# Comparative Analysis of Butternut (Juglans cinerea) and Japanese Walnut (Juglans ailantifolia) Chloroplast Genomes

**Aziz Ebrahimi**

aebrahi@purdue.edu

Purdue University

**Mojtaba Zamani Faradonbeh**

Purdue University

**Abdullah .**

Tianjin University of Traditional Chinese Medicine

**Samarth Mathur**

The Ohio State University

**Martin Williams**

Natural Resources Canada

**Anna O. Conrad**

USDA Forest Service, Northern Research Station

**Douglass F. Jacobs**

Purdue University

**Research Article**

**Keywords:** butternut, whole chloroplast genome, conservation genetics, genetic diversity

**Posted Date:** May 7th, 2025

**DOI:** https://doi.org/10.21203/rs.3.rs-6206868/v1



**Additional Declarations:** No competing interests reported.

**Version of Record:** A version of this preprint was published at BMC Plant Biology on December 8th, 2025. See the published version at https://doi.org/10.1186/s12870-025-07678-1.

# Abstract

*Juglans cinerea* L. (2n = 16), commonly known as butternut, is a diploid species native to the eastern forests of the United States and Canada. It plays a crucial ecological role as a food source for wildlife and is traditionally used as a medicine for tribal communities. Butternut populations have severely declined over 50 years due to butternut canker disease (BCD). The ability of butternut to hybridize with Japanese walnut, which exhibits resistance to BCD, complicates hybrid identification based on morphological traits. In this study, we assembled the chloroplast genome of butternut using next-generation sequencing, resulting in a genome size of 160,289 bp. The genome comprises 112 genes, including 78 protein-coding genes, 30 tRNA genes, and four rRNA genes, with a GC content of 36%. Comparative analysis revealed 1,156 SNPs, including 500 species-specific SNPs between 13 butternut and 6 Japanese walnut accessions, with 367 SNPs located in coding regions, 32 of which had unique SNPs leading to codon and amino acid variations between the two species. We developed 12 species-specific markers using the *matK* and *ycf1* genes because of their higher variability compared with other genes located in coding regions that can be used to assess the genetic background of *Juglans* species. The *ycf1* gene displayed the highest variability, harboring 38 unique SNPs, many resulting in non-synonymous mutations. The newly sequenced chloroplast genome of *J. cinerea* and the identified SNP resources provide valuable genetic tools for characterizing butternut's genetic diversity and distinguishing it from Japanese walnut. These findings enhance our understanding of *Juglans* species and contribute significantly to plant genetics and genomics by comprehensively analyzing chloroplast genomes and SNP variation, addressing a critical gap in previous studies.


# Introduction

*Juglans cinerea* L. (2n = 16), commonly known as butternut, is a diploid species native to the eastern United States and Canada. This species produces valuable timber for industry and plays an important ecological role by providing a rich food source for wildlife. Additionally, butternuts hold significant cultural importance for First Nations and tribal communities, who have used them as a food source, in medicine, and for artwork [1]. Unfortunately, like many North American forest trees (such as elm, ash, and chestnut), butternut populations have declined due to an invasive pathogen. Butternut canker disease (BCD), caused by the fungus *Ophiognomonia clavigignenti-juglandacearum*, was first identified in the late 1960s [2] and has since been identified across the entire range of the species. BCD infections have led to *J. cinerea* being listed as globally endangered, including in Canada [3, 4], and at risk in many U.S. states. With no cure available and minimal evidence of resistance within the species, efforts are ongoing in Canada and the U.S. to develop strategies for protecting North American butternut from extinction.

One promising approach involves taking advantage of naturally occurring butternut hybrids with Japanese walnut (*Juglans ailantifolia*), which was introduced as an ornamental in the late 1800s. Butternut and Japanese walnut hybridize naturally in contact zones, producing offspring with varying resistance to BCD. However, because butternut and Japanese walnut hybridize so readily, distinguishing between pure species and hybrids based on morphology alone can be challenging. Successful conservation and

restoration of butternut will require precise genetic characterization of individual trees for future breeding efforts. Understanding these trees' genetic background at the nuclear and chloroplast levels is essential to identifying pure butternut from hybrids, particularly for advanced backcrosses (those with more than 90% butternut genome). Due to their maternal inheritance, the chloroplast genome can screen the maternal background of butternuts and their hybrids and help characterize the ancestral hybridization between butternuts and other *Juglans* species.

Presently, efforts are ongoing to conserve and restore this species, which requires a greater understanding of butternut and hybrid trees' genetic background (i.e., population structure, gene flow, and marker-assisted breeding) to characterize their hybrid status and screen for resistant traits. Numerous studies have used nuclear-genome to evaluate the genetic diversity of *Juglans* species [5, 6]. Sequence data from chloroplast genomes have been used extensively in plant biology, including phylogenetic analysis [7, 8], biotechnological applications [9], and species identification [10, 5]. While structurally conserved, chloroplast genomes contain highly polymorphic regions facilitating accurate species identification [11, 12]. Additionally, their high copy number enhances the robustness of barcode region amplification. Chloroplast genome divergence is not high within species but higher at an interspecies-specific level [13]. Although useful for species identification, the genome's maternal inheritance needs to be considered since chloroplast capture events can happen during hybridization between closely related congeners, which could lead to misidentification.

Several studies have shown that many chloroplast genes have been lost throughout an evolutionary event in the history of plants or have been functionally transferred to the nuclear genome [13]. For example, genes such as *tuf*A, *fts*H, and *odp*B have moved from the chloroplast to the nuclear genome in certain plant species [13]. Conversely, genes like *ycf1* have been reportedly lost in grasses and cranberries (Ericaceae) despite their essential roles in the import of TOC (translocon on the outer envelope of chloroplasts) and TIC (translocon on the inner envelope of chloroplasts) [14]. These gene transfers have resulted in the expansion, contraction, or even loss of genetic content and a decrease in the size of the chloroplast genome, with these evolutionary events being either species-specific or, in some cases, affecting entire plant orders [15, 16, 17]. Additionally, *ycf1* was the first plastid-encoded protein identified as essential for the survival of *Chlamydomonas reinhardtii* and tobacco (*Nicotiana tabacum*), although its specific function remained unclear at the time [6]. Due to its interspecific variation, *ycf1* is currently listed as one of land plants' most promising plastid DNA barcodes compared to *matK* and *rbcL* [18], previously used for species identification. Using species-specific markers is a practical approach for accurate species identification and population genetic studies. This strategy enhances resolution in distinguishing closely related taxa and assessing genetic diversity within populations [19]. However, a comprehensive comparative analysis of the chloroplast between the North American butternut species and its more resistant Asian hybrid congener, the Japanese walnut, has yet to be conducted.

In this study, we had four objectives: 1) assemble and characterize the chloroplast genome of butternut, 2) evaluate the intraspecific SNP variation within each species and compare the interspecific SNP variation between species, 3) investigate whether the variation observed within the coding regions could

lead to changes at the protein level, and 4) use the fixed variation identified to design new CAPS based markers for use as a taxonomic barcoding tool. This study provides a reference chloroplast genome for butternut and identifies CDS-derived SNP markers that differentiate the two species based on codon and amino acid variations. These CAPS markers that have been developed can be used as barcoding markers to determine the maternal lineage of hybrids, study population genetics, and conduct phylogenetic analyses.

## Materials and Methods

# Chloroplast Assembly of Butternut

Fresh butternut leaves (*Accession #1351*) were collected from the butternut orchard at Purdue University, West Lafayette, Indiana, USA. DNA was extracted using Doyle and Doyle methods [20] with minor modifications. Whole-genome sequencing was performed using the Illumina short-read technology with paired-end libraries on the HiSeq 2500 sequencing platform in a single lane previously described in Ebrahimi et al [21]. The chloroplast *de-novo* genome assembly was completed using the GetOrganelle version (1.7.7.0) [22], and the assembled chloroplast was deposited in NCBI (*Accession* # OR134831) (Fig. 1). The genes in the chloroplast genome were predicted using GeSeq Annotation of Organellar Genomes (23, 24], and the chloroplast map was drawn using Chloroplot [25]. To analyze the loss or presence/absence of genes in butternut (*Juglans cinerea*, Accession #OR134831) compared with Japanese walnut (*Juglans ailantifolia*, Accession #NC_046433.1) and black walnut (*Juglans nigra*, Accession #NC_035967), we use the Feature Tables downloaded from NCBI and the presence/absence genes listed in Supplementary Table 1.

## Phylogeny Analysis

The chloroplast genomes of various *Juglans* (walnuts) and *Carya* (hickories) species were downloaded from NCBI. This includes the assembled chloroplast genome of butternut (*J. cinerea*, #O13R4831). Phylogenetic analysis was conducted using the Maximum Likelihood method, implemented in the MAFFT software, with the Jukes-Cantor substitution model [16). Bootstrap resampling was set to 1000 to ensure robust results. The phylogenetic trees and data were visualized using the tools available at ngphylogeny.fr and iTOL (Interactive Tree of Life). To conduct this analysis, we included the chloroplast genome of *J. cinerea* (accession number: OR134831) alongside four complete chloroplast genomes of other *Juglans* species. Additionally, chloroplast genome sequences from various other plant species were included for comparative purposes, such as *Nicotiana undulata* (tobacco, NC016068), *Cannabis sativa* (hemp, NC026562), *Morus indica* (mulberry, NC008359), *Populus euphratica* (Euphrates poplar, NC024747), *Castanea sativa* (sweet chestnut, NC054204), *Malus domestica* (apple, MK_434916), *Alnus glutinosa* (black alder, NC_039930.1), *Betula platyphylla* (white birch, NC_039994.1), *Carya aquatica* (water hickory, NC_069581.1), *Carya cathayensis* (Chinese hickory, NC_046572.1), *Carya illinoinensis* (pecan, NC_041449.1), *Carya kweichowensis* (Guizhou hickory, NC_040864.1), *Fagus sylvatica* (European beech, NC_041437.1), and *Quercus shennongii* (Shennong oak, NC_068538.1). This broad selection of species

allowed for a thorough phylogenetic analysis (Fig. 2), helping to clarify the evolutionary relationships within and between the *Juglans* and *Carya* genera and other related species.

## Variant Calling

Paired-end sequences of 13 *J. cinerea* and 6 *J. ailantifolia* were downloaded from NCBI (deposited by 8, 25). Then, reads were aligned with the *J. cinerea* reference chloroplast. We processed the reads using BWA-MEM [26] for alignment and Picard tools (https://github.com/broadinstitute/picard) to remove duplicates [27]. We used GATK [28] for genotyping, applying quality filters to minimize errors and converting the *.hmp format (Supplementary Table 2). The genotype frequency of A, T, G, C, and *InDel* for each genotype extracted by vcftools is listed in Fig. 3 and Supplementary Table 5. Based on gene annotation results, the CDS regions start and stop for each gene extracted, and each CDS SNP variant was identified in the *.vcf file using vcftools (Supplementary Table 3). Coding region SNPs were screened to identify unique variants specific to each species and subsequently analyzed to evaluate if the variants impacted codons and subsequent amino acid substitution (Table 1, Supplementary Table 4).

Table 1

Summary of SNPs and unique variations across genome and coding region sequences of **matK** and **ycf1** for butternut (JC) and Japanese walnut (JA) used for evaluating codon and amino acid variations (The information related to other genes listed in Supplementary Table 3). In the "amino acid type based on R group: JA-JC" column, N, P, A, B, and S are abbreviations for Nonpolar, Polar, Acidic, Basic, and Synonymous, respectively.

| ID | Total SNPs | SNPs unique | Unique SNP positions: genome | Unique SNP positions: CDS | JA/JC | Codon in JA | Codon in JC | Amino acid: JA/JC | amino acid type based on R group: JA-JC* |
|---|---|---|---|---|---|---|---|---|---|
| *matK* | 40 | 5 | 2069 | 1398 | C/T | GAG | GAA | Glu/Glu | S |
| | | | 2206 | 1261 | G/T | CAA | AAA | Gln/Lys | P-B |
| | | | 2801 | 666 | G/T | TTC | TTA | Phe/Leu | N-N |
| | | | 3177 | 290 | C/A | GGA | GTA | Gly/Val | N-N |
| | | | 3350 | 117 | C/T | AGG | AGA | Arg/Arg | S |
| *ycf1* | 84 | 37 | 129813 | 5671 | C/A | GAT | TAT | Asp/Tyr | A-P |
| | | | 130038 | 5446 | G/A | CAT | TAT | His/Tyr | B-P |
| | | | 130126 | 5358 | A/C | TTT | TTG | Phe/Leu | N-N |
| | | | 130600 | 4884 | T/G | AAA | AAC | Lys/Asn | B-P |
| | | | 130656 | 4828 | T/G | AAA | CAA | Lys/Gln | B-P |
| | | | 130733 | 4751 | G/T | ACT | AAT | Thr/Asn | P-P |
| | | | 130744 | 4740 | G/C | TCC | TCG | Ser/Ser | S |
| | | | 130783 | 4701 | G/T | AGC | AGA | Ser/Arg | P-B |
| | | | 130847 | 4637 | A/G | GTC | GCC | Val/Ala | N-N |
| | | | 131446 | 4038 | C/T | ATG | ATA | Met/Ile | N-N |
| | | | 131584 | 3900 | A/G | TCT | TCC | Ser/Ser | S |
| | | | 131825 | 3659 | C/T | TGT | TAT | Cys/Tyr | P-P |
| | | | 131995 | 3489 | A/T | TTT | TTA | Phe/Leu | N-N |
| | | | 132095 | 3389 | G/A | TCA | TTA | Ser/Leu | P-N |
| | | | 132259 | 3225 | A/G | ATT | ATC | Ile/Ile | S |
| | | | 132282 | 3202 | T/G | ATA | CTA | Ile/Leu | N-N |
| | | | 132387 | 3097 | T/C | AAA | GAA | Lys/Glu | B-A |

| ID | Total SNPs | SNPs unique | Unique SNP positions: genome | Unique SNP positions: CDS | JA/JC | Codon in JA | Codon in JC | Amino acid: JA/JC | amino acid type based on R group: JA-JC* |
|---|---|---|---|---|---|---|---|---|---|
| | | | 132833 | 2651 | G/C | ACA | AGA | Thr/Arg | P-B |
| | | | 132976 | 2508 | T/A | ATA | ATT | Ile/Ile | S |
| | | | 133011 | 2473 | G/T | CTT | ATT | Leu/Ile | N-N |
| | | | 133120 | 2364 | T/G | AGA | AGC | Arg/Ser | B-P |
| | | | 133170 | 2314 | G/A | CAT | TAT | His/Tyr | B-P |
| | | | 133192 | 2292 | C/A | AAG | AAT | Lys/Asn | B-B |
| | | | 133287 | 2197 | G/C | CCC | GCC | Pro/Ala | N-N |
| | | | 133616 | 1868 | C/G | GGA | GCA | Gly/Ala | N-N |
| | | | 133681 | 1803 | A/C | TGT | TGG | Cys/Trp | P-N |
| | | | 133710 | 1774 | G/T | CTT | ATT | Leu/Ile | N-N |
| | | | 133756 | 1728 | T/G | TTA | TTC | Leu/Phe | N-N |
| | | | 133770 | 1714 | A/G | TCA | CCA | Ser/Pro | P-N |
| | | | 133787 | 1697 | A/C | ATC | AGC | Ile/Ser | N-P |
| | | | 133906 | 1578 | A/C | TTT | TTG | Phe/Leu | N-N |
| | | | 133942 | 1542 | A/C | ACT | ACG | Thr/Thr | S |
| | | | 134109 | 1375 | G/A | CTA | TTA | Leu/Leu | S |
| | | | 134137 | 1347 | A/G | AAT | AAC | Asn/Asn | S |
| | | | 134304 | 1180 | T/G | AAA | CAA | Lys/Gln | B-P |
| | | | 134334 | 1150 | T/C | AAA | GAA | Lys/Glu | B-A |
| | | | 134374 | 1110 | T/G | CGA | CGC | Arg/Arg | S |

## Development of High-Specificity Markers Using CAPS Methods

CAPS primers for the *matK* and *ycf1* genes were designed based on species-specific SNPs identified within these genes. To identify suitable restriction enzymes for each SNP and its surrounding nucleotide sequence, we utilized the web-based tool CAPS Finder 2.0 (http://helix.wustl.edu/CAPS/). NEBcutter 2.0 (http://tools.neb.com/NEBcutter) was subsequently employed to analyze restriction enzyme sites for

each gene. Based on the CAPS Finder and NEBcutter results, primers were designed to amplify SNP-containing segments by PCR. For primer design, we used the PrimerQuest™ Tool (https://www.idtdna.com/pages/tools/primerquest), ensuring that the CAPS assay cut sites would be appropriate for separation and visualization on agarose gels.

## Results

# Chloroplast Assembly of Butternut

Paired-end genome data was used for the de novo genome assembly of butternut. The chloroplast genome has 112 genes, of which 78 are protein-coding genes, 30 are tRNA genes, and 4 are rRNA genes. The GC content is 36%, while the LSC (Large Single Copy) length is 89,805 bp, the IR (Inverted Repeat) length is 26,035 bp, and the SSC (Small Single Copy) length is 18,415 bp. The complete chloroplast genome length was 160,288 bp (Fig. 1).

*Phylogenomic Analysis of* Juglans *Species*

Phylogenomic analysis revealed that the *J. cinerea* chloroplast is most closely related to *Juglans nigra*, while other *Juglans* species from Asia clustered together (Fig. 2). All *Juglans* and *Carya* species formed a cluster consistent with their classification in the Juglandaceae family. The chloroplast genome of the Juglandaceae family was found to be more closely related to those of the *Alnus* and *Betula* genera compared to the more distantly related *Fagus*, *Castanea*, and *Quercus* genera.

## Gene Loss Based on the Whole Chloroplast Genome

Comparing butternut chloroplast with Japanese walnut and black walnut revealed several key differences. Specific genes such as *trnM-CAU, trnT-GGU, trnP-GGG, infA*, *ycf15*, and *rps12* are present in Japanese walnut but absent in both butternut and black walnut. The *ycf1* gene has two copies in Japanese walnut, while butternut and black walnut have only one copy. The length of ycf1 genes is 5,675 bp with significant sequence diversity and variation. The *rpl2* gene is present in Japanese walnut and butternut but is absent in black walnut (Supplementary Table 1).

## SNPs in the Chloroplast Genome of Butternut and Japanese Walnut

We identified 1,156 SNPs in the chloroplast genomes of 13 butternut and 6 Japanese walnut samples (Supplementary Table 2). Among these, 366 SNPs were in coding regions (CDS) (Supplementary Table 3). Most SNPs in these coding regions belong to *ycf1* and *mat*K genes, with 84 and 40 SNPs, respectively. Other notable SNP counts include 27 in *psa*A, 19 in *rpo*C2, and 17 in *rbc*L. The *ycf1* and *matK* genes were the largest coding-region genes identified. Of 47 CDS genes, 32 had both synonymous and non-synonymous SNP variations, and the other 15 CDS were not species-specific (Table 1; Supplementary

Table 4). When comparing allele frequencies across all SNPs, we observed that the A, G, and C allele ratio was higher in butternut than in Japanese walnut (Fig. 3).

# Codon and Amino Acid Variations in Chloroplast Genomes of Butternut and Japanese Walnut

The analysis revealed notable codon and amino acid variations between butternut and Japanese walnut across 32 genes in their chloroplast genomes, comprising 37.4% synonymous and 62.6% nonsynonymous mutations (Fig. 4; Table 1; Supplementary Table 4). Five unique SNPs were identified in the *matK* gene. Two of the five SNPs have synonymous effects. Among the other three remaining substitutions, only one is a radical substitution, where glutamine is substituted for a lysine (CAA → AAA).

The *ycf1* gene showed the highest variability, with 37 unique SNPs resulting in 8 synonymous and 29 non-synonymous mutations. Fifteen of the 29 nonsynonymous mutations are considered radical and include GAT → TAT (aspartate to tyrosine), AAA → AAC (lysine to asparagine), and TTT → TTA (phenylalanine to leucine). According to our SNP analysis results for this gene, none of the amino acid substitutions were located in the alpha helix coding region of *ycf1*.

*Optimized CAPS Marker Selection and Primer Design for* matK *and* ycf1 *Genes*

For CAPS marker selection, we focused on *matK* and *ycf1* genes consistently used for plant barcoding. For *matK*, 3 out of 5 SNPs were selected for CAPS design. For *ycf1*, 9 out of 37 SNPs were chosen. CAPS marker amplicons are typically shorter than 300 bp. When using the CAPS Finder tool, setting the nucleotide mismatch to zero provides greater flexibility in primer design. However, for SNP position 666 in *matK*, a mismatch count of two was required, causing the reverse primer to overlap the SNP. Primers designed for this SNP amplify a 191 bp segment, and digestion with BsaJI in *J. aliantifolia* samples is expected to yield two fragments of 166 bp and 25 bp. Since fragments shorter than 50 bp are challenging to detect on agarose gels, we prioritized other SNPs that required no mismatches during the initial analysis. For the remaining 11 CAPS markers, the smallest expected fragment post-digestion is 79 bp (for SNP 117 in *matK*), ensuring that all fragments are within a detectable range for agarose gel analysis (Supplementary Table 6).

## Discussion

# Chloroplast Genome Assembly and Gene Content

In this study, the chloroplast genome of butternut was assembled and annotated using paired-end reads data that has provided a comprehensive overview of its genetic architecture, revealing a highly conserved characteristic in terms of genomic numbers and GC content. The complete chloroplast genome of butternut is 160,288 base pairs (bp) in length and consists of 112 genes, including 78 protein-coding genes, 30 transfer RNA (tRNA) genes, and four ribosomal RNA (rRNA) genes. The genome exhibits a typical quadripartite structure, with a Large Single Copy (LSC) region of 89,805 bp, an Inverted Repeat (IR)

region of 26,035 bp, and a Small Single Copy (SSC) region of 18,415 bp. The overall GC content is 36.00%, consistent with other chloroplast genomes in the Juglandaceae family and other angiosperms [8, 29, 30].

# Comparative Cp Genomics Identifies Gene Loss and Non-synonymous SNP Variants in Genes of Interest

Comparative analysis of the butternut chloroplast genome with those of Japanese walnut and black walnut reveals several significant differences in gene content, which provide insights into the evolutionary divergence within the Juglandaceae family. Specifically, genes such as *trnM-CAU*, *trnT-GGU*, *trnP-GGG*, *inf*A, *ycf15*, and *rps12* are present in Japanese walnut but absent in both butternut and black walnut. Additionally, *ycf1* has two copies in Japanese walnut, while only functional copies exist in butternut and black walnut. One of the *ycf1* proteins in butternut is fragmented, and this copy is missing in black walnut. Our web research in NCBI shows that *ycf1* duplication is also present in other *Juglans* species from Asia, such as *J. cathayensis* (NC_033893.1), *J. mandshurica* (NC_033892.1), and *J. regia* (NC_028617.1). Like black walnut, other members of the *Rhysocaryon* section of the Juglandaceae, such as California walnut (J. hindsii, NC_035965.1) and Arizona walnut (J. major, NC_035966.1), also have only one copy of *ycf1*. Interestingly, another North American congener species, the Texas walnut (*J. microcarpa*, NC_046432.1), which also belongs to the *Rhysocaryon* section, has two copies of the *ycf1* gene, like eastern Asia *Juglans* species.

Comparative analysis of the butternut chloroplast genome with those of Japanese walnut and black walnut reveals several significant differences in gene content, which provide insights into the evolutionary divergence within the Juglandaceae family. These genes play various roles in chloroplast function, including protein synthesis and photosynthesis, suggesting potential differences in chloroplast functionality and efficiency between these species. Gene loss or pseudogenes in chloroplast genomes have lost their ability to encode a protein [31]. Although lacking the ability to produce proteins, the remaining gene regions have the potential to act like regulatory functions akin to those found in other non-coding regions [32]. Pseudogenes can also be significant in normal physiological and pathological conditions [33]. Notably, our research shows that the *ycf1* and *infA* genes are either pseudogenes or absent in butternut, while they are present in the chloroplast of Japanese walnut. It has previously been reported that the loss of *ycf1* and *infA* genes has already been observed in the Liliaceae family, as documented by She et al. [34], with seemingly no detrimental impacts, suggesting possible functional compensation through nuclear-encoded genes or alternative mechanisms. The *infA* gene, absent in butternut, encodes a component of ribosomal protein L23 (*rpl23*), which in turn is involved in the synthesis of translation initiation factor 1 (*IF1*) and is found near other soluble translation factors coding regions in the chloroplast genome [35]. Among the bacterial genes encoding translation initiation factors, only *infA* has a homolog in the plastome of higher plants [Millen et al., 2001]. The *yfc1* gene, predicted as a pseudogene, has an unusual reading frame for plastid genes, typically located at the junction of the inverted repeat (IR) and small single copy (SSC) regions [36]. Although the exact function of the *ycf1* protein is unknown, it is crucial for plant viability [37]. The size of *ycf1* genes in both butternut and Japanese walnut is notably long (5676 bp) for the size of the genes, and they have 1891 amino acids. This aligns with findings that specific lineages encode *ycf1* proteins exceeding 1,500 amino acids; sometimes,

these proteins can be longer than 3,000 [14]. Furthermore, the *ycf1* gene is involved in photosynthesis and has been identified as a critical marker in phylogenetic studies [38, 30]. The presence of two copies in Japanese walnut could indicate a duplication event that may confer an evolutionary advantage, possibly related to enhanced photosynthetic efficiency or adaptability to different environmental conditions. Another gene of interest, the *rpl2* gene, essential for ribosome biogenesis, is present in Japanese walnut and butternut but absent in black walnut. The loss of this gene in black walnut could have significant implications for its chloroplast ribosome assembly or overall protein synthesis capacity. These gene loss events and structural variations are crucial for understanding the evolutionary pressures or adaptations that have shaped the chloroplast genomes of these species [38, 39].

# Butternut Reveals Distinct Allele Frequency Patterns Compared to Japanese Walnut

Our allele frequency analysis revealed distinct patterns between butternut and Japanese walnut, with higher ratios of alleles A, G, and C in butternut. This finding is consistent with the allele frequency patterns observed in other chloroplast genome studies, such as those conducted on pomegranate and Ficus species, where specific allele distributions were linked to environmental adaptation and evolutionary history [38, 30]. These allele frequency differences highlight the importance of SNP analysis in understanding the genetic basis of phenotypic traits and environmental responses. Some genes, such as *matK* and *ycf1*, were highly variable in butternut and Japanese walnut. Of all the variations identified, 31.7% were located within CDS regions and found in most chloroplast genes (68%). Interestingly, 33% of that SNP variation was found in the CDS regions of only two genes, *ycf1*, and *matK*, with 84 and 40 SNPs, respectively. This high variability within these genes was previously reported in the chloroplast of eastern Asia *Juglans* species [8].

# Codon and Amino Acid Variations of Butternut and Japanese Walnut

Codon and amino acid variations in the chloroplast genomes of butternut and Japanese walnut provide critical insights into species differentiation and evolutionary dynamics. Genes such as *matK*, *atpA*, *rps4*, *rps8*, *rpl14*, *rpl22*, and *ycf1* exhibit synonymous and non-synonymous mutations, potentially affecting protein function and species-specific traits. The ycf1 gene shows the highest variability, with 38 unique SNPs, making it a key marker for distinguishing butternut from Japanese walnut and consistent for use as a barcode for divergence in other plants. Non-synonymous mutations in these genes suggest potential functional protein differences, which could contribute to distinct phenotypic traits. These variations are vital for species authentication, particularly in conservation efforts and hybridization studies, where precise identification is essential to clarify genetic lineages. Amino acid mutations are classified as conservative or radical, with radical substitutions more likely to affect protein stability [40]. Although amino acid substitutions can potentially alter protein function, SNP analysis in this study revealed no substitutions within the alpha-helix region of *ycf1*, critical for maintaining the protein's secondary and tertiary structure. This region comprises six alpha-helix domains (amino acids 106–326) that anchor the complex [22]. Additionally, key Tic214 residues (K1939, K1880, T1795), which are critical for protein

function, showed no variation. The Tic214 protein plays a vital role in the TOC-TIC super-complex, facilitating chloroplast protein import by bridging the TOC and TIC complexes and spanning the intermembrane space [22].

*Application of CAPS Markers for Authentication of* Juglans *Species*

DNA barcoding uses short DNA sequences from organelle genomes to differentiate plant species and is widely applied for species-specific identification, biodiversity assessment, and evolutionary studies (41). Accurate identification is crucial, particularly when congeners with distinct interspecific traits (i.e., disease resistance and adaptive traits) can hybridize naturally. For example, hybridization between butternut and Japanese walnut over the past century complicates the identification of genetic background, as morphological markers are often insufficient to distinguish between species. This is especially important given the threat of BCD to butternut populations, highlighting the need for practical conservation tools. Chloroplast markers, with their high conservation, copy numbers, and sufficient interspecific divergence, are ideal for developing DNA barcodes for species verification [42, 12]. This study identified 32 species-specific SNP markers from chloroplast coding sequences (CDS) in butternut and Japanese walnuts. We developed 12 unique markers for each species targeting matK and ycf1 using comparative chloroplast genome analysis. These CAPS markers are highly distinctive and provide practical tools for species authentication and conservation, particularly distinguishing closely related species, such as other Asian walnuts.

# Conclusion

This study comprehensively analyzed chloroplast genomes in butternut and Japanese walnut, revealing structural variations, SNP patterns, and differences in gene content. The assembled butternut chloroplast genome demonstrates high structure and GC content conservation, consistent with other Juglandaceae species. Comparative analyses identified 32 species-specific SNP markers, exhibiting significant variability in the ycf1 and matK genes, which are important for species differentiation. Gene loss and the presence of pseudogenes, such as infA and ycf1, further highlight the evolutionary divergence in these species. The developed species-specific CAPS markers provide practical tools for species authentication and conservation, supporting efforts to protect butternut populations and mitigate the impacts of hybridization. These findings enhance understanding of chloroplast genome dynamics, supporting biodiversity and conservation genetics.

# Declarations

### Abbreviations

NA

### Ethics approval and consent to participate

Approved

### Consent for publication

Approved

### Availability of data and materials

All the data included in manuscript and the genome deposited in NCBI: chloroplast genome of butternut (*J. cinerea*, accession number: O13R4831)

### Competing Interests

NA

### Funding

The USDA NIFA (Award Number #: 2023-67012-39608) and Fred M. van Eck Foundation funded this research.

### Authors' contributions

A.E. assembled the chloroplast genome, designed the research, analyzed the data, and wrote the manuscript. M.Z.F., S.M., and A. participated in data analysis and edited the paper. M.W., A.O.C., and D.F.J. revised the manuscript.

### Acknowledgements

We acknowledge the Genomics Center at Purdue University for Providing Services for Genome sequencing.

## References

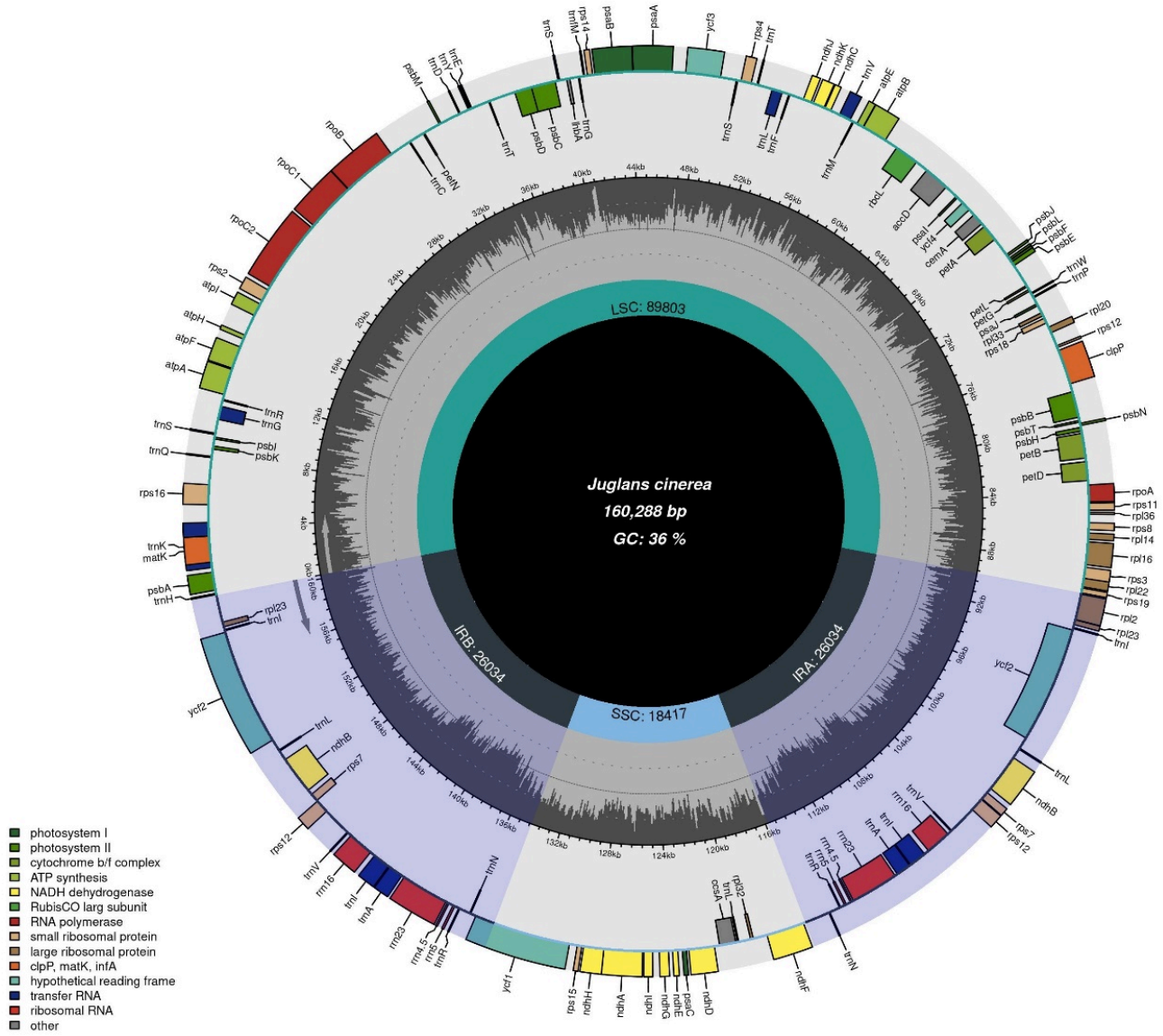
1. Moerman DE. Native American Ethnobotany. Portland: Timber; 1998.
2. Renlund DW. Forest pest conditions in Wisconsin. In: Annual Report 1971. Madison: Wisconsin Department of Natural Resources; 1971. pp. 26–28.
3. Environment Canada. Species at Risk Public Registry. 2010. Available from: https://www.canada.ca/en/environment-climate-change/services/species-risk-public-registry.html
4. Stritch L, Barstow M. *Juglans cinerea*, Butternut. The IUCN Red List of Threatened Species 2019: e.T62019689A62019696. 2019. Available from: https://www.iucnredlist.org/species/62019689/62019696
5. Zhao P, Zhou HJ, Potter D, Hu YH, Feng XJ, Dang M, et al. Population genetics, phylogenomics and hybrid speciation of *Juglans* in China determined from whole chloroplast genomes, transcriptomes, and genotyping-by-sequencing (GBS). Mol Phylogenet Evol. 2018;126:250–65.
6. Drescher A, Ruf S, Calsa T, Carrer H, Bock R. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. 2000;22(2):97–104.

7. Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. From algae to angiosperms—inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol Biol. 2014;14:1–27.
8. Zhou H, Hu Y, Ebrahimi A, Liu P, Woeste K, Zhao P, et al. Whole genome based insights into the phylogeny and evolution of the Juglandaceae. BMC Ecol Evol. 2021;21(1):1–16.
9. Sabir J, Schwarz E, Ellison N, Zhang J, Baeshen NA, Mutwakil M, et al. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. Plant Biotechnol J. 2014;12:743–54.
10. Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JM, et al. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. Am J Bot. 2012;99:320–9.
11. Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. Plant DNA barcoding: from gene to genome. Biol Rev. 2015;90(1):157–66.
12. Giang VNL, Waminal NE, Park HS, Kim NH, Jang W, Lee J, et al. Comprehensive comparative analysis of chloroplast genomes from seven *Panax* species and development of an authentication system based on species-unique single nucleotide polymorphism markers. J Ginseng Res. 2020;44(1):135–44.
13. Eckardt NA. Genomic hopscotch: gene transfer from plastid to nucleus. Plant Cell. 2006;18(12):2865–8.
14. de Vries J, Sousa FL, Bölter B, Soll J, Gould SB. YCF1: a green TIC. Plant Cell. 2015;27(7):1827–33.
15. Turmel M, Otis C, Lemieux C. The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. Mol Biol Evol. 2006;23(6):1324–38.
16. Wicke S, Schneeweiss GM, Depamphilis CW, Müller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 2011;76:273–97.
17. Mohanta TK, Mishra AK, Khan A, Hashem A, Abd_Allah EF, Al-Harrasi A. Gene loss and evolution of the plastome. Genes. 2020;11(10):1133.
18. Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S. 2015. ycf1, the most promising plastid DNA barcode of land plants. Scientific reports, 5(1), pp.1–5.
19. Ahmed I, Lockhart PJ, Agoo EM, Naing KW, Nguyen DV, Medhi DK, et al. Evolutionary origins of taro (*Colocasia esculenta*) in Southeast Asia. Ecol Evol. 2020;10(23):13530–43.
20. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987;19:11–5.
21. Ebrahimi A, Mathur S, Lawson SS, LaBonte NR, Lorch A, Coggeshall MV, et al. Microsatellite borders and micro-sequence conservation in *Juglans*. Sci Rep. 2019;9(1):3748.
22. Jin JJ, Yu WB, Yang JB, Song Y, DePamphilis CW, Yi TS, et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020;21:1–31.
23. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq—versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45(W1):W6–11.

24. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 2019;47:W59–64.
25. Zheng S, Poczai P, Hyvönen J, Tang J, Amiryousefi A. Chloroplot: an online program for the versatile plotting of organelle genomes. Front Genet. 2020;11:576124.
26. Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30:2843–2851.
27. Wysoker A, Tibbetts K, Fennell T. Picard tools version 1.90. 2013. Available from: https://github.com/broadinstitute/picard
28. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
29. Liu X, Luo J, Zhang M, Wang Q, Liu J, Wu D, Fu Z. 2023. Phylogenomic analysis of two species of Parasenecio and comparative analysis within tribe Senecioneae (Asteraceae). Diversity, 15(4), p.563.
30. Huang Y, Li J, Yang Z, An W, Xie C, Liu S, Zheng X. 2022. Comprehensive analysis of complete chloroplast genome and phylogenetic aspects of ten Ficus species. BMC plant biology, 22(1), p.253.
31. Prade VM, Gundlach H, Twardziok S, Chapman B, Tan C, Langridge P, et al. The pseudogenes of barley. Plant J. 2018;93(3):502–14.
32. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumor biology. Nature. 2010;465(7301):1033–8.
33. Lu X, Yang J, Zhao S, Zhang X, Meng Y, Li M, et al. Pseudogene RNA: a novel functional molecule in plants. Int J Mol Sci. 2015;16(11):25352–63.
34. She R, Zhao P, Zhou H, Yue M, Yan F, Hu G, et al. Complete chloroplast genomes of Liliaceae (sl) species: comparative genomic and phylogenetic analyses. Nord J Bot. 2020;38(1):1–12.
35. Sijben-Muller EA, Neuteboom LW, de Vries PM, van den Berg H. Translation initiation factor IF1 in Escherichia coli. Eur J Biochem. 1986;154(2):455–9.
36. Raubeson LA, Jansen RK. Chloroplast genomes of plants. Plant Diversity and Evolution: Genotypic and Phenotypic Variation in Higher Plants. Cambridge: Cambridge University Press; 2005. pp. 45–68.
37. Drescher A, Ruf S, Calsa T, Carrer H, Bock R. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. 2000;22(2):97–104.
38. Singh NV, Patil PG, Sowjanya RP, Parashuram S, Natarajan P, Babu KD, et al. Chloroplast genome sequencing, comparative analysis, and discovery of unique cytoplasmic variants in pomegranate (*Punica granatum* L). Front Genet. 2021;12:704075.
39. Li B, Huang K, Chen X, et al. Comparative and phylogenetic analysis of chloroplast genomes from four species in *Quercus* section *Cyclobalanopsis*. BMC Genom Data. 2024;25:57.
40. Zhang J. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J Mol Evol. 2000;50(1):56–68.
41. Zuo Y, Chen Z, Kondo K, Funamoto T, Wen J, Zhou S. DNA barcoding of *Panax* species. Planta Med. 2011;77(02):182–7.

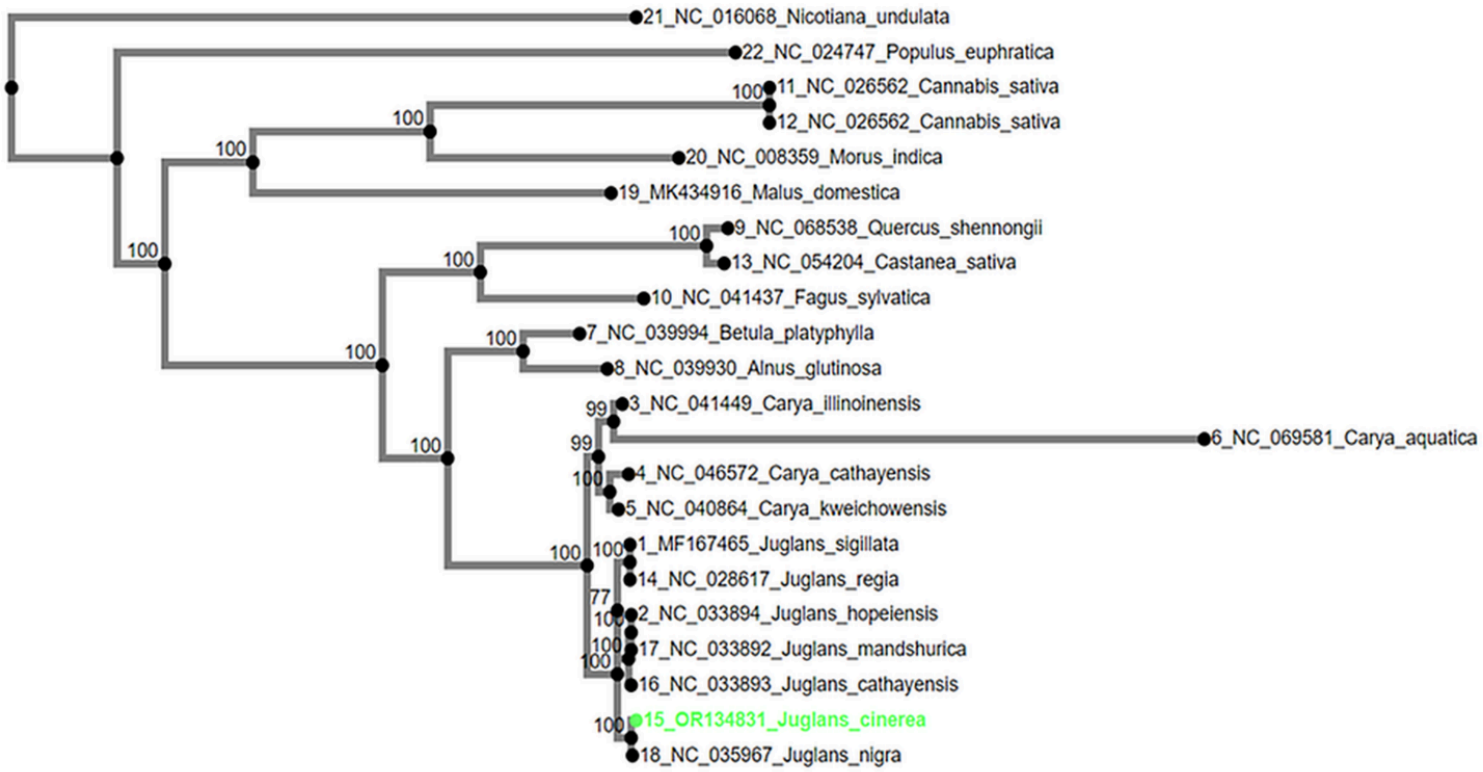
42. Nguyen VB, Park HS, Lee SC, Lee J, Park JY, Yang TJ. Authentication markers for five major *Panax* species developed via comparative analysis of complete chloroplast genome sequences. J Agric Food Chem. 2017;65(30):6298–306.

# Figures



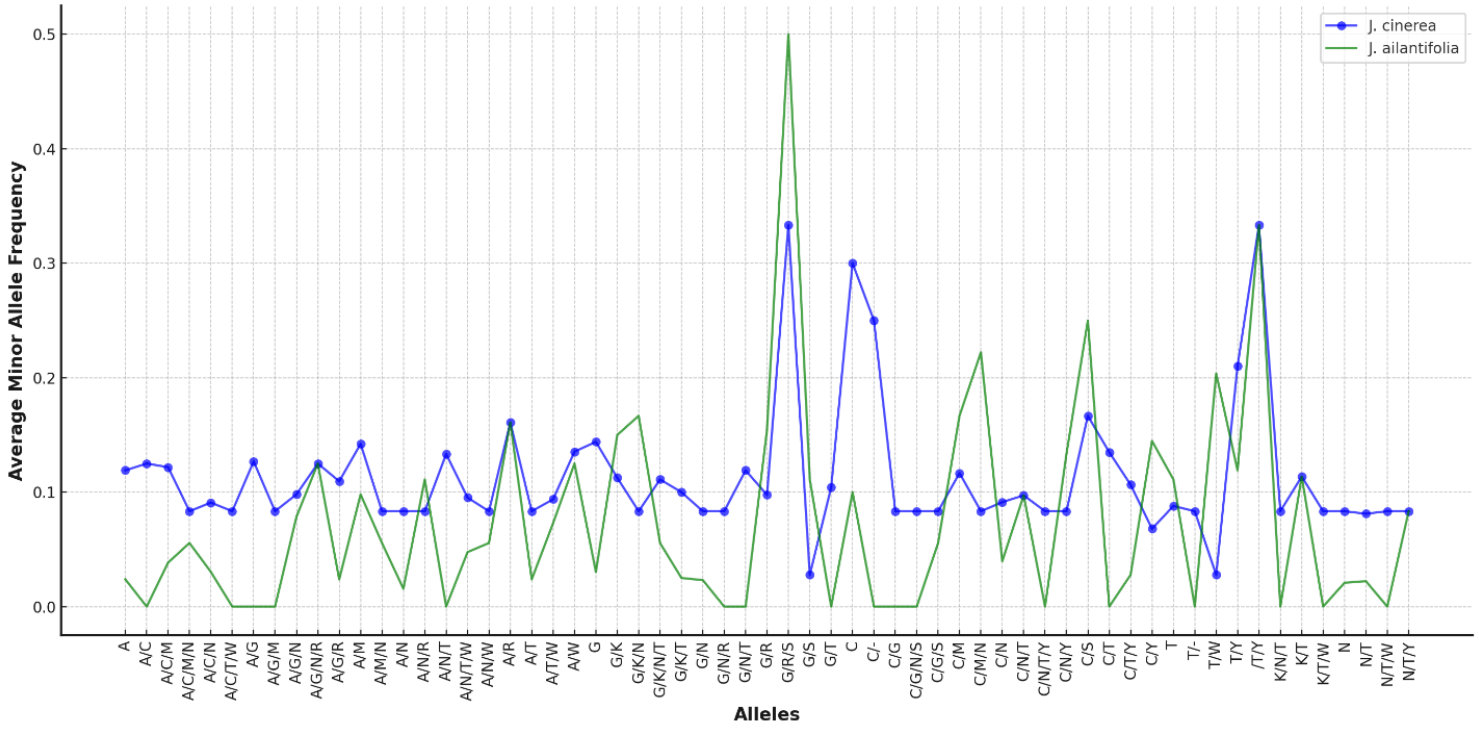

## Figure 1

A Circular map displaying the *Juglans cinerea* cp genome with all the genes annotated. The genes transcribed clockwise are shown within the circle, while those transcribed anticlockwise are conducted outside. LSC, SSR, IRA, and IRB define the chloroplast genome's borders. Each gene is coded according to its specific function. A dashed grey color represents the inner circle showing GC content, while the lighter grey represents AT range.




## Figure 2

Maximum Likelihood phylogenetic tree based on chloroplast of *Juglans cinerea*, other species like *Juglans*, *Carya*, *Engelhardia, and other* hardwood species belonging to the *Juglandaceae*family.




## Figure 3

Comparison of Minor Allele Frequencies for *J. cinerea* and *J. ailantifolia*. The blue circles represent data for *J. cinerea*, while the green crosses correspond to *J. ailantifolia*. Alleles are arranged starting with A, G, C, and T, followed by alphabetical order for the remaining alleles.

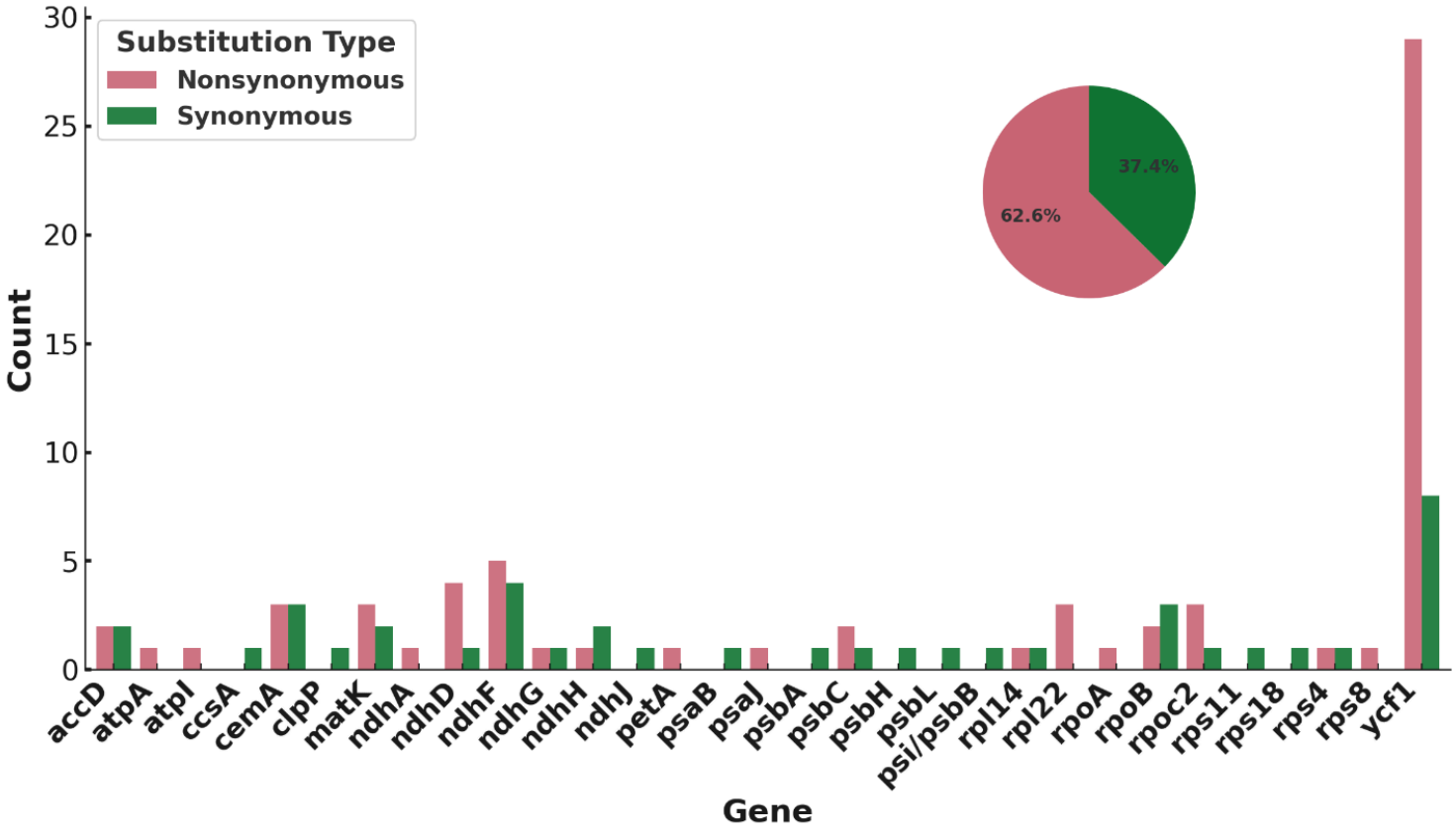



**Figure 4**

Distribution of Synonymous and Nonsynonymous Substitutions in butternut and Japanese walnut. The bar chart shows the counts of synonymous (green) and nonsynonymous (red) substitutions across genes in butternut and Japanese walnut. The pie chart summarizes the overall proportions of substitution types.

## Supplementary Files

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

- SupplementaryTable1a.xlsx
- SupplementaryTable2.csv
- SupplementaryTable3.csv
- SupplementaryTable4.xlsx
- SupplementaryTable5.csv
- SupplementaryTable6.docx