Comparative analysis of complete *Ilex* (Aquifoliaceae) chloroplast genomes: insights into evolutionary dynamics and phylogenetic relationships

Kewang Xu¹, Chenxue Lin¹, Shiou Yih Lee², Lingfeng Mao¹* and Kaikai Meng³*

**Abstract**

**Background:** *Ilex* (Aquifoliaceae) are of great horticultural importance throughout the world for their foliage and decorative berries, yet a dearth of genetic information has hampered our understanding of phylogenetic relationships and evolutionary history. Here, we compare chloroplast genomes from across *Ilex* and estimate phylogenetic relationships.

**Results:** We sequenced the chloroplast genomes of seven *Ilex* species and compared them with 34 previously published *Ilex* plastomes. The length of the seven newly sequenced *Ilex* chloroplast genomes ranged from 157,182 bp to 158,009 bp, and contained a total of 118 genes, including 83 protein-coding, 31 rRNA, and four tRNA genes. GC content ranged from 37.6 to 37.69%. Comparative analysis showed shared genomic structures and gene rearrangements. Expansion and contraction of the inverted repeat regions at the LSC/IRA and IRA/SSC junctions were observed in 22 and 26 taxa, respectively; in contrast, the IRb boundary was largely invariant. A total of 2146 simple sequence repeats and 2843 large repeats were detected in the 41 *Ilex* plastomes. Additionally, six genes (*psaC*, *rbcL*, *trnQ*, *trnR*, *trnT*, and *ycf1*) and two intergenic spacer regions (*ndhC-trnV* and *petN-psbM*) were identified as hypervariable, and thus potentially useful for future phylogenetic studies and DNA barcoding. We recovered consistent phylogenetic relationships regardless of inference methodology or choice of loci. We recovered five distinct, major clades, which were inconsistent with traditional taxonomic systems.

**Conclusion:** Our findings challenge traditional circumscriptions of the genus *Ilex* and provide new insights into the evolutionary history of this important clade. Furthermore, we detail hypervariable and repetitive regions that will be useful for future phylogenetic and population genetic studies.

**Keywords:** Aquifoliaceae, Chloroplast genome, Hypervariable regions, Phylogenomics, Relationship

---

**Introduction**

*Ilex* L., comprised of ca. 600 evergreen or deciduous tree and shrub species, is the only genus in the family Aquifoliaceae [1]. Members of the genus are mostly distributed in the tropics, with centers of species diversity located in tropical America and southeast Asia, but also extending into temperate regions [2, 3]. Most species of *Ilex*, including *I. cornuta* Lindl. et Paxt., *I. purpurea* Hassk., *I. paraguariensis* A. St.-Hil., and *I. rotunda* Thunb.,
have economic and horticultural value [4–8] and relatively broad ranges, although many species are narrowly endemic. To date, as many as 250 species of *Ilex* have been classified as endangered and placed on the International Union for Conservation of Nature (IUCN) red list [9].

In the past two decades, advances in sequencing technology and analytical methods have contributed to greater phylogenetic resolution within *Ilex*. Several loci from both the nuclear and plastid genomes, including *rbcL*, *trnL-trnF*, atpB-*rbcL*, nuclear ribosomal DNA internal transcribed spacers (nrITS), and chloroplast glutamine synthetase (*negGS*), have been used to estimate phylogenetic relationships within the genus [10–17]. However, a broad and representative sample of *Ilex* species has not yet been achieved in any phylogenetic study; thus the phylogeny of *Ilex* remains largely unresolved [13, 16]. Furthermore, recent phylogenetic studies have revealed substantial incongruence between the nuclear and plastid topologies [10, 13–15]. Recent molecular phylogenies did not support traditional classifications of *Ilex* based on morphological features [18, 19]; however, these studies used only a few plastid or nuclear gene fragments and had generally poor resolution due to high conservation of plastid genes. At present, the phylogenetic relationships among lineages in genus *Ilex* remain uncertain, thus, further investigations are needed to reconstruct the evolutionary history of this clade.

Complete chloroplast genomes have been relatively more successful than short sequence fragments in resolving the relationships of many land plant clades at different taxonomic levels [20–22]. In general, land plant chloroplast genomes are relatively stable and contain different taxonomic levels [20–22]. In general, land plant chloroplast genomes are relatively stable and contain different taxonomic levels [20–22]. In general, land plant chloroplast genomes are relatively stable and contain different taxonomic levels [20–22]. In general, land plant chloroplast genomes are relatively stable and contain different taxonomic levels [20–22].

The diversity of nucleotide variability (Pi) for the seven newly assembled plastomes, combined with 34 plastomes obtained from GenBank, ranged from 0.0000 to 0.0128, with an average of 0.0028. Based on the cutoff value of Pi ≥ 0.009, eight highly variable regions (807 bp + *trnR*~UCU~ + 384 bp, 579 bp + *psaC* + 382 bp, *ycf1* (3378 bp–4798 bp), 136 bp + *trnT*~GGU~ + 801 bp, *rbcL* (335 bp–1134 bp), *ndhC-trnV*~AUC~, 1449 bp + *trnQ*~UG~ + 24 bp, and *petN*-psbM) were identified; six of which (*rbcL*, *trnQ*, *trnK*, *trnT*, *ndhC-trnV*, and *petN*-psbM) were located in the LSC region, while two (*psaC* and *ycf1*) were from the SSC region (Fig. 2, Additional file 1: Table S1). The Pi value of the eight hypervariable loci ranged from 0.00754 (807 bp + *trnR*~UCU~ + 384 bp) to 0.00955 (*petN*-psbM) (Table 4). At least four distinct gaps were observed in the chloroplast genome alignment, all located in the LSC region (Additional file 2: Fig. S1) within intergenic spacer regions, including *cemA*-ycf4, *petA*-psbI, *rpoB-trnC*, and *trnL-trnT*. Four species (*I. championii*, *I. fukiensis*, *I. hanceana*, and *I. lohauensis*) had a gap at the *rpoB-trnC* region, while three species (*I. polyneura*, *I. pubeascens*, and *I. rotundifolia*) had a gap at the *petA-psbI* region. Species that contained gaps at the *cemA*-ycf4 region also contained gaps at the *trnL-trnT* region, which
included *I. cinerea*, *I. cornuta*, *I. dabieshanensis*, *I. ficoidea*, *I. graciliflora*, *I. intermedia*, *I. latifolia*, *I. zhejiangensis*, and *Ilex* sp. However, two species, *I. delavayi*, and *I. integra* only had one of these gaps, which was at the *cemA-ycf4* region. Upon manual checking, these variations represented indels, ranging from about 210 bp (*petA-psbJ*) to 379 bp (*rpoB-trnC*) in length. Genome synteny of the 41 chloroplast genomes revealed no large gene rearrangement events (Additional file 2: Fig. S2).

Expansion and contraction of the IR regions

Comparative sequence analysis of the *Ilex* species showed that chloroplast genome structure and the number and sequence of genes were highly conserved. However, some structure and size variations at the IR boundaries were detected. The lengths of IRs among all *Ilex* species analyzed were relatively consistent: *I.
vomitoria had the shortest (26,005 bp), while I. rotunda had the longest (26,121 bp). About half (22/41) of the Ilex plastomes had LSC/IRA junctions located in rps19, with 4 to 5 bp crossing into the IRa region, which indicated an expansion of the IR in these species (Fig. 3). The majority of IRa/SSC junctions were located adjacent to ycf1 and ndhF, and overlap of 22 to 61 bp between ndhF and ycf1 was detected in 26 species. However, in I. dasyphylla, I. fukienensis, I. lohfaensis, I. venusta, I. viridis, I. yunnanensis, and I. zhejiangensis,
Table 2 List of annotated genes in the chloroplast genomes of the Ilex species

| Function of Genes                        | Group of Genes | Gene Name |
|-----------------------------------------|----------------|-----------|
| Protein synthesis and DNA-replication  | Transfer RNAs  | trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnM-CAU, trnG-GCC, trnG-UCG, trnH-GUG, trnK-UUU, trnL-UAA, trnM-CAU, trnQ-UG, trnP-UGG, trnP-UAG, trnR-UCA, trnS-GCU, trnS-GGA, trnT-GGU (x 2), trnT-UGU, trnV-UAC, trnW-CCA, trnY-GUA, trnA-UGC, petB |
| Ribosomal RNAs                          |                | ndhB, ndhA, rpoA, rpoB, rpoC1, rpoC2 |
| Ribosomal protein large subunit         |                | ndhF, ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhI, ndhJ, ndhK |
| Ribosomal protein small subunit         |                | psaA, psaB, psaC, psal, psaI |
| Subunits of RNA polymerase              |                | atpA, atpB, atpE, atpF, atpH, atpI |
| Photosynthesis                          | photosystem I  | rbcL |
|                                        | Photosystem II | psbA, psbB, psbC, psbD, psbE, psbF, psbG, psbH, psbI, psbJ, psbl, psbK, psbL, psbM, psbN, psbP, lhbA |
|                                        | ATP synthase   | atpA, atpB, atpE, atpF, atpH, atpI |
|                                        | Large subunit  | rbcL |
|                                        | Rubisco        | |
|                                        | Cytochrome b/f complex | petA, petB, petD, petG, petL, petN |
|                                        | NADH-dehydrogenase | ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhI, ndhJ, ndhK |
| Other genes                             | Translation initiation factor | infA |
|                                        | Cytochrome c biogenesis | ccsA |
|                                        | ATP-dependent protease | ctpP |
|                                        | Maturase       | matK |
|                                        | Inner membrane protein | cemA |
|                                        | Acetyl-CoA carboxylase | accD |
| Genes of unknown function              | Conserved hypothetical gene | orf42 (x 2), orf56 (x 2), orf188, ycf3B, ycf4, ycf1, ycf2 (x 2), ycf15 (x 2), ycf68 (x 2) |

Note: (x 2) indicates the number of repeat units is 2; *Gene contains a single intron; **Gene contains two introns

Table 3 Genes with introns in the chloroplast genome of Ilex species

| Gene | Location | Exon I (bp) | Intron I (bp) | Exon II (bp) | Intron II (bp) | Exon III (bp) |
|------|----------|-------------|---------------|--------------|----------------|---------------|
| rpl2 | Ira + Irb | 393         | 661           | 435          |                |               |
| rps12| LSC + IRs | 114         | 543           | 232          | 602            | 26            |
| clpP | LSC       | 69          | 819           | 291          | 408            | 78            |
| atpF | LSC       | 159         | 681           | 408          |                |               |
| rpoC1| LSC       | 456         | 756           | 1635         |                |               |
| ndhA | SSC       | 552         | 1140          | 540          |                |               |
| ndhB | IRA       | 777         | 679           | 756          |                |               |
| petB | LSC       | 6           | 745           | 657          |                |               |
| tmA-UGC | Ira + Irb | 38         | 807           | 35           |                |               |
| tmI-GAU | Ira + Irb | 42         | 934           | 35           |                |               |
| tmL-UAA | Ira + Irb | 37         | 490           | 50           |                |               |
| tmN-UAC | Ira + Irb | 39         | 579           | 37           |                |               |
| tmG-GCC | LSC       | 23         | 703           | 48           |                |               |
| tmI-UUU | LSC       | 37         | 2562          | 35           |                |               |
| ycf3 | LSC       | 126         | 727           | 228          | 749            | 153           |

Note: The number indicates the count of mononucleotide repeats.

ndhF and ycf1 were absent from the IRa/SSC junction. In all analyzed Ilex chloroplast genomes, the SSC/IRb junction was located in ycf1, with an extension into the IRb region ranging from 1047 bp (I. lohfuensis) to 1166 bp (I. dumosa) (Fig. 3).

SSR polymorphisms and long repeat sequence analysis
A total of 2146 simple sequence repeats (SSRs) were detected among the 41 Ilex chloroplast genomes, ranging from 10 to 168 bp (Fig. 4, Additional file 1: Table S2). Mononucleotide repeats were most abundant (1771),
while tetranucleotide repeats were rarest (49). The number of di-, trinucleotide, and compound repeats were 109, 79, and 138, respectively. Of the mononucleotide repeats, A/T repeats were most frequent (1769), while C/G repeats were only detected from two taxa (*I. asprella var. tapuensis* and *I. micrococca*). Dinucleotide repeats were represented by only the AT/TA motif; while tri- and tetranucleotides contained motifs AAT/ATT, CAG/CTG, and TTC/GAA, as well as AAAG/CTTT, ATAA/TTAT, ATTT/AAAT, TATT/AATA, and TCTT/AAGA repeats, respectively. Most SSRs were located in LSC regions (1649), followed by IR (275), and SSC (222) regions. We detected a total of 2843 large repeats between the 41 species (Fig. 5, Additional file 1: Table S3); *I. crenata* had the highest (79), while *I. latifolia* the fewest (62), large repeats. All species involved had forward, palindromic, and tandem repeats, but only 11 had complementary and/or reverse repeats.

**Phylogenomic analyses**

We reconstructed phylogenetic relationships from 52 complete chloroplast genomes and 75 protein-coding genes using both maximum likelihood (ML) and Bayesian inference (BI) methods, and used the closely related species *Helwingia himalaica* (NC031370) as an outgroup [26]. The total alignment lengths of the complete plastome and the protein-coding gene matrices were 157,836 bp and 68,601 bp, respectively. The complete plastome matrix contained 8869 variable and 1735 parsimony informative sites, while the protein-coding gene matrix contained 2247 and 458 variable and parsimony informative sites, respectively. The backbones of trees constructed using ML and BI methods were almost identical for each sequence matrix and supported the monophyly of *Ilex* (Fig. 6; ML BS: 100%; BI PP: 1.00); thus, we present only the ML tree here, with posterior probability (PP) values shown (Fig. 6, Additional file 2: Fig. S3).

### Table 4 Variable site analyses in the chloroplast genomes of *Ilex* species

| Region | Total number of sites | Polymorphic sites | Singleton variable sites | Parsimony informative sites | Nucleotide diversity |
|--------|-----------------------|-------------------|--------------------------|----------------------------|---------------------|
| LSC    | 88,362                | 2182              | 1200                     | 982                        | 0.00384             |
| IRa    | 26,162                | 94                | 57                       | 37                         | 0.00055             |
| SSC    | 18,460                | 582               | 319                      | 263                        | 0.00498             |
| IRb    | 26,167                | 89                | 54                       | 35                         | 0.00050             |
| Plastome | 159,151              | 2947              | 1630                     | 1317                       | 0.00286             |

**Fig. 2** Sliding-window analysis showing the nucleotide diversity (Pi) values of the aligned *Ilex* chloroplast genomes.
Based on our phylogenetic analyses, and with consideration of macro-morphological and distribution information, we recognize five highly supported clades within *Ilex* (clades A–E) that were well resolved (Fig. 6; ML BS: 100%; BI PP: 1.00). Clade A comprises one species (*I. micrococca*) of sect. *Micrococca*, two species (*I. asprella* and *I. chapaensis*) and one variety (*I. asprella var. tapuensis*) of sect. *Prinoides*, and seven species (*I. championii*, *I. fukienensis*, *I. hanceana*, *I. lohfauensis*, *I. memecylifolia*, *I. pubescens*, and *I. wilsonii*) of sect. *Pseudoaquifolium*. Clade B is sister to clade A, and includes three species (*I. polyneura*, *I. pubescens*, and *I. rotunda*). Clade C contains five species (*I. dasyphylla*, *I. kwangtungensis*, *I. lancilimba*, *I. purpurea*, and *I. suaveolens*) from sect. *Prinoides*, and seven species (*I. championii*, *I. fukienensis*, *I. hanceana*, *I. lohfauensis*, *I. memecylifolia*, *I. pubescens*, and *I. wilsonii*) of sect. *Pseudoaquifolium*. Clade B is sister to clade A, and includes three species (*I. polyneura*, *I. pubescens*, and *I. rotunda*). Clade C contains five species (*I. dasyphylla*, *I. kwangtungensis*, *I. lancilimba*, *I. purpurea*, and *I. suaveolens*) from sect. *Prinoides*, and seven species (*I. championii*, *I. fukienensis*, *I. hanceana*, *I. lohfauensis*, *I. memecylifolia*, *I. pubescens*, and *I. wilsonii*) of sect. *Pseudoaquifolium*. Clade D includes members from sect. *Aquifolium*, and is sister to Clade E, which only contains three species (*I. dumosa*, *I. paraguariensis*, and *I. vomitoria*). Only sect. *Aquifolium* was resolved as monophyletic, while the other five sections (*Lioprinus*, *Micrococca*, *Paltoria*, *Prinoides*, and *Pseudoaquifolium*) and six series (*Denticulatae*, *Hanceanae*, *Longecaudatae*, *Prinifoliae*, *Repandae*, and *Stigmatophorae*) were not. Interspecific relationships within each clade were generally well resolved with high support.

**Discussion**

**Comparison *Ilex* chloroplast genomes**

We found that *Ilex* possesses typical, quadripartite chloroplast genomes at sizes consistent with most land plants [23]. The 41 chloroplast genomes analyzed here had highly conserved structure, with minor variation between species. Expansion and contraction events at SC/IR boundaries often give rise to variation in chloroplast genome length [27], but *Ilex* plastomes varied by at most...
Fig. 4 Analysis of simple sequence repeats (SSR) in the 41 chloroplast genomes of *Ilex* species. A Number of different SSR types detected in the 41 genomes; B Number of different SSR types in LSC, SSC and IR regions.

Fig. 5 Analysis of long repeats in 41 chloroplast genomes of *Ilex* showing the number of complementary, forward, palindromic, reverse, and tandem long repeats.
Fig. 6 Phylogenetic trees inferred from maximum likelihood (ML) and Bayesian inference (BI) analyses based on the complete chloroplast genomes. Numbers near the nodes are ML bootstrap support values (BS, left of the slashes) and Bayesian posterior probabilities (PP, right of the slashes). 100% BS or 1.00 PP are indicated by asterisks. Incongruences between the BI and ML trees are indicated by dashes. Hu's classification is illustrated by color graphic pattern. Recognized groups (major clades) were also marked by the right-hand black bar.
901 bp in length. Although we detected small variations around IR junctions, the IR regions of the *Ilex* chloroplast genomes examined showed only modest expansions or contractions; IR regions varied from 25,080 to 26,121 bp, while LSC regions varied by about 900 bp (Table 1).

Variation in intergenic spacer regions, as well as gene loss and gain, also play important roles in shaping plant chloroplast genomes [23, 28]. In the seven newly sequenced chloroplast genomes, except for *I. dasyphylla*, all species lacked the gene *psbI*. Plastid gene loss has been previously documented in *Ilex*—specifically, deletions in the *trnT-trnL* and *ycf4-cemA* spacers of *I. graciliflora* [29]—which suggests that gene loss may be a relatively more common force influencing *Ilex* plastome architecture.

**Repetitive sequence analysis**

Chloroplast simple sequence repeats (SSRs) are commonly employed in population genetics and evolutionary studies because of their high rate of polymorphism and abundant variation at the species level [30]. We identified a total of 2146 SSR loci from the 41 *Ilex* chloroplast genomes. Few population genetic studies have used SSRs in *Ilex*, and these newly identified loci will facilitate future research into genomic diversity, structure, and phylogeography at the population, intraspecific, and cultivar levels in *Ilex*.

Long repeat sequences with lengths greater than 30 bp play important roles in creating insertion/deletion mismatches and rearrangements that lead to genomic variation [31–34]. We found that the number of long repeat sequences in *Ilex* is high compared to other angiosperm clades (e.g., 364 long repeats in Oxalidaceae [35]; 403 in *Veratrum* [36]; 32 in *Orestitrophe rupifraga*, and 34 in *Mukdenia rossiiand* [37]). Among these long repeats, forward, palindromic, and tandem repeats were rather common, accounting for 33.84, 30.81, and 34.44% of the total number of repeats, respectively, while complementary and reverse repeats were quite rare, only accounting for 0.42 and 0.49%, respectively.

**Hypervariable regions**

Hypervariable regions often provide a wealth of phylogenetic information and can be used to delimit closely related taxa [38, 39]. In general, IR regions are more highly conserved than SSC and LSC regions [40]. We identified eight hypervariable regions in *Ilex* plastomes, including four genes and four genes with flanking regions. Consistent with angiosperm-wide patterns of plastomes variability [32, 33], all hypervariable loci were distributed in the SC regions, while IR regions exhibited low variation.

To date, phylogenetic analyses of *Ilex* have been based on a handful of plastid markers (mainly *atpB-rbcL, psbA-trnH, rbcL*, and *trnL-trnF*), which could not resolve many interspecific relationships [1, 2, 10, 13, 15, 41–43]. When comparing these markers to the highly variable regions identified here, only one (*rbcL*) has been used to construct phylogenies. We believe that these eight highly variable regions will be useful for phylogenetic inference and DNA barcoding in *Ilex*. However, further studies are required to evaluate the strength of these regions for identifying and delimiting species.

**Phylogenetic inference**

There have been numerous attempts to resolve relationships amongst major *Ilex* lineages and test the consistency between molecular phylogenetics and traditional taxonomic systems based on morphology evidence [10–15, 26, 41]. A dearth of genetic data has resulted in poor resolution at the species level and weak support at most nodes in the *Ilex* phylogeny [10, 12–14, 26, 41]. These limitations can be addressed by using longer and more variable DNA sequences [44], such as complete chloroplast genomes [16, 21, 29, 45].

We present a well resolved and highly supported phylogeny of *Ilex*, and—in combination with macro-morphological and distribution information—suggest five clades (A–E) that are not generally congruent with traditional taxonomic systems. Clades A–E were largely consistent with previous plastid phylogenies, but relationships among clades differed significantly [10, 13, 15]. Our results showed that the American groups (Clade E) and the Eurasia groups (Clade F) were sister, and together formed the earliest diverging *Ilex* lineage, sister to a large clade containing the mostly Asian Clades A–C. In contrast, Manen [13] found the American (Group 3) and Eurasia (Group 4) groups to be among the most recently diverged lineages. The discordance between these results likely stems from the choice of loci included in analyses; previous studies have generally used less variable regions that led to low resolution among major clades [10, 13].

Our results highlight inconsistencies between molecular phylogenetics and traditional taxonomic systems. Almost all traditionally recognized subgenera, sections, and series included in our analysis were paraphyletic (all but sect. *Aquifolium*). Although the resolution of earlier phylogenetic trees was quite low, they indicated significant cyto-nuclear discordance, with nuclear phylogenies generally more consistent with traditional morphological classifications [13]. We confirmed the incongruences between plastid data and morphological systems by improving the resolution of the plastid phylogeny using complete chloroplast genomes.
Species found in close geographic proximity are often assumed to be closely related. This is accurate for most of the Ilex species in our study, including I. cornuta, I. dasyphylla, I. latifolia, and I. integra. However, both I. pubescens and I. lohfuensis were non-monophyletic in our analysis: the two accessions of I. pubescens were placed in two distinct clades (A and B), while the two accessions of I. lohfuensis were paraphyletic with respect to I. champignon. Three samples of I. viridis were placed with the morphologically similar species I. trifloral. Non-monophyletic species may result from chloroplast capture or hybridization events [13, 41, 43], or stem from misidentification. Further phylogenetic studies are needed to continue to clarify relationships and taxonomy in Ilex.

Conclusions
We conducted comparative and phylogenetic analyses of 41 Ilex chloroplast genomes, including seven newly sequenced taxa. To reach a more complete understanding of the evolutionary history of the clade, future studies should focus on phylogenetic reconstructions based on nuclear DNA. We suggest using low-copy nuclear genes from genome-skimming data, which can provide better resolution than traditional, short nuclear DNA markers (e.g., ITS). Incorporating nuclear phylogenies with existing phylogenies based on complete chloroplast genomes, as well as morphology, with enhance our understanding of the complex evolutionary history of Ilex.

Materials and methods
Taxon sampling, DNA extraction, and sequencing
Seven species of Ilex (I. dasyphylla, I. fukienensis, I. lohfuensis, I. venusta, I. viridis, I. yunnanensis, I. zhejiangensis, I. fukienensis, I. venusta, and I. zhejiangensis) were collected from their native ranges in China. Fresh leaf tissues were collected in the field and stored in silica gel prior to DNA extraction. Voucher specimens were prepared and deposited at the herbarium of Nanjing Forestry University (NF). In addition, 34 complete chloroplast genomes of Ilex species that are publicly available in NCBI GenBank were downloaded with annotations (Additional file 1: Table S4). Based on the classification in NCBI GenBank were downloaded with annotations and morphology, as well as morphology, with enhance our understanding of the complex evolutionary history of Ilex.

Chloroplast genome assembly and annotation
Raw reads were filtered with fastp v.0.20.0 software [46] to remove low-quality reads. The filtered data were then fed into the NOVOPlasty 2.6.3 [47] pipeline for genome assembly, with the rbcL gene sequence of I. latifolia (Accession number: KX897017) as the seed sequence and the chloroplast genome sequence of I. latifolia (Accession number: MN688228) as reference genome. A contig was obtained at the end of the process, and annotation was conducted using Plann [48], in which the annotated chloroplast genome of I. latifolia (Accession number: MN688228) was set as reference. Start and stop codons in the chloroplast genomes were manually corrected using DOGMA [49], and tRNA genes were verified with tRNA scan-SE v2.0.3 within in GeSeq [50] using default parameters. Circular chloroplast genome maps were visualized using OrganellarGenomeDRAW [51].

Comparative genomic analyses
Sequence alignment of the 41 complete chloroplast genomes was carried out using MAFFT v.7 [52] and the alignment was further trimmed using trimAl v1.2 using the “-gappyout” setting [53]. The expansions and contractions of IR regions were visualized using IRscope [54] online and then was manually checked. The nucleotide diversity (π) was estimated using DnaSP v.5 [55] with a step size of 200 bp and a window length of 800 bp. Tandem Repeat Finder [60] was used to analyze tandem repeat sequences with the default parameters. Compound SSRs were detected by identifying independent SSRs that were separated by less than 100 nucleotides and were combined into one.

Repeat sequence identification
The number of large repeats, including forward, palindromic, reverse, and complementary repeats were identified using onlineREPuter [59] according to the following criteria: sequence identities of 90%, cutoff point at ≥30 bp, Hamming distance set at 3, and a minimum repeat size of 30 bp. Tandem Repeat Finder [60] was used to analyze tandem repeat sequences with the default parameters. SSRs were identified using web-MISA [61], with minimum repeat number set at 10, 5, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta-, and hexanucleotides, respectively. Compound SSRs were detected by identifying independent SSRs that were separated by less than 100 nucleotides and were combined into one.

Phylogenetic analyses
Phylogenetic analyses were conducted using 52 complete chloroplast genomes and 75 protein-coding genes. A total of 39 Ilex species from six sections and 11 series
were included in the phylogenetic analyses. Based Yao et al. [26], *Helwingia himalaica* (Accession number: NC031370) was used as the outgroup. Genome alignment was carried out using MAFFT v.7 [52] and then trimmed using trimAl v1.2 with the “-gappyout” setting [53].

Maximum likelihood (ML) analyses were conducted using IQ-tree [62] with 10,000 ultrafast bootstrap (UFBS) replicates [63]. According to Bayesian information criterion (BIC), the best fitting substitution models that were estimated using ModelFinder [64] were GTR + F + I + G4 for the complete chloroplast genome sequences and GY + F + R3 for the protein-coding genes, respectively. Bayesian inference (BI) analysis was carried out using MrBayes version 3.2 [65], as implemented in CIPRES [66]. The Markov chain Monte Carlo analysis was executed for 2,000,000,000 generations, with four chains (one cold and three heated), each starting with a random tree, and sampled at every 1000 generations. Convergence of runs was accepted when the average standard deviation (d) of split frequencies was < 0.01. The first 25% of the trees were discarded as burn-in, and the remaining trees were used to construct majority-rule consensus trees. The final trees from both analyses were visualized using FigTree v.1.4.2 [67].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08397-9.

Acknowledgments

We thank Wanyi Zhao, Zhongcheng Liu of Sun Yat-sen University, Guangzhou, China, for their assistance in sample collection; Yubing Zhou (Jierui Biotech, Guangzhou, China) for data analysis of chloroplast genomes; and thank Dr. Ian Gilman at Yale University for his assistance with English language and grammatical editing.

Authors’ contributions

Conceptualization, K.X. and K.M; methodology, K.X. and K.M; formal analysis, K.X. and K.M; investigation, K.X., K.K. and J.K.; resources, K.X.; data curation, K.K. and K.M; writing—original draft preparation, K.X.; writing—review and editing, K.X., S.Y. Lee, K.M and K.M; visualization, K.X. and K.M; supervision, K.K., L.M. and K.M; project administration, K.X. and L.M.; funding acquisition, K.K. and L.M. All authors have read and agreed to the published version of the manuscript.

Funding

The project was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31000000), the Natural Science Foundation of Jiangsu Province (BK20210612), the National Natural Science Foundation of China (32101067 and 31970506), and the Nanjing Forestry University project funding (163108093).

Availability of data and materials

All data generated or analyzed in this study were included in this published article and the Additional files. The complete chloroplast genomes of the seven newly sequenced *Ilex* species were submitted to GenBank and the accession numbers can be found in Additional file 1. Table S4. All raw reads are available in the short sequence archive under accession no. PRJNA678933. All complete genome sequences used in this study were downloaded from NCBI (https://www.ncbi.nlm.nih.gov), and the accession numbers can be found in Additional file 1. Table S4.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflict of interest.

Author details

1Co-Innovation Center for Sustainable Forestry in Southern China, College of Biology and the Environment, Nanjing Forestry University, Nanjing 210075, China. 2Faculty of Health and Life Sciences, INTI International University, 71800 Nilai, Malaysia. 3State Key Laboratory of Biocontrol and Guangdong Provincial Key Laboratory of Plant Resources, School of Life Sciences, Sun Yat-sen University, Guangzhou, China.

Received: 23 September 2021 Accepted: 17 February 2022

Published online: 14 March 2022

References

1. Loizeau PA, Barriera G, Manen JF, Broeninnmann O. Towards an understanding of the distribution of *Ilex* L. (Aquifoliaceae) on a world-wide scale. Biol Skr. 2005;55:501–20.
2. Powell M, Savolainen V, Cuenoud P, Manen JF, Andrews S. The mountain holly (*Nemopanthus mucronatus* Aquifoliaceae) revisited with molecular data. Kew Bull. 2000;55:341–7. https://doi.org/10.2307/4115646.
3. Loizeau PA, Savolainen V, Andrews S, Spichiger R. Aquifoliaceae. In: Kubitzki K, editor. Flowering plants. Eudicots, the families and genera of vascular plants. Berlin: Springer; 2016. p. 31–6.
4. Filip R, Lópezb P, Giberti G, Cossio J, Ferraro G. Phenolic compounds in seven south American *Ilex* species. Fitoterapia. 2001;72(7):774–8.
5. Tang ZX, Zhou Y, Zeng YK, Zang SL, He PG, Fang YZ. Determination of active ingredients of *Ilex* purpurea Hassk and its medicinal preparations by capillary electrophoresis with electrochemical detection. J Pharm Biomed Anal. 2006;39:2861–75.
6. Yi F, Zhao XL, Peng Y, Xiao PG. Genus *Ilex*: phytochemistry, ethnopharmacology, and pharmacology. Chin Herb Med. 2016;8:209–30.
7. Yao X, Zhang F, Corlett RT. Utilization of the hollies (*Ilex* L. spp.): A Review. Forests. 2022;13(1):94.
8. Yao X, Lu Z, Song Y, Hu XD, Corlett RT. A chromosome-scale genome assembly for the holly (*Ilex polyantha*) provides insights into genomic adaptations to elevation in Southwest China. Hortic Res. 2022;9:uhab049.
9. International Union for Conservation of nature and natural resources (IUCN). The IUCN red list of threatened species. 2021. https://www.iucnredlist.org/. Accessed 11 Aug 2021.

10. Cuello P, del Pero Martínez MA, Loizou PA, Spichiger H, Andrews S, Manen JF. Molecular phylogeny and biogeography of the genus Ilex L. (Aquifoliaceae) Ann Bot (Oxford). 2000;85:111–22.

11. Setoguchi H, Watanabe I. Intersectional gene flow between insular endemics of Ilex (Aquifoliaceae) on the Bonin Islands and the Ryukyu Islands. Amer J Bot. 2000;87:793–810.

12. Manen JF, Bouller MC, Naciri-Graven Y. The complex history of the genus Ilex (Aquifoliaceae): evidence from the comparison of plastid and nuclear DNA sequences and from fossil data. Pl Syst Evol. 2002;235:79–98.

13. Manen JF, Barretta G, Loizou PA, Naciri Y. The history of extant Ilex species (Aquifoliaceae): evidence of hybridization within a Miocene radiation. Molec Phylogen Evol. 2010;57:961–77. https://doi.org/10.1016/j.ympev.2010.09.006.

14. Gottlieb AM, Giberti GC, Poggiolo L. Molecular analyses of the genus Ilex (Aquifoliaceae) in southern South America, evidence from AFLP and ITS sequence data. Amer J Bot. 2005;92:352–69.

15. Jhang L, Xu K, Fan Q, Peng H. A new species of Ilex (Aquifoliaceae) from Jiangxi Province, China, based on morphological and molecular data. Phytotaxa. 2017;298:147–57.

16. Yao X, Tan HH, Liu YY, Song Y, Yang J, Corlett RT. Chloroplast genome structure in Ilex (Aquifoliaceae). Sci Rep. 2016;6:1–10.

17. Yao X, Liu YY, Tan YH, Song Y, Corlett RT. The complete chloroplast genome sequence of Helwingia himalaica (Helwingiaceae, Aquifoliidae) and a chloroplast phylogenetic analysis of the Campanulaceae. PeerJ. 2016;4:e2734.

18. Hu S. The evolution and distribution of the species of Aquifoliaceae in the Pacific area (1). Jap J Bot. 1967;42:13–27.

19. Loesener T. Monographia aquifoliacearum. Part I. Nova Acta Acad Caes Leop-Carol German Nut Cur. 1901;78:1–589.

20. Zhang SD, Jin JJ, Chen SY, Chase MW, Soltis DE, Li HT, et al. Diversification analysis and phylogenetic investigation of Hong Kong Ilex chloroplast genomes. Sci Rep. 2021;11(1):1–13.

21. Zhang S, Jin J, Chen S, Wang H, Zhang E, Wang Z, et al. The complete chloroplast genome sequence of Ilex edgeworthii (Aquifoliaceae) from the Nanling Mountains, China. Sci Rep. 2015;5:9808–13. https://doi.org/10.1038/srep09808.

22. Meng KK, Chen SF, Xu KW, Zhou RC, Li MW, Dhamala MK, et al. Phylogeny, biogeography, and applications to a wide array of plant species. Mol Ecol Resour. 2009;9:673–90.

23. Weng M, Blazier JC, Govindu M, Jansen RK. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Mol Ecol Resour. 2009;9:673–90.

24. Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. Phylogenetic analysis and phylogenetic investigation of Hong Kong Ilex chloroplast genomes. Sci Rep. 2015;5:9808–13. https://doi.org/10.1038/srep09808.

25. Chen SK, Ma HY, Feng YX. Aquifoliaceae. In: Wu ZY, Raven PH, Hong DY, editors. Flora of China. Beijing and St. Louis: Science Press and Missouri Botanical Garden Press. 2008. p. 449–61.

26. Yao X, Song Y, Yang J, Tan YH, Corlett RT. Phylogeny and biogeography of the hollies (Ilex L. Aquifoliaceae). J Syst Evol. 2021;59(1):73–82.

27. Wiecke JL. Structural diversity among plastid genomes of Ilex (Aquifoliaceae). Plant Syst Evol. 2016;302(1):1–10.

28. Wolfe KH, Morden CW, Palmer JD. Function and evolution of a minimal plastid gene family. Science. 1990;247:580–585.

29. Wang M, Blazier JC, Govindu M, Jansen RK. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Mol Ecol Resour. 2009;9:673–90.

30. Weng M, Blazier JC, Govindu M, Jansen RK. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Mol Ecol Resour. 2009;9:673–90.
54. Amiryousefi A, Hyvönen J, Poczai P. IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics. 2018;34(17):3030–1.
55. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25(11):1451–2.
56. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004;32(Suppl 2):273–9.
57. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14:1394–403.
58. Kears M, Moir R, Wilson A, Stonex-Havas S, Cheung M, Sturrock S, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647–9.
59. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. RP-puter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29:4633–42.
60. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.
61. Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor Appl Genet. 2003;106:411–22.
62. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
63. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018;35:518–22. https://doi.org/10.1093/molbev/msx281.
64. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587–9. https://doi.org/10.1038/nmeth.4285.
65. Ronquist F, Teslenko M, Van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61:539–42.
66. Miller MA, Pfeiffer W, Schwartz T. Proceedings of the Gateway Computing Environments Workshop (GCE), 14 November 2010, New Orleans, LA: Creating the CIPRES science gateway for inference of large phylogenetic trees. 2010. p. 1–8.
67. Rambaut A. FigTree V1.4.2. 2012. http://tree.bio.ed.ac.uk/software/figtree/.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.