Vol.68: e25240954, 2025
https://doi.org/10.1590/1678-4324-2025240954
ISSN 1678-4324 Online Edition

BABT
Brazilian Archives of Biology and Technology

Article - Biological and Applied Sciences

An Integrated Resource for Genetics and Breeding of Erva/Yerba Mate (*Ilex paraguariensis*)

Patricia Mabel Aguilera[1]
https://orcid.org/0000-0002-5063-9350

Mauro Grabiele[1*]
https://orcid.org/0000-0001-7386-4924

[1]Instituto de Biología Subtropical (UNaM-CONICET), Universidad Nacional de Misiones, Posadas, Misiones, Argentina.

Editor-in-Chief: Paulo Vitor Farago
Associate Editor: Jane Manfron

Received: 11-Nov-2024; Accepted: 20-Jul-2025

*Correspondence: maurograbiele@conicet.gov.ar (MG)

HIGHLIGHTS

- We assembled the yerba mate nuclear genome, plastome, and mitogenome to chromosome level, and also the nuclear rDNAs, to achieve a comprehensively annotated original set of high-quality genetic markers.
- These valued original sets of sequences and markers are now openly available to support genetic analyses, breeding programs, and genomics research in this significant crop species.

Abstract: The erva/yerba mate tree *Ilex paraguariensis* is a South American crop significant to the industry and local economies of producer countries. Here, we provide a leading omics integral framework of high-quality annotated markers in yerba mate, notable for its exhaustive employed resources and rigorous curation and validation steps. We assembled the yerba mate nuclear genome (872.9 Mb; BUSCO: 97.3%) using a reference-guided approach, plastome (157.6 Kb), and mitogenome (521.3 Kb) to chromosome level, in addition to nuclear rDNA units 5S (511 bp) and 18S-25S (10,760 bp). The nuclear assembly exhibited 64.9% repetitiveness, while 23.1% corresponded to 41,922 protein-coding genes (PCGs) (BUSCO: 97.1%), with 93.5% mapped to 20 chromosomes and 99.1% supported by joint evidence from RNA-seq (89.1%), orthology in *Ilex* (84.7%), and functional annotation (92.9%); additionally, 23,089 alternative splice variants for 10,446 genes, 78,992 UTRs for 18,100 genes, and 500 tRNA genes were annotated. Genes were further categorized through description labels, GO terms, KEGG identifiers, EC numbers, structural signatures, pathway identifiers, subcellular localization, and membership in transcription factor families and kinase gene groups. The plastome conformed to the typical *Ilex* structure, contained 811 repeats and 131 genes (86 PCGs, 37 tRNAs, 8 rRNAs), and featured a single transcript with 53 PCG editing sites. The mitogenome comprised two chromosomes, 2,439 repeats, 65 genes (40 PCGs, 22 tRNAs, 3 rRNAs), and multiple transcripts with 578 PCG editing sites; *Ilex* mitogenomes showed comparable sequences but lacked synteny. Nuclear rDNA units were further annotated for genes, spacers, transcription-associated sites, and repeats. This comprehensive approach is anticipated to be pivotal for genetics, breeding programs, and downstream genomics research in yerba mate.

Keywords: chromosome-level genome; plastome; mitogenome; rDNA units; high-quality annotation.

INTRODUCTION

The hot infusion known as "mate" or "chimarrão" and its cold version, "tereré," are commonly prepared from the dried leaves and twigs of the yerba mate tree, *Ilex paraguariensis* A. St.-Hil. (Aquifoliaceae), and are widely consumed in Argentina, Brazil, Paraguay, and Uruguay [1]. The yerba mate industry is significant in the producer regions of Argentina, Brazil, and Paraguay, encompassing 280 thousand planted hectares and yielding over 900 thousand tons of raw material annually, deeply influencing their economies [1, 2]. Recently, yerba mate has received increased attention for its bioactive compounds, which may expand the possibilities for technological applications of this plant species [3].

Our group established the foundational omics approach in *Ilex* with the high-throughput RNA sequencing (RNA-seq), assembly, and annotation of the leaf transcriptome of the elite yerba mate cultivar INTA-EEA Cerro Azul Pg538 from Misiones, Argentina. This initial exploration identified 32,355 genes, including caffeine synthase, provided insights into chloroplast, mitochondrion, and ribosomal sequences at the RNA level, and further enabled the in-depth characterization of the DNAJ gene family in this species, among other novel findings [4-8]. Subsequently, a tissue-specific RNA-seq approach in yerba mate sequenced 13 libraries from leaves, roots, and plantlets from diverse populations in Northeast Argentina for detailed transcriptome characterization [9]. Additionally, 12 more RNA-seq libraries from leaves and roots were generated to conduct a leading drought-stress analysis in the elite yerba mate material SI-49 from Corrientes, Argentina [10, 11]. Furthermore, supported by the ProMateAr sequencing project of the elite yerba mate cultivar INTA-EEA Cerro Azul 8/74 -the male progenitor of Pg538- [12], a high-quality set of 10,611 genomic scaffolds, totaling 1,064.8 Mb, was achieved and released to NCBI (https://ncbi.nlm.nih.gov) in September 2023 under accession GCA_963454935.2 [13], predated by GCA_905181385.1 (1,421.2 Mb) in January 2021. Notably, Garberoglio and coauthors [14] measured the DNA content of yerba mate using flow cytometry and reported 2C=1.716±0.065 pg (n=20 chromosomes), predicting a genome size of 839.1±31.8 Mb at the sequence level for this species. These draft genomic sequences of yerba mate facilitated the recent cloning of caffeine biosynthesis genes [15, 16] and our characterization of 156 families of transposable elements in Repbase between 2021 and 2024 (https://www.girinst.org/repbase/update/search.php?query=paraguariensis&querytype=Taxonomy).

Despite a decade of increasing sequencing data in yerba mate from various research groups, an original, integrated, chromosome-level resource annotated with high-quality markers is still lacking for this species, unlike other sequenced *Ilex* species [17-19]. This comprehensive, reliable localization of genetic markers, superior to the scaffold level, is crucial for integrating omics data with genetics and breeding efforts in yerba mate. To address this current information gap, we first generated and validated chromosome-level assemblies for the nuclear, chloroplast, and mitochondrion genomes of yerba mate, along with supported assemblies of nuclear ribosomal units. These assemblies will then undergo exhaustive and reliable structural and functional annotation of genes and other valuable markers, considering all available omics resources and state-of-the-art tools and databases. This integrated approach aims to be a pivotal framework for genetic analysis and breeding programs, accelerating the traditional characterization of cultivars with updated selection tools, as well as facilitating downstream genomics analysis for the broader yerba mate research community.

MATERIAL AND METHODS

Data of *Ilex* used in bioinformatic analyses

We employed *I. paraguariensis* sequences available at NCBI, including the genomic scaffolds of GCA_963454935.2 with its associated 66.3 Gb of short-read DNA-seq paired-end data libraries ERR12037359 and ERR12040303, and a 21.6 Gb long-read DNA-seq library ERR12708757. In addition, we utilized 26 RNA-seq paired-end data libraries, totaling 127 Gb, from leaves (16), roots (9), and a seedling (1) of this species: SRR1411087, SRR5740154-SRR5740159, SRR3584943-SRR3584948, SRR3584950-SRR3584952, SRR3584954-SRR3584957, and SRR24273717-SRR24273722. We also employed available genome sequences and annotated sets of transcripts and proteins from other *Ilex* taxa: from NCBI, *I. aquifolium* L. (GCA_951799425.1; 20 pseudomolecules + 86 scaffolds; 800.7 Mb) [20] and *I. asprella* (Hook. & Arn.) Champ. ex Benth. (GCA_023539305.1; 19 pseudomolecules + 48,755 scaffolds; 804.1 Mb; 45,449 proteins); and from NGDC (https://ngdc.cncb.ac.cn/), *I. latifolia* Thunb. (GWHBIST00000000; 20 pseudomolecules + 335 contigs; 766 Mb; 35,217 transcripts; 35,217 proteins) and *I. polyneura* (Hand.-Mazz.) S.Y. Hu (GWHBDNW00000000; 20 pseudomolecules + 42 contigs; 727.1 Mb; 32,253 transcripts; 32,253 proteins). Furthermore, we utilized all non-redundant chloroplast (93) and mitochondrion (5) genome sequences of *Ilex* species available at NCBI.

A variety of bioinformatic analyses in *Ilex* were performed using the Galaxy (https://galaxyproject.org/) [21] and Gensas (https://www.gensas.org/) [22] servers, as well as the Geneious platform (Dotmatics). The latest version of each tool was employed, using default parameters unless otherwise specified in parenthesis. BUSCO analyses [23; galaxy 5.5.0] to assess the completeness of genome assemblies, transcripts, and annotated proteins were conducted on overall *Ilex* sequences, considering both the Embryophyta_odb10 and Eudicots_odb10 gene sets. The genomic mode of BUSCO was executed via Augustus, followed by Metaeuk to improve the Fragmented and Missing scores of the former gene predictor. Sequence reads were processed with Trimmomatic [24; galaxy 0.38.0] to remove adapters and low-quality bases (sliding window trimming; bases: 4; quality: 20), and only paired-end data were considered for further analysis. Ploidy level and sub genomic structures in yerba mate were inferred using Smudgeplot [25; galaxy 0.2.5] with the total forward reads from paired-end DNA-seq data. Phylogenetic relationships and divergence times within *Ilex* were determined following the criteria established by Yao and coauthors [26].

Chromosome-level reference assembly and validation in yerba mate

The 10,611 genomic scaffolds of yerba mate were scanned using the Purge Dups tool [27; galaxy 1.2.6] (mode: purge_dups; minimum chaining score for a match: 7000; get_seqs mode: trim end and middle duplications) to identify and remove potential heterozygous duplications in the primary assembly (1,064.8 Mb, with 1,043.9 Mb ungapped), considering the expected genome size of 839.1±31.8 Mb for this species. This process identified and removed 191.5 Mb, consisting of 88.5% purely haplotigs and 11.5% heterozygous overlaps. Plastid, mitochondrion, and nuclear ribosomal DNAs, previously characterized in yerba mate (see below), were subsequently identified based on homology using Minimap2 [28; galaxy 2.28] and also removed from the remaining primary assembly, totaling 1.0 Mb. Care was taken to avoid removing the 5S rRNA-derived SINE loci from the remaining primary assembly, the consensus sequence of which we characterized under Repbase ID SINE3-1_Ipa. The resulting 6,256 genomic scaffolds (872.3 Mb) were then subjected to reference-guided scaffolding using RagTag [29; galaxy 2.1.0] (mode: scaffold; aligner: Minimap2; Asm20: up to 20% sequence divergence; gaps: 100 nt) with the 20 pseudomolecules of *I. aquifolium* (787.3 Mb) as reference sequences. This latter taxon was chosen over other sequenced *Ilex* species due to its largest assembled fraction into 20 pseudomolecules and superior genomic BUSCO quality. The resulting chromosome-level assembly of yerba mate, hereafter termed YMM1 (Yerba Mate Misiones version 1), was initially validated by the corresponding RagTag confidence scores (grouping, location, and orientation; range 0-1, with 1 being the highest value) for each scaffold, and its genomic BUSCO scores. Further validation was obtained from species-specific read mapping analysis against YMM1: total RNA-seq reads were mapped using HISAT2 [30; galaxy 2.2.1], and total DNA-seq short reads were mapped using Minimap2 (mode: without splicing). For both types of data, separate mapping runs were performed for each library, and the resulting BAM files were then merged into single files using the Merge BAM tool in Galaxy. These merged BAM datasets were submitted to Samtools stats [31; galaxy 2.0.5] to determine the percentage of mapped reads and other statistics. Furthermore, the chromosome-level sequences of YMM1 and other *Ilex* species were pairwise compared using Chromeister [32; galaxy 1.5.a] to evaluate the RagTag assembly and to detect block collinearity at different levels, as well as structural rearrangements within other *Ilex* species. This analysis generated a dot plot graph, a list of detected events, and a scoring metric for each comparison (range 0-1, where 0 indicates the highest value). Additionally, a library of repetitive consensus sequences was constructed for YMM1 using RepeatModeler [33; galaxy 2.0.5]. This library was used for whole self-genome identification of repetitive DNA loci using RepeatMasker [34; galaxy 4.1.1] (cutoff: 250, 100% real matches), and also against GCA_963454935.2 and other *Ilex* genomes in the same manner. Statistics derived from RepeatMasker were used to compare genome structures among taxa to further support YMM1. To achieve the same goal, a non-redundant library of LTR retrotransposons was constructed for YMM1 using the Dante_LTR protocol [35; galaxy 4.0.4.1] for subsequent use by RepeatMasker against all aforementioned taxa.

Genome annotation and validation in yerba mate

The Gensas pipeline for whole-genome structural and functional annotation of eukaryotes was followed, comprising a series of steps that ultimately yielded weighted consensus gene models and their subsequent functional characterization.

Repeat masking. YMM1 was subjected to a soft repeat masking step using RepeatMasker (cutoff: 250; "no low" option, suitable for gene annotation) with the species-specific repeat library generated by

RepeatModeler. These masked sequences were then used for all subsequent downstream analyses, including species-specific RNA-seq mapping, transcript and protein alignments, and gene prediction.

Transcript reconstruction. A hybrid approach to transcript reconstruction was applied according to PASA [36; gensas 2.4.1]. Initially, each of the 26 RNA-seq libraries was individually mapped to the genome using HISAT2 (mode: tailored) and subsequently subjected to *de novo* genome-guided transcript assembly using StringTie [37; galaxy 2.2.3]. The resulting 1.2 million transcripts were merged into a non-redundant set of 76,749 sequences using the StringTie_merge function. Additionally, separate transcriptomes were assembled *de novo* for each RNA-seq library using rnaSPAdes [38; galaxy 3.15.5], yielding a total of 3.5 million sequences. These sequences were then collapsed into a set of 301,582 non-redundant transcripts that mapped to the genome using the assembly-clustering function of PASA. A total transcriptome *de novo* assembly was also performed using Trinity [39; galaxy 2.15.1], where reads from all RNA-seq libraries were used to generate 574,621 sequences. These sequences were subsequently collapsed into a non-redundant set of 160,045 transcripts that mapped to the genome using PASA, as described above. To obtain a comprehensive non-redundant set of transcripts while preserving essential isoforms, the three initial sets of transcripts were combined using the FASTA Merge Files and Filter Unique Sequences tool in Galaxy (galaxy 1.2.0), resulting in 518,252 sequences. Moreover, this set of unique sequences was then subjected to a final PASA assembly-clustering step, yielding 354,840 unique transcripts. The BUSCO analysis of this final transcriptome set showed a completeness score of C:93.7% [S:6.3%, D:87.4%], F:3.1%, M:3.2% (Embryophyta_odb10; n:1614) and C:93.7% [S:8.6%, D:85.1%], F:2.4%, M:3.9% (Eudicots_odb10; n:2326).

Gene predictions. Funannotate [40; galaxy 1.8.15] was applied using only protein evidence from UniProtKb/Swiss-Prot (https://www.uniprot.org/uniprotkb) (BUSCO models: Eudicots_odb10; BUSCO species: tomato; maximum intron length: 350,000; repeats2evm: yes). The resulting sets of annotated transcripts and proteins were subsequently used in a second run of this tool, in combination with other evidence from yerba mate (total mapped RNA-seq libraries and the unique transcript set) and other *Ilex* species (overall annotated transcripts and proteins). This approach yielded 39,563 gene models and associated sets of annotated transcripts and proteins. Maker [41; galaxy 2.31.11] was also applied, directly using the Funannotate outputs and to train SNAP (http://korflab.ucdavis.edu/software.html; version 2013_11_29), along with transcripts from other *Ilex* species, resulting in 53,132 gene models. Various gene predictors were also used independently, including GeneMarkES [42; gensas 4.48], GlimmerM [43; gensas 2.5.1], and SNAP, the latter two trained with *Arabidopsis thaliana*. Additionally, a run of BRAKER2 [44; gensas 2.1.5] using overall mapped RNA-seq evidence and a run of BRAKER3 [45; galaxy 3.0.8] that also incorporated the final annotated protein set from Funannotate were performed. Further gene predictions were generated through six separate runs with Miniprot [46; galaxy 0.13] and Metaeuk [47; galaxy 5.34c21f2], utilizing the yerba mate protein set derived from PASA gene models (see below), the final protein set from Funannotate, and those of other *Ilex* species, respectively. Moreover, three separate runs were conducted with Augustus [48; gensas 3.4.0] (mode: complete genes; strands: both), with the first trained using tomato hints, the second trained with the gene models derived from the initial Funannotate run, and the third trained with the final annotated transcript and protein sets from Funannotate. In total, 1,633,351 independent gene predictions were generated.

Transcript alignments. pBLAT transcript alignments [49; gensas 2.5] with a minimum identity of 95% were generated using evidence from the NCBI RefSeq Plants section and the data described above for yerba mate and other *Ilex* species, resulting in 2.1 million matches. PASA alignments were performed using the unique transcript set of yerba mate and that of other *Ilex* species as well, generating 316,060 and 21,198 gene models, respectively.

Protein alignments. Protein evidence alignments were also performed using Diamond [50; gensas 2.0.6] (mode: more-sensitive), which included the UniProtKb/Swiss-Prot dataset and the data described above for yerba mate and other *Ilex* species, resulting in 3.8 million matches.

All the aforementioned data were finally integrated into EvidenceModeler (EVM) [51; gensas 1.1.1] following the Gensas pipeline to generate weighted consensus gene structures, considering the following weight parameters: gene predictions: 1; gene annotations: 1; transcript alignments: 10; protein alignments: 5.The open reading frame (ORF) of each EVM gene model was checked and curated to ensure that it started with ATG, ended with TAA/TAG/TGA, and lacked internal stop codons. Gene models were further validated through species-specific total RNA-seq read mapping analysis using Kallisto quant (support: 100 bootstrap replicates) [52; galaxy 0.48.0] and StringTie (mode: reference), considering a minimum coverage value greater than 1.

Subsequently, the resulting EVM protein set was used throughout the annotation validation process. These analyses included BUSCO assessment and global comparisons with *A. thaliana* to identify orthologs and orthogroups using Proteinortho [53; galaxy 6.3.1], with a non-redundant set of 27,409 nuclear genome proteins (Araport11_genes.201606.pep.1 set). The percentage of shared essential genes was determined by comparison with the DEG database (http://origin.tubic.org/deg/public/index.php; *A. thaliana* set: 356 sequences) [54] and that of KOGs (cluster of eukaryotic orthologous groups) by comparison with the KOGs database (https://ftp.ncbi.nlm.nih.gov/pub/COG/KOG/; *A. thaliana* set: 3,273 KOGs of 13,459 sequences) [55]. Orthology comparisons were also performed within *Ilex* using Proteinortho and OrthoFinder [56; galaxy 2.5.5]. Furthermore, a lineage-dependent proteome quality assessment, using the Asterids clade as a reference (a set of 11,198 conserved hierarchical orthologous groups -HOGs-), was performed across *Ilex* species using OMark [57; galaxy 0.3.0].

Furthermore, functional sequence annotation of EVM proteins by orthology was performed using EggNOG-Mapper (http://eggnog-mapper.embl.de/) [58] (mode: realign queries to the whole PFAM DB). Additional functional features were identified using InterPro [59; 5.53-87.0] and PFAM (1.6), SignalP (5.0; mode: eukarya) and TargetP (2.0; mode: plant) tools at Gensas, as well as DeepSig [60; 1.2.5] (mode: eukaryotes), NoD [61; 0.1.1], WoLF PSORT [62; 0.0.12] (mode: plant), iep [63; 5.0.0.1], and pepstats (5.0.0) tools in Galaxy. Further characterization was obtained from protein alignments using Diamond against diverse databases such as the NCBI RefSeq Plant section, UniProtKb/Swiss-Prot, NCBI Nr section, and Darwin Tree of Life (all proteins) (https://www.darwintreeoflife.org/). Finally, a genome-wide annotation of transcription factor (TF) families was achieved following the classification criteria of PlantTFDB (https://planttfdb.gao-lab.org/prediction.php) [64] and iTAK (https://itak.feilab.net/cgi-bin/itak/index.cgi) [65], with the latter also annotating transcriptional regulators and kinase genes.

Subsequently, the EVM consensus gene models were submitted to PASA using the unique transcript set of yerba mate to obtain alternative splice (AS) variants and UTRs, each of which was subsequently filtered to yield locus non-redundant sequences. The ORF of each AS variant was checked for integrity following the same criteria as EVM genes and discarded if those criteria were not met. The protein set derived from the AS variants was subjected to the same functional annotation as the EVM genes. The genomic loci of transfer RNAs (tRNAs) were predicted using tRNAscan-SE [66; gensas 2.0.7] and then compared with PltRNAdb (https://bioinformatics.um6p.ma/) [67], as well as with overall eukaryote tRNAs and our characterized 24 tRNA-derived SINEs (SINE2) from yerba mate in Repbase, to avoid false positives and curation if necessary.

Chloroplast and mitochondrion genomes of yerba mate: assembly, annotation and validation

The plastid genome and mitogenome were assembled *de novo* using GetOrganelle [68; galaxy 1.7.7.1] from yerba mate short DNA-seq reads and corresponding *Ilex* sequences as seeds. These assemblies were then reference-annotated [69, 70] via genome-wide mapping of corresponding protein-coding, tRNA, and rRNA genes using the Geneious mapper (cutoff: 10% mismatches; map to all sites). Contig validation was performed by mapping species-specific non-redundant short (cutoff: 3% mismatches) and long (>10,000 nt; cutoff: 30% mismatches) DNA-seq reads. C-to-T editing sites in protein-coding genes (PCGs) of both genomes were annotated by mapping non-redundant reads from the 26 RNA-seq libraries of yerba mate (cutoff: 3% mismatches) and using the SNP identification tool in Geneious. The occurrence of repeats, ranging from mononucleotide to 100 nt units, was investigated in both genomes using the Phobos tool in Geneious. Similarity regions within (direct and inverted repeats) and between both genomes were identified using the MAUVE alignment tool in Geneious. The yerba mate plastid genome was compared with 93 non-redundant plastid sequences of *Ilex* species at NCBI via reference mapping (cutoff: 30% mismatches). The yerba mate mitogenome was compared with 5 mitogenome sequences of *Ilex* species at NCBI via MAUVE and Chromeister alignments. For further comparison, the set of non-redundant short DNA-seq reads previously mapped to the yerba mate mitogenome were mapped against the mitogenomes of those other *Ilex* species.

5S and 18S-25S ribosomal units of yerba mate: assembly, annotation and validation

Scaffolds containing ribosomal sequence signatures were identified using BLASTn (E-value cutoff: 1e-05) analysis with our reference rDNA sequences [6, 71]. Consensus sequences were generated for each ribosomal unit using the MAFFT sequence alignment tool in Geneious and subsequently validated by mapping short (cutoff: 3% mismatches) and long (>10,000 nt; cutoff: 30% mismatches) DNA-seq reads using the Geneious mapper. Genomic ribosomal units -genes and spacers- were reference-annotated [6, 71], and additional repetitive DNA features were identified through self-dot plot analysis. The resulting sequences were deposited in Repbase under the IDs 5S_rDNA_Ipa and 18S-25S_rDNA_Ipa.

RESULTS AND DISCUSSION

Chromosome-level reference assembly and validation in yerba mate

Two sub genomic structures were identified in the yerba mate nuclear genome: a diploid region comprising 85% of the genome and a tetraploid region comprising 15% (Figure 1a), thus proposing it as a predominantly diploid species consistent with cytological analysis [72]. This duplicated region appears to be a remnant of paleo-polyploidization events common in the evolution of *Ilex*, as reported in *I. asprella*, *I. latifolia*, and *I. polyneura* [17-19]. A major consequence of these findings for yerba mate geneticists and breeders is likely the loci exhibiting polysomic inheritance within the species, which complicates downstream analysis.

The yerba mate nuclear genome assembly YMM1 spanned 872.9 Mb in length, including 0.6 Mb of novel gaps. Of this assembly, 799.3 Mb (91.6%) were incorporated into 20 chromosomes (scaffold0000001-scaffold0000020), ranging in size from 27 to 71.7 Mb (Table 1; Figure 1b; File S1). The remaining 8.4%, which was not assembled to the reference, was grouped into the artificial sequence scaffold0000021 (Table 1; File S1). This reported genome size at the sequence level is 96.1% of the mean DNA content measured by Garberoglio and coauthors [14] for yerba mate, highlighting the value of YMM1 for the species, and surpassing the genome sizes of the other four sequenced diploid *Ilex* species (727.1-804.1 Mb) [17-20].

All sequences of YMM1 were annotated at intervals with the original names of the scaffolds used, along with their assembly coordinates. Furthermore, these 20 reference-assembled sequences also display the three distinct RagTag confidence score values at each interval, which is highly useful for detailed inspection of each region of the assembly (Files S2A and S2B). With respect to this, the mean confidence score for the YMM1 assembly was 0.84 (mean grouping: 0.81; mean location: 0.75; mean orientation: 0.97) (File S3), a high value -suggesting a low number of expected misassembled regions- considering the divergence time from *I. aquifolium* [26] and emphasizing the worth of this first chromosome-level assembly for the species. Overall, chromosome-level comparisons among *I. aquifolium* (n=20), *I. asprella* (n=19), *I. latifolia* (n=20), and *I. polyneura* (n=20) via Chromeister revealed nearly linear dot plot graphs, low divergence scores (mean 0.15), and a low number of detected rearrangements among them (Figure S1; File S10). This demonstrated an almost complete block collinearity at different levels, that is, synteny and microsynteny events, across those species. Hence, despite the divergence time of 30.4-32.7 million years for this group of taxa, part of a genus that is 84.7 million years old [26], a high correlation in genome structure is still evident. Considering the phylogenetic position of yerba mate within *Ilex* [26], it is expected that it shares the main karyotype features with the other four chromosome-level assembled species in the genus [17-20]. In this context, further comparisons of the 20 chromosome-level sequences of YMM1 to its reference *I. aquifolium* [20] and the other *Ilex* species [17-19] via Chromeister positively evaluated the ability of RagTag to replicate the typical chromosome structures in *Ilex* (Figure 1c; File S10).

The quality of a plant sequence assembly can also be assessed by comparing its predicted gene content to known Embryophyta and Eudicots gene sets, and then comparing those indices among more closely related species [23]. The genomic BUSCO completeness of the yerba mate assembly YMM1 reached 97.3%, with a low level of duplicated loci (4.6%) (Table 1). All genomic BUSCO scores observed in yerba mate were comparable to those of four other chromosome-level assembled *Ilex* species [17-20], characterized by mean 6.2% duplicated loci (File S4), highlighting the quality of the assembly YMM1. These scores at YMM1 showed a marked improvement compared to GCA_963454935.2 [13], which exhibits a mean of 22.2% duplicated loci (File S4), partly due to the purging step applied in this study (Files S5 and S6). This purging process did not negatively impact genome completeness or the repetitive repertoire (File S7). Furthermore, the percentage of species-specific reads mapped to YMM1 was also high, with 90.7% for RNA-seq and 98.1% for DNA-seq libraries, respectively (Files S8 and S9), indicating the quality of the assembly [73].

Masking analysis using the YMM1 library derived from RepeatModeler revealed 64.9% repetitiveness in the yerba mate assembly YMM1, a genome predominantly composed of LTR retrotransposons, as is typical in plants [35], along with a small fraction of non-LTR retrotransposons and Class II transposons (File S7). This library masked 64.6% of GCA_963454935.2 [13], which has a genome size 18% larger but a similar repetitive background to YMM1 (File S7). This emphasizes the representativeness of the chromosome-level assembly YMM1 regarding the repetitive repertoire of the original scaffold-level sequences [13], despite the purging step carried out. This library also masked a mean of 61.4% of genomic sequences in the other four

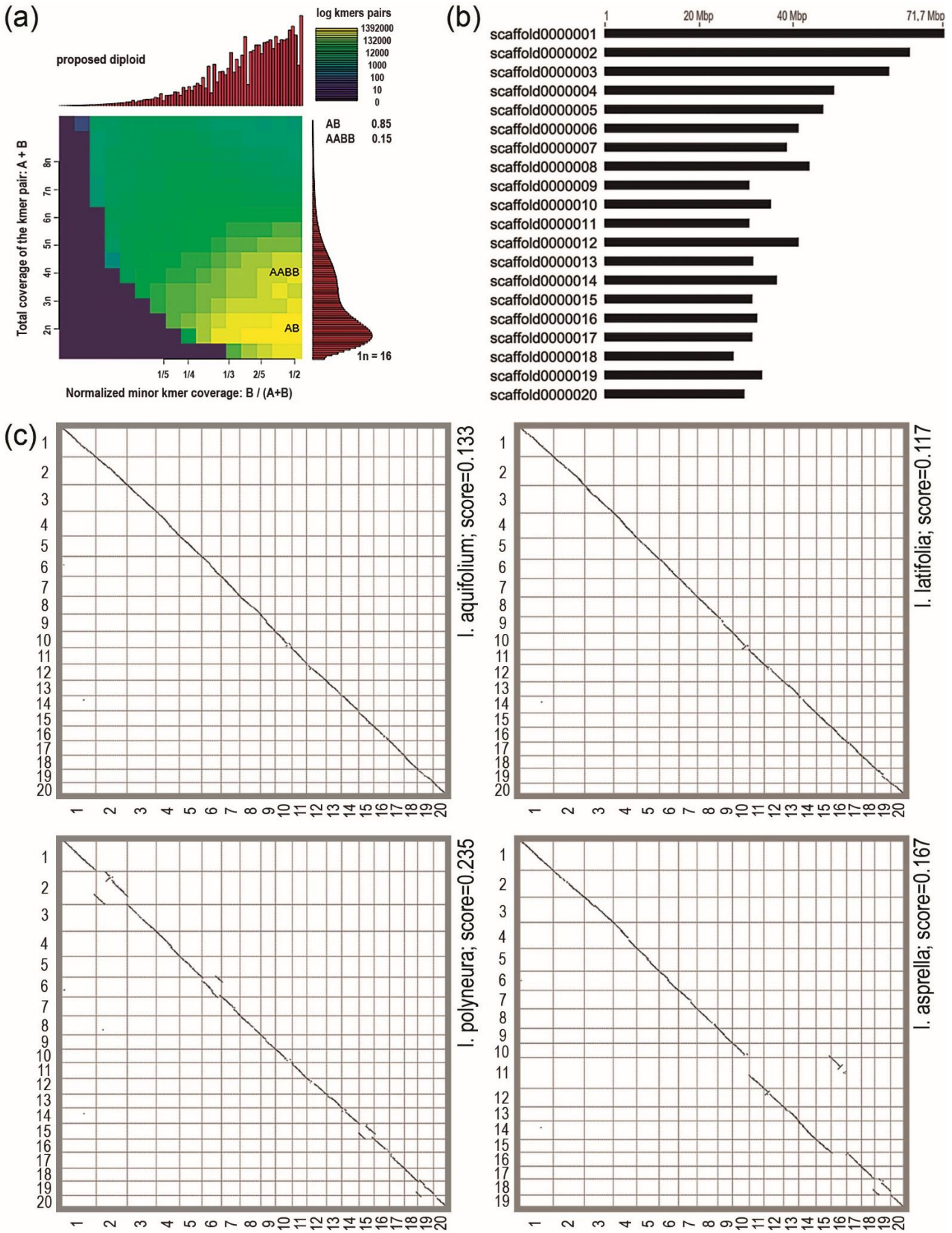

Figure 1. Nuclear genome characterization in yerba mate. (a) Smudgeplot graph showing ploidy level and sub genomic structures in yerba mate; (b) YMM1 20 chromosome-level sequences; (c) Chromeister dot plot graphs and scores from chromosome-level comparisons of YMM1 (x-axis) and other *Ilex* species (y-axis); note the high collinearity of chromosomes and respective supporting low-value scores; also note that chromosome 11 of *I. asprella* (n=19) matches chromosomes 11 and 16 of YMM1, a trait also evident when compared to other *Ilex* species (Figure S1).

chromosome-level assembled *Ilex* species [17-20], which shared a similar repetitive repertoire among themselves and with respect to YMM1 (File S7). Furthermore, masking analysis using the YMM1 library derived from Dante_LTR revealed that LTRs constitute 48.1% of the YMM1 genome. Within these LTR retrotransposons, gypsy elements accounted for 84.1%, while copia elements comprised 15.9% (File S7). The major LTR lineages in yerba mate, including Tekay, Athila, and Retand, collectively represented 80.2% of the gypsy clade in YMM1 (File S7). Additional masking analysis using the same YMM1 library revealed a similar LTR repertoire among the other four chromosome-level assembled *Ilex* species and GCA_963454935.2 [13, 17-20] with respect to YMM1 (File S7), thus enhancing the coherence of the achieved yerba mate assembly.

Table 1. Summary of core features of YMM1 nuclear assembly.

Genome size	872.9 Mbp
Assembly level	20 chromosomes and 1 artificial scaffold
Genomic BUSCO Embryophyta_odb10	C:97.4%[S:92.9%,D:4.5%],F:0.7%,M:1.9%,n:1614
Genomic BUSCO Eudicots_odb10	C:97.1%[S:92.5%,D:4.6%],F:0.5%,M:2.4%,n:2326

C: complete; S: single; D: duplicated; F: fragmented; M: missing; n: number of genes in the set.

Genome annotation and validation in yerba mate

The YMM1 nuclear assembly was annotated with 41,922 EVM-weighted consensus curated protein-coding gene models (Figure 2; Table 2; Files S11A-S11E), collectively occupying 23.1% of the genome space, with 93.5% distributed across 20 chromosomes, 76.1% containing introns, and 65.7% composed of 2 to 10 exons. A high number of these gene structures were further validated via RNA-seq analysis involving available species-specific libraries from leaves, roots, and plantlets (File S12), totaling 37,344 (89.1%). The number of annotated genes in YMM1 surpassed by 22.8% the 32,355 genes primarily identified in yerba mate at the leaf transcriptome level [4]. Yerba mate has a mean of 48 genes per megabase of its genome, a number consistent with those of other annotated *Ilex* species [17-19]: 48.3 in *I. asprella*, 46 in *I. latifolia*, and 44.4 in *I. polyneura*.

The quality of the YMM1 genome annotation was assessed employing a comprehensive approach involving multiple sources. Protein BUSCO completeness of YMM1, considering Embryophyta and Eudicots gene sets, reached 97.1%, with a low level of duplicated loci (6.3%) (Table 2; File S13). The consistency of genomic and protein BUSCO scores in YMM1 indicates that the genome annotation achieved the expected gene content. In addition, all protein BUSCO scores observed in yerba mate were comparable to those of three other annotated *Ilex* species [17-19], which are characterized by a mean of 7.6% duplicated loci (File S13), highlighting the quality of the YMM1 genome annotation. These scores for YMM1 showed a noticeable improvement compared to GCA_963454935.2 [13], which exhibits a mean of 77.5% complete and 19% duplicated loci, a significant difference from the mean genome annotation in *Ilex* (File S13).

The annotated gene set of YMM1 was further compared with the proteome of *A. thaliana*. Global analysis using Proteinortho identified 20,493 orthologous genes distributed into 12,224 orthogroups between them (File S14). These numbers are consistent with mean 20,873 orthologous genes to *A. thaliana* distributed into mean 12,441 orthogroups considering also the other three annotated species of *Ilex* [17-19] (File S14). The percentage of essential genes shared between the YMM1 proteome and *A. thaliana*, according to the DEG database [54], reached 88.5%, consistent with the percentages found when considering the proteomes of other sequenced *Ilex* species (File S15). In addition, the percentage of KOGs shared among the gene set of YMM1 and *A. thaliana*, according to the KOGs database [55], reached 96.1%, which is also consequent with the percentages found when the proteomes of other sequenced *Ilex* species [17-19] are considered (*Ilex* mean = 96.6%) (File S15).

Furthermore, a lineage-dependent proteome quality assessment of YMM1 and the other three annotated *Ilex* species [17-19] was carried out using OMark. Primarily, the proteomes were identified as belonging to the Asterids clade, to establish a reference [74]. This analysis found 93.6% conserved HOGs supported by 31,002 genes in YMM1 (File S16). These numbers are consistent with a mean of 94.5% conserved HOGs with respect to Asterids, supported by a mean of 29,910 genes when also considering the proteomes of the other three annotated *Ilex* species [17-19] (File S16). The considered proteomes in the genus average 28.9% duplicated HOGs compared to their reference Asterids (File S16), likely related to the whole-genome duplication events common in the evolution of *Ilex*, as commented previously [17-19].

Orthology comparisons were also performed considering the proteomes of YMM1 and the other three annotated *Ilex* species [17-19] using Proteinortho and OrthoFinder. The former analysis classified the overall genes in *Ilex* into 27,303 orthology groups, including groups with two (18.4%), three (23.2%), or four (58.4%)

species each (File S17). In this sense, yerba mate participates in 83.2% of those orthology groups, involving 25,930 genes (File S17). When the other *Ilex* species are considered [17-19], the annotated genes in the genus participate in a mean of 85% of those orthology groups, involving a mean of 25,833 genes (File S17), demonstrating coherence in the YMM1 annotation. In turn, a more detailed analysis by OrthoFinder restricted the classification of the whole genes in *Ilex* to 23,650 orthology groups, including groups with two (15.1%), three (18.4%), or four (66.5%) species each (File S18). Hence, the YMM1 proteome participates in 88.1% of those orthology groups, involving 35,131 genes (File S18). Considering the other *Ilex* species [17-19], the annotated genes in the genus participate in a mean of 87.8% of those orthology groups, involving a mean of 33,419 genes (File S18), in agreement with the YMM1 annotation. Finally, 82% of the orthology groups in YMM1 involved one (54.8%), two (19%), three (5.4%), four (1.9%), or five (0.9%) gene members, respectively, and contained 76.8% of its genes (File S18). Comparable parameters to YMM1 were found across the genus, where a mean of 79.7% of the orthology groups per species involved one (56.9%), two (16.2%), three (4.4%), four (1.6%), or five (0.7%) gene members, respectively, and contained 80.8% of their genes (File S18). Altogether, comparative orthology analysis in *Ilex* validated 35,491 (84.7%) YMM1 EVM genes (File S19).

Table 2. Summary of core features of YMM1 nuclear assembly annotation

# EVM gene models	41,922
# Alternative splice variants	23,089
# Genes with alternative splice variants	10,446
# Genes with UTRs	18,100
# tRNA genes	500
Protein BUSCO Embryophyta_odb10	C:97.1%[S:91.3%,D:5.8%],F:1.0%,M:1.9%,n:1614
Protein BUSCO Eudicots_odb10	C:97.0%[S:90.3%,D:6.7%],F:0.7%,M:2.3%,n:2326

#: number; C: complete; S: single; D: duplicated; F: fragmented; M: missing; n: number of genes in the set.

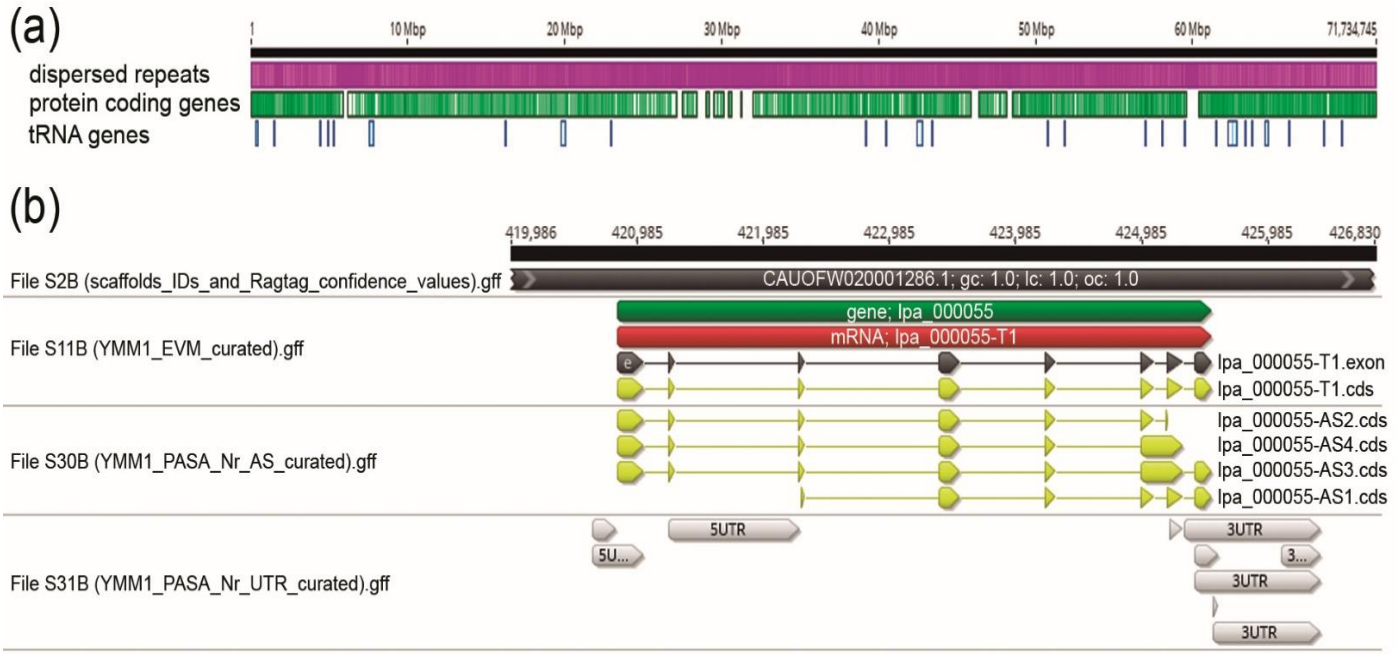

Figure 2. YMM1 nuclear assembly annotation. (a) Representation of scaffold0000001 and respective density maps of dispersed repeats, protein-coding genes, and tRNA genes; (b) Selected region of scaffold0000001 displaying annotation tracks and respective annotated features. Ipa_000055 is a predicted peptidase gene from the M48 family, an activator of leaf senescence under normal and stress conditions; its suppression delays leaf aging.

Functional annotation of the YMM1 EVM proteome utilized diverse and complementary tools and databases (Table 3), which together annotated and validated 38,940 genes (92.9%) (File S19). A molecular characterization of the EVM proteins was also carried out using iep and pepstats tools, considering their molecular weight, charge, isoelectric point, and other parameters (Files S20 and S21). The updated and widely used orthology-based functional sequence annotation protocol EggNOG-Mapper [58] was applied to the YMM1 EVM proteome and the proteomes of the other three annotated *Ilex* species [17-19] for further comparison. This analysis in-depth characterized 32,085 YMM1 EVM genes (76.5%) (File S22). Considering those other species of the genus, a mean of 31,581 genes (85.2%) in *Ilex* found orthology at the EggNOG database, in agreement with the mentioned mean of 29,910 annotated genes (80.7%) in the genus orthologous to Asterids (File S16). This 15-19% proportion of non-orthologous genes in *Ilex* likely reflects that Aquifoliaceae and related species are still underrepresented in reference gene sets from major tools and databases. In this regard, yerba mate holds more non-orthologous genes than other annotated *Ilex* species [17-19], correlated with a larger genome size and gene content. In particular to the classification of the YMM1 EVM proteome, EggNOG-Mapper extracted description labels and GO terms for 29,676 genes (*Ilex* mean = 29,313) and 16,228 genes (*Ilex* mean = 15,498), respectively, and classified 7,052 enzymes by EC number (*Ilex* mean = 6,801) (File S22). This analysis also extracted KO terms and BRITE identifiers for 15,800 genes (*Ilex* mean = 15,277) and pathway identifiers for 9,907 genes (*Ilex* mean = 9,549) from KEGG databases, and annotated 28,790 genes via the PFAM database (*Ilex* mean = 28,444), among others (File S22). This consistency in the classification of genes by EggNOG-Mapper, considering yerba mate and the proteomes of the other three annotated *Ilex* species [17-19], reinforces the quality of the YMM1 genome annotation observed throughout the validation process.

A subsequent functional classification of the YMM1 EVM proteins into families and the prediction of domains and important sites were achieved using InterPro through signatures provided by different member databases [59]. In the Galaxy environment, this integrated database and diagnostic tool extracted 285,866 signatures from 17 distinct databases for 33,656 genes, and also provided description labels for 33,538 genes, GO terms for 20,091 genes, and pathway identifiers for 23,178 genes (File S23). In turn, in the Gensas environment, InterPro extracted 242,268 signatures from 10 diverse databases for 30,806 genes, and linked description labels for 34,367 genes and GO terms for 20,682 genes (File S24). In addition, a functional classification of proteins was achieved using the PFAM database alone in the Gensas environment, which extracted 151,045 signatures for 29,848 yerba mate genes (File S24). All these resources combined finally annotated 35,719 (85.2%) yerba mate genes (File S19), a good number considering that within the model species *A. thaliana*, 4.5% of protein-coding loci remain functionally unannotated [75].

To further characterize YMM1 EVM proteins, the signal peptide prediction tools SignalP, DeepSig, and TargetP were used in combination to finally identify 4,625 genes (File S25) of the secretory pathway [77]. TargetP also identified 2,955 proteins in yerba mate possessing transit peptides targeting chloroplasts (54.8%), mitochondria (40.1%), or thylakoid lumen (5.2%) (File S25). Furthermore, WoLF PSORT predicted the subcellular localization of 11,641 proteins with a score >10 to twelve different compartments, encompassing 24 sites including dual-localized proteins. These sites included the nucleus (38.5%), chloroplast (33.2%), cytosol (16%), plasma membrane (6.2%), and extracellular space (2.5%), among others (File S25). A nucleolar localization signal (NoLS) is a short amino acid sequence that directs a protein to the nucleolus [61]. In this sense, the nucleolar localization sequence detector NoD predicted 7,696 proteins with NoLS in yerba mate (File S25).

Further characterization of the YMM1 EVM proteins was achieved through comparison with the most important global databases, such as NCBI RefSeq Plant, UniProtKb/Swiss-Prot, NCBI Nr, and Darwin ToL. These combined approaches identified 34,125 (81.4%) yerba mate proteins homologous to proteins in other species (Files S19 and S26), surpassing the percentage of annotated proteins in corresponding analyses of yerba mate at the transcriptome level [4, 9], and in agreement with *I. polyneura* [19].

Transcription factors (TFs) are proteins that regulate the expression of target genes by binding to specific cis-elements in promoter regions [65]. To further achieve a genome-wide annotation and classification of TFs in the YMM1 EVM proteome, two distinct but complementary approaches were followed: PlantTFDB [64] and iTAK [65]. PlantTFDB identified 1,928 TF genes and classified them into 58 families, while iTAK identified 1,991 TF genes and classified them into 70 families, with bHLH, MYB, ERF, NAC, and C2H2 as the largest clusters, each including more than 100 members in yerba mate (File S27). Hence, the number of TF genes in yerba mate is consistent with the mean of 1,959 TF genes found in Eudicot genomes [64], highlighting the completeness of the achieved YMM1 genome annotation in this valuable crop species. Most important to yerba mate culture are those TF families associated with stress, such as NAC, MYB, ERF, WRKY, BZIP, bHLH, and HSF [78, 79], which altogether reached 898 genes considering both classification approaches

(File S27). In addition, transcriptional regulators (TRs) are proteins that also regulate the expression of target genes, but indirectly [65]. In this regard, iTAK annotated 436 TR genes and classified them into 24 families in the YMM1 EVM proteome (File S27), in agreement with the 428 TR genes found in *A. thaliana* [65]. Kinases are another type of regulatory proteins that function in signal transduction pathways and alter the activity of target proteins by phosphorylating them [65]. Considering the YMM1 EVM annotation, iTAK identified 1,179 kinase genes, consistent with the 1,028 kinase genes of *A. thaliana* [65]. These yerba mate kinases were further distributed into 20 supported groups, including RLK-Pelle (File S27). Their genes are significant to yerba mate due to their involvement in pathogen recognition and plant disease resistance processes, among others [80].

To summarize, when considering overall evidence from diverse analyses such as RNA-seq, comparative orthology in *Ilex* and functional annotation, 72% of YMM1 EVM genes matched 18 to 10 sources, with 374 genes solely annotated by EVM, and validation reached to 41,548 gene models (99.1%) (File S19).

Table 3. Summary of the functional annotation of YMM1 EVM genes

Source	# genes	Source	# genes
InterPro Gensas	34,381	WoLF PSORT score >10	11,641
NCBI Nr	33,885	NoD	7,696
InterPro Galaxy	33,656	TargetP	7,215
NCBI Refseq Plants	33,054	SignalP	3,125
Eggnogmapper	32,085	DeepSig	2,626
Darwin ToL all proteins	32,016	iTAK-TF-TR	2,427
Pfam Gensas	29,848	PlantTFDB	1,928
UniprotKB/Swiss-Prot	27,145	iTAK-PK	1,179

#: number; TF-TR: transcription factors/transcriptional regulators; PK: protein kinases.

Furthermore, supported by the achieved set of unique transcripts of yerba mate (Files S28A-S28C and S29) and the YMM1 EVM gene models, a PASA locus non-redundant curated set of alternative splice variants and UTR sequences was generated (Files S30A-S30D and S31A-S31B), an original approach among the other three annotated *Ilex* species [17-19]. In brief, 78,992 UTR sequences, including 36,407 5' UTRs and 42,945 3' UTRs, were annotated for 18,100 genes (Table 2). Besides, 23,089 alternative splice variants were annotated for 10,446 genes (Table 2). The corresponding protein set was submitted to the same functional annotation steps as the EVM genes, and 22,987 (99.6%) could be further labeled and molecularly characterized (Files S32-S37).

Moreover, 500 tRNA genomic loci were also annotated in the YMM1 nuclear assembly (Table 2; Files S38A and S38C). These tRNA genes involved 56 diverse anticodons predicted to target the 20 standard amino acids, in addition to a suppressor tRNA (File S39), in agreement with previous findings in plants [67]. The number of tRNA loci in yerba mate is also consistent with the 494 and 448 tRNA genes found in *I. asprella* and *I. latifolia*, respectively [17, 18].

Chloroplast and mitochondrion genomes of yerba mate: assembly, annotation and validation

The YMM1 chloroplast genome was assembled into a single contiguous molecule of 157,591 bases in length (9,518.3X coverage), a length that slightly differs from those obtained using other RNA and DNA assembly approaches [4, 69]. High-coverage mapping of short (404.2X) and long (85.2X) DNA-seq reads validated this sequence (Figure 3; Table 4; Files S40-S43). Compared to 93 non-redundant plastid sequences of *Ilex* species at NCBI, the YMM1 plastome exhibits a typical structure for the genus in terms of length (mean 157,668 bases), nucleotide composition (mean 99.2% pairwise identity), number of sub genomic regions (4), number of repeat loci (811), and number of annotated genes (131), which are distributed into protein-coding genes with known function (86), tRNA genes (37), and rRNA genes (8) (File S44). RNA-seq mapping at 1,039.1X coverage revealed a single contiguous transcript matching its genomic region and 53 C-to-T RNA editing sites distributed within protein-coding genes (File S45). Furthermore, 14 different regions transferred from the chloroplast to the mitochondrial genome (see below), totaling 6,789 bases, were detected in the large single-copy region of the plastome (File S41).

The YMM1 mitochondrial genome was initially assembled into a single contiguous molecule of 521,271 bases in length (14,387.9X coverage) and subsequently validated by 233.8X coverage of mapped short reads. However, mapping of long reads (23.4X coverage) revealed two separate sub genomic molecules of 322,393 bases (A) and 198,878 bases (B) in length, respectively (Figure 4; Table 4; Files S46-S49), consistent with findings in diverse plants other than *Ilex* [81, 82]. In contrast to animals, plant mitogenomes are typically larger and composed of multiple chromosomes capable of rearranging to generate diverse isoforms [83, 84].

The YMM1 mitogenome was annotated with 65 genes, including protein-coding genes with known function (40), tRNA genes (22), and rRNA genes (3), in addition to 2,439 low complexity and SSR loci (File S47). MAUVE comparisons further revealed five inverted repeat loci and three direct repeat loci in each sub genome, with the latter constituting validated A-B junction sites based on short-read mapping alone (File S47). Specifically, sub genome A contained transferred regions from the chloroplast, totaling 4,907 bases and including three tRNA genes (File S47). RNA-seq mapping at 192.9X coverage revealed multiple primary transcripts along almost the entire mitogenome length (excluding 623 bases) and 578 C-to-T RNA editing sites distributed within protein-coding genes (File S50). Multipartite transcription and trans-splicing phenomena, recently described for the *Arabidopsis* mitogenome [81], could help explain the transcriptional landscape observed in yerba mate, including the processing of the *nad1* and *nad2* genes with large introns and exons distributed across both sub genomes, a phenomenon also reported in rice [82]. The mitogenomes of the five *Ilex* species at NCBI showed comparable lengths (mean 535,627 bases) and numbers (mean 62) and categories of annotated genes (mean 39 protein-coding, mean 20 tRNA, mean 3 rRNA) to YMM1 mitogenome. According to mapped reads, a mean of 69.1% of the YMM1 mitogenome targeted sequences of already characterized *Ilex* species (File S51). Moreover, whole mitogenome comparisons in *Ilex* using Chromeister and MAUVE revealed corresponding sequence similarity but weak synteny among species (Files S52-S53; Figures S2-S7). In this regard, further analyses will be needed to assess the currently captured variability at the assembly level in *Ilex* mitogenomes.

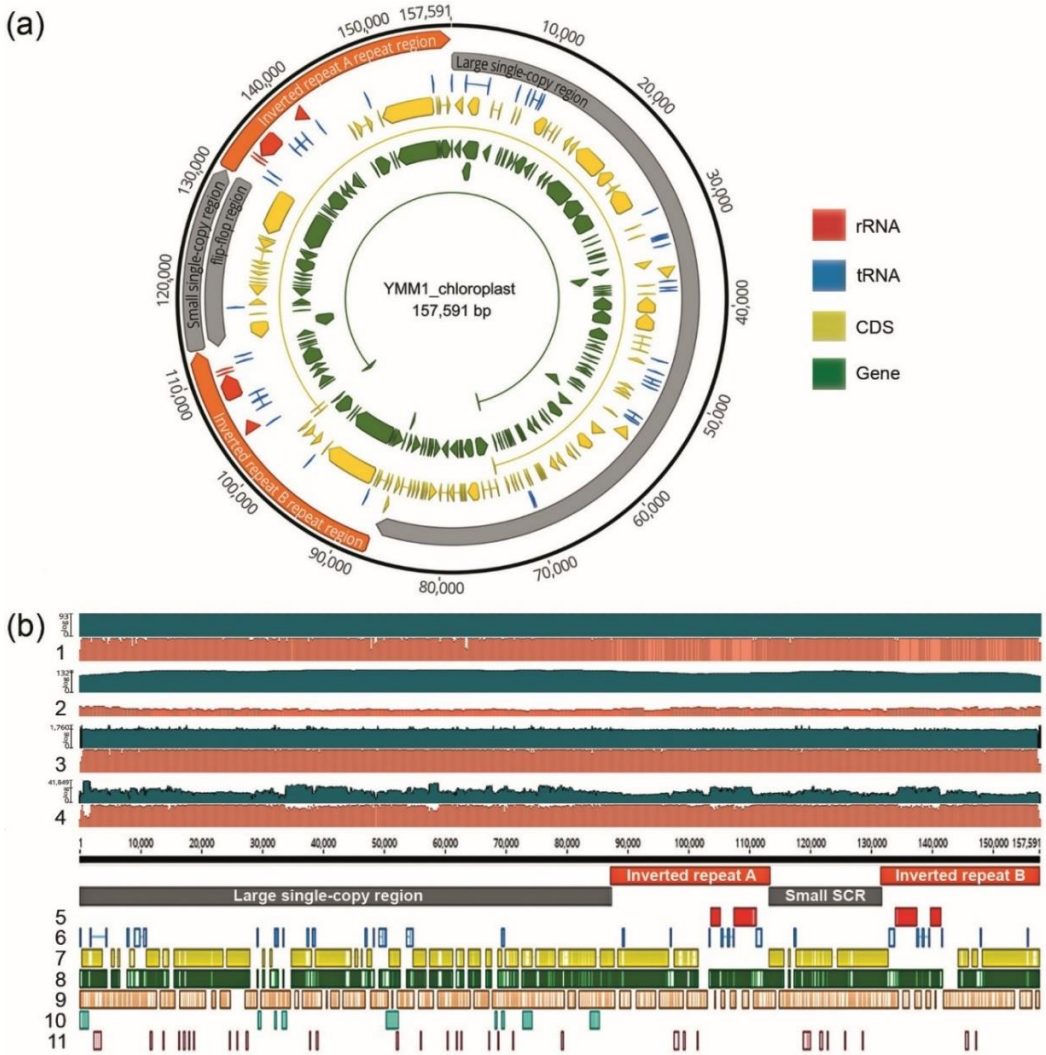

Figure 3. YMM1 plastome characterization. (a) Circular map of the chloroplast genome displaying the four sub genomic regions and annotated gene features; note that the small single-copy region can flip-flop between two configurations validated by read mapping; (b) Linear representation of the plastome and sub genomic regions, with graphs above showing sequence coverage (blue) and identity (orange) corresponding to the mapping of 93 non-redundant *Ilex* plastomes (1), long reads (2), short reads (3), and RNA-seq data (4) to the YMM1 structure; below are density maps displaying rRNA genes (5), tRNA genes (6), PCGs (7), overall genes (8), repeats (9), regions shared with the mitogenome (10), and RNA editing sites in PCGs (11).

Table 4. Summary of core features of YMM1 plastome and mitogenome.

	plastome	mitogenome
Genome size	157,591 bases	521,271 bases
Assembly level	1 chromosome	2 chromosomes
# Overall genes	131	65
# Protein-coding genes	86	40
# tRNA genes	37	22
# rRNA genes	8	3
# RNA editing sites at PCG	53	578
# Low complexity and SSR loci	811	2,439

#: number; PCG: protein-coding genes.

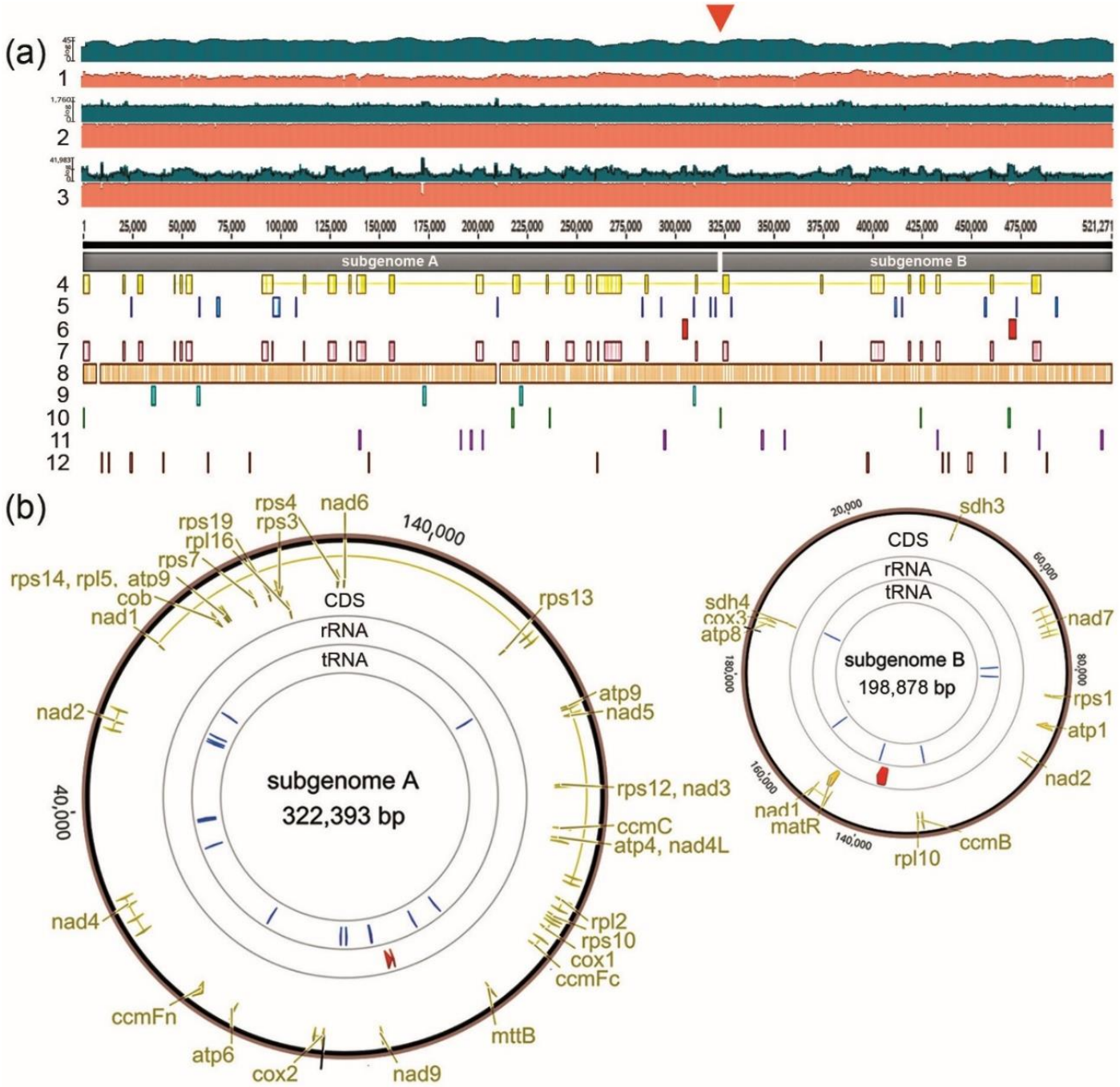

Figure 4. YMM1 mitogenome characterization. (a) Linear representation of the full mitochondrial genome with actually separated sub genomes A and B (red arrowhead), with graphs above showing sequence coverage (blue) and identity (orange) corresponding to the mapping of long reads (1), short reads (2), and RNA-seq data (3) to the YMM1 structure; below are density maps displaying PCGs (4), tRNA genes (5), rRNA genes (6), RNA editing sites in PCGs (7), repeats (8), regions shared with the plastome (9), A-B direct repeats (10), A-B inverted repeats (11), and regions lacking mapped RNA (12); (b) Circular maps of YMM1 mitogenome sub genomes A and B and respective annotated gene features.

5S and 18S-25S ribosomal units of yerba mate: assembly, annotation and validation

The YMM1 5S and 18S-25S ribosomal nuclear sequences were assembled into corresponding units and then validated via high-coverage mapping of DNA reads: short reads at 86,242X and 19,407X coverage, and long reads at 515.7X and 7,161.9X coverage, respectively (Figure 5; Files S54-S57).

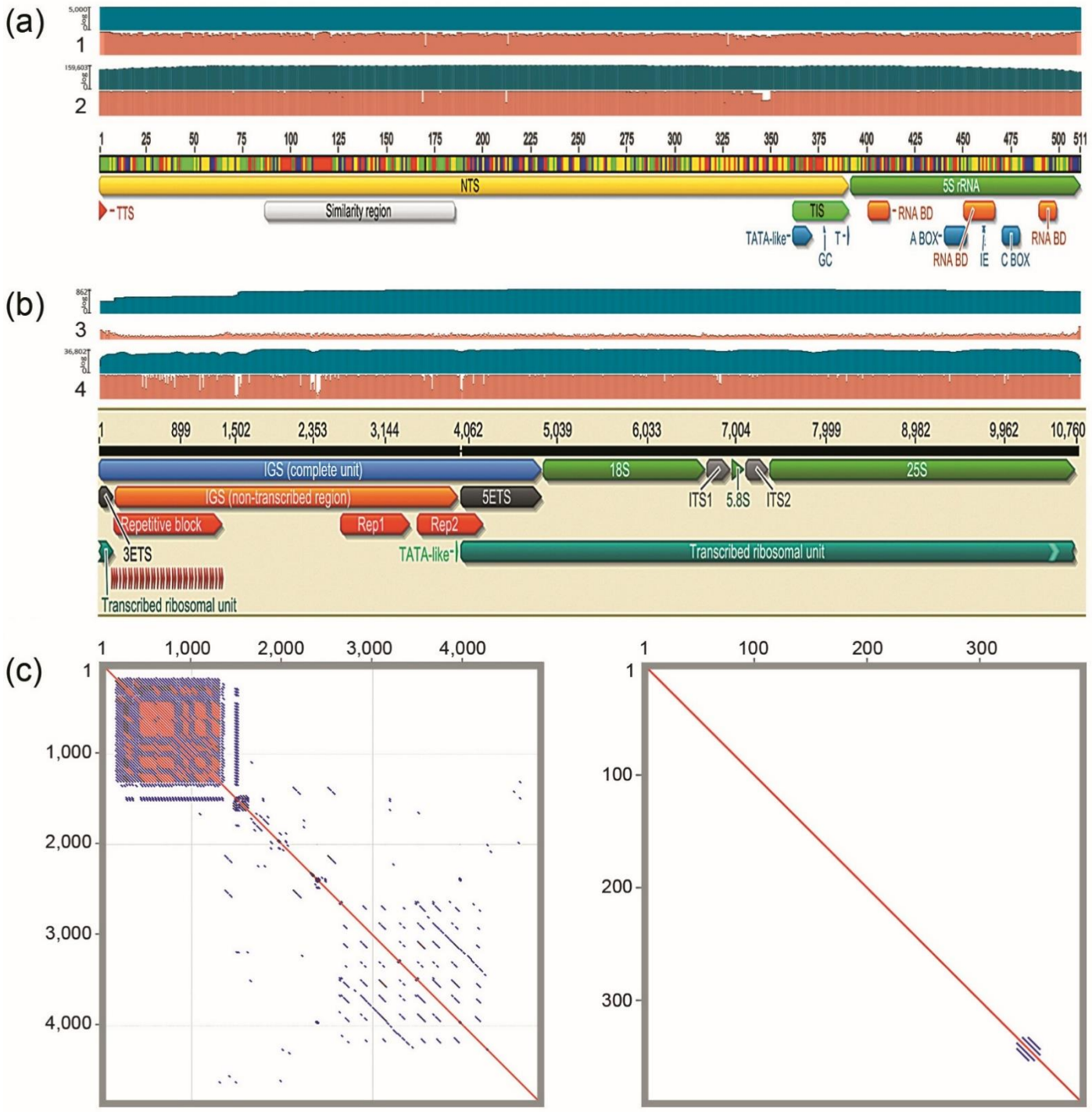

Figure 5. YMM1 nuclear ribosomal units' characterization. (a) Linear representation of the 5S unit with respective annotated features; (b) Linear representation of the 18S-25S unit with respective annotated features. Above each ribosomal unit illustration are graphs showing sequence coverage (blue) and identity (orange) corresponding to the mapping of long reads (1, 3) and short reads (2, 4) to each structure, respectively; (c) Self-dot plot graphs of each ribosomal unit, 18S-25S (left) and 5S (right); note the two major repetitive regions in the IGS of the 18S-25S ribosomal unit.

The 5S rDNA unit was 511 bases in length, with 391 bases corresponding to the non-transcribed spacer (NTS) and 120 bases to the rRNA gene (Table 5). Possible transcription initiation sites (TIS), including the TATA-like sequence (-30), the GC dinucleotide (-14), and the C nucleotide (-1), as well as a transcription

Table 5. Summary of core features of YMM1 5S and 18S-25S ribosomal nuclear sequences.

	5S rDNA unit	18S-25S rDNA unit
Ribosomal unit size	511 bases	10,760 bases
Gene types (# bases)	5S (120)	18S (1,809), 5.8S (159), 25S (3,375)
Spacer types (# bases)	NTS (391)	IGS (4,883), 5ETS (909), 3ETS (175), ITS1 (269), ITS2 (265)

#: number.

termination site (TTS), a poly-T stretch of 5 bases, were recognized in the NTS (Figure 5a). Typical RNA binding domains (RNA BD) and transcription-associated features for RNA polymerase III at the 5S rRNA gene, such as the A Box, the intermediate element (IE), and the C Box, were also predicted (Figure 5a). Notably, a 100 bases stretch in the middle region of the NTS showed 83% similarity to the internal portion of a gypsy retrotransposon from yerba mate, but the entire NTS lacked any major repetitive feature (Figure 5a, c). In addition, the 18S-25S rDNA unit was 10,760 bp in length, with 4,883 bases corresponding to the intergenic spacer (IGS) and 5,877 bases to the rRNA genes 18S, 5.8S, and 25S, along with the internal transcribed spacers (ITS1 and ITS2) (Table 5; Figure 5b). The transcribed region from this genomic sequence, which also included the external transcribed spacers (5ETS and 3ETS) from the IGS, was found to be 99.9% identical to the reported RNA sequence in yerba mate (File S58) [6]. Two major repetitive regions were revealed in the IGS, near its 5' end: a 1,211 bp block composed of 38 repetitions of 32 bases each with 96% similarity; and close to its 3' end: two direct repeat stretches of 791 and 750 bases that shared 76.4% identity and contained the TATA-like sequence (Figure 5b, c).

Ribosomal units, along with plastome and mitogenome sequences and annotated features, constitute a significant source of useful markers for comparative analysis within yerba mate, *Ilex*, and beyond, as is common for these regions throughout plants [86].

CONCLUSION

Original chromosome-level assemblies were achieved for the nuclear genome, plastome, and mitogenome of yerba mate, in addition to assembling the nuclear ribosomal units. These genomic sequences were further characterized in structure and function to generate a set of high-quality annotated markers now openly available. In particular, 93.5% of characterized nuclear protein-coding genes were annotated onto the 20 linkage groups of *Ilex paraguariensis*, and this chromosomal map constitutes a milestone for this valuable crop species. Yerba mate is significant to producer countries that demand increased yield, tolerance to diseases, and adaptation to stress conditions, in addition to finding technological applications for its bioactive compounds. This valued original set of sequences and markers is intended to support future genetic analyses, improve selection strategies in breeding programs, and serve as a platform for genomics research in yerba mate.

Funding: This research received no external funding.
Acknowledgments: We are grateful to those teams of the Universidad Nacional de Misiones, Universidad Nacional del Nordeste, Universidad de Buenos Aires and ProMateAr (Argentina) to generate and release the nucleotide sequences of yerba mate valuable to this contribution. We are indebted with teams of Galaxy and Gensas to made openly available their servers and bioinformatic tools and also with Repbase to their assist in curate repetitive sequences. We also want to thank to the Consejo Nacional de Investigaciones Científicas y Técnicas (Argentina), from which we both are associate researchers, for continuous support. This contribution is dedicated to our beloved Pamela.
Conflicts of Interest: The authors declare no conflict of interest.
Supplementary Material: 10.17632/8v9ws627tg.1 [87], 10.17632/kgz37pgpxw.1 [88], and https://ngdc.cncb.ac.cn/ (accession OMIX007912, BioProject PRJCA032295) host 104 supplementary files comprising *Ilex paraguariensis* genome assemblies (fasta, gbff), corresponding annotation tracks (tabular, GFF, gbff), nucleotide and translated sequences, unique transcripts with genome coordinates, as well as validation and functional annotation datasets in multiple formats, all accompanied by detailed file legends.
Data Availability Statement: Research data are available in the repositories https://data.mendeley.com/datasets/8v9ws627tg/1, https://data.mendeley.com/datasets/kgz37pgpxw/1 and https://ngdc.cncb.ac.cn/omix/release/OMIX007912.

REFERENCES

1. Croge CP, Cuquel FL, Pintro PTM. Yerba mate: cultivation systems, processing and chemical composition. A review. Sci Agric. (Piracicaba, Braz) [Internet]. 2021 [cited 2024 Oct 5];78(5):e20190259. Available from: https://doi.org/10.1590/1678-992X-2019-0259
2. Gortari F, Londero WO, Rocha P, Niella F. Growth and physiological responses of yerba mate seedlings and mini-cuttings under drought stress. CERNE [Internet]. 2020 Jul [cited 2024 Oct 5];26(3):341-8. Available from: https://doi.org/10.1590/01047760202026032740
3. Gerber T, Nunes A, Moreira BR, Maraschin M. Yerba mate (*Ilex paraguariensis* A. St.-Hil.) for new therapeutic and nutraceutical interventions: A review of patents issued in the last 20 years (2000–2020). Phytother Res. 2023 Feb;37(2):527-48.
4. Debat HJ, Grabiele M, Aguilera PM, Bubillo RE, Otegui MB, Ducasse DA, et al. Exploring the genes of yerba mate (*Ilex paraguariensis* A. St.-Hil.) by NGS and de novo transcriptome assembly. PLoS One [Internet]. 2014 Oct [cited 2024 Oct 5];9(10):e109835. Available from: https://doi.org/10.1371/journal.pone.0109835
5. Debat HJ, Grabiele M, Aguilera PM, Bubillo RE, Zapata PD, Marti DA, et al. The complete genome of a putative endornavirus identified in yerba mate (*Ilex paraguariensis* St. Hil.). Virus Genes. 2014 Oct;49:348-50.
6. Aguilera PM, Grabiele M, Debat HJ, Bubillo RE, Marti DA. The 18S–25S ribosomal RNA unit of yerba mate (*Ilex paraguariensis* A. St.-Hil.). Plant Biosyst. 2015 Mar 11;150(6):1240-8.
7. Aguilera PM, Debat HJ, Grabiele M. Dataset of the first transcriptome assembly of the tree crop “yerba mate” (*Ilex paraguariensis*) and systematic characterization of protein coding genes. Data Brief [Internet]. 2018 Feb 10 [cited 2024 Oct 5];17:1036-40. Available from: https://doi.org/10.1016/j.dib.2018.02.015
8. Aguilera PM, Debat HJ, Castrillo ML, Bich GA, Grabiele M. The DNAJ gene family in yerba mate (*Ilex paraguariensis*): genome-wide identification, structural characterization, orthology based classification and expression analysis. Rodriguésia [Internet]. 2023 [cited 2024 Oct 5];74:e00492022. Available from: https://doi.org/10.1590/2175-7860202374020
9. Fay JV, Watkins CJ, Shrestha RK, Litwiñiuk SL, Talavera Stefani LN, Rojas CA, et al. Yerba mate (*Ilex paraguariensis* A. St.-Hil.) de novo transcriptome assembly based on tissue specific genomic expression profiles. BMC Genomics [Internet]. 2018 Dec [cited 2024 Oct 5];19:891. Available from: https://doi.org/10.1186/s12864-018-5240-6
10. Acevedo RM, Avico EH, González S, Rodrigues Salvador A, Rivarola M, Paniego N, et al. Transcript and metabolic adjustments triggered by drought in *Ilex paraguariensis* leaves. Planta. 2019 Aug;250:445-62.
11. Avico EH, Acevedo RM, Duarte MJ, Rodrigues Salvador A, Nunes-Nesi A, Ruiz OA, et al. Integrating transcriptional, metabolic, and physiological responses to drought stress in *Ilex paraguariensis* roots. Plants (Basel) [Internet]. 2023 Jun [cited 2024 Oct 5];12(13):2404. Available from: https://doi.org/10.3390/plants12132404
12. Aguilera PM, Grabiele M. ProMateAr: *Ilex paraguariensis* sequencing project. A summary [dataset]. 2024 May 2 [cited 2024 Oct 5]. Mendeley Data. Available from: https://data.mendeley.com/datasets/3pp9jkw9rk/1 Referenced in doi: 10.17632/3pp9jkw9rk.1
13. NCBI, National Center for Biotechnology Information [Internet]. [GCA_963454935.2]. Genome assembly ILEXP v3. Bathesda, Maryland, USA; 2024 Mar 5 [cited 2024 Oct 5]. Available from: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/963/454/935/GCA_963454935.2_ILEXP_v3/
14. Garberoglio MJ, González GE, Kryvenki MA, Gottlieb AM. [Genome size variation of southern South American species of *Ilex* (Aquifoliaceae)]. Darwiniana [Internet]. 2023 Jul 31 [cited 2024 Oct 5];11(1):167-179. Spanish. Available from: https://www.ojs.darwin.edu.ar/index.php/darwiniana/article/view/1106
15. Rossi SM, Ferri CA, Zapata PD. First report of pseudogenes of SABATH enzymes from xanthines metabolic pathway in *Ilex paraguariensis*. RECyT. 2023 Dec 18;40(1):14-24.
16. Vignale FA, Hernandez Garcia A, Modenutti CP, Sosa EJ, Defelipe LA, Oliveira RRM, et al. Yerba mate (*Ilex paraguariensis*) genome provides new insights into convergent evolution of caffeine biosynthesis. bioRxiv [Preprint]. 2023 [cited 2023 Oct 5]: [45 p.] Available from: https://www.biorxiv.org/content/10.1101/2023.09.08.556846v1?versioned=true
17. Kong BL, Nong W, Wong KH, Law ST, So WL, Chan JJ, et al. Chromosomal level genome of *Ilex asprella* and insight into antiviral triterpenoid pathway. Genomics [Internet]. 2022 May [cited 2024 Oct 5];114(3):110366. Available from: https://doi.org/10.1016/j.ygeno.2022.110366
18. Xu K-W, Wei X-F, Lin C-X, Zhang M, Zhang Q, Zhou P, et al. The chromosome-level holly (*Ilex latifolia*) genome reveals key enzymes in triterpenoid saponin biosynthesis and fruit color change. Front Plant Sci. [Internet]. 2022 Aug 22 [cited 2024 Oct 5];13:982323. Available from: https://doi.org/10.3389/fpls.2022.982323
19. Yao X, Lu Z, Song Y, Hu X, Corlett RT. A chromosome-scale genome assembly for the holly (*Ilex polyneura*) provides insights into genomic adaptations to elevation in Southwest China. Hortic Res. [Internet]. 2022 Jan 5 [cited 2024 Oct 5];9:uhab049. Available from: https://doi.org/10.1093/hr/uhab049
20. NCBI, National Center for Biotechnology Information [Internet]. [GCA_951799425.1]. Genome assembly drIleAqui2.1. Bathesda, Maryland, USA; 2023 Jun 4 [cited 2024 Oct 5]. Available from: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/951/799/425/GCA_951799425.1_drIleAqui2.1/
21. galaxyproject.org [Internet]. The Galaxy Community; 2022 [cited 2024 Oct 5]. Available from: https://galaxyproject.org/

22. Humann JL, Lee T, Ficklin S, Main D. Structural and functional annotation of eukaryotic genomes with GenSAS. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York: Humana Press; 2019. p. 29-51.
23. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021Oct;38(10):4647-54.
24. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20.
25. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. [Internet]. 2020 Mar 18 [cited 2024 Oct 5];11(1):1432. Available from: https://doi.org/10.1038/s41467-020-14998-3
26. Yao X, Song Y, Yang J-B, Tan Y-H, Corlett RT. Phylogeny and biogeography of the hollies (*Ilex* L., Aquifoliaceae). J Syst Evol. 2021Jan;59(1):73-82.
27. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020 May 1;36(9):2896-8.
28. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094-3100.
29. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. [Internet]. 2022 Dec 15 [cited 2024 Oct 5];23:258. Available from: https://doi.org/10.1186/s13059-022-02823-7
30. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015 Mar 09;12(4):357-60.
31. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience [Internet]. 2021 Feb [cited 2024 Oct 5];10(2):giab008. Available from: https://doi.org/10.1093/gigascience/giab008
32. Pérez-Wohlfeil E, Diaz del Pino S, Trelles O. Ultra-fast genome comparison for large-scale genomic experiments. Sci Rep. [Internet]. 2019 Jul 16 [cited 2024 Oct 5];9(1):10274. Available from: https://doi.org/10.1038/s41598-019-46773-w
33. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA. 2020 Apr 16;117(17):9451-7.
34. Smit A, Hubley R, Green, P. RepeatMasker Open-4.0. [Internet]. 2013 [cited 2024 Oct 5]. Available from: http://www.repeatmasker.org
35. Neumann P, Novák P, Hoštáková N, Macas J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mob DNA [Internet]. 2019 Jan 3 [cited 2024 Oct 5];10:1. Available from: https://doi.org/10.1186/s13100-018-0144-1
36. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003 Oct 1;31(19):5654-66.
37. Pertea M, Pertea GM, Antonescu CM1, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015 March;33(3):290-5.
38. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience [Internet]. 2019 Sep [cited 2024 Oct 5];8(9):giz100. Available from: https://doi.org/10.1093/gigascience/giz100
39. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011 May 15;29(7):644-52.
40. Palmer JM, Stajich J. Funannotate v1.8.1: Eukaryotic genome annotation [dataset]. 2020 Sep 28 [cited 2024 Oct 5]. Zenodo. Available from: https://doi.org/10.5281/ZENODO.4054262 Referenced in doi: 10.5281/zenodo.1134477
41. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008 Jan;18(1):188-96.
42. Lomsadze A, Ter-Hovhannisyan V, Chernoff Y, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005 Nov;33(20):6494-506.
43. Pertea M, Salzberg SL. Using GlimmerM to find genes in eukaryotic genomes. Curr Protoc Bioinformatics. 2002 Nov;Chapter 4:Unit 4.4.
44. Bruna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. [Internet]. 2021 Jan 6 [cited 2024 Oct 5];3(1):lqaa108. Available from: https://doi.org/10.1093/nargab/lqaa108
45. Gabriel L, Bruna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024 Jun 25;34(5):769-77.
46. Li H. 2023. Protein-to-genome alignment with miniport. Bioinformatics [Internet]. 2023 Jan 1 [cited 2024 Oct 5];39(1):btad014. Available from: https://doi.org/10.1093/bioinformatics/btad014
47. Karin EL, Mirdita M, Söding J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome [Internet]. 2020 Apr 3 [cited 2024 Oct 5];8(1):48. Available from: https://doi.org/10.1186/s40168-020-00808-x

48. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008 Mar 1;24(5):637-44.
49. Wang M, Kong L. pblat: a multithread blat algorithm speeding up aligning sequences to genomes. BMC Bioinformatics [Internet]. 2019 Jan 15 [cited 2024 Oct 5];20:28. Available from: https://doi.org/10.1186/s12859-019-2597-8
50. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59-60.
51. Haas, BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. [Internet]. 2008 Jan 11 [cited 2024 Oct 5];9:R7. Available from: https://doi.org/10.1186/gb-2008-9-1-r7
52. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016 Apr 4;34(5):525-7.
53. Klemm P, Stadler PF, Lechner M. Proteinortho6: pseudo-reciprocal best alignment heuristic for graph-based detection of (co-)orthologs. Front Bioinform. [Internet]. 2023 Dec 13 [cited 2024 Oct 5];3:1322477. Available from: https://doi.org/10.3389/fbinf.2023.1322477
54. Luo H, Lin Y, Liu T, Lai FL, Zhang CT, Gao F, et al. DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. Nucleic Acids Res. 2021 Jan 8;49(D1):D677-D686.
55. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. [Internet]. 2004 Jan [cited 2024 Oct 5];5:R7. Available from: http://genomebiology.com/2004/5/2/R7
56. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. [Internet]. 2015 Aug 6 [cited 2024 Oct 5];6:157. Available from: https://doi.org/10.1186/s13059-015-0721-2
57. Nevers Y, Rossier V, Train CM, Altenhoff A, Dessimoz C, Glover N. Multifaceted quality assessment of gene repertoire annotation with OMArk. BioRxiv 2022.11.25.517970 [Preprint]. 2022 [cited 2024 Oct 5]: [17 p.] Available from: https://doi.org/10.1101/2022.11.25.517970
58. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021 Dec 9;38(12):5825-9.
59. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. Nucleic Acids Res. 2023 January 6;51(D1):D418-D427.
60. Savojardo C, Martelli PL, Fariselli P, Casadio R. DeepSig: deep learning improves signal peptide detection in proteins. Bioinformatics. 2018 May;34(10)1690-6.
61. Scott MS, Troshin PV, Barton GJ. NoD: a Nucleolar localization sequence detector for eukaryotic and viral proteins. BMC Bioinformatics [Internet]. 2011 Aug 3 [cited 2024 Oct 5];12:317. Available from: https://doi.org/10.1186/1471-2105-12-317
62. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res. [Internet]. 2007Jul [cited 2024 Oct 5];35(Suppl2):W585-W587. Available from: https://doi.org/10.1093/nar/gkm259
63. Blankenberg D, Taylor J, Schenck I, He J, Zhang Y, Ghent M, et al. A framework for collaborative analysis of ENCODE data: Making large-scale analyses biologist-friendly. Genome Res. 2007;17(6):960-4.
64. Jin JP, Tian F, Yang DC, Meng YQ, Kong L, Luo JC, et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 2017;45(D1):D1040-D1045.
65. Zheng Y, Jiao C, Sun H, Rosli HG, Pombo MA, Zhang P, et al. iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. Mol Plant. 2016 Dec 5;9(12):1667-70.
66. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021 Sep 20;49(16):9077-96.
67. Mokhtar MM, EL Allali A. PltRNAdb: Plant transfer RNA database. PLoS One [Internet]. 2022 May 23 [cited 2024 Oct 5];17(5):e0268904. Available from: https://doi.org/10.1371/journal.pone.0268904
68. Jin JJ, Yu WB, Yang JB, Song Y, de Pamphilis CW, Yi TS, et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. [Internet]. 2020 [cited 2024 Oct 5];21:241. Available from: https://doi.org/10.1186/s13059-020-02154-5
69. Cascales J, Bracco M, Garberoglio MJ, Poggio L, Gottlieb AM. Integral phylogenomic approach over *Ilex* L. species from southern South America. Life [Internet]. 2017 [cited 2024 Oct 5];7:47. Available from: 10.3390/life7040047
70. Wang Y, Sun N, Shi W, Ma Q, Sun L, Hao M, et al. Assembly and comparative analysis of the complete mitochondrial genome of *Ilex macrocarpa*. Forests [Internet]. 2023 [cited 2024 Oct 5];14:2372. Available from: https://doi.org/10.3390/f14122372
71. Aguilera PM, Debat HJ, Scaldaferro MA, Martí DA, Grabiele M. FISH-mapping of the 5S rDNA locus in chili peppers (*Capsicum*-Solanaceae). An Acad Bras Cienc. 2016 Mar;88(1):117-25.
72. Barral G, Poggio L, Giberti GC. Chromosome numbers and DNA content from *Ilex argentina* (Aquifoliaceae). Bol Soc Argent Bot. 1995;30:243-8.

73. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. [Internet]. 2016 Jan 26 [cited 2024 Oct 5];17:3. Available from: https://doi.org/10.1186/s13059-016-0881-8
74. Group Angiosperm Phylogeny. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc Lond. [Internet]. 2016 May [cited 2024 Oct 5];181(1):1-20. Available from: https://doi.org/10.1111/boj.12385
75. Broeck LVD, Bhosale DK, Song K, Lima CFF, Ashley M, Zhu T, et al. Functional annotation of proteins for signaling network inference in non-model species. Nat Commun. [Internet]. 2023 Aug 3 [cited 2024 Oct 5];14(1):4654. Available from: https://doi.org/10.1038/s41467-023-40365-z
76. Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, Heijne GV, Elofsson A, et al. Detecting sequence signals in targeting peptides using deep learning. Life Sci Alliance [Internet]. 2019 Sep 30 [cited 2024 Oct 5];2(5):e201900429. Available from: https://www.life-science-alliance.org/content/2/5/e201900429
77. Nielsen H, Tsirigos KD, Brunak S, Heijne GV. A brief history of protein sorting prediction. Protein J. 2019 Jun;38(3):200-16.
78. Meraj TA, Fu J, Raza MA, Zhu C, Shen Q, Xu D, et al. Transcriptional factors regulate plant stress responses through mediating secondary metabolism. Genes [Internet].2020 [cited 2024 Oct 5];11:346. Available from: https://doi.org/10.3390/genes11040346
79. Wang X, Tan NWK, Chung FY, Yamaguchi N, Gan ES, Ito T. Transcriptional regulators of plant adaptation to heat stress. Int J Mol Sci. [Internet]. 2023 [cited 2024 Oct 5];24:13297. Available from: https://doi.org/10.3390/ijms241713297
80. Gish LA, Clark SE. The RLK/Pelle family of kinases. Plant J. 2011 April;66(1):117-27.
81. Kazama T, Toriyama K. Whole mitochondrial genome sequencing and reexamination of a cytoplasmic male sterility associated gene in boro-taichung-type cytoplasmic male sterile rice. PLoS One [Internet]. 2016 [cited 2024 Oct 5];11(7):e0159379. Available from: https://doi.org/10.1371/journal.pone.0159379
82. Khachaturyan M, Reusch TBH, Dagan T. Worldwide population genomics reveal long-term stability of the mitochondrial genome architecture in a keystone marine plant. Genome Biol Evol. [Internet]. 2023 Sep 4 [cited 2024 Oct 5];15(9):evad167. Available from: https://doi.org/10.1093/gbe/evad167
83. Kolesnikov AA, Gerasimov ES. Diversity of mitochondrial genome organization. Biochemistry (Moscow). 2012;77(13):1424-35.
84. Kozik A, Rowan BA, Lavelle D, Berke L, Schranz ME, Michelmore RW, et al. The alternative reality of plant mitochondrial DNA: One ring does not rule them all. PLoS Genet. [Internet]. 2019 Aug 30 [cited 2024 Oct 5];15(8):e1008373. Available from: https://doi.org/10.1371/journal.pgen.1008373
85. Garcia LE, Sanchez-Puerta MV. Transcriptional landscape and splicing efficiency in *Arabidopsis* mitochondria. Cells [Internet]. 2021 Aug 11 [cited 2024 Oct 5];10:2054. Available from: https://doi.org/10.3390/cells10082054
86. Letsiou S, Madesis P, Vasdekis E, Montemurro C, Grigoriou ME, Skavdis G, et al. DNA barcoding as a plant identification method. Appl Sci. [Internet]. 2024 Feb [cited 2024 Oct 5];14:1415. Available from: https://doi.org/10.3390/app14041415
87. Aguilera PM, Grabiele M. Data for: An integrated resource for genetics and breeding of erva/yerba mate (*Ilex paraguariensis*) [dataset] 2024 Nov 6 [cited 2024 Nov 9]. Mendeley Data. Available from: https://data.mendeley.com/datasets/8v9ws627tg/1 Referenced doi: 10.17632/8v9ws627tg.1
88. Aguilera PM, Grabiele M. Associated data to the annotation of the YMM1 nuclear genome assembly of *Ilex paraguariensis*: gene predictions-annotations and transcripts [dataset] 2024 Dec 9 [cited 2025 Mar 18]. Mendeley Data. Available from: https://data.mendeley.com/datasets/kgz37pgpxw/1 Referenced doi: 10.17632/kgz37pgpxw.1