Comparative Chloroplast Genome Analyses of Ilex (Aquifoliaceae): Insights Into Evolutionary Dynamics and Phylogenetic Relationships

Kewang Xu  
Nanjing Forestry University
Chenxue Lin  
Nanjing Forestry University
Shiou Yih Lee  
INTI International University
Lingfeng Mao (maolingfeng2008@163.com)  
Nanjing Forestry University
Kaikai Meng  
Sun Yat-sen University

Research Article

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

Posted Date: October 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-930025/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background

Despite many species of *Ilex* (Aquifoliaceae) are of horticultural importance and are widely grown in parks and gardens throughout the world for their foliage and decorative berries, limited genetic information has greatly hampered our understanding of the chloroplast genome evolution and phylogenetic relationships within the genus. This study attempted to address these problems by comparing the chloroplast genomes and analyzing phylogenetic relationships within the genus.

Results

In this study, analyses of chloroplast genome structure, codon usage, GC content, gene rearrangement, nucleotide diversity, inverted repeats (IR) boundary, repeat sequence, and SSR component were conducted by comparing 41 chloroplast genomes of *Ilex*. The results showed that these *Ilex* chloroplast genomes were evolutionary conserved at the genome level and no rearrangement of the complete cp genome in the 41 *Ilex* genomes was recorded. On the contrary, there were still a few mutational hotspots identified from these chloroplast genomes, which were considered as hypervariable regions useful for future phylogenetic studies and DNA barcoding. Using the complete chloroplast genome sequences, we reconstructed a highly supported phylogeny of *Ilex* and well-resolved the complicated relationships among the different lineages within *Ilex*.

Conclusion

The present study increased our understanding of the chloroplast genome evolution and phylogenetic relationships within *Ilex*. The availability of these genetic resources will be helpful for the future studies in DNA barcoding, species delimitation, phylogenetic reconstruction as well as the hybridization and introgression events between distantly related lineages within *Ilex*.

Introduction

*Ilex* L., comprises 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, but also extend into some temperate regions, making its center of species diversity to be located at the tropical America and the south-east Asia regions [2, 3]. Most species of *Ilex* such as, *I. comuta* Lindl. et Paxt., *I. purpurea* Hassk., *I. paraguariensis* A. St.-Hil., and *I. rotunda* Thunb., contain important economic values [4–6], while many of them contain high endemism and are limited to certain geographic distributions. To date, as much as 250 species of *Ilex* have been classified as endangered species based on the International Union for Conservation of Nature (IUCN) red list [7].

In the past two decades, the advances in sequencing technology and analytical methods have contributed to the phylogenetic reconstruction of the genus *Ilex* using several plastid DNA fragments, including *rbcL*, *trnL-trnF*, and *atpB-rbcL* as well as the nuclear ribosomal DNA internal transcribed spacers (nrITS) and chloroplast glutamine synthetase gene (*nepGS*) sequences [8–14]. Although a large taxon sampling of *Ilex* has been achieved, the phylogeny of *Ilex* is still not completely resolved [11, 14]. Substantial incongruence between the nuclear phylogeny and the plastid phylogeny of the genus *Ilex* has been found in recent phylogeny studies, but only a few plastid or nuclear gene fragments were used in these studies and the resolution was generally poor due to their relatively conserved features in some plastid genes [8, 11–13]. Furthermore, the findings obtained from molecular phylogenetic analysis did not complement to those previously proposed traditional classification of *Ilex* based on morphological features [15, 16]. At present, the phylogenetic relationships among lineages in genus *Ilex* remain uncertain, thus, further investigations are requisite to decipher the status and evolution pattern of this genus.

In addition to the commonly used short plastid gene sequences that have been useful in analyzing the phylogenetic relationship in most land plants, the complete chloroplast genome sequences have been tested to be more promising in resolving the relationships of many genera at different taxonomic levels [17–19]. In general, the chloroplast genome is relatively stable, and usually contains four extremely evolutionarily conserved parts: a pair of inverted repeat regions (IRa and IRb), a large single-copy region (LSC), and a small single-copy region (SSC) in most land plants [20]. At the same time, it contains a large amount of effective information with a suitable mutation rate that are sufficient for phylogeny reconstruction and species delimitation [21].
To date, the complete chloroplast genome sequences of a total of 34 *Ilex* species have been made available so far in the NCBI GenBank database (accessed on 1 August 2021). Yet, this number only accounts for ca. 5.7% of the total number of species recorded in the genus. Despite the limited number of the complete chloroplast genome reported, phylogenetic analysis of *Ilex* based on the complete chloroplast genome sequence could still provide some insights to the relationship between species of the genus. In the effort to expand the current genetic resource of *Ilex* at chloroplast genome-scale, herein, we newly sequenced the chloroplast genomes of seven species of *Ilex*, including *I. dasyphylla*, *I. fukienensis*, *I. lohfuensis*, *I. venusta*, *I. viridis*, *I. yunnanensis*, and *I. zhejiangensis*. Three of them, *Ilex fukienensis*, *I. venusta*, and *I. zhejiangensis*, were known to have a very narrow distribution in China [13, 22], while the other four species are widely distributed in China and adjacent regions. We aimed: (i) to investigate the structural and compositional variations of the chloroplast genomes in genus *Ilex* by involving more lineages; (ii) to identify highly variable regions in the chloroplast genomes that could be used as potential markers in resolving the interspecific relationship and species delimitation of *Ilex*; and (iii) to reconstruct a high resolution phylogenetic tree based on the complete chloroplast genome sequences that is further tested for its cyto-nuclear discordance by comparing to existing nuclear-based phylogenetic tree of the genus *Ilex.*

**Results**

**Chloroplast genome structure of *Ilex* species**

All the seven newly sequenced chloroplast genomes obtained through this study possess a typical quadripartite structure of a vascular plant, which consisted of two single-copy regions (LSC and SSC) that are separated by a pair of inverted repeats (IRa and IRb) (Figure 1). By taking account of the 34 chloroplast genomes downloaded from the GenBank, the length of the chloroplast genome of *Ilex* ranged from 157,119 bp (*I. graciliora*) to 158,020 bp (*I. kwangtungensis*). The length of the LSC region was ranged from 86,506 bp (*I. graciliora*) to 87,414 bp (*I. crenata*), the SSC regions was ranged from 18,228 bp (*I. yunnanensis*) to 18,447 bp (*I. lohfuensis*), and the IR regions ranged from 26,005 bp (*I. vomitoria*) to 26,121 bp (*I. rotunda*) (Table 1). The GC content was from 37.60–37.69% (Table 1). All the chloroplast genomes obtained in this study contained 116 genes, including 82 protein-coding, 31 tRNA, and four rRNA genes, except for *I. dasyphylla*, which had 83 protein-coding genes (Table 2 and 3). All chloroplast genomes had the same gene order and gene arrangement.
| Taxon                     | Accession number | Gene number | Length (bp) | GC(%) | AT(%) |
|--------------------------|------------------|-------------|-------------|-------|-------|
|                           |                  | PCG       | tRNA | rRNA | Full | Plastome | LSC | IRA/IRB | SSC |
| *I. asprella*             | NC_045274        | 94        | 37   | 8    | 139  | 157,856  | 87,265 | 26,075   | 18,441 | 37.62  | 62.38  |
| *I. asprella* var.       | MT767004         | 87        | 37   | 8    | 132  | 157,671  | 87,161 | 26,065   | 18,380 | 37.65  | 62.35  |
| *tapuensis*              |                  |           |      |      |      |           |       |          |      |       |       |
| *I. championii*          | MT764248         | 87        | 37   | 8    | 132  | 157,468  | 86,878 | 26,074   | 18,442 | 37.64  | 62.36  |
| *I. chapaensis*          | MT764251         | 87        | 37   | 8    | 132  | 157,665  | 87,155 | 26,065   | 18,380 | 37.65  | 62.35  |
| *I. cinerea*             | MT764247         | 87        | 37   | 8    | 132  | 157,215  | 86,601 | 26,094   | 18,426 | 37.69  | 62.31  |
| *I. coruta*              | MT764252         | 87        | 37   | 8    | 132  | 157,216  | 86,607 | 26,091   | 18,427 | 37.69  | 62.31  |
| *I. crenata*             | MW528027         | 88        | 38   | 8    | 134  | 157,988  | 87,414 | 26,076   | 18,422 | 37.65  | 62.35  |
| *I. dabieshanensis*      | MT435529         | 90        | 37   | 8    | 135  | 157,218  | 86,723 | 26,034   | 18,427 | 37.69  | 62.31  |
| *I. dasyphylla*          | This study       | 92        | 40   | 8    | 140  | 158,009  | 87,388 | 26,105   | 18,411 | 37.62  | 62.38  |
| *I. dariophylla*         | MT764253         | 87        | 37   | 8    | 132  | 158,009  | 87,388 | 26,105   | 18,411 | 37.62  | 62.38  |
| *I. delavayi*            | KX426470         | 95        | 40   | 8    | 143  | 157,671  | 87,077 | 26,078   | 18,438 | 37.65  | 62.35  |
| *I. dumosa*              | KP016927         | 86        | 37   | 8    | 131  | 157,732  | 87,033 | 26,087   | 18,415 | 37.62  | 62.27  |
| *I. ficoidea*            | MT764243         | 87        | 37   | 8    | 132  | 157,536  | 86,922 | 26,094   | 18,426 | 37.64  | 62.36  |
| *I. fukienensis*         | This study       | 91        | 40   | 8    | 139  | 157,474  | 86,886 | 26,074   | 18,440 | 37.64  | 62.36  |
| *I. graciliflora*        | MT764249         | 87        | 37   | 8    | 132  | 157,119  | 86,506 | 26,093   | 18,427 | 37.61  | 62.39  |
| *I. hanceana*            | MT764246         | 87        | 37   | 8    | 132  | 157,478  | 86,889 | 26,074   | 18,441 | 37.63  | 62.37  |
| *I. integrata*           | NC_044417        | 86        | 37   | 8    | 131  | 157,548  | 86,936 | 26,093   | 18,426 | 37.64  | 62.36  |
| *I. intermedia*          | MT471320         | 89        | 37   | 8    | 134  | 157,577  | 87,083 | 26,034   | 18,426 | 37.63  | 62.37  |
| *I. kwangtungensis*      | MT764241         | 87        | 37   | 8    | 132  | 158,020  | 87,400 | 26,104   | 18,412 | 37.62  | 62.38  |
| *I. lanciliamba*         | MT767005         | 87        | 37   | 8    | 132  | 157,998  | 87,382 | 26,105   | 18,406 | 37.62  | 62.38  |
| *I. latifolia*           | NC_047291        | 95        | 40   | 8    | 143  | 157,601  | 87,020 | 26,077   | 18,427 | 37.63  | 62.37  |
| *I. lohfuensis*          | This study       | 91        | 40   | 8    | 139  | 157,464  | 86,875 | 26,071   | 18,447 | 37.64  | 62.36  |
| *I. lohfuensis*          | MT764240         | 87        | 37   | 8    | 132  | 157,470  | 86,873 | 26,078   | 18,441 | 37.64  | 62.36  |
| *I. memecyllifolia*      | MT764250         | 87        | 37   | 8    | 132  | 157,842  | 87,249 | 26,076   | 18,441 | 37.62  | 62.38  |
| *I. microcoa*            | MN830251         | 89        | 37   | 8    | 134  | 157,782  | 87,200 | 26,074   | 18,434 | 37.64  | 62.36  |
| *I. paraguariensis*      | NC_031207        | 86        | 37   | 8    | 131  | 157,614  | 87,154 | 26,076   | 18,308 | 37.63  | 62.35  |
| *I. polynoeura*          | KX426468         | 95        | 40   | 8    | 143  | 157,621  | 87,140 | 26,061,26,980 | 18,434 | 37.60  | 62.40  |
| *I. pubescens*           | KX426467         | 95        | 40   | 8    | 143  | 157,741  | 87,143 | 26,081   | 18,436 | 37.61  | 62.39  |
| *I. purpurea*            | MT471318         | 89        | 37   | 8    | 134  | 157,885  | 87,289 | 26,104   | 18,388 | 37.62  | 62.38  |
| *I. rotunda*             | MW292559         | 88        | 37   | 8    | 133  | 157,743  | 87,069 | 26,121   | 18,432 | 37.62  | 62.38  |
| Taxon            | Accession number | Gene number | Length (bp) | GC(%) | AT(%) |
|------------------|------------------|-------------|-------------|-------|-------|
|                  |                  | PCG tRNA    | rRNA Full   | Plastome | LSC | IRA/IRB | SSC |       |       |
| I. sp.           | KX426469         | 95 40 8     | 143         | 157,611 | 87,137| 26,020 | 18,434| 37.62 | 62.38 |
| I. suaveolens    | MN830249         | 89 37 8     | 134         | 157,857 | 87,255| 26,102 | 18,398| 37.65 | 62.35 |
| I. szechwanensis | KX426466         | 95 40 8     | 143         | 157,822 | 87,281| 26,053 | 18,435| 37.65 | 62.35 |
| I. triflora      | MT764242         | 87 36 8     | 131         | 157,706 | 87,183| 26,065 | 18,393| 37.67 | 62.33 |
| I. venusta       | This study       | 91 40 8     | 139         | 157,860 | 87,290| 26,079 | 18,412| 37.66 | 62.34 |
| I. viridis       | This study       | 91 40 8     | 139         | 157,673 | 87,150| 26,065 | 18,393| 37.68 | 62.32 |
| I. viridis       | MN830250         | 89 37 8     | 134         | 157,701 | 87,177| 26,065 | 18,394| 37.67 | 62.33 |
| I. vomitoria     | MT471319         | 90 36 8     | 134         | 157,328 | 86,920| 26,005 | 18,398| 37.66 | 62.34 |
| I. wilsonii      | KX426471         | 95 40 8     | 143         | 157,918 | 87,341| 26,073 | 18,431| 37.62 | 62.38 |
| I. yunnanensis   | This study       | 91 40 8     | 139         | 157,833 | 87,389| 26,108 | 18,228| 37.65 | 62.35 |
| I. zhejiangensis | This study       | 91 40 8     | 139         | 157,182 | 86,575| 26,092 | 18,423| 37.69 | 62.31 |
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, trnFM-CAU, trnG-GCC*, trnG-UCC, trnH-GUG, trnK-UUU*, trnL-UAA*, trnM-CAU, trnQ-UUG, trnP-GG, trnP-UGG, trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU (× 2), trnT-UGU, trnV-UAC*, trnW-CCA, trnY-GUA, trnA-UGC* (× 2), trnL-CAU (× 2), trnL-CAA (× 2), trnL-UAG, trnN-GUU (× 2), trnR-ACG (× 2), trnV-GAC (× 2), trnM-CAU |
| Ribosomal RNAs | rm4.5 (× 2), rm5 (× 2), rm16 (× 2), rm23 (× 2) |
| Ribosomal protein large subunit | rpl33, rpl20, rpl36, rpl14, rpl16, rpl22, rpl32, rpl2* (× 2), rpl23 (× 2) |
| Ribosomal protein small subunit | rps2, rps14, rps4, rps18, rps11, rps8, rps3, rps19, rps12* (× 2), rps7 (× 2) |
| Subunits of RNA polymerase | rpoA, rpoB, rpoC1*, rpoC2 |
| **Photosynthesis** | photosystem I | psaA, psaB, psaC, psal, psaJ |
| Photosystem II | psbA, psbB, psbC, psbD, psbE, psbF, psbG, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, lhbA |
| ATP synthase | atpA, atpB, atpE, atpF*, atpH, atpI |
| Large subunit Rubisco | rbcL |
| Cytochrome b/f complex | petA, petB*, petD, petG, petL, petN |
| NADH-dehydrogenase | ndhA*, ndhB* (× 2), ndhC, ndhD, ndhE, ndhG, ndhH, ndhI, ndhJ, ndhK |
| **Other genes** | Translation initiation factor | infA |
| Cytochrome c biogenesis | ccsA |
| ATP-dependent protease | clpP** |
| Maturase | matK |
| Inner membrane protein | cemA |
| Acetyl-CoA carboxylase | accD |
| Genes of unknown function | Conserved hypothetical gene | orf42 (× 2), orf56 (× 2), orf188, ycf3**, ycf4, ycf1, ycf2 (× 2), ycf15 (× 2), ycf68 (× 2) |

Note: (× 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          | 26             |               |
| clpP  | LSC       | 69          | 819           | 291          | 602            | 78            |
| atpF  | LSC       | 159         | 681           | 408          |                |               |
| rpoC1 | LSC       | 456         | 756           | 1,635        |                |               |
| ndhA  | SSC       | 552         | 1,140         | 540          |                |               |
| ndhB  | IRA       | 777         | 679           | 756          |                |               |
| petB  | LSC       | 6           | 745           | 657          | 657            |               |
| tmA-UGC | IRa+IRb  | 38          | 807           | 35           |                |               |
| tml-GAU | IRa+IRb  | 42          | 934           | 35           |                |               |
| tml-UAA | IRa+IRb  | 37          | 490           | 50           |                |               |
| tmV-UAC | IRa+IRb  | 39          | 579           | 37           |                |               |
| tmG-GCC | LSC      | 23          | 703           | 48           |                |               |
| tmK-UUU | LSC      | 37          | 2,562         | 35           |                |               |
| ycf3  | LSC       | 126         | 727           | 228          | 749            | 153           |

**Comparative analysis of genomic divergence and genome rearrangement**

The diversity of nucleotide variability (Pi) for the 41 chloroplast genomes was between 0.000 and 0.01266, while the average Pi value was 0.00286. 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 + *tmr*GGU + 801 bp, *rbcl* (335 bp - 1134 bp), *ndhC-trnV*JAC, 1449 bp + *tronQ*JUG + 24 bp and *petN-psbM*) were identified, in which six of them (*rbcl, trnQ, trnR, trnT, ndhC-trnV*, and *petN-psbM*) were located at the LSC region, while two of them (*psaC* and *ycf1*) were from the SSC region (Figure 2, Additional files: 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, in which all of them were located in the LSC region (Figure 3). The gaps were located at the intergenic spacer regions, including *cemA-ycf4, petA-psbJ, rpoB-tmC*, and *tmL-tmT*. Species that have gap at the *rpoB-tmC* region were *I. championii, I. fukienensis, I. hanceana*, and *I. lohauensis*; while three species, *I. polyneura, I. pubescens*, and *I. rotunda*, had gap at the *petA-psbJ* region. Species that contained gaps at the *cemA-ycf4* region also contained gaps at the *tmL-tmT* region, which include *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 the gap located at the *cemA-ycf4* region. Upon manual checking, these variations were in forms of indels, ranging from about 210 bp (*petA-psbJ*) to 379 bp (*rpoB-tmC*) in length. Genome synteny of the 41 chloroplast genomes revealed no presence of large gene rearrangement event (Figure 4). In addition, a total of 2,947 polymorphic, 1,630 singleton variable, and 1,317 parsimony-informative sites were detected in the 41 chloroplast genome sequences.
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    | 26162                 | 94                | 57                       | 37                          | 0.00055             |
| SSC    | 18460                 | 582               | 319                      | 263                         | 0.00498             |
| IRb    | 26167                 | 89                | 54                       | 35                          | 0.00050             |
| Plastome | 159151               | 2947              | 1630                     | 1317                        | 0.00286             |

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 of those chloroplast genomes in *Ilex* were detected. The length of IRs among all *Ilex* species included for analysis was relatively consistent, in which the IRs of *I. vomitoria* were the shortest (26,005 bp), while those of *I. rotunda* were the longest (26,121 bp). For a total of 41 genomes analyzed in this study, the LSC/IRa junction of 22 species were located in the *rps19* gene, with 4 to 5 bp spanning into the IRa region, indicating an expansion event of the IR in these species (Figure 5). The IRa/SSC junctions of majority species were located adjacent to genes *ycf1* and *ndhF*. Overlap with 22 to 61 bp between the *ndhF* and *ycf1* genes was detected in 26 *Ilex* species. However, in *I. dasyphylla*, *I. fukienensis*, *I. lohfauensis*, *I. venusta*, *I. viridis*, *I. yunnanensis*, and *I. zhejiangensis*, *ndhF* and *ycf1* genes were absent from the IRa/SSC junction. In all *Ilex* chloroplast genomes, the SSC/IRb junction is located in the *ycf1* gene, and the extension in sequence length into the IRb region ranged from 1,047 bp (*I. lohfauensis*) to 1,166 bp (*I. dumosa*) (Figure 5).

SSR Polymorphisms and Long Repeat Sequence Analysis

A total of 2,146 simple sequence repeat (SSR) loci were detected among the 41 *Ilex* chloroplast genomes, with a length variation from 10 to 168 bp (Figure 6, Additional files 2: Table S2). The mononucleotide repeats were most abundant (1,771), while tetranucleotide repeats had the least occurrence (49). In contrast, the number of di-, trinucleotide and compound repeats were 109, 79, and 138, respectively. For mononucleotide repeats, the A/T repeats occurred the most (1769), while C/G repeats were only detected from two taxa, including *I. asprella* var. *tapuensis* and *I. micrococca*. For dinucleotide repeats were represented with only the AT/TA motif; while trinucleotides and tetranucleotides contained motifs AAT/ATT, CAG/CTG, and TTC/GAA, as well as AAAG/CTTT, ATAA/TTAT, ATT/AAAT, and TCTT/AAGA repeats, respectively. Most of the SSRs were primarily located at the LSC region (1649), followed by the IR regions (275), and the SSC region (222).

A total of 2843 large repeats were detected (Figure 7, Additional files 3: Table S3), in which *I. crenata* contained the highest number of large repeats (79), while *I. latifolia* had the least (62). All species involved in the long repeat sequence analysis had forward, palindromic, and tandem repeats. While 11 chloroplast genomes contained all four types of large repeats, 30 of them did not come with any complementary and/or reverse repeats.

Codon usage bias

From the total number of 23,092 (*I. asprella* and *I. dumosa*) to 23,137 (*I. dabieshanensis*) encoded codons analyzed in this study, the codon encoded amino acid that recorded the highest frequency was leucine (vary from 10.51–10.57% in different species), while the stop codon had the least occurrence (77 codons, 0.33% in each species) (Additional files 4: Table S4). The alanine-encoded GCU had the highest relative synonymous codon usage (RSCU) value (1.83–1.86), while both the serine-encoded AGC and alanine-encoded GCG were recorded with the lowest RSCU value (0.34–0.36) (Figure 8). Codon bias was not detected in methionine-encoded AUG or tryptophan-encoded UGG and was detected in other amino acid encoded codons (Figure 8, Additional files 4: Table S4). For stop codon, the amino acid UAA was like preferred when compared to UAG. Overall, codons with A/U-ends were more preferred than those with G/C-ends among the amino acids in *Ilex* species (Figure 8, Additional files 4: Table S4).
Phylogenomic analyses

The phylogenetic trees of both the maximum likelihood (ML) and Bayesian inference (BI) were reconstructed based on the 52 complete chloroplast genomes. The closely related species Helwingia himalaica (NC031370) was selected as outgroup [23]. The total alignment length of the dataset used was 157,836 bp after trimmed. A total of 8,869 variable sites and 1,735 parsimony informative sites were detected in the alignment. The phylogenetic trees based on the ML and BI displayed similar topologies; thus, the posterior probability (PP) values were merged into the ML tree for clearer presentation (Figure 9). The monophyly of Ilex was well supported by the two analyses (Fig. 8; ML BS: 100%; BI PP: 1.00).

Based on our phylogenetic analysis and with consideration of macro-morphological and distribution information, the genus Ilex was categorized into five highly supported clades (Figure 9; ML BS: 100%; BI PP: 1.00). The relationships among the five major clades were well resolved with high resolution (ML BS: 100%; BI PP: 1.00) in the present phylogenetic study. Members of clade D and clade E were sister to the rest, followed by the clade C, which is sister to a clade containing clade A and clade B. The relationships within each clade were also well resolved with high support values.

Discussion

Comparison of the chloroplast genome in genus Ilex

The chloroplast genomes of Ilex were found to possess the typical quadripartite structure with similar genome size of most land plant genomes [20]. At genus level, all 41 chloroplast genomes showed high conservatism in term of genome structure, but still exhibited moderate variations among different species. The genome length of the chloroplast genes of Ilex was similar among different species with a maximum 901 bp difference in length. Expansion and contraction events at the SC/IR boundaries usually give rise to the change of the chloroplast genome length [24]. Although small variations around the IR junctions were detected in the present study, the IR regions of the chloroplast genomes of these Ilex species showed a modest expansion or contraction and the length of IR regions varied from the longest 26,121 bp to the shortest 25,080 bp.

In addition, variation in gene spacer region, the loss or gain of genes and introns also play an important role in determining the chloroplast genome size of most plants [20, 25]. In the seven newly sequenced chloroplast genomes, except for I. dasyphylla, all species lacked the gene psbl. The event of gene loss in genus Ilex (e.g., deletions in the tmT-trnL and ycf4-cemA spacers of I. graciliflora) has been identified in an earlier report [26], which suggested that the gene loss in Ilex could be a common phenomenon. The length of the SSC and the IR regions were very similar, while that of LSC region showed a difference of about 900 bp (Table 1).

SSRs and simple sequence repeat analyses

SSRs derived from the plastid region were commonly known to be the markers of choice by population genetics and evolutionary studies owing to their high polymorphism rate and abundant variation at the species level [27]. In the present study, a total of 2146 SSR loci were identified from the 41 Ilex chloroplast genomes (Figure 6, Additional files 2: Table S2). Few studies have involved in the population genetics using data of SSRs in genus Ilex, and these SSR loci will contribute to further studies of genetic diversity and polymorphisms at the population, intraspecific and cultivar levels, as well as phylogeography for a certain species in Ilex.

Long repeat sequences with length longer than 30 bp are important because they might play important roles in the insertion/deletion mismatches and rearrangements that led to genomic variation [28–30] and the increase in the genetic diversity of populations [31]. We found that the number of long repeat sequences in Ilex is comparably higher to some other angiosperm species (e.g., 364 long repeats in Oxalidaceae [32], 403 in Veratrum [33]; 32 in Oresitrophe rupifraga, and 34 in Mukdenia rossiiand [34]). Among these long repeats, the 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 the complementary and reverse repeats were quite rare, only accounting for 0.42% and 0.49%, respectively.

Hypervariable regions for future studies of phylogeny and DNA barcoding
Hypervariable regions are usually used as effective evidence for delimitation of closely related taxa as well as to provide a wealth of phylogenetic information [35, 36]. In general, sequence divergence in IR regions is more conserved than that in the SSC and the LSC regions [37]. In the present study, the eight hypervariable regions including four genes and four gene with their adjacent and overlapped regions were detected and selected as the most divergent loci (Table 4). All the hypervariable loci were distributed in the SC regions, and also showed lower levels of variation in the IR regions, which was consistent with most angiosperm genera [32, 33].

Previous phylogenetic analyses of *Ilex* that were based on a handful of plastid markers (mainly including *atpB-rbcL*, *psbA-trnH*, *rbcL*, and *tmL-trnF*) did not generate well-resolved interspecific relationships within genus *Ilex* [1, 2, 8, 11, 13, 38–40]. When comparing these markers to those of highly variable regions identified from this study, only one of them, which is the *rbcL* region, was present. Therefore, we proposed the use of the eight highly variable regions reported in this study to be used as potential markers for reconstruction of the phylogenetic tree based on short plastid regions. Furthermore, with the high resolution at species level, a powerful DNA barcode can be obtained for *Ilex* that is suitable for species identification and delimitation. However, further studies are required to evaluate the delimitation strength in these regions prior to conducting studies on the DNA barcoding of *Ilex*.

### Phylogenetic inference

Researchers have attempted to resolve the phylogenetic relationships among major lineage of *Ilex* and to test the consistency between the molecular phylogeny and the traditional taxonomic systems based on morphology evidence [8–13, 23, 38]. However, these studies failed to elucidate the phylogenetic relationships effectively, especially at the species level. The presences of poor species resolution and weak support at most branch nodes were due to lack of sufficient variable information sites when using short DNA markers [8, 10–12, 23, 38]. Undoubtedly, this limitation can be addressed by using longer DNA sequences that contain an abundant of variable sites [41]. The complete chloroplast genome sequences containing sufficient informative sites and can obviously resolve the phylogenetic relationships at almost any taxonomic level [18]. Earlier phylogenetic analyses using the complete chloroplast genome sequences based on a small quantity of samples achieved the same conclusion [14, 26, 42].

At present, the relationships within the genus *Ilex* were well resolved based on the current phylogeny, and five major clades (lineages), clade A to clade E, were identified with consideration of macro-morphological and distribution information (Figure 9). The main lineages were largely consistent with the previous plastid phylogeny, while the relationships among these major lineages were largely discrepant [8, 11, 13]. Our results showed that the American groups (clade E) and the Eurasia groups (clade F) were sister to each other, while both of them formed the basal clade of the genus. These two basal lineages were sister to the rest (clade A–C), which were mostly from the Asia region. This finding was not consistent to the previous results produced by Manen [11], of which the American groups (group 3) and the Eurasia groups (group 4) were the most recent diversified groups. The discordance between our results and the previous conclusions mainly owed to the fact that lacking genes from hypervariable regions by their phylogenetic studies and the low resolution among major clades [8, 11].

The inconsistency between the plastid phylogeny and the traditional taxonomic systems was also revealed by the present studies (Figure 9). Almost all the subgenera, sections and series included in the analysis were support as non-monophly (both the two subgenera, five out of the six sections, and six out of eleven series). We did not rule out the possibility that more non-monophy groups will be present in this plastid-based phylogenetic tree when more taxa are involved. In addition, though the resolution of phylogenetic tree in early studies was quite low, they indicated that the nuclear and plastid phylogeny were highly incongruent and the nuclear phylogeny was more consistent with traditional morphological classification than the plastid one [11]. We here confirmed their conclusions by improving the resolution of plastid phylogenetic tree using the complete chloroplast genome data.

In general, species of the same origin should be clustered next to each other in the same clade due to them having the same inheritance. This is proven accurate for some of the *Ilex* species present in our study, including *I. comuta*, *I. dasyphylla*, *I. latifolia*, and *I. integra*. However, the accessions from *I. pubescens* and *I. lohafuensis* displayed otherwise; the two accessions of *I. pubescens* were resolved in two distinct clades (clade A and clade B), and the two accessions of *I. lohafuensis* were separated from each other, with *I. championii* placed in between them. Three samples of *I. viridis* were clustered with morphologically similar species *I. triflora*. Some of the non-monophy species might be cause by genomic capture or hybridization events, as well as misidentification caused by taxonomic confusion among different taxa. Howsoever, our well resolved phylogeny will help to clarify the taxonomic confusion among closely related species in the genus.
Conclusions

In this study, the characterization and phylogenetic analysis of the chloroplast genome of Ilex have been carried out, and as a result, an in-depth understanding on the evolution of Ilex at genome-scale level was obtained. However, we proposed the future studies on Ilex should focus on the phylogenetic reconstruction based on the nuclear region. We propose the capturing of low-copy nuclear genes from the genome-skimming data, which is thought to provide a better resolution for most plant species when compared to short nuclear DNA markers, i.e. ITS, and incorporate it with the phylogenetic tree based on the complete chloroplast genomes as well as morphology analyses, to enhance our understanding of the complex evolutionary history of genus Ilex.

Materials And Methods

Taxon sampling, DNA extraction, and sequencing

Seven species of Ilex, including I. dasyphylla, I. fukienensis, I. lohfaensis, I. venusta, I. viridis, I. yunnanensis, I. zhejiangensis, Ilex fukienensis, I. venusta, and I. zhejiangensis, were collected from their natural accessions in China. Fresh leaf specimens were collected in the field and stored in silica gels prior to be sent to the laboratory for 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 the NCBI GenBank were downloaded for their sequences in fasta format and kept for subsequent comparative analyses (Additional files 5: Table S5). Based on the classification of Ilex that is generally accepted [22], the current dataset comprised of species from six sections and 11 series of the genus Ilex.

Total genomic DNA was extracted using the Plant Genomic DNA Kit (Tiangen Biotech, China) following the manufacturer's protocol. The DNA extract was inspected with agarose gels, and quantified using Qubit 2.0 (Life Technologies, Country) for its integrity, purity, and concentration. The qualified DNA ($\geq$ 50 ng) was used to construct a paired-end (2×150 bp) library, and sequencing was conducted on a HiSeq X Ten platform (Illumina, USA). The generated raw data were filtered with fastp v.0.20.0 software [43] to remove low-quality reads.

Chloroplast genome assembly and annotation

The filtered raw data were fed into the NOVOPlasty 2.6.3 [44] pipeline for genome assembly, with the rbcL gene sequence of Ilatifolia (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 [45], in which the annotated chloroplast genome of Ilatifolia (Accession number: MN688228) was set as reference. Start and stop codons in the chloroplast genomes were manually corrected using DOGMA [46], while the tRNA genes were verified using tRNA scan-SE v2.0.3 embedded in GeSeq [47] based on default parameters. The circular chloroplast genome maps were visualized using OrganellarGenomeDRAW [48].

Genome comparative analyses

Sequence alignment of the 41 complete chloroplast genomes was carried out using MAFFT v.7 [49] and the alignment was further trimmed using trimAI v1.2 using the “-gappyout” setting [50]. The expansion and contraction of the IR regions were visualized using online IRscope [51] and then was manually checked. The nucleotide diversity (Pi) was estimated using DnaSP v.5 [52] with a step size of 200 bp and a window length of 800 bp. The genome variability across the 41 species of Ilex was assessed using mVISTA [53], under Shuffle-LAGAN mode. Mauve version 2.3.1 [54] plug-in that was available in Geneious version 11.0.3 [55] was used to identify locally collinear blocks presented among the chloroplast genomes based on default parameters.

Repeat sequence identification

The number of large repeats, including forward, palindromic, reverse, and complement repeats were identified using the onlineREPuter program [56] according to the following criteria: sequence identities of 90%, cutoff point at $\geq$ 30 bp, Hamming distance set at 3, and a minimum repeat size of 30 bp. Tandem Repeat Finder [57] was used to analyze tandem repeat sequences with the default parameters. SSRs were identified using web-MISA [58], 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 are separated by less than 100 nucleotides and were combined into one.
**Codon Usage Analysis**

RSCU is defined as the ratio of the observed frequency of codons to the expected frequency. The sequences of 77 shared protein-coding genes were extracted and concatenated using PhyloSuite v1.2.1 [59]. The value of RSCU was calculated using MEGA X [60]. A dendrogram based on the value of RSCU was constructed using pheatmap, which is available in R package (https://cran.r-project.org/web/packages/pheatmap/).

**Phylogenetic Analyses**

Phylogenetic analysis was conducted based on 51 complete chloroplast genome sequences representing 39 *Ilex* species from six sections and 11 series. Based on the finding from Yao et al. [23], *Helwingia himalaica* (Accession number: NC031370) was selected as the outgroup. Genome alignment was carried out using MAFFT v.7 [49] and then trimmed using trimAI v1.2 with the “-gappyout” setting [50].

Maximum likelihood (ML) analyses were conducted using IQ-tree [61] under 10,000 ultrafast bootstrap support values (UFBS) replicates [62]. According to Bayesian information criterion (BIC), the best fitting substitution model that was estimated using ModelFinder [63] was GTR+F+I+G4. Bayesian inference (BI) analysis was carried out using MrBayes version 3.2 [64] that was implemented in CIPRES [65]. The Markov chain Monte Carlo analysis was executed for 20,000,000 generations, with four chains (one cold and three heated), each starting with a random tree, and readings were sampled at every 1000 generations. Convergence of runs was accepted when the average standard deviation of split frequencies achieved <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 tree from both analyses was visualized using FigTree v.1.4.2 [66].

**Declarations**

**Acknowledgments**

We thank Wanyi Zhao, Zhongcheng Liu for their assistance in sample collection, Yubing Zhou (Jierui Biotech, Guangzhou, China) for data analysis of chloroplast genomes.

**Authors’ contributions**

Conceptualization, K.X. and K.M; methodology, K.X. and K.M; formal analysis, K.X. and K.M; investigation, K.X.; resources, K.X.; data curation, K.X. 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.X., L.M. and K.M; project administration, K.X. and L.M.; funding acquisition, K.X. 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 (32100167), 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 seven newly sequenced *Ilex* species were submitted to GenBank and the accession numbers could be found in Additional file 5. All raw reads are available in the short sequence archive under accession no. PRJNA768933. All of the complete genome sequences used in this study were downloaded from NCBI (https://www.ncbi.nlm.nih.gov), and the accession numbers could also be found in Additional file 5.

**Ethics approval and consent to participate**
Not applicable. No specific permits were required for the collection of specimens for this study. All materials used in the study were collected in public areas of China in compliance with the relevant laws of China. The formal identification of the plant material was carried out by Kewang Xu. Voucher specimens were prepared and deposited at the herbarium of Nanjing Forestry University (NF) and their collection numbers could be found in Additional file 5.

Consent for publication

Not applicable.

Author details

1 Co-Innovation Center for Sustainable Forestry in Southern China, College of Biology and the Environment, Nanjing Forestry University, Nanjing; 510275, China. 2 Faculty of Health and Life Sciences, INTI International University, 71800 Nilai, Malaysia. 3 State Key Laboratory of Biocontrol and Guangdong Provincial Key Laboratory of Plant Resources, School of Life Sciences, Sun Yat-sen University, Guangzhou; China.

Competing interests

The authors declare that they have no conflict of interest.

References

1. Loizeau PA, Barriera G, Manen JF, Broennimann O. Towards an understanding of the distribution of *Ilex* L. (Aquifoliaceae) on a world-wide scale. Biol Skr. 2005;55:501–520.

2. Powell M, Savolainen V, Cuénoud P, Manen JF, Andrews S. The mountain holly (*Nemopanthus mucronatus*: Aquifoliaceae) revisited with molecular data. Kew Bull. 2000;55:341–347. [https://doi.org/10.2307/4115646](https://doi.org/10.2307/4115646)

3. Loizeau PA, Savolainen V, Andrews S, Spichiger R. Aquifoliaceae. In: Kubitzki K, editors. Flowering plants. Eudicots, The families and genera of vascular plants. Berlin: Springer; 2016. p. 31–36.

4. Filip R, López P, Giberti G, Coussio J, Ferraro G. Phenolic compounds in seven South American *Ilex* species. Fitoterapia. 2001;72(7):774–778.

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;39:2861–2875.

6. Yi F, Zhao XL, Peng Y, Xiao PG. Genus *Ilex* L.: phytochemistry, ethnopharmacology, and pharmacology. Chin Herb Med. 2016;8:209–230.

7. International Union for Conservation of Nature and Natural Resources (IUCN): The IUCN Red List of Threatened Species. [https://www.iucnredlist.org/](https://www.iucnredlist.org/) (2021). Accessed 11 Aug 2021.

8. Cuénoud P, del Pero Martinez MA, Loizeau PA, Spichiger R, Andrews S, Manen JF. Molecular phylogeny and biogeography of the genus *Ilex* L. (Aquifoliaceae). Ann Bot (Oxford). 2000;85:111–122.

9. 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.

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

11. Manen JF, Barriera G, Loizeau PA, Naciri Y. The history of extant *Ilex* species (Aquifoliaceae): evidence of hybridization within a Miocene radiation. Molec Phylogen Evol. 2010;57:961–977. [https://doi.org/10.1016/j.ympev.2010.09.006](https://doi.org/10.1016/j.ympev.2010.09.006)

12. Gottlieb AM, Giberti GC, Poggio L. Molecular analyses of the genus *Ilex* (Aquifoliaceae) in southern South America, evidence from AFLP and ITS sequence data. Amer J Bot. 2005;92:352–369.

13. Jiang 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;298:147–157.
14. Yao X, Tan YH, Liu YY, Song Y, Yang JB, Corlett RT. Chloroplast genome structure in *Ilex* (Aquifoliaceae). Sci Rep. 2016;6(1):1–10.

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

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

17. Zhang SD, Jin JJ, Chen SY, Chase MW, Soltis DE, Li HT, et al. Diversification of Rosaceae since the Late Cretaceous based on plastid phylogenomics. New Phytol. 2017;214:1355–1367.

18. Li HT, Yi TS, Gao LM, Ma PF, Zhang T, Yang JB, et al. Origin of angiosperms and the puzzle of the Jurassic gap. Nat Plants. 2019;5:455–456.

19. Meng KK, Chen SF, Xu KW, Zhou RC, Li MW, Dhamala MK, et al. Phylogenomic analyses based on genome-skimming data reveal cyto-nuclear discordance in the evolutionary history of *Cotoneaster* (Rosaceae). Molec Phylogen Evol. 2021;158:107083.

20. Mower JP, Vickrey TL. Structural diversity among plastid genomes of land plants. In: Chaw SM, Jansen RK, editors. Plastid Genome Evolution. Amsterdam (The Netherlands) and New York: Academic Press; 2018. p. 2e382.

21. Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc Natl Acad Sci USA. 2010;107:4623–4628. https://doi.org/10.1073/pnas.0907801107.

22. Chen SK, Ma HY, Feng YX. Aquifoliaceae. In: Wu ZY, Raven PH, Hong DY, editors. Flora of China. Beijing: Science Press, St. Louis: Missouri Botanical Garden Press; 2008. p. 449–461.

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

24. Wicke S, Schneeweiss GM, De Pamphilis CW, Müller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 2011;76:273e297.

25. Wolfe KH, Morden CW, Palmer JD. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci USA. 1992;89(22):10648–10652.

26. Kong BLH, Park HS, Lau TWD, Lin Z, Yang TJ, Shaw PC. Comparative analysis and phylogenetic investigation of Hong Kong *Ilex* chloroplast genomes. Sci Rep. 2021;11(1):1–13.

27. Ebert D, Peakall R. Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. Mol Ecol Resour. 2009;9:673–690.

28. 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 Biol Evol. 2014;31:645e659.

29. Asaf S, Khan AL, Khan MA, Shahzad R, Lubna, Kang SM, et al. Complete chloroplast genome sequence and comparative analysis of loblolly pine (*Pinus taeda* L.) with related species. PLoS One. 2018;13:e0192966. https://doi.org/10.1371/journal.pone.0192966.

30. Huang YY, Yang ZR, Huang S, An WL, Li J, Zheng XS. Comprehensive analysis of *Rhodomyrtus tomentosa* chloroplast genome. Plants. 2019;8:89.

31. Qian J, Song JY, Gao HH, Zhu YJ, Xu J, Pang XH, et al. The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. PLoS One. 2013;8:e57607.

32. Li XP, Zhao YM, Tu XD, Li CR, Zhu YT, Zhong, H. Comparative analysis of plastomes in Oxalidaceae: Phylogenetic relationships and potential molecular markers. Plant Divers. 2021;43:281–291.

33. Zhang YM, Han LJ, Yang CW, Yin ZL, Tian X, Qian ZG, Li GD. Comparative chloroplast genome analysis of medicinally important *Veratrum* (Melanthiaceae) in China: insights into genomic characterization and phylogenetic relationships. Plant Divers. 2021; https://doi.org/10.1016/j.pld.2021.05.004

34. Liu LX, Wang YW, He PZ, Li P, Lee J, Soltis DE, et al. Chloroplast genome analyses and genomic resource development for epilithic sister genera *Oresitrophe* and *Mukdenia* (Saxifragaceae), using genome skimming data. BMC Genomics. 2018;19:235. https://doi.org/10.1186/s12864-018-4633-x

35. Li JW, Yang Y, Henry RJ, Rossetto M, Wang YT. Chen SL. Plant DNA barcoding: from gene to genome. Biol Rev Camb Phil Soc. 2015;90:157e166. https://doi.org/10.1111/brv.12104.
36. Zeng CX, Hollingsworth PM, Yang J, He ZS, Zhang ZR, Li DZ, et al. Genome skimming herbarium specimens for DNA barcoding and phylogenomics. Plant Methods. 2018;14(1):1–14.

37. Luo C, Huang WL, Sun HY, Yer H, Li XY, Li Y, et al. Comparative chloroplast genome analysis of *Impatiens* species (Balsaminaceae) in the karst area of China: insights into genome evolution and phylogenomic implications. BMC Genomics. 2021;22:571.

38. Manen JF. Are both sympatric species *Ilex perado* and *Ilex canariensis* secretly hybridizing? Indication from nuclear markers collected in Tenerife. BMC Evol Biol. 2004;4(1):1–12.

39. Selbach-Schnadelbach A, Cavalli SS, Manen JF, Coelho GC, de Souza-Chies TT. New information for *Ilex* phylogenetics based on the plastid *psbA-trnH* intergenic spacer (Aquifoliaceae). Bot J Linn Soc. 2009;159:182–193.

40. Shi L, Li NW, Wang SQ, Zhou YB, Huang WJ, Yang YC, et al. Molecular evidence for the hybrid origin of *Ilex dabieshanensis* (Aquifoliaceae). PLoS One. 2016;11(1):e0147825.

41. Philippe H, Brinkmann H, Lavrov DV, Manuel MG, Wörheide G, et al. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 2011;9:e1000602.

42. Su T, Zhang MR, Shan ZY, Li XD, Zhou BY, Wu H, et al. Comparative survey of morphological variations and plastid genome sequencing reveals phylogenetic divergence between four endemic *Ilex* species. Forests. 2020;11(9):964.

43. Chen SF, Zhou YQ, Chen YR, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:1884–1890.

44. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 2016;45:e18–e18.

45. Huang DJ, Cronk QC. Plann: A command-line application for annotating plastome sequences. Appl Plant Sci. 2015;3(8):1500026.

46. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004;20(17):3252–3255.

47. 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:W6–W11.

48. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3: 1: expanded toolkit for the graphical visualization of organelle genomes. Nucleic Acids Res. 2019;47:W59–W64.

49. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–780.

50. Capella-Gutiérrez S, Silla-Martinez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

51. Amiryousefi A, Hyvönen J, Poczai P. IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics. 2018;34(17):3030–3031.

52. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25(11):1451–2.

53. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004;32 Suppl 2:273–279.

54. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14:1394–1403.

55. Kearse M, Moir R, Wilson A, Stones-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):1647–1649.

56. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29:4633–4642.

57. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.

58. 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;106:411–422.
59. Zhang D, Gao FL, Jakovlic I, Zou H, Zhang J, Li WX, Wang GT. PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour. 2020;20:348–355. https://doi.org/10.1111/1755-0998.13096.

60. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molec Biol Evol. 2018;35(6):1547.

61. 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.

62. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018;35:518–522. https://doi.org/10.1093/molbev/msx281.

63. 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.

64. 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;61:539–542.

65. Miller, M.A.; Pfeiffer, W.; Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE), 14 November 2010, New Orleans, LA, 2010, pp. 1–8.

66. Rambaut A. FigTree V1.4.2. 2012. http://tree.bio.ed.ac.uk/software/figtree/.

**Figures**
Figure 1

Gene circle maps of seven newly sequenced Ilex species. The colored bars indicate different functional groups. Thick lines of the large circle indicate the extent of the inverted repeat regions (IRa and IRb), which separate the genome into small (SSC) and large (LSC) single copy regions. Genes on the inside and outside of the large circle are transcribed clockwise and counterclockwise, respectively. The darker gray columns in the inner circle correspond to the GC content, the light gray to AT content.
Figure 2

Sliding-window analysis showing the nucleotide diversity (Pi) values of the aligned Ilex chloroplast genomes.

Figure 3
Figure 3

Sequence alignment of 41 Ilex chloroplast genomes using the mVISTA program with I. szechwanensis as a reference. The vertical scale indicates the percent identity, ranging from 50% to 100%. The horizontal axis indicates the location within the plastomes. Genome regions are color-coded as exon, intron and untranslated regions (UTRs).

Figure 4

Mauve multiple alignment of 41 Ilex chloroplast genomes revealing no interspecific rearrangements.
Figure 5

Comparison of the SC/IR junctions among the 41 Ilex chloroplast genomes. JLA, LSC/IRa boundary; JSA, SSC/IRa boundary; JSB, SSC/IRb boundary; JLB, LSC/IRb boundary.
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.
Figure 7

Analysis of long repeats in 41 chloroplast genomes of Ilex showing the number of complement, forward, palindromic, reverse and tandem long repeats.
Figure 8

Heatmap analysis of relative synonymous codon usage (RSCU) values among the 41 Ilex chloroplast genomes.
Figure 9

Phylogenetic trees inferred from maximum likelihood (ML) and Bayesian inference (BI) analyses of 52 accessions 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 were indicated by asterisks. Incongruences between the BI and ML trees were indicated by dashes. Hu's classification was illustrated by color graphic pattern. Recognized groups (major clades) were also marked by the right-hand black bar.

Supplementary Files

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

- TableS1.xls
- TableS2.xls
- TableS3.xlsx
- TableS4.xls
TableS5.xlsx