P Tripodi, A Veres and C C Garcia (2019). *Capsicum*—An abbreviated compendium. J. Am. Soc. Hortic. Sci 144(1): 3-22. https://doi.org/10.21273/JASHS04446-18.
33. Jo Y D, Y Choi, D H Kim, B D Kim and B C Kang (2014). Extensive structural variations between mitochondrial genomes of CMS and normal peppers (*Capsicum annuum* L.) revealed by complete nucleotide sequencing. BMC Genomics 15(1): 561. https://doi.org/10.1186/1471-2164-15-561.
34. Kan S-L, T-T Shen, P Gong, J-H Ran and X-Q Wang (2020). The complete mitochondrial genome of *Taxus cuspidata* (Taxaceae): eight protein-coding genes have transferred to the nuclear genome. BMC Evol. Biol 20: 1-17. https://doi.org/10.1186/s12862-020-1582-1.
35. Li D, C Long, X Pang, D Ning, T Wu, M Dong, X Han and H Guo (2020). The newly developed genomic-SSR markers uncover the genetic characteristics and relationships of olive accessions. PeerJ 8: e8573. https://doi.org/10.7717/peerj.8573.
36. Li Q, X Su, H Ma, K Du, M Yang, B Chen, S Fu, T Fu, C Xiang and Q Zhao (2021). Development of genic SSR marker resources from RNA-seq data in *Camellia japonica* and their application in the genus *Camellia*. Sci. Rep 11(1): 1-11. https://doi.org/10.1038/s41598-021-89350-w.
37. Liu Y C, S Liu, D C Liu, Y X Wei, C Liu, Y M Yang, C G Tao and W S Liu (2014). Exploiting EST databases for the development and characterization of EST-SSR markers in blueberry (Vaccinium) and their cross-species transferability in *Vaccinium* spp. Sci. Hortic 176: 319-329. https://doi.org/10.1016/j.scienta.2014.07.026.
38. Ma P F, Y X Zhang, Z H Guo and D Z Li (2015). Evidence for horizontal transfer of mitochondrial DNA to the plastid genome in a *bamboo* genus. Sci Rep 5(1): 11608. https://doi.org/10.1038/srep11608.

39. Ma Q, S Li, C Bi, Z Hao, C Sun and N Ye (2017). Complete chloroplast genome sequence of a major economic species, *Ziziphus jujuba* (Rhamnaceae). Curr Genet 63(1): 117-129. https://doi.org/10.1007/s00294-016-0612-4.
40. Ma Q, Y Wang, S Li, J Wen, L Zhu, K Yan, Y Du, J Ren, S Li, Z Chen, C Bi and Q Li (2022). Assembly and comparative analysis of the first complete mitochondrial genome of *Acer truncatum* Bunge: a woody oil-tree species producing nervonic acid. BMC Plant Biol 22(1): 29. https://doi.org/10.1186/s12870-021-03416-5.
41. Magdy M and B Ouyang (2020). The complete mitochondrial genome of the chiltepin pepper (*Capsicum annuum* var. glabriusculum), the wild progenitor of *Capsicum annuum* L. Mitochondrial DNA B Resour 5(1): 683-684. https://doi.org/10.1080/23802359.2020.1714496.
42. Mahapatra K, S Banerjee, S De, M Mitra, P Roy and S Roy (2021). An insight into the mechanism of plant organelle genome maintenance and implications of organelle genome in crop improvement: an update. Front. Cell Dev. Biol 9: 671698. https://doi.org/10.3389/fcell.2021.671698.
43. Martin W, B Stoebe, V Goremykin, S Hapsmann, M Hasegawa and K V Kowallik (1998). Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393(6681): 162-165. https://doi.org/10.1038/30234.
44. Mower J P (2005). PREP-Mt: predictive RNA editor for plant mitochondrial genes. BMC Bioinf 6: 96. https://doi.org/10.1186/1471-2105-6-96.
45. Notsu Y, S Masood, T Nishikawa, N Kubo, G Akiduki, M Nakazono, A Hirai and K Kadowaki (2002). The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. Mol Genet Genomics 268: 434-445. https://doi.org/10.1007/s00438-002-0767-1.
46. Oda K, K Yamato, E Ohta, Y Nakamura, M Takemura, N Nozato, K Akashi and K Ohyama (1992). Transfer RNA genes in the mitochondrial genome from a liverwort, *Marchantia polymorpha*: the absence of chloroplast-like tRNAs. Nucleic Acids Res 20(14): 3773-3777. https://doi.org/10.1093/nar/20.14.3773.
47. Odahara M, K Nakamura, Y Sekine and T Oshima (2021). Ultra-deep sequencing reveals dramatic alteration of organellar genomes in Physcomitrella patens due to biased asymmetric recombination. Commun. Biol 4(1): 633. https://doi.org/10.1038/s42003-021-02141-x.
48. Ogihara Y, Y Yamazaki, K Murai, A Kanno, T Terachi, T Shiina, N Miyashita, S Nasuda, C Nakamura, N Mori, S Takumi, M Murata, S Futo and K Tsunewaki (2005). Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. Nucleic Acids Res 33(19): 6235-6250. https://doi.org/10.1093/nar/gki925.
49. Palombo N E and C Carrizo García (2022). Geographical patterns of genetic variation in Locoto chile (*Capsicum pubescens*) in the americas inferred by genome-wide data analysis. Plants 11(21): 2911. https://doi.org/10.3390/plants11212911.
50. Qiao Y, X Zhang, Z Li, Y Song and Z Sun (2022). Assembly and comparative analysis of the complete mitochondrial genome of *Bupleurum chinense* DC. BMC Genomics 23(1): 1-17.

https://doi.org/10.1186/s12864-022-08892-z.

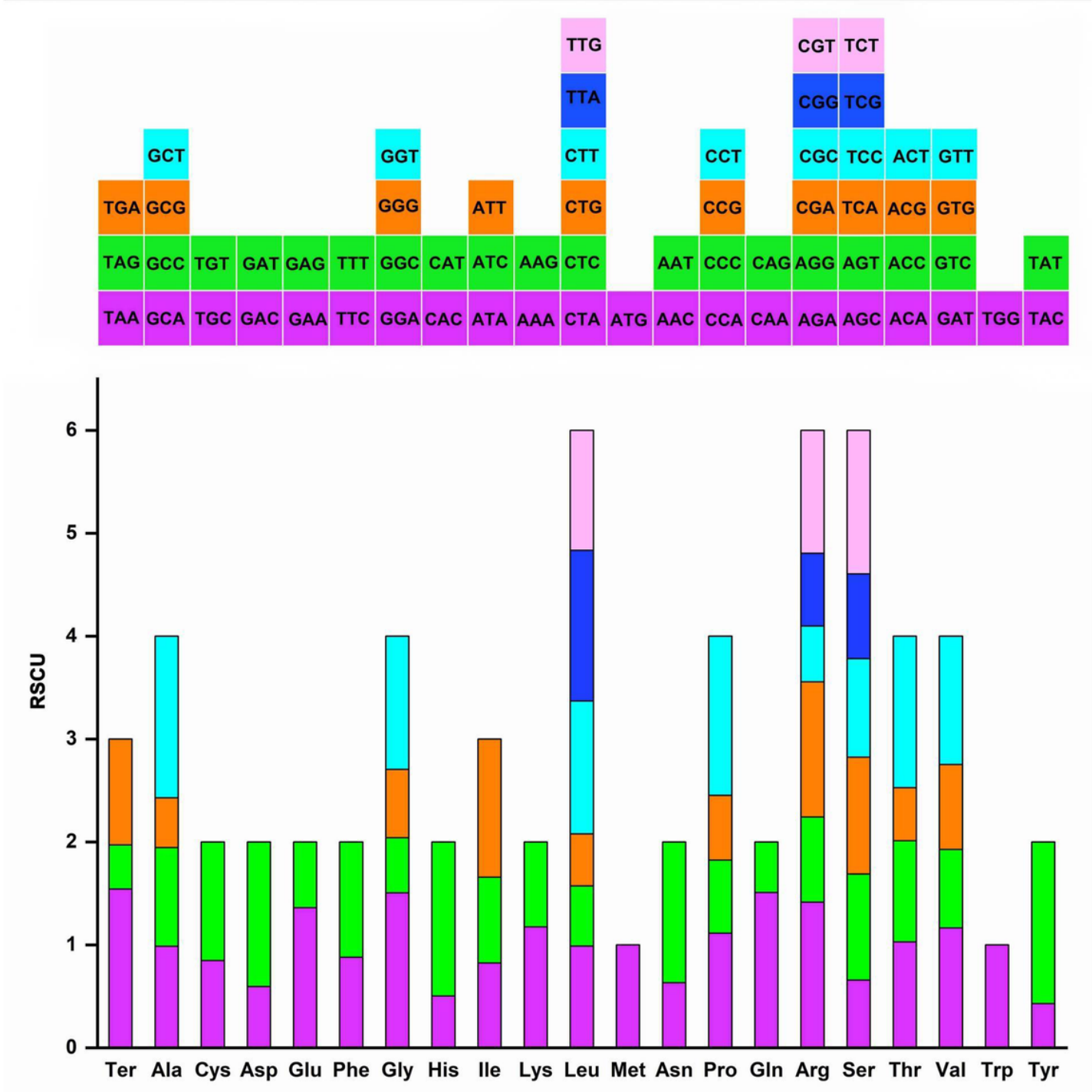
51. Rabah S O, C Lee, N H Hajrah, R M Makki, H F Alharby, A M Alhebshi, J S M Sabir, R K Jansen and T A Ruhlman (2017). Plastome Sequencing of Ten Nonmodel Crop Species Uncovers a Large Insertion of Mitochondrial DNA in Cashew. Plant Genome 10(3): plantgenome2017.2003.0020. https://doi.org/10.3835/plantgenome2017.03.0020.
52. Richardson A O and J D Palmer (2007). Horizontal gene transfer in plants. J Exp Bot 58(1): 1-9. https://doi.org/10.1093/jxb/erl148.
53. Richardson A O, D W Rice, G J Young, A J Alverson and J D Palmer (2013). The "fossilized" mitochondrial genome of *Liriodendron tulipifera*: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. BMC Biol 11(1): 29. https://doi.org/10.1186/1741-7007-11-29.
54. Rodríguez-Burruezo A, J Prohens, M D Raigón and F Nuez (2009). Variation for bioactive compounds in ají (*Capsicum baccatum* L.) and rocoto (*C. pubescens* R. & P.) and implications for breeding. Euphytica 170: 169-181. https://doi.org/10.1007/s10681-009-9916-5.
55. Savage L J, K M Imre, D A Hall and R L Last (2013). Analysis of essential Arabidopsis nuclear genes encoding plastid-targeted proteins. PLoS One 8(9): e73291. https://doi.org/10.1371/journal.pone.0073291.
56. Shen X, S Song, C Li and J Zhang (2022). Synonymous mutations in representative yeast genes are mostly strongly non-neutral. Nature 606(7915): 725-731. https://doi.org/10.1038/s41586-022-04823-w.
57. Shukla R P, G J Tiwari, B Joshi, K Song-Beng, S Tamta, N M Boopathi and S N Jena (2021). GBS-SNP and SSR based genetic mapping and QTL analysis for drought tolerance in upland cotton. Physiol Mol Biol Plants 27(8): 1731-1745. https://doi.org/10.1007/s12298-021-01041-y.
58. Skippington E, T J Barkman, D W Rice and J D Palmer (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(27): E3515-E3524. https://doi.org/10.1073/pnas.1504491112.
59. Sloan D B, A J Alverson, J P Chuckalovcak, M Wu, D E McCauley, J D Palmer and D R Taylor (2012). Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol 10(1): e1001241. https://doi.org/10.1371/journal.pbio.1001241.
60. Smith D R (2011). Extending the limited transfer window hypothesis to inter-organelle DNA migration. Genome Biol Evol 3: 743-748. https://doi.org/10.1093/gbe/evr068.
61. Sperisen C, U Buchler, F Gugerli, G Matyas, T Geburek and G G Vendramin (2001). Tandem repeats in plant mitochondrial genomes: application to the analysis of population differentiation in the conifer *Norway spruce*. Mol Ecol 10(1): 257-263. https://doi.org/10.1046/j.1365-294x.2001.01180.x.
62. Straub S C, R C Cronn, C Edwards, M Fishbein and A Liston (2013). Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (apocynaceae). Genome Biol Evol 5(10): 1872-1885. https://doi.org/10.1093/gbe/evt140.

63. Sullivan A R, Y Eldfjell, B Schiffthaler, N Delhomme, T Asp, K H Hebelstrup, O Keech, L Öberg, I M Møller and L Arvestad (2020). The mitogenome of *Norway spruce* and a reappraisal of mitochondrial recombination in plants. Genome Biol. Evol 12(1): 3586-3598. https://doi.org/10.1093/gbe/evz263.
64. Wang H (2005). pepReap: a peptide identification algorithm using support vector machines. J. Comp. Res. Dev 42(9): 555-564. https://doi.org/10.1360/crad20050909.
65. Wang X, L L Li, Y Xiao, X Y Chen, J H Chen and X S Hu (2021). A complete sequence of mitochondrial genome of *Neolamarckia cadamba* and its use for systematic analysis. Sci Rep 11(1): 21452. https://doi.org/10.1038/s41598-021-01040-9.
66. Wang Y, S Chen, J Chen, C Chen, X Lin, H Peng, Q Zhao and X Wang (2023). Characterization and phylogenetic analysis of the complete mitochondrial genome sequence of *Photinia serratifolia*. Sci. Rep 13(1): 770. https://doi.org/10.1038/s41598-022-24327-x.
67. Woloszynska M and D Trojanowski (2009). Counting mtDNA molecules in Phaseolus vulgaris: sublimons are constantly produced by recombination via short repeats and undergo rigorous selection during substoichiometric shifting. Plant Mol. Biol 70(5): 511-521. https://doi.org/10.1007/s11103-009-9488-8.
68. Wu K, M Yang, H Liu, Y Tao, J Mei and Y Zhao (2014). Genetic analysis and molecular characterization of Chinese sesame (*Sesamum indicum* L.) cultivars using insertion-deletion (InDel) and simple sequence repeat (SSR) markers. BMC Genet 15(1): 35. https://doi.org/10.1186/1471-2156-15-35.
69. Wynn E L and A C Christensen (2019). Repeats of unusual size in plant mitochondrial genomes: identification, incidence and evolution. G3 (Bethesda) 9(2): 549-559. https://doi.org/10.1534/g3.118.200948.
70. Xiong Y, Q Yu, Y Xiong, J Zhao, X Lei, L Liu, W Liu, Y Peng, J Zhang, D Li, S Bai and X Ma (2021). The complete mitogenome of *Elymus sibiricus* and insights into its evolutionary pattern based on simple repeat sequences of seed pant mitogenomes. Front Plant Sci 12: 802321. https://doi.org/10.3389/fpls.2021.802321.
71. Ye N, X Wang, J Li, C Bi, Y Xu, D Wu and Q Ye (2017). Assembly and comparative analysis of complete mitochondrial genome sequence of an economic plant *Salix suchowensis*. PeerJ 5: e3148. https://doi.org/10.7717/peerj.3148.
72. Zhang X, W Zhou, X Cheng, X Chen, X Hu, Y Hu and X Wang (2019). Assessing the ecological and evolutionary processes underlyingcytonuclear interactions. Scientia Sinica Vitae 49(8): 951-964. https://doi.org/10.1360/ssv-2019-0049.

# Figures

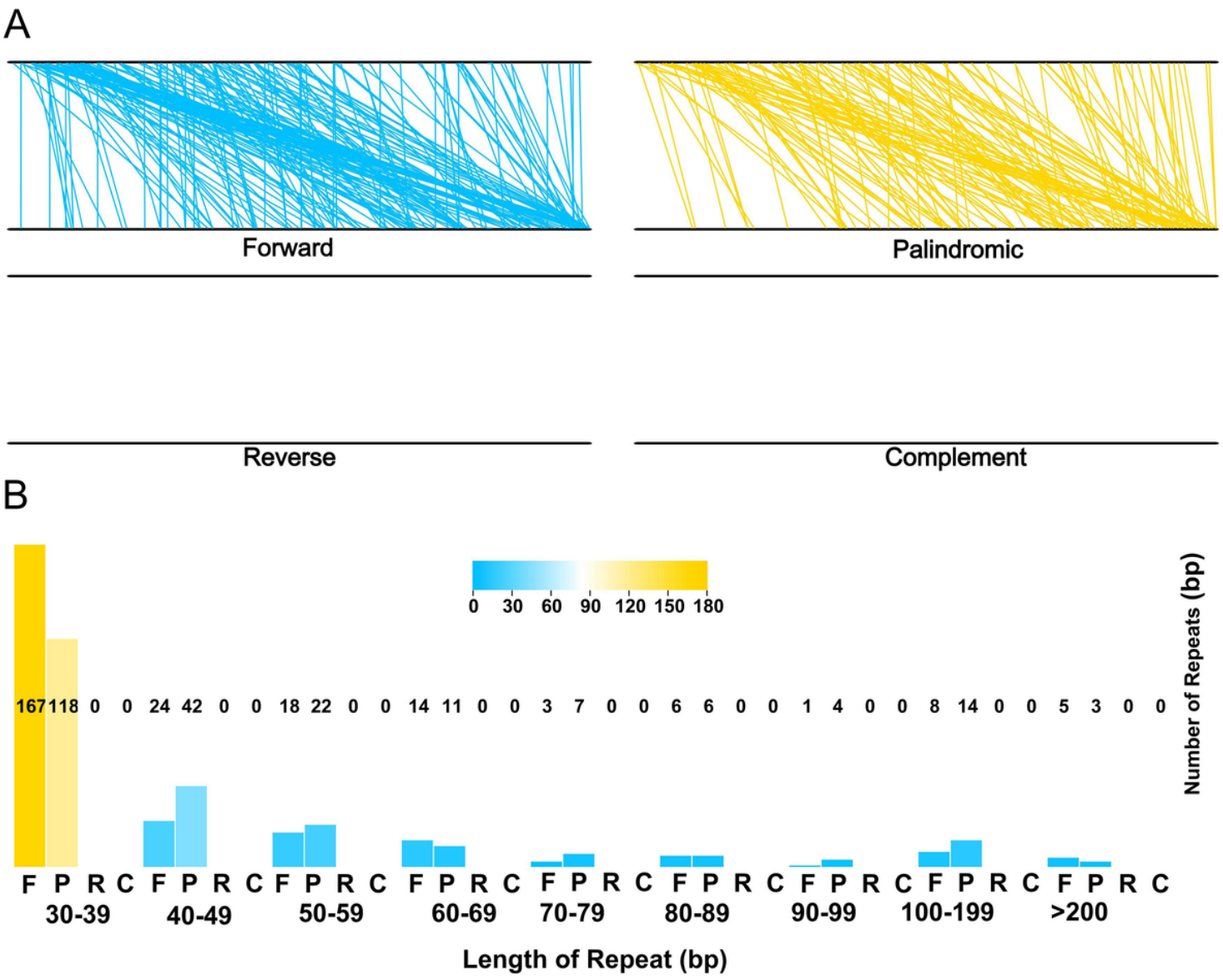

## Figure 1

The circular map of *C. pubescens* mitogenome. The forward coding genes are placed outside the circle, while the reverse coding genes are nestled on the inside, and the inner gray circle represents the GC content. The circular map of the *C. pubescens*mitogenome is created using the OGDRAW online tool (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html).




## Figure 2

Relative synonymous codon usage (RSCU) in the *C. pubescens* mitogenome. The lower squares represent all codons encoding each amino acid, and the height of the upper columns represents the sum of all codon RSCU values.




## Figure 3

The distribution of the four types of interspersed repeats on the genome and the statistics of the number of repeats. (a) Two black lines represent the genome, with identical repeats connected by line segments; (b) F, P, R, and C represent forward, palindromic, reverse, and complementary repeats. The abscissa represents the type and length of interspersed repeats, and the ordinate represents the number of interspersed repeats.

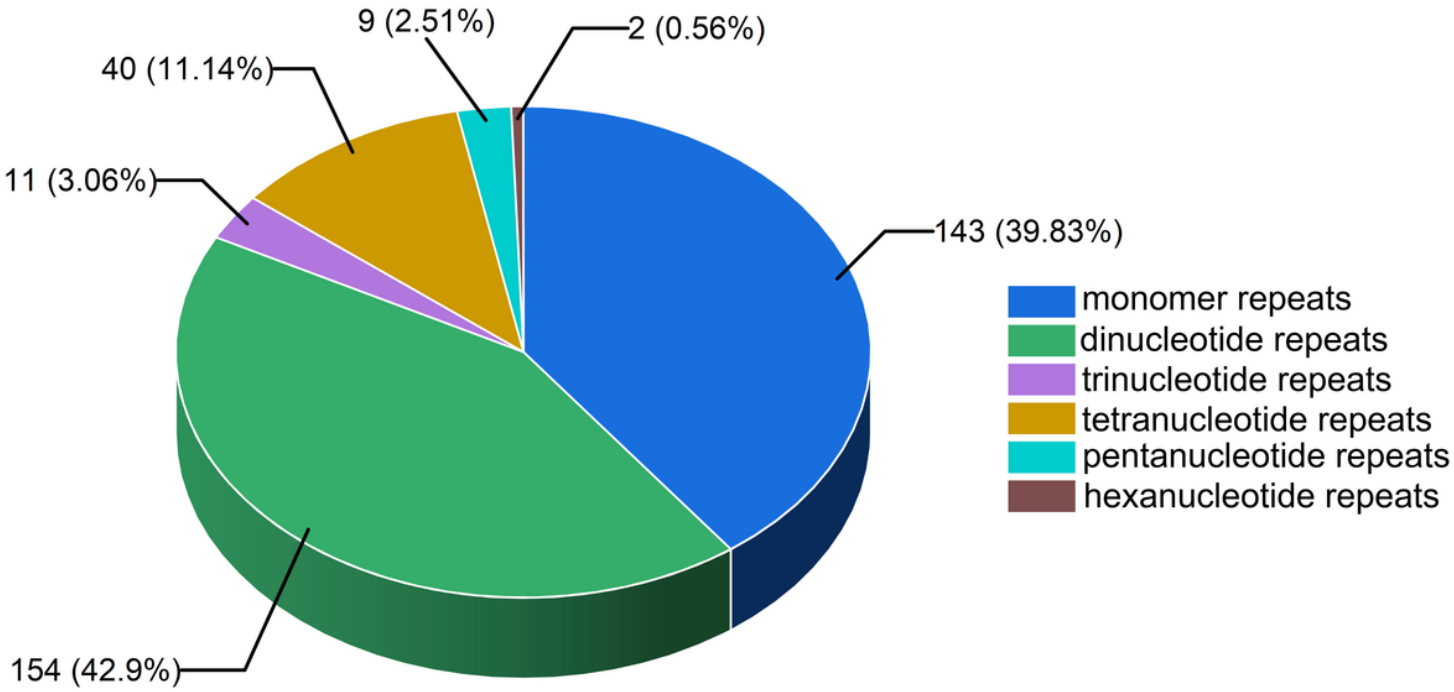
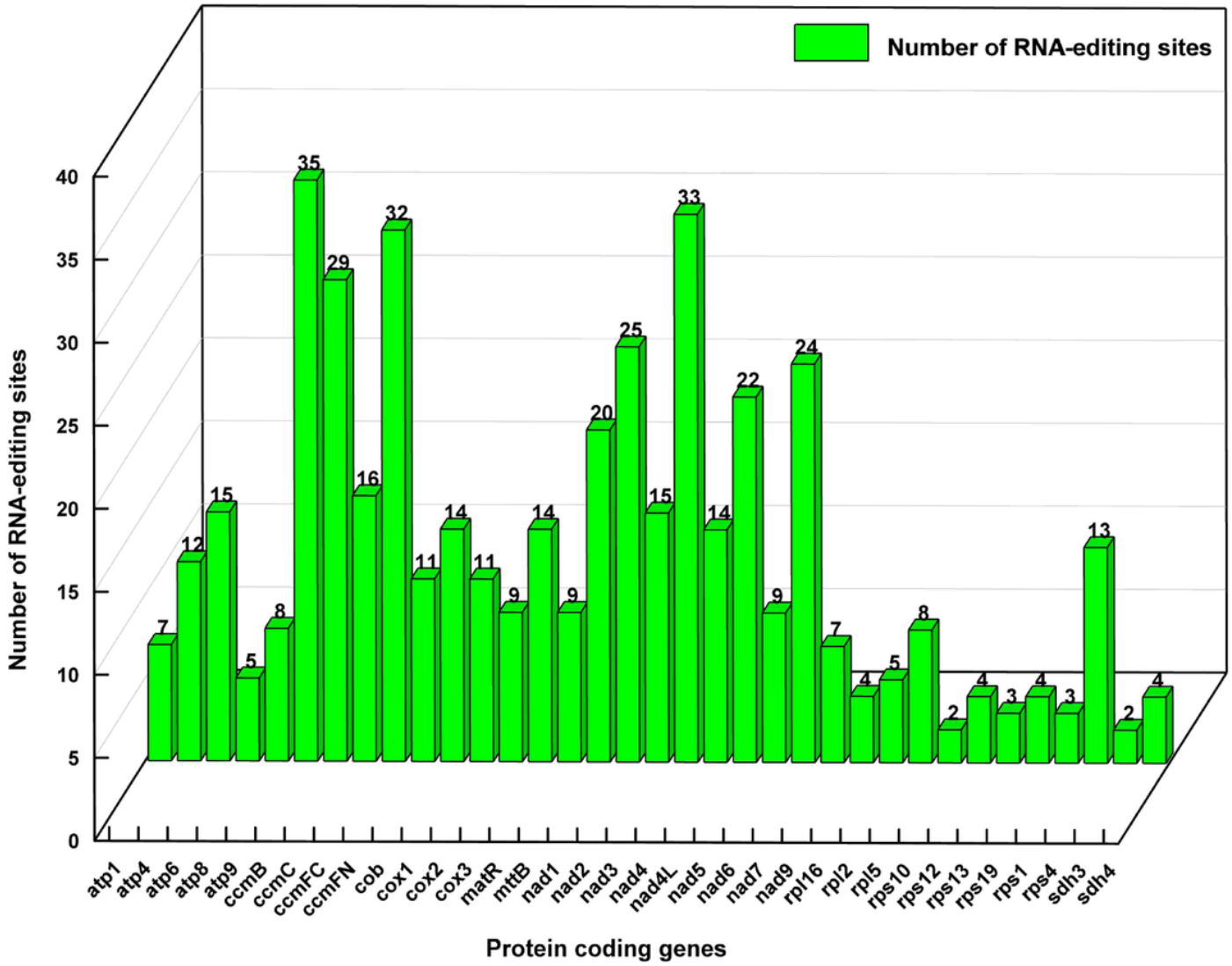


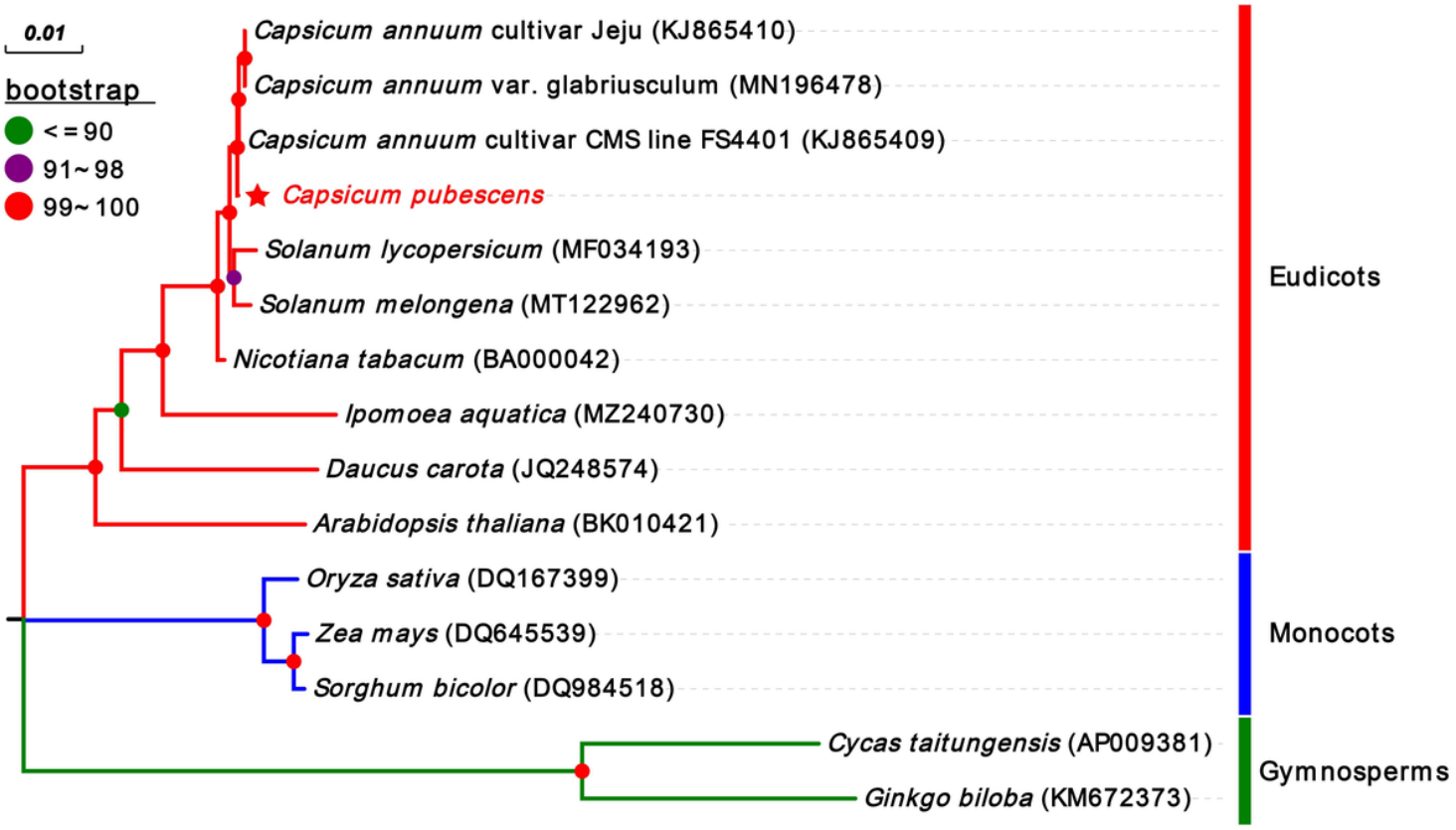

**Figure 4**

Statistics on the type and number of simple sequence repeats (SSR).



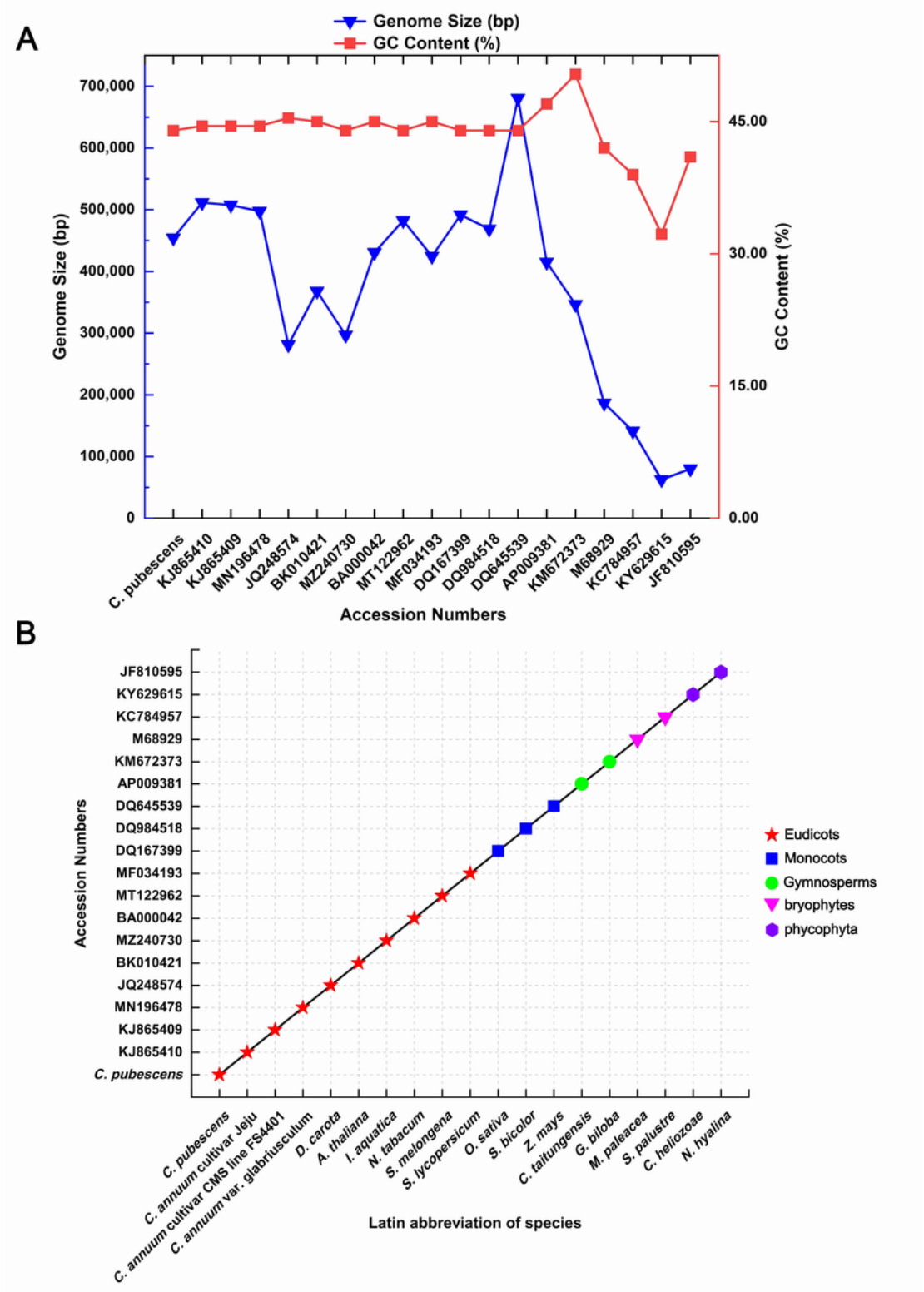

## Figure 5

Distribution of RNA editing sites in PCGs of the *C. pubescens* mitogenome. The online website (http://prep.unl.edu/) is used to make predictions of RNA editing sites.




## Figure 6

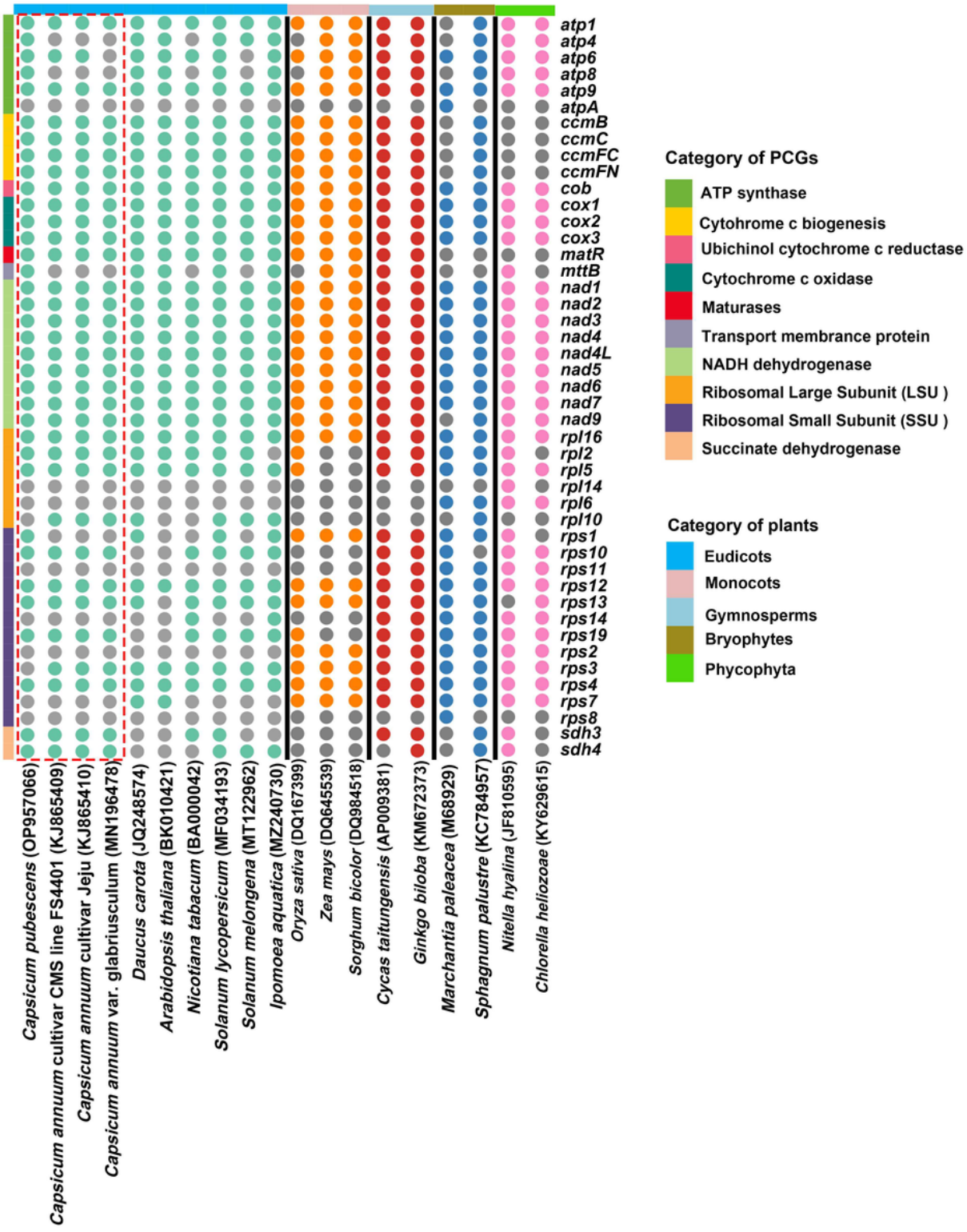
Phylogenetic tree of 23 common conserved PCGs of the *C. pubescens* and 14 other species mitogenomes (9 eudicots, 3 monocots, and 2 gymnosperms). MAFFT v7.490 software (default parameters) is used to perform multiple sequences alignment, and MEGA 11 is used to construct a phylogenetic tree (Maximum-likelihood phylogenetic tree, Bootstrap value set to 1,000).




## Figure 7

Comparison of mitogenomes size and GC content of *C. pubescens* with other species. **(A)** The abscissa represents the GenBank accession number of the mitogenomes, and the ordinate represents the size and GC content of the mitogenomes; **(B)** The abscissa represents the abbreviated Latin name of the species, and the ordinate represents the GenBank accession number of the corresponding mitogenomes.

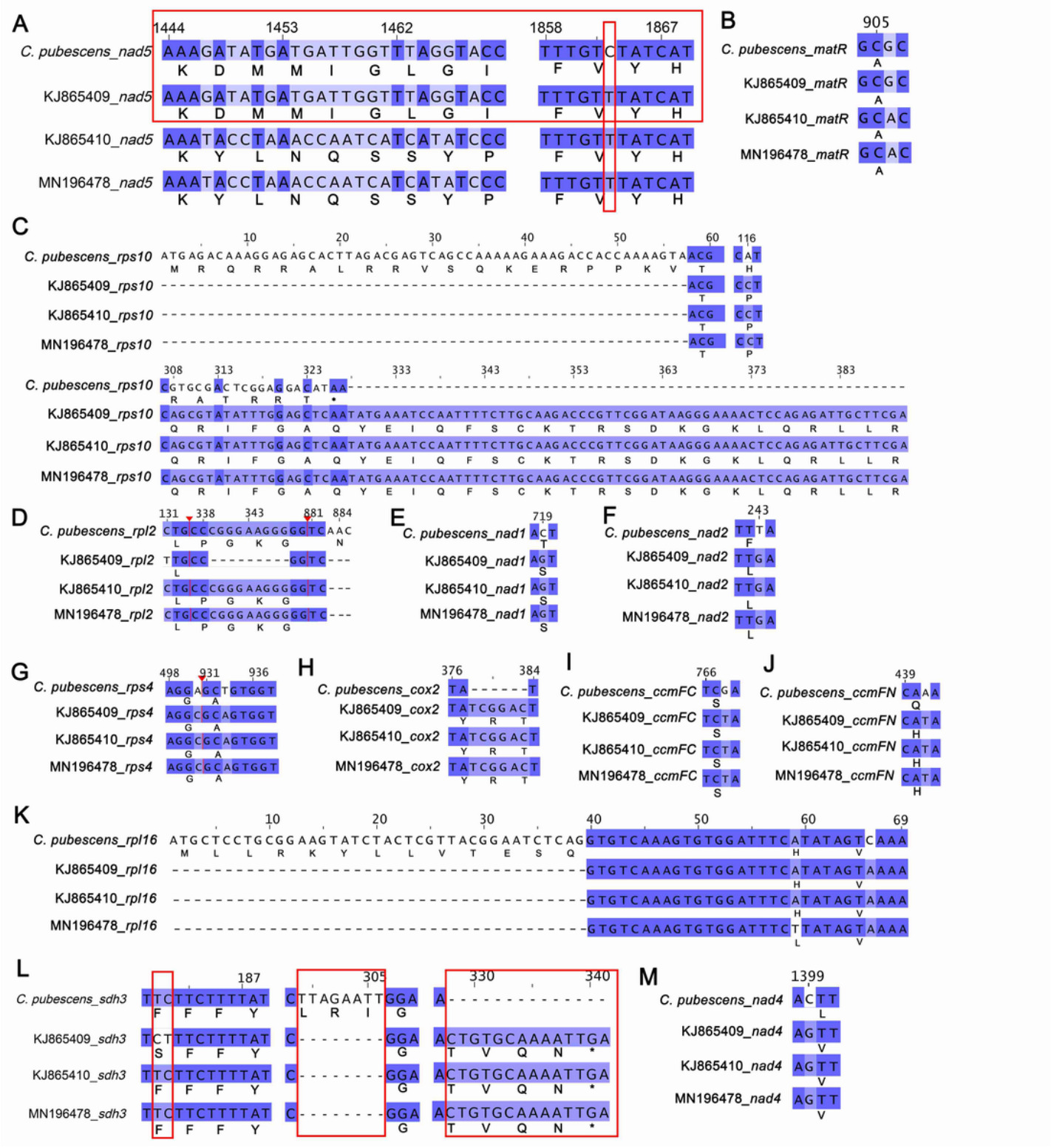
Pentagrams, squares, circles, triangles, and hexagons represent eudicots, monocots, gymnosperms, bryophytes, and phycophyta, respectively.




## Figure 8

Distribution of PCGs in plant mitogenomes. The gray circles indicate that the gene does not exist in the corresponding mitogenomes. The light green, orange, rosered, light blue, and pink circles indicate that

PCGs exist in the mitogenomes, corresponding to eudicots, monocots, gymnosperms, bryophytes, and phycophyta, respectively.




## Figure 9

Alignments of the nucleotide and amino acid sequences of PCGs from four pepper mitogenomes in *C. pubescens* and *C. annuum*. We perform multiple sequence alignments using MAFFT v7.490 software,

using the default parameters. **(A-M)** Represented as *nad5*, *matR*, *rps10*, *rpl2*, *nad1*, *nad2*, *rps4*, *cox2*, *ccmFC*, *ccmFN*, *rpl16*, *sdh3*, and *nad4* genes, respectively. KJ865409, KJ865410, and MN196478 are the mitogenomes of *C. annuum*.

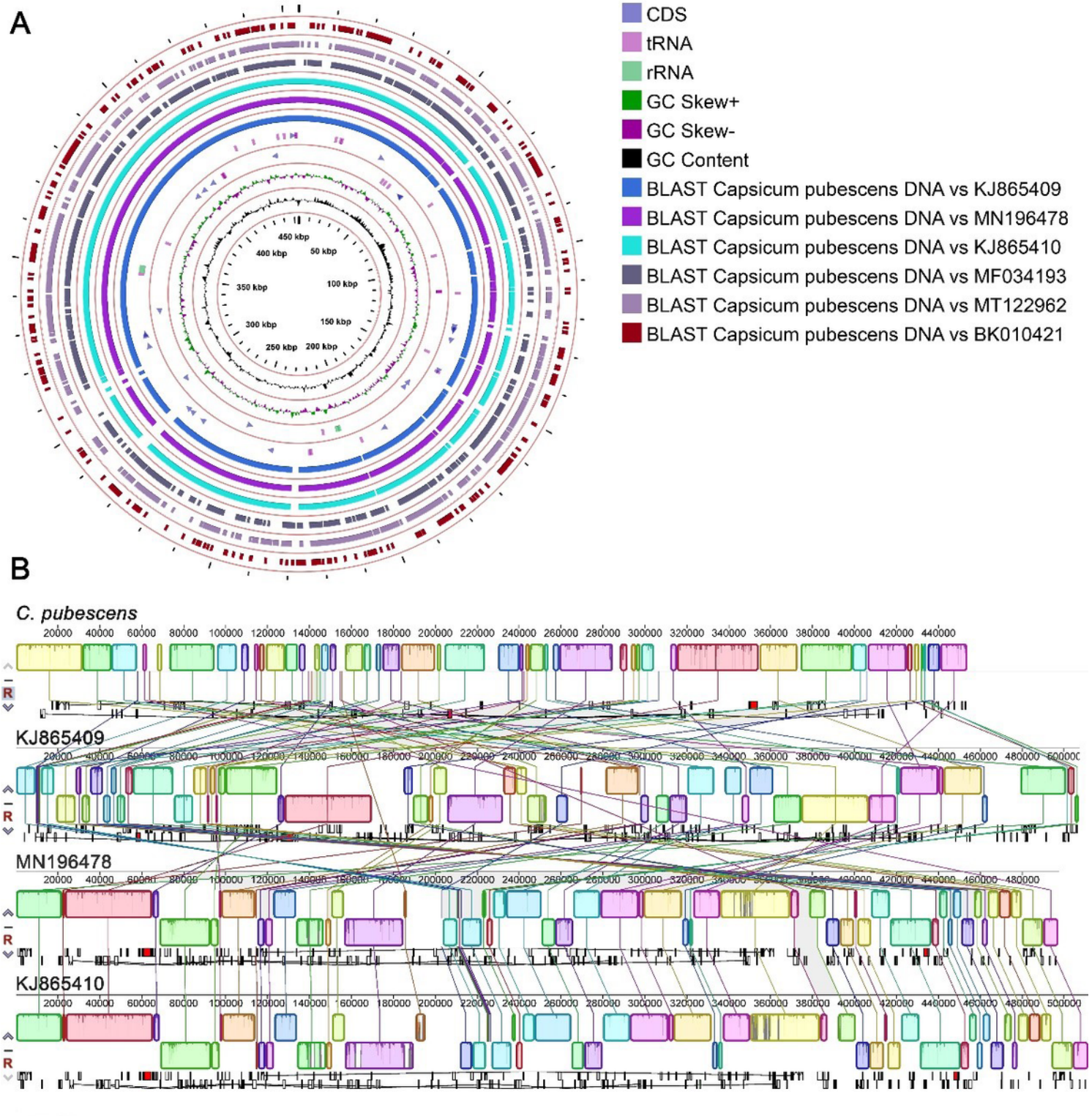
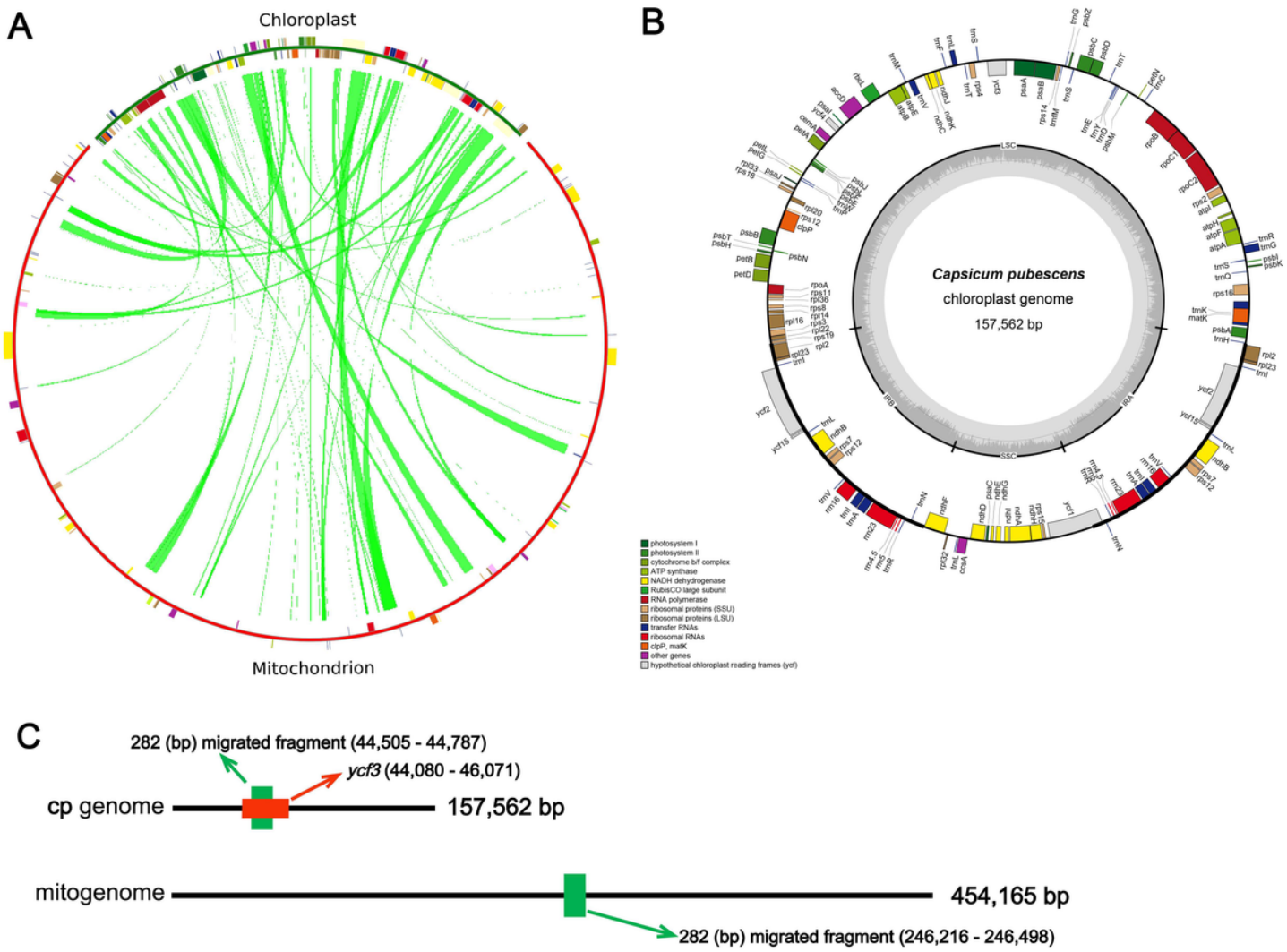



**Figure 10**

Comparison of mitogenome structure and collinearity analysis. **(A)** The outermost six tracks represent the similarity results of the mitogenome alignment of *C. pubescens* and several other species. Further inwards, the four tracks represent positive-strand genes, minus-strand genes, GC Skew, and GC content on the mitogenome of *C. pubescens*, respectively. The innermost circle represents the *C. pubescens* genome size; **(B)** The rectangles represent the similarity between genomes, and the lines between the rectangles represent a collinear relationship. The short white squares represent CDS, the short green squares represent tRNA, and the short red squares represent rRNA. The two graphs are drawn using Proksee (https://proksee.ca/) online tool and Mauve (https://darlinglab.org/mauve/mauve.html) software, respectively.




## Figure 11

Homologous sequence analysis of cp genome and mitogenome. **(A)** Homologous fragments of chloroplast and mitochondrial sequences. "Chloroplast" represents chloroplast sequences, and "Mitochondrion" represents mitochondrial sequences. Genes from the same complex are marked with the same color, and the middle green line joined the homologous sequences; **(B)** The circular map of the cp genome. The genes coded by the upper strand are located outside the circle, while the genes encoded by

the lower strand are on the inside. The inner gray circle represents the GC content; **(C)** The schematic diagram illustrates the migration of the cp genome's *ycf3* gene to the mitogenome. The upper and lower black lines represent the cp genome and mitogenome, respectively. For visual observation, the two genomes are depicted as linear. The Circos v0.69-5 software is used to visualize and map the homologous sequences between chloroplasts and mitochondria. Additionally, the circular map of the chloroplast genome is drawn using the OGDRAW online site (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html).

## Supplementary Files

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

- SupplementaryFiles.zip