Conserving a threatened North American walnut: a chromosome-scale reference genome for butternut (*Juglans cinerea*)

Authors: Cristopher R. Guzman-Torres[*1,3], Emily Trybulec[*2,3], Hannah LeVasseur[*1,3], Harshita Akella[1,3], Maurice Amee[1,3], Emily Strickland[1], Nicole Pauloski[2,3], Martin Williams[4], Jeanne Romero-Severson[5], Sean Hoban[6], Keith Woeste[7], Carolyn C. Pike[7], Karl C. Fetter[**1,3], Cynthia N. Webster[**1,3], Michelle L. Neitzey[**2,3], Rachel J. O'Neill[**2,3], Jill L. Wegrzyn[**1,3]

[1]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA
[2]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA
[3]Biodiversity and Conservation Genomics Center, Institute for Systems Genomics, University of Connecticut, Storrs, CT USA
[4]Atlantic Forestry Center, Natural Resources Canada, Fredericton, NB, Canada
[5]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA
[6]The Center for Tree Science, The Morton Arboretum, Lisle, IL, USA
[7]USFS Eastern Region, Lafayette, IN, USA

[*]First co-authors (undergraduate team)
[**]Corresponding co-authors (mentors)

Abstract

With the advent of affordable and more accurate third generation sequencing technologies, and the associated bioinformatic tools, it is now possible to sequence, assemble, and annotate more species of conservation concern than ever before. *Juglans cinerea*, commonly known as butternut or white walnut, is a member of the walnut family, native to the Eastern United States and Southeastern Canada. The species is currently listed as Endangered on the IUCN Red List due to decline from an invasive fungus known as *Ophiognomonia clavigignenti-juglandacearum* (O-cj) that causes butternut canker. O-cj creates visible sores on the trunks of the tree which essentially starves and slowly kills the tree. Natural resistance to this pathogen is rare. Conserving butternut is of utmost priority due to its critical ecosystem role and cultural significance. As part of an integrated undergraduate and graduate student training program in biodiversity and conservation genomics, we describe the first reference genome for *Juglans cinerea*. This chromosome-scale 539 Mb assembly was generated from over 100X coverage of Oxford Nanopore long reads and scaffolded with the *Juglans mandshurica* genome. Comparisons with sequenced Juglandaceae revealed high levels of synteny and further supported *J. cinerea's* recent phylogenetic placement. Comparative assessment of gene family evolution revealed a significant number of contracting families, including several associated with biotic stress response.

Introduction

Temporal analysis of biodiversity loss indicates that we are in the midst of the Earth's sixth mass extinction event (Ceballos et al. 2017). Habitat destruction and degradation, overexploitation, pollution, invasive species, climate change, and other anthropogenic factors are the greatest threats to Earth's flora and fauna. A recent report from the International Union for the Conservation of Nature (IUCN) Red List noted that nearly 30% of the species assessed are at risk of extinction (IUCN 2022). The Royal Botanic Gardens, Kew, estimates that 1 in every 5 species of plant is threatened with extinction.

Preserving adaptive potential through conservation genetics has been widely recognized as important to ensuring long-term survival, especially in the face of climate change (Kardos et al 2021). Conservation managers and genomicists can work together to monitor population genetic diversity and develop restoration plans including assisted breeding. However, several studies have acknowledged that a "conservation genetics gap" exists and this likely stems from, among other things, a lack of training, barriers to collaboration, discipline-specific terminology, as well as access to resources (Taylor et al. 2017; Sandstrom et al. 2019). The slow advancement of tool development and implementation in conservation is further impacted by the lack of genomic resources for the vast majority of threatened species, with less than 1% of all eukaryotic species sequenced to date (Lewin et al. 2018). As such, the implementation of genetic tools in conservation and/or restoration often requires the creation of that first reference genome.

Butternut (*Juglans cinerea* L.) is a deciduous, outcrossing, and wind-pollinated canopy tree native to eastern North America. Once valued for its edible nuts and wood. Today, it is threatened by butternut canker, a disease caused by the fungus *Ophiognomonia clavigignenti-juglandacearum* (Nair et al. 1979), introduced in the late 1960s and now found throughout butternut's range. The disease has killed approximately 80% of the trees and is threatening its overall survival. Diseased trees die within 5 to 10 years, and little resistance has been identified outside of hybrids with the introduced Japanese walnut (*Juglans ailantifolia*). The severity of butternut canker as well as Armillaria root rot is greater in low-elevation sites with heavy soils and in humid sites with heavy canopies (LaBonte et al. 2015). Dramatic declines in butternut populations over the past few decades have resulted in the listing of butternut as endangered in Canada (Nielsen et al. n.d.) as well as an Endangered on the IUCN Red list in 2018.

Butternut shows high genetic diversity (moderate cline from south to north) (Hoban et al 2008; 2010). In more isolated populations, lower genetic diversity and inbreeding are observed. Assessments from microsatellites and chloroplast markers have revealed population-level genetic differentiation across an east–west gradient that may have originated from multiple Ice Age refugia (Laricchia et al 2015). Butternut is sympatric with *Juglans nigra* through much of its range however it cannot hybridize with this species, and exclusively forms hybrids with non-native *J. ailantifolia.* Hybrids with *Juglans ailantifolia* are known as buartnuts and are more tolerant of butternut canker, grow faster/larger, and are more productive (Boraks and Broders 2014, Brennan et al. 2020). Given the limited resistance observed in the wild among pure *J. cinerea* individuals, a plan proposed by practitioners seeks to conserve native germplasm and eventually introduce genes for resistance from *J. ailantifolia* (Pike et al. 2020), similar to work on American chestnut. The development of reliable markers and the materials for forward and/or reverse breeding approaches will rely on a high quality reference genome.

To address the conservation genetics gap, undergraduate and graduate students trained in the newly formed Biodiversity and Conservation Genomics Center at the University of Connecticut worked with practitioners to generate the first reference genome for the threatened *Juglans cinerea*. This study represents the intersection of industry (Oxford Nanopore), federal research agencies (USDA Forest Service), forest tree conservation (Morton Arboretum), and academia. The chromosome-scale reference was annotated with existing transcriptomic (RNA-Seq) resources and examined in an evolutionary context to recently sequenced members of the Juglandaceae from around the globe.

Materials & Methods

Plant Material, Extraction & Sequencing

Leaf material was collected from a single individual from a wild population of trees in Sheffield, New Brunswick, Canada (45.88 N, -66.36 W) (Fig. 1A) that exhibit signs of healthy growth (Fig. 1B). A vouchered specimen was deposited at the Connell Memorial Herbarium, University of New Brunswick (UNB #69382). Plant material was flash frozen in liquid nitrogen, and stored at -80°C until extraction. Leaf tissue weighing 750mg was bulked from a single accession and ground with a mortar and pestle in liquid nitrogen. High molecular weight (HMW) genomic DNA extraction was then conducted following Vaillancourt and Buell CTAB-Genomic-tip protocol (Vaillancourt et al. 2019). In brief, the aqueous layer containing the DNA was collected and mixed with isopropanol to precipitate the DNA, which was then resuspended in Buffer G2 and purified using a Qiagen Genomic-tip column. After being eluted in Buffer QF, the purified DNA was further purified by precipitation with isopropanol, followed by a wash with ethanol and elution in Buffer EB overnight at room temperature. DNA was stored at 4°C until shipped to University of Connection for library preparation and sequencing. The DNA was sheared to approximately 25 kb using a Covaris® G-tube (#520079, Covaris, LLC, Woburn, MA, USA). The sequencing library was prepared using the ligation sequencing DNA v14 method (SQK-LSK114, Oxford Nanopore Technologies, Oxford, UK). Three separate loads of the library were run on one flow cell (FLO-PRO114M) over 90 hours, resulting in 5.2 million reads with MinKNOW® (v. 22.08.6).

Assembly

A haploid chromosome-scale assembly was generated using Oxford Nanopore PromethION long reads at 100x coverage, based on the estimated genome size of 570Mb (Fig. 1C; Fig. S1). The raw reads were basecalled by Guppy (v.6.3.8) in two modes: (1) high accurate (Q>9) duplex and (2) super accurate (Q>10) simplex. Kmer-freq and GCE were run to estimate genome size at K-mer size 17, taking the long reads (FASTQ) as input (Wang et al. 2020; Liu et al. 2013).

To remove potential contaminants from both read sets, including bacteria and fungal associations, we used Centrifuge (v.1.0.4-beta) (Table S1) (Kim et al. 2016). Quality assessment was performed with NanoPlot before and after contaminant removal (De Coster et al. 2018). We evaluated two assembly methods, Flye (v.2.9) and Canu (v.2.2) (Kolmogorov et al. 2019; Koren et al. 2017). Following *de novo* assembly, Medaka (v.1.7.2) was used for error-correcting (polishing) the Flye-generated assemblies (*medaka*). Polishing was not independently run outside of Canu's approach. Completeness and contiguity were assessed at

each stage with BUSCO (v.5.0.0) against the Embryophyta single-copy ortholog database (OrthoDB v.10.0), QUAST (v.5.0.2), Minimap2 (v.2.15), and PycoQC (v.2.5.2) (Manni et al. 2021; Kriventseva et al. 2019; Gurevich et al. 2013; Li 2018, Leger and Leonard 2019). These metrics led to the selection of the super accuracy simplex-based assemblies. Following error-correction, Purge Haplotigs (v.1.1.2) was applied to the Canu and Flye assemblies to reduce allele-based duplications (Roach et al. 2018). Both assemblies were subsequently scaffolded with the chromosome-scale *Juglans mandshurica* reference (BioProject: PRJCA006358) using RagTag (v.2.1.0) in *scaffold mode* (Li et al. 2022; Alonge and Leonard 2021). The assemblies were then filtered to include the 16 longest scaffolds that aligned to the 16 chromosomes. Using the same assessments listed above, the scaffolded Canu assembly was selected for downstream annotation and analysis.

Annotation

Repeat identification on the selected Canu assembly was performed with Repeat Modeler (v.2.0.4) which provided a set of consensus sequences that served as input to Repeat Masker (v.4.1.4) (Table S2) (Flynn et al. 2020). Regions identified as repeats were softmasked in preparation for genome annotation. Six public Illumina RNA-Seq libraries were accessed from NCBI (PRJNA413418) for *J. cinerea* leaf tissue (Table S3). Each library was adapter trimmed with FastP (v.0.23.2) (Chen et al. 2018). The adapter trimmed reads were aligned to the softmasked reference with HISAT2 (v.2.2.1) to provide evidence for gene prediction with 85% mapping rate cutoff, leading us to exclude two libraries (Kim et al. 2019). The RNA alignments and the softmasked genome served as the input to both BRAKER (v2.1.6) and TSEBRA (v.1.0.1) to generate reference annotations (Bruna et al. 2021; Gabriel et al. 2021). BRAKER was run with transcriptomic alignments and viridiplantae protein input from OrthoDB (v.10) separately. Both sets of predictions were fed into TSEBRA. This resulted in one transcriptomic set of predictions from BRAKER and one combined run from TSEBRA. The resulting GFF files were summarized with AGAT (v.1.0.0) (Dainat 2022) and functionally annotated with EnTAP (v.0.10.8) (Hart et al. 2020). This provided a reciprocal BLAST estimate against NCBI's Plant RefSeq Plant Protein 208 database as well as annotations derived from EggNOG and Gene Ontology. Both sets of predictions were filtered by the functional annotations; candidates without an EggNOG assignment were removed. The final annotation was selected from a combination of BUSCO, AGAT statistics, and the annotation rate reported by EnTAP.

Methylation & Circos Visualization

Because CpG and non-CpG methylation are central to plant gene regulation, 5mC methylation was called in all contexts. Bonito-remora (v.0.6.2) with rerio modified base model 5mC_all_context_sup_r1041_e82 was used to estimate methylation (*bonito*, *rerio*). The resulting calls were summarized with modbam2bed (v0.9.5) and filtered to >= 10X coverage per site (*modbam2bed*). CpG content was quantified with fastaRegexFinder (v.0.2.0) (Beraldi). For visualization with ShinyCircos, 5mC methylation was averaged across 100Kb windows, and CpG, repeat, and gene coverage were summed across 100Kb, 10Kb, and 10Kb windows, respectively (Yu et al. 2018).

Synteny

A macrosynteny analysis was conducted using MCscan (Python version) and jcvi (v.1.3.3) with available *Juglans* species, including *Juglans cinerea* (PRJEB56451), *Juglans nigra* (Zhou et al. 2023), *Juglans regia* (PRJNA291087), and *Juglans mandshurica* (PRJCA006358) (Tang et al. 2008; Tang et al. 2015) (Table 1). If not pre-filtered, any unclassified scaffolds were eliminated from both the genome and annotation files, resulting in a set of 16 chromosomes for each *Juglans* species. CDS sequences from chromosome-scale assemblies along with their respective annotations (GFF) were used as input for MCscan. The jcvi format module was used to first filter input FASTA files by longest isoform and convert GFF files to BED format. Next, a pairwise synteny search was conducted across *J. cinerea* and *J. mandshurica*, *J. mandshurica* and *J. regia* and *J. regia* and *J. nigra* respectively. The resulting anchors were then clustered into synteny blocks with a minimum span of 30 and visualized by the jcvi karyotype graphical module. The final chromosome order was determined by length increasing from left to right. *J. regia* had eight noticeable inversions when compared to *J. mandshurica*; however, upon further inspection we found these chromosomes were reverse complemented.

Gene Family Analysis

Orthologous gene families were identified with OrthoFinder (v2.5.4) (Emms and Kelly 2019) (Table S4) using 10 species with reference genomes among *Juglans (J. cinerea, J. regia, J. sigillata, J. microcarpa, J. hindsii, J. mandshurica, J. cathayensis)*, *Carya (Carya illinoinensis)*, and *Pterocarya (Pterocarya stenoptera)* as an outgroup. The species tree, based on resulting gene trees, was constructed in OrthoFinder and rooted using the species tree inference from all genes (STAG) methodology (Emms and Kelly 2018). Support values for internal nodes reflect the proportion of orthogroup trees where the bipartition is observed, and ranges from 0 to 1. Orthogroups only representing one species and orthogroups with oversized representatives from a single species (>100 genes) were removed. The remaining 17,471 orthogroups and ultrametric tree were provided to CAFE v5 (Mendes et al. 2021) to identify gene families with evidence of significant expansion or contraction at p-value < 0.01 using the base model. The longest gene from each orthogroup was used to assign function from EnTAP. Gene families associated with retroelements were filtered from the final analysis.

Results & Discussion

Assembly

A total of 59.7 Gb of high accuracy-basecalled (HAC) duplex reads and 59.9 Gb of super-basecalled (SUP) simplex Oxford Nanopore reads were generated using a PromethION sequencer, which each provided 105x coverage at an estimated genome size of 570 Mb (Fig. 1C). Contamination filtering removed primarily bacteria, with the highest proportion of reads associated with *Acinetobacter baumannii*. Filtering resulted in 58.0 Gb of HAC reads with an N50 length of 19.5 Kb, and 58.0 Gb of SUP reads with an N50 length of 22.4 Kb (Fig. S2).

With the SUP reads, Flye initially produced a slightly less complete but more contiguous assembly than Canu; 98.9% vs. 99.2% BUSCO completeness and N50 of 2.55 Mb (1,320 scaffolds) vs. N50 of 2.39 Mb (2,286 contigs), respectively (Table S5; Fig. 1D). Polishing of Flye with Medaka yielded no increase in overall BUSCO completeness, but N50 increased slightly by

2.55 Kb. Purge Haplotigs increased the contiguity of both assemblies while largely maintaining the completeness (N50 of 2.59 Mb in 667 scaffolds for Flye and N50 of 3.02 Kb in 345 contigs for Canu). RagTag scaffolding with the *J. mandshurica* assembly drastically increased the contiguity of both assemblies, leaving Flye with 175 scaffolds and Canu with 58 scaffolds. We extracted 16 scaffolds from each assembly that were determined to be chromosome-level based on genome size, completeness, and read alignment. The assemblies were fairly similar in terms of completeness, contiguity, and overall accuracy (Table S5). The scaffolded Canu assembly was selected for downstream analyses and distinguished only by fewer gaps when compared to the scaffolded Flye assembly (9.18 vs. 5.32 Ns per 100 Kb). The final 16 pseudo-chromosome assembly reported is 539 Mb in length with an N50 of 34.86 Mb and an embryophyta BUSCO completeness of 99.0% (Table S5; Fig. 1D).

This exclusively nanopore-assembled genome benefitted from high quality base calls and >100X coverage to generate a genome of 539 Mb (slightly less than the estimated 570 Mb). The initial genome size estimates were produced from uncorrected long-reads and may not be as reliable of an estimate but were consistent with C-value estimates from Genomes on a Tree (GoaT) (Challis et al. 2023). *Juglans cinerea* is the 10th member of the Juglandaceae to have its whole genome sequenced. Comparing our final *J. cinerea* assembly with that of *J. mandshurica* to which it was scaffolded (Li et al. 2022), *J. cinerea* exhibits a smaller scaffold N50 size of 34.86 Mb, higher BUSCO completeness score of 99.0%, larger genome size of 539 Mb, and higher GC content of 36.64% (Table 1).

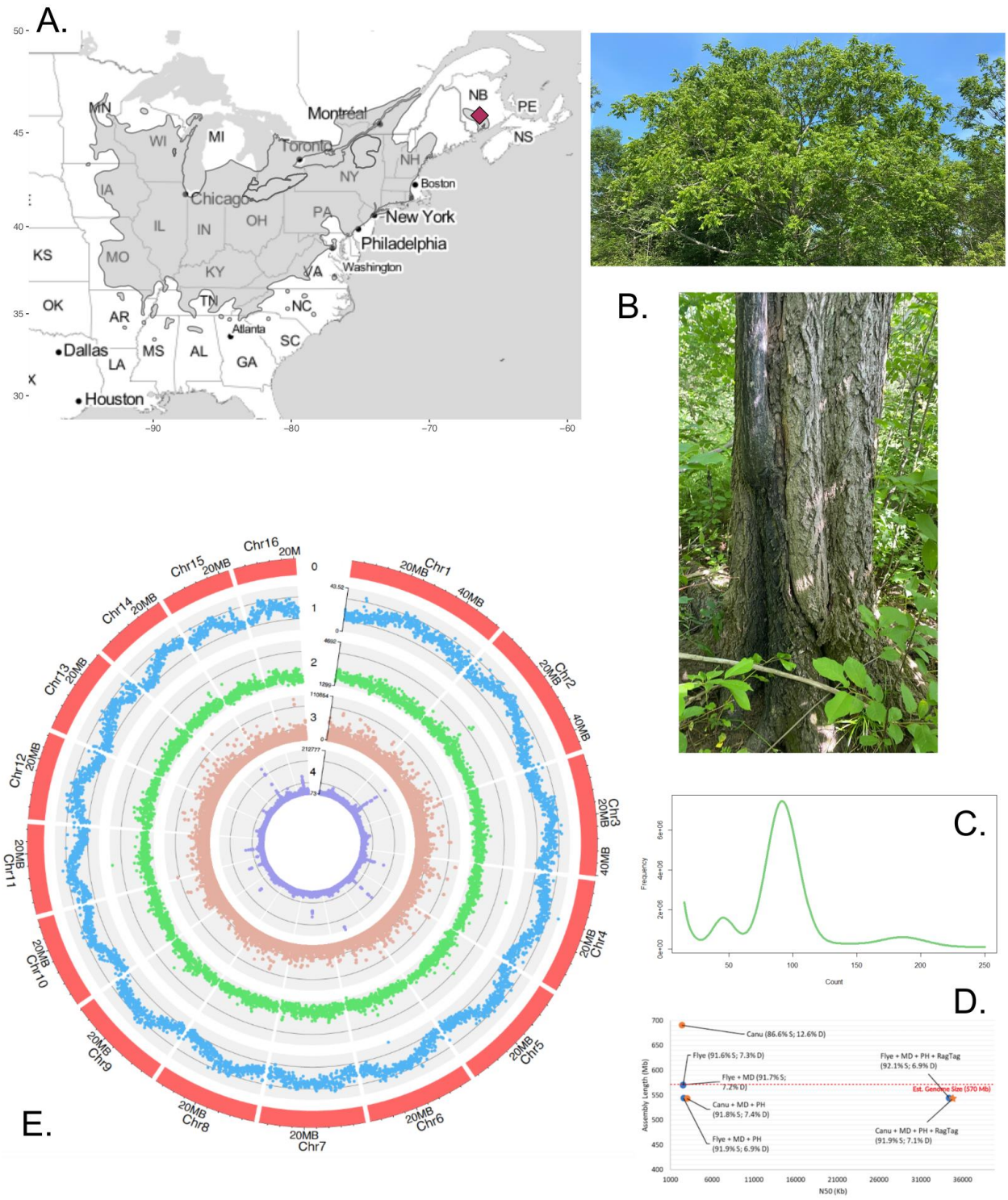

Figure 1. *Range map (adapted from Little et al. 1971) and sample location of the Juglans cinerea reference tree (UNB #69382) and photographs from the collection site* ***(A)****. Photographs of the sampled reference tree from the Sheffield, NB population exhibiting a healthy crown and trunk absent of cancers* ***(B)****. k-mer plot (length=17bp) of k-mer count (x-axis) versus frequency (y-axis). Genome size from long-read data initially estimated at 570 Mb* ***(C)****. Scatter plot comparing assembly length (Mb) and N50 (Kb) obtained from QUAST for Flye and Canu*

throughout the assembly process. Assemblies represented by color (Canu=yellow; Flye=blue) and shape (drafts=circles; selected final assembly pre-chromosome filtering=star) with abbreviated assembly software (MD=Medaka; PH=Purge Haplotigs). BUSCO single (S) and duplicated (D) completeness scores in parentheses ***(D)****. Circos plot representing 0) assembly chromosomes 1) average 5mC methylation in all sequence contexts (CG, CHG, and CHH) over 100kb windows 2) CpG site count over 100kb windows 3) gene coverage over 10kb windows 4) repeat coverage over 10kb windows* ***(E)****.*

Annotation

Prior to structural annotation, repeat sequences in the assembly were identified, resulting in 2,353 RepeatScout/RECON families and 1,117 LTRPipeline families. A total of 298 of the LTR families were determined to be redundant and removed, resulting in 3,172 unique repeats. Repeatmasker used this library to softmask 302 Mb (56.02%) of the butternut genome (Fig. 1E). Among the softmasked regions, 28.72% represented retroelements, most of which (22.46%) were predicted to be long terminal repeats (LTRs). Gypsy elements accounted for 8.13% of the genome, whereas Copia elements made up 11.12%. As expected, a large proportion of the repeats (22.42%) were unclassified (Table S2). *J. cinerea* exhibited repeat content that is consistent with other sequenced *Juglans*, with a value that fell between *J. sigillata* (50.06%) and *J. mandshurica* (60.99%) (Ning et al. 2020, Li et al., 2022). Retroelements–mostly LTRs–compose the highest classified repeat content in all three genomes. Like the *J. regia* x *Juglans microcarpa* hybrid and *J. regia* genomes, we also identified telomeric repeat CCCTAAA at the ends of each chromosome (Zhu et al. 2019, Marrano et al. 2020).

After trimming, the six RNA-Seq libraries had an average mapping rate of 91.2% against the assembled *J.cinerea* genome (Table S3). Gene prediction was attempted with both BRAKER and TSEBRA on the softmasked assembly (Table S6). BRAKER provided a high quality annotation with only transcriptomic data, resulting in 38,377 predicted genes, 41,357 transcripts, a mono:multi ratio of 0.286, a BUSCO completeness score of 97.2% and an annotation rate of 77.1%. TSEBRA, which combined transcriptomic and protein BRAKER runs, resulted in a slightly lower quality annotation with 35,437 genes, 38,928 transcripts, 0.540 mono:multi, BUSCO completeness score of 96.8% and 86.2% annotation. After filtering with EggNOG, the quality of both runs increased, and the number of predicted genes and transcripts was reduced to 30,526 and 33,275 respectively in BRAKER, and 30,328 and 33,694 in TSEBRA. Both programs resulted in a 98.1% annotation rate after filtering, although the BRAKER annotation was closer to the expected mono:multi ratio of plants at 0.21, outperforming that of TSEBRA, 0.36 (Vuruputoor et al. 2022). As a result, BRAKER was chosen to represent the final structural annotation for *J. cinerea*. These annotation results are similar to other *Juglans* species, including the *J.mandshurica* genome released in 2022, which has 39,466 predicted genes, 47,355 transcripts, a mono:multi ratio of 0.347, 95.2% BUSCO completeness, and an annotation rate of 84.9% in chromosome-level scaffolds. Likewise, *J.nigra* and *J. regia*, which have 29,949 and 30,709 genes respectively, 29,949 and 45,959 transcripts, BUSCO completeness of 96.4% and 99.6%, mono:multi of 0.210 and 0.206, and an annotation rate of 96.3% and 99.9% (Table 1). The impressive BUSCO completeness score of the first version of

J. cinerea, ranking second to version three of *J. regia*, is a testament to the effectiveness of our methodology.

5mC all context methylation revealed a central peak pattern to each chromosome that did not correlate with CpG content (Fig. 1E). However, it does appear to correspond with small peaks of repeat content, possibly representing centromeres. Future analyses should analyze methylation by type (CpG, CHG, CHH) in association with repeats and across gene lengths.

Table 1. Comparison of Juglans chromosome-level genome assemblies and annotations.
**Assembly and annotations not yet publicly available.*

	J. cinerea	***J. mandshurica***	***J. nigra***	***J. regia***
Bioproject	PRJEB56451	PRJNA730664	*	PRJNA291087
Biosample	SAMEA111438462	SAMN19238346	*	SAMN03704235
GenBank	*	GCA_022457165.1	*	GCA_001411555.2
Sequence method	Oxford Nanopore PromethION	PacBio HiFi, Hi-C, Illumina HiSeq	Illumina HiSeq; Pacbio SMRTbell; Hi-C, Illumina NovaSeq6000	Illumina HiSeq2500; Oxford Nanopore MinION; Hi-C, Illumina HiSeq4000
Assembly				
Genome size (Mb)	539.43	528.15	530.24	545.78
Number of chromosomes	16	16	16	16
N50 (Mb)	34.86	35.38	35.14	37.11
GC content	36.64%	36.49%	36.36%	36.08%
Complete BUSCO	99.0%	98.8%	99.0%	99.0%
Annotation				
Number of genes	30,526	39,466	29,949	30,709
Number of transcripts	33,275	47,355	29,949	45,959
Average gene length (bp)	4,187	3,748	5,175	5,309
Average exon length (bp)	209	287	297	291
Complete BUSCO	97.2%	95.2%	96.4%	99.6%
Mono:multi ratio	0.207	0.347	0.210	0.206
Annotation rate	98.1%	84.9%	96.3%	99.9%

Synteny

The success of our genome assembly is in part due to the quality of the *J. mandshurica* genome, which was used to scaffold our *de novo* assembly. However, utilizing the *scaffold mode* of RagTag and not the *correction mode* allowed some rearrangements to be retained in the *J. cinerea* assembly, as highlighted by the pairwise macrosynteny analysis (Fig. 2). Conserved synteny at chromosome-level was observed in a phylogenetic context, comparing *J. cinerea* to *J. mandshurica*, *J. mandshurica* to *J. nigra,* and *J. nigra* to *J. regia*. A total of 52,993 (NR:38,556) seed synteny blocks (anchors) were found in 294 clusters between *J. cinerea* and *J. mandshurica*, 46,251 (NR:37,316) anchors were identified in 297 clusters between *J.mandshurica* and *J. nigra*, and 47,683 (NR:37,354) anchors were found in 247 cluster across *J.nigra* and *J. regia*. Considering the total number of protein-coding genes in each pairwise comparison, respectively, 68.1%, 66.6%, and 78.6% of compared genomes were collinear, indicating strong support for ancestral collinearity (Li et al. 2022).

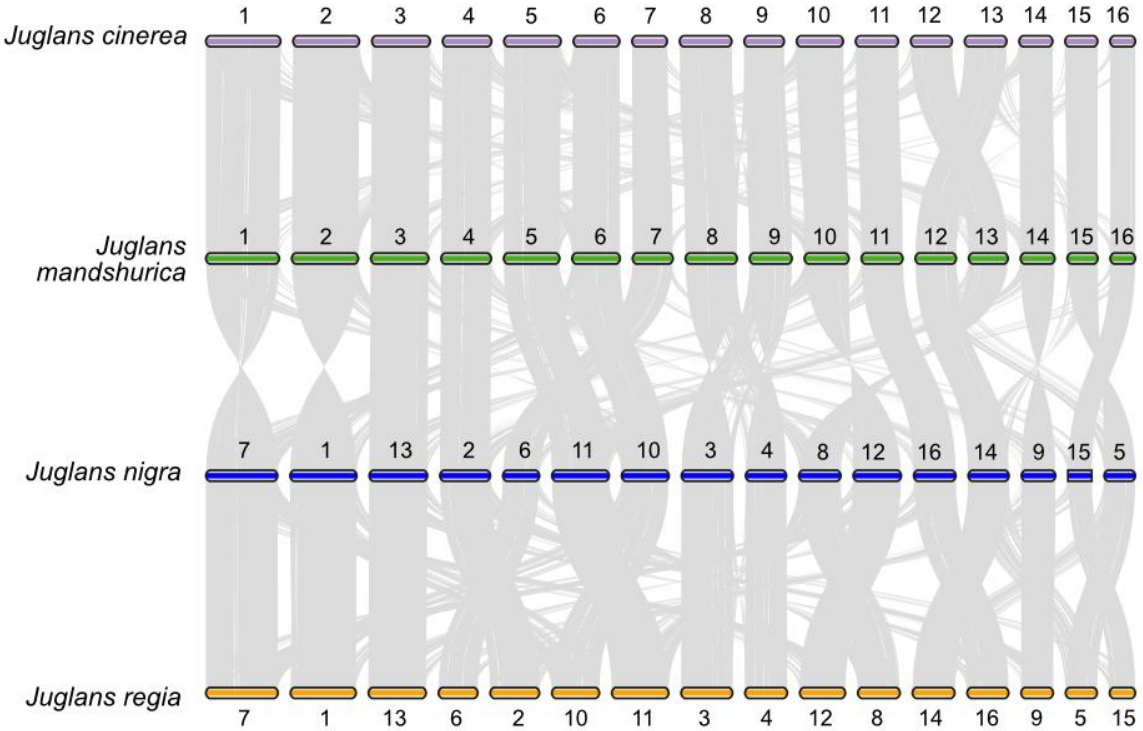

Figure 2. *Pairwise macro-synteny analysis in phylogenetic order comparing chromosomes (boxes) and syntenic clusters (grey lines) in J. cinerea (purple) and J. mandshurica (green), J. mandshurica and J. nigra (blue) and J. nigra and J. regia (orange). Published chromosome identifiers and orientation from annotation are maintained, but chromosomes are re-ordered by size (largest to smallest).*

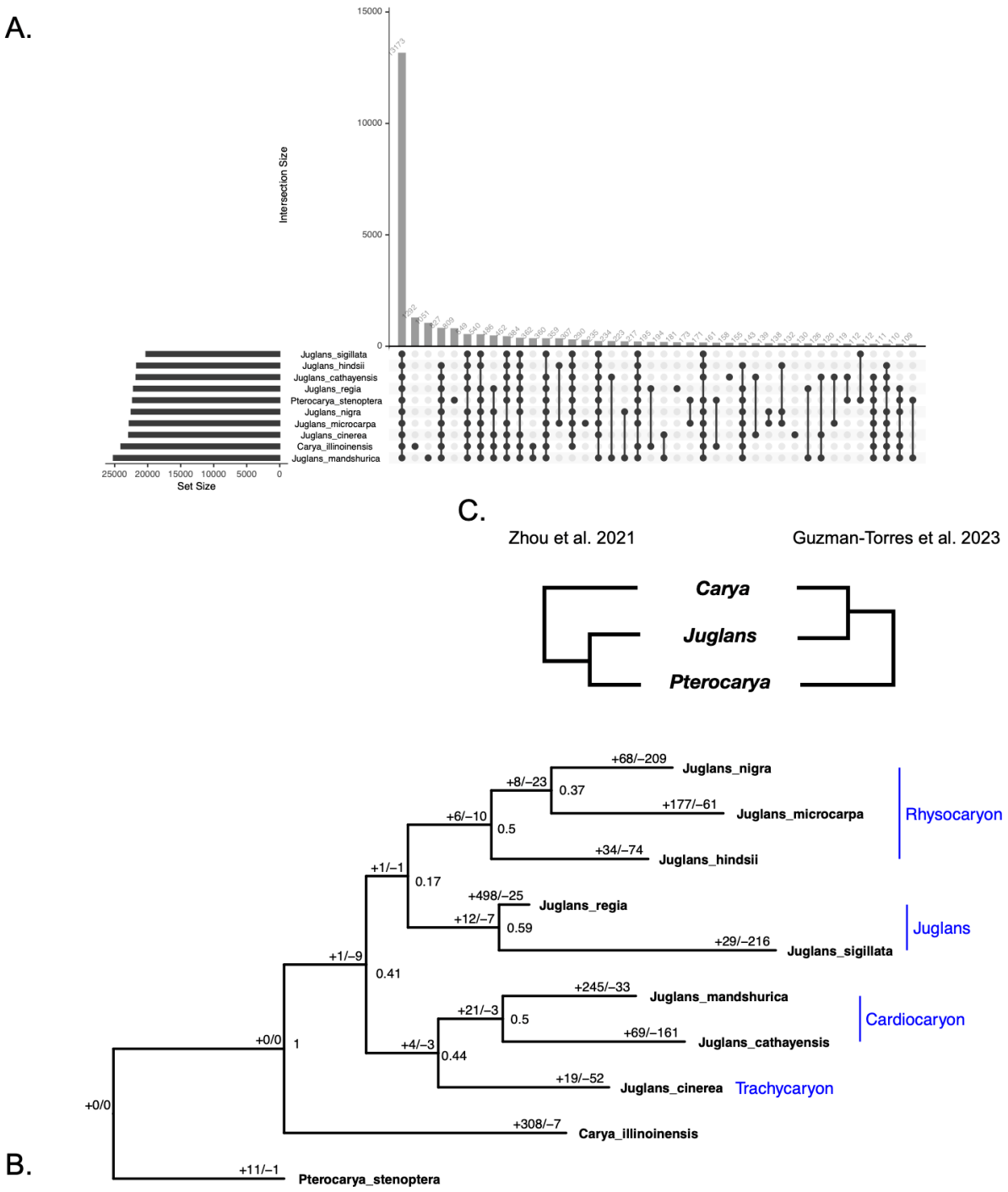

Figure 3. *Results of OrthoFinder and the phylogenomic analysis of 33,922 orthrogroup gene trees. Upset plot of the numbers of genes shared by different intersections of 10 species. Intersections smaller than 200 orthogroups are omitted for clarity* ***(A)****. Consensus phylogeny from 26,240 orthogroup gene trees. STAG support values at internal nodes and the number significantly expanding or contracting gene families (p-value < 0.01) on branches* ***(B)****. Alternate*

hypotheses of relationships between Carya, Juglans, and Pterocarya using data from RADseq SNPs (Zhou et al. 2021) versus whole genome sequencing and orthogroup gene tree analysis ***(C)***.

Gene Family Analysis

The OrthoFinder analysis conducted with ten Juglandaceae assigned 384,404 genes out of 401,209 (95.8%) into 33,922 orthogroups. A total of 13,173 orthogroups (38.8%) were observed in all species and 4,076 orthogroups (12%) were specific to a single species. There were 807 single-copy orthogroups and 16,805 unassigned genes. In *J. cinerea*, 32,743 genes (98.4%) were assigned to an orthogroup. Of the total number of orthgroups found across the 10 species, 22,863 (67.4%) included *J. cinerea*. A total of 132 orthogroups (0.39%) containing 328 genes were unique to *J. cinerea*, and only 532 genes (1.6%) could not be assigned to an orthogroup (Table S4; Fig. 3A). *J. cinerea* shares more one-to-one orthologs outside of the *Trachycaryon-Cardiocaryon* clade than within the clade, possibly reflecting isolation and divergence of the ancestral *Cardiocaryon* population to the ancestral *J. cinerea* population (Fig. 3A, Table S4).

Members of the Juglandaceae assessed here belong to one of four sections. The North American species *J. nigra*, *J. hindsii* and *J. microcarpa* are members of sect. *Rhysocaryon*. The sampled Eurasian species (sect. *Juglans)* include *J. sigillata and J. regia*, the latter being the predominant cultivated species for walnut production. *Juglans cathayensis and J. mandshurica* are members of sect. *Cardiocaryon*. *Juglans ailantifolia*, also in section *Cardiocaryon,* but not included here, is sometimes described as a sub-species of *J. mandshurica*. This is of note given the ability of *J. cinerea* to hybridize with *J. ailantifolia* but not with the sympatric *J. nigra* (Brennan et al. 2020). Section *Trachycaryon* contains only *J. cinerea* whose placement within *Juglans* has varied by marker and data type (Mu et al. 2020).

Phylogenetic relationships among Juglandaceae remain enigmatic, despite multiple sequence-based and morphological analysis in recent decades (Ding et al. 2023). The rooted consensus phylogeny constructed from orthogroups using the STAG methodology places sect. *Rhysocaryon* sister to *Juglans*, which in turn, are sister to a clade containing sections *Cadiocaryon* and *Trachycaryon* (Fig. 3B). The recovered STAG phylogeny supports topologies within *Juglans* identified in recent whole genome sequence-based (Trouern-Trend et al. 2020) and RADseq-based phylogenies (Mu et al. 2020). Our topology of Juglandaceae is in contrast to a RADseq-derived phylogeny using SNPs mapped to the *J. regia* genome (Zhou et al. 2021). In particular, the SNP-based tree has the *Trachycaron-Cardiocaron* clade sister to *Juglans*, which is in turn sister to *Rhysocaron*. Importantly, Zhou et al. (2021) nest *Pterocarya* within a *Carya-Juglans* clade, while we observe *Carya* as sister to *Juglans* and *Pterocarya* as the outgroup (Fig. 3C). The allopolyploid origin of the Juglandaceae (Ding et al. 2023) lends itself to phylogenetic analysis using structural genomic comparative methods. Using micro-syntenic phylogenetic methods, Ding et al. (2023) identify *Carya* as sister to *Juglans*, while their alignment-based methods do not. Clearly, there is a pressing need for additional research in this area using multiple phylogenetic approaches.

The overall gene family analysis with CAFE identified a mean gene family birth-death rate (λ) for the whole tree of 0.88, with a tri-modal distribution of probability of gene family expansion/contraction (Fig. S3). The species with the most rapidly evolving gene families (significant at p-value < 0.01) was *J. regia* with 523 gene families. *J. cinerea* was the species

within *Juglans* with the lowest number of significantly evolving gene families, with a total of 71, and the majority of these families were contracting (N = 52, Fig. 3B). After removing orthogroups consisting of retroelements, a total of 54 orthogroups were identified from CAFE as rapidly expanding (9) and rapidly contracting (45) in *J. cinerea* (Fig. 4; Table S7). Among the small set expanding, COG categories broadly described them as cytoskeleton-related, transcription regulation, post-translational modification, cell cycle control, and ion transport. Among those contracting, functions related to transcription regulation, signal transduction, secondary metabolite biosynthesis, carbohydrate transport, disease resistance, and cytoskeleton-related were represented. Examining expanded gene families related to biotic-stress (pathogen/pest resistance), revealed two gene families, one is associated with Auxin response factors (ARFs) and the other with Ripening-related proteins. The pathogenesis-related contracting families were far more numerous and included: NB-LRRs, Wall-associated receptor kinases (WAK), Germin-like proteins, Cytochrome P450s, glycosyl hydrolases, lipolytic acyl hydrolase (LAH), mTERFs, Class III peroxidases (PRX3), UPF0481 proteins, and PAN-like domain proteins. Of these, the following are actually absent in *J. cinerea*: one NB-LRR gene family, one of the three contracting WAK gene families, one cytochrome P450, one of three contracting glycosyl hydrolases gene families, lipolytic acyl hydrolase (LAH), one of two PRX3s, and both families associated with UPF0481 proteins.

Wall-associated receptor kinases (WAKs) are increasingly associated with plant immunity, including neo- and sub-functionalization in response to changing selection pressure (Stephens et al. 2022; Zhou et al. 2023). Among *Juglans*, they are believed to be absent in *J. hindsii* and *J. sigillata* (Trouern-Trend et al. 2020), rapidly contracting in *J. mandshurica*, and expanding in *J. regia* (Yan et al. 2021). Evaluation of WAK gene families in *J. mandshurica* and *J. regia* found they are under purifying selection (Yan et al. 2021). Expression levels of WAK gene family members were observed to be higher in anthracnose-resistant varieties of *J. regia* (Li et al. 2022). In *J. cinerea*, we identified a strong signature of contraction across two WAK gene families and absence in the third. Further investigation of WAKs, and other pathogenesis-related families, in a population context, may reveal more about their role in resistance to specific fungal pathogens.

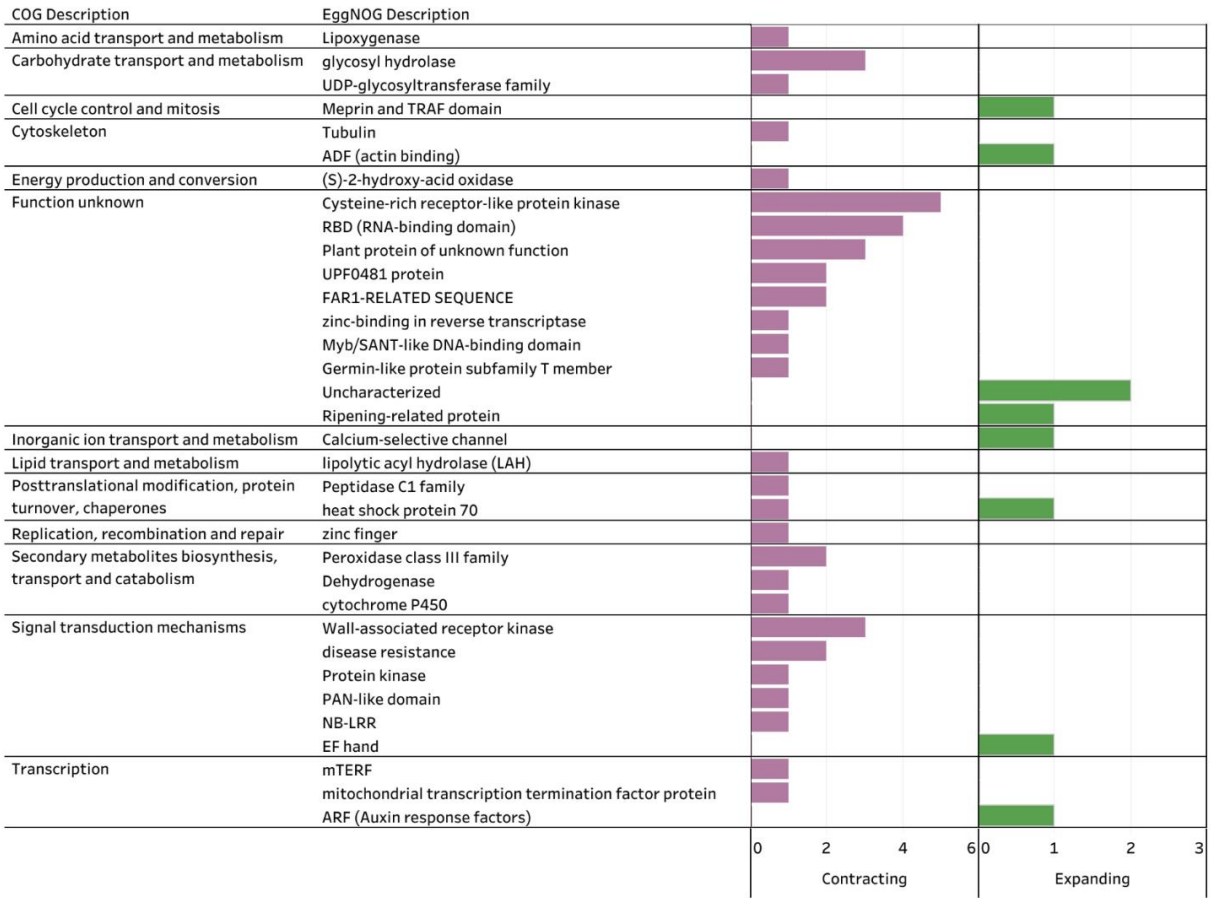

Figure 4. *Annotations of J. cinerea significantly contracting (pink) and expanding (green) gene families, excluding retroelement families.*

Conclusions

The present study reports *J. cinerea*'s first high-quality reference genome, generated using Oxford Nanopore technology and a combination of assembly tools. The *J. mandshurica* genome enabled pseudo-chromosome generation and chromosome-level analyses. A preliminary comparative view of *J. cinerea* was presented in context to other species in the Juglandaceae family to assess both synteny and gene family evolution. The construction of the genome was connected to a year-long training program initiated with high molecular weight extraction of the target tree. Students worked closely with graduate students and faculty members to understand the challenges associated with genome assembly and annotation. In doing so, the genome benefitted from robust methods testing. Finally, the program intersected with practitioners who provided input on the reference accession and a framework that will guide the application of this high quality reference in the future. Future studies expanded to population-level analyses and *J. ailantifolia* hybrid resistance to O-cj will be needed to develop conservation strategies informed by genetics.

Data Availability Statement
The ONT raw reads are available on the European Nucleotide Archive (PRJEB56451). Code, genome assembly, and genome annotation files are available on GitLab: https://gitlab.com/PlantGenomicsLab/butternut-genome-assembly.

Funder Information
Funding was provided by the University of Connecticut, College of Liberal Arts and Sciences, through the Earth and its Future initiative that enabled the initiation of the Biodiversity and Conservation Genomics Training Program. Funding was also provided through National Science Foundation Awards DBI-1943371. Support for nanopore sequencing of the IUCN-listed *Juglans cinerea* was provided by the OrgOne project by Oxford Nanopore Technologies.

Acknowledgements
We thank the Institute for Systems Genomics Center for Genome Innovation at the University of Connecticut for molecular resources and support. We would also like to thank the Computational Biology Core for access to HPC resources. Additional thanks to Patrick G.S. Grady for genome assembly consultation.

SUPPLEMENTAL:

Table S1: Comprehensive distribution of bacterial and fungal contaminants identified in raw sequence reads. All bacterial, fungal, viral and archaea contaminants identified by Centrifuge and removed are categorized by name, taxonomic ID (taxID), rank (taxRank), genome size, number of reads, and number of unique reads.

Table S2: Repeat content identified by RepeatMasker. Using a repeat library generated by RepeatModeler, RepeatMasker generated classified output for genome masking. General repeat categories are identified, along with select retroelement classes and families.

Table S3: RNA-Seq library filtering and mapping. Six short-read *Juglans cinerea* (leaf tissue) RNA libraries with a read mean length of 100 PE were collected for genome annotation analysis from SRA. Each accession was run through QC with FastP, as shown by total bases before and after adapter trimming. HISAT2 alignment rates above 85% were considered high quality.

Table S4: Summary of OrthoFinder statistics and sample information. Ten Juglandaceae reference genomes were analyzed for orthologous gene families with OrthoFinder. Eight species of *Juglans* were analyzed, and one species each of *Carya* and *Pterocarya* were included as outgroups. Project and Sample IDs for the genome accessions are provided.

Table S5: Genome assembly process statistics. Genomes generated with high accurate duplex and super accurate simplex reads were benchmarked at each assembly stage by measuring BUSCO completeness (embryophyta_odb10) and QUAST metrics.

Table S6: Genome annotation process statistics. For each benchmarked annotation method (BRAKER w/ protein, BRAKER w/RNA, and TSEBRA), BUSCO completeness, number of genes, mono:multi ratio, EnTAP annotation rate, and mean gene and exon length were calculated pre- and post-filtering with EggNOG.

Table S7: Rapidly expanding and contracting gene families. The resulting orthogroups from ten Juglandaceae reference genomes were provided to CAFE v5 to identify evidence of significant expansion (positive base change) or contraction (negative base change) at p-value < 0.01. Fifty-four gene families without retrotransposons were annotated with both COG and EggNOG. *J. cinerea* was present in twenty-three.

Figure S1: Overview of the genome assembly process. This figure illustrates the step-by-step process that was employed to generate the high-quality reference genome for *J. cinerea*. Red boxes indicate the specific software programs utilized at various stages of the assembly process and green boxes outline key stages and outputs obtained. This figure was created in Biorender.com

Figure S2. Sankey diagram of Centrifuge contaminants. Major bacterial and fungal contaminants identified in Nanopore reads, which were subsequently removed using Centrifuge. Generated using the Pavian R package (version 1.2.0).

Figure S3. Distribution of rapid evolution probabilities for orthogroups. Combined distribution of all probabilities for the null hypothesis of no significant expansion or contraction for all 10 species. These data are output of CAFE.

REFERENCES:

Alonge, M., & Leonard, A. (2021). *malonge/RagTag: RagTag v2.1.0*. https://doi.org/10.5281/zenodo.5634263

Beraldi, D. (n.d.). *fastaRegexFinder at master · dariober/bioinformatics-cafe*. Github. Retrieved May 4, 2023, from https://github.com/dariober/bioinformatics-cafe

bonito: A PyTorch Basecaller for Oxford Nanopore Reads. (n.d.). Github. Retrieved May 4, 2023, from https://github.com/nanoporetech/bonito

Boraks, A., & Broders, K. D. (2014). Butternut (Juglans cinerea) health, hybridization, and recruitment in the northeastern United States. *Canadian Journal of Forest Research. Journal Canadien de La Recherche Forestiere*, *44*(10), 1244–1252.

Brennan, A. N., McKenna, J. R., Hoban, S. M., & Jacobs, D. F. (2020). Hybrid Breeding for Restoration of Threatened Forest Trees: Evidence for Incorporating Disease Tolerance in Juglans cinerea. *Frontiers in Plant Science*, *11*, 580693.

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, *3*(1), lqaa108.

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(30), E6089–E6096.

Challis, R., Kumar, S., Sotero-Caio, C., Brown, M., & Blaxter, M. (2023). Genomes on a Tree (GoaT): A versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. *Wellcome Open Research*, *8*, 24.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* , *34*(17), i884–i890.

Dainat, J., Hereñú, D., Davis, E., Crouch, K., LucileSol, Agostinho, N., pascal-git, & tayyrov. (2022). *NBISweden/AGAT: AGAT-v1.0.0*. https://doi.org/10.5281/zenodo.7255559

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* , *34*(15), 2666–2669.

Ding, Y.-M., Pang, X.-X., Cao, Y., Zhang, W.-P., Renner, S. S., Zhang, D.-Y., & Bai, W.-N. (2023). Genome structure-based Juglandaceae phylogenies contradict alignment-based phylogenies and substitution rates vary with DNA repair genes. *Nature Communications*, *14*(1), 617.

Dumschott, K., Schmidt, M. H.-W., Chawla, H. S., Snowdon, R., & Usadel, B. (2020). Oxford Nanopore sequencing: new opportunities for plant genomics? *Journal of Experimental Botany*, *71*(18), 5313–5322.

Emms, D. M., & Kelly, S. (2018). STAG: Species Tree Inference from All Genes. In *bioRxiv* (p. 267914). https://doi.org/10.1101/267914

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9451–9457.

Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M., & Stanke, M. (2021). TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics*, *22*(1), 1–12.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* , *29*(8), 1072–1075.

Hart, A. J., Ginzburg, S., Xu, M. S., Fisher, C. R., Rahmatpour, N., Mitton, J. B., Paul, R., & Wegrzyn, J. L. (2020). EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Molecular Ecology Resources*, *20*(2), 591–604.

Hoban, S., Anderson, R., McCleary, T., Schlarbaum, S., & Romero-Severson, J. (2008). Thirteen nuclear microsatellite loci for butternut (Juglans cinerea L.). *Molecular Ecology Resources*, *8*(3), 643–646.

Hoban, S. M., Borkowski, D. S., Brosi, S. L., McCleary, T. S., Thompson, L. M., McLachlan, J. S., Pereira, M. A., Schlarbaum, S. E., & Romero-Severson, J. (2010). Range-wide distribution of genetic diversity in the North American tree Juglans cinerea: a product of range shifts, not ecological marginality or recent population decline. *Molecular Ecology*, *19*(22), 4876–4891.

IUCN 2021 : International Union for Conservation of Nature annual report. (2022). https://portals.iucn.org/library/node/49944

Kardos, M., Armstrong, E. E., Fitzpatrick, S. W., Hauser, S., Hedrick, P. W., Miller, J. M., Tallmon, D. A., & Funk, W. C. (2021). The crucial role of genome-wide genetic variation in conservation. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(48). https://doi.org/10.1073/pnas.2104642118

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915.

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729.

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736.

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, *47*(D1), D807–D811.

LaBonte, N. R., Ostry, M. E., Ross-Davis, A., & Woeste, K. E. (2015). Estimating heritability of disease resistance and factors that contribute to long-term survival in butternut (Juglans cinerea L.). *Tree Genetics & Genomes*, *11*(3), 63.

Laricchia, K. M., McCleary, T. S., Hoban, S. M., Borkowski, D., & Romero-Severson, J. (2015). Chloroplast haplotypes suggest preglacial differentiation and separate postglacial migration paths for the threatened North American forest tree Juglans cinerea L. *Tree Genetics & Genomes*, *11*(2). https://doi.org/10.1007/s11295-015-0852-3

Leger, A., & Leonardi, T. (2019). pycoQC, interactive quality control for Oxford Nanopore Sequencing. *Journal of Open Source Software*, *4*(34), 1236.

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, *115*(17), 4325–4333.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* , *34*(18), 3094–3100.

Li, M., Ma, J., Liu, H., Ou, M., Ye, H., & Zhao, P. (2022). Identification and Characterization of Wall-Associated Kinase (WAK) and WAK-like (WAKL) Gene Family in Juglans regia and Its Wild Related Species Juglans mandshurica. *Genes*, *13*(1). https://doi.org/10.3390/genes13010134

Li, Q., Hu, A., Qi, J., Dou, W., Qin, X., Zou, X., Xu, L., Chen, S., & He, Y. (2020). CsWAKL08, a pathogen-induced wall-associated receptor-like kinase in sweet orange, confers resistance to citrus bacterial canker via ROS control and JA signaling. *Horticulture Research*, *7*, 42.

Li, X., Cai, K., Zhang, Q., Pei, X., Chen, S., Jiang, L., Han, Z., Zhao, M., Li, Y., Zhang, X., Li, Y., Zhang, S., Chen, S., Qu, G., Tigabu, M., Chiang, V. L., Sederoff, R., & Zhao, X. (2022). The Manchurian Walnut Genome: Insights into Juglone and Lipid Biosynthesis. *GigaScience*, *11*. https://doi.org/10.1093/gigascience/giac057

Little (Jr.), E. (1971). *Conifers and Important Hardwoods*. U.S. Government Printing Office.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., & Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. In *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1308.2012

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654.

Marrano, A., Britton, M., Zaini, P. A., Zimin, A. V., Workman, R. E., Puiu, D., Bianco, L., Pierro, E. A. D., Allen, B. J., Chakraborty, S., Troggio, M., Leslie, C. A., Timp, W., Dandekar, A., Salzberg, S. L., & Neale, D. B. (2020). High-quality chromosome-scale assembly of the walnut (Juglans regia L.) reference genome. *GigaScience*, *9*(5). https://doi.org/10.1093/gigascience/giaa050

medaka: Sequence correction provided by ONT Research. (n.d.). Github. Retrieved May 4, 2023, from https://github.com/nanoporetech/medaka

Mendes, F. K., Vanderpool, D., Fulton, B., & Hahn, M. W. (2021). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* , *36*(22-23), 5516–5518.

modbam2bed. (n.d.). Github. Retrieved May 4, 2023, from https://github.com/epi2me-labs/modbam2bed

Mu, X.-Y., Tong, L., Sun, M., Zhu, Y.-X., Wen, J., Lin, Q.-W., & Liu, B. (2020). Phylogeny and divergence time estimation of the walnut family (Juglandaceae) based on nuclear RAD-Seq and chloroplast genome data. *Molecular Phylogenetics and Evolution*, *147*, 106802.

Nair, V. M. G., Kostichka, C. J., & Kuntz, J. E. (1979). Sirococcus Clavigignenti-Juglandacearum: An Undescribed Species Causing Canker on Butternut. *Mycologia*, *71*(3), 641–646.

Nielsen, C., Cherry, M., Boysen, B., & Hopkin, A. (n.d.). COSEWIC assessment and status report on the butternut (Juglans cinerea) in Canada. *Committee on the Status of*.

Ning, D.-L., Wu, T., Xiao, L.-J., Ma, T., Fang, W.-L., Dong, R.-Q., & Cao, F.-L. (2020). Chromosomal-level assembly of Juglans sigillata genome using Nanopore, BioNano, and Hi-C analysis. *GigaScience*, *9*(2). https://doi.org/10.1093/gigascience/giaa006

Pike, C. C., Williams, M., Brennan, A., Woeste, K., Jacobs, J., Hoban, S., Moore, M., & Romero-Severson, J. (2020). Save Our Species: A Blueprint for Restoring Butternut (Juglans cinerea) across Eastern North America. *Journal of Forestry*, *119*(2), 196–206.

R. Taylor, H., Dussex, N., & van Heezik, Y. (2017). Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners. *Global Ecology and Conservation*, *10*, 231–242.

rerio: Research release basecalling models and configurations. (n.d.). Github. Retrieved May 4, 2023, from https://github.com/nanoporetech/rerio

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*(1), 460.

Sandström, A., Lundmark, C., Andersson, K., Johannesson, K., & Laikre, L. (2019). Understanding and bridging the conservation-genetics gap in marine conservation. *Conservation Biology: The Journal of the Society for Conservation Biology*, *33*(3), 725–728.

Stephens, C., Hammond-Kosack, K. E., & Kanyuka, K. (2022). WAKsing plant immunity, waning diseases. *Journal of Experimental Botany*, *73*(1), 22–37.

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., & Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science*, *320*(5875), 486–488.

Tang, H., Krishnakumar, V., & Li, J. (2015). *jcvi: JCVI utility libraries*. https://doi.org/10.5281/zenodo.31631

Trouern-Trend, A. J., Falk, T., Zaman, S., Caballero, M., Neale, D. B., Langley, C. H., Dandekar, A. M., Stevens, K. A., & Wegrzyn, J. L. (2020). Comparative genomics of six Juglans species reveals disease-associated gene family contractions. *The Plant Journal: For Cell and Molecular Biology*, *102*(2), 410–423.

Vaillancourt, B., & Robin Buell, C. (2019). High molecular weight DNA isolation method from diverse plant species for use with Oxford Nanopore sequencing. In *bioRxiv* (p. 783159). https://doi.org/10.1101/783159

Vuruputoor, V. S., Monyak, D., Fetter, K. C., Webster, C., Bhattarai, A., Shrestha, B., Zaman, S., Bennett, J., McEvoy, S. L., Caballero, M., & Wegrzyn, J. L. (2022). Welcome to the big leaves: best practices for improving genome annotation in non-model plant genomes. In *bioRxiv* (p. 2022.10.03.510643). https://doi.org/10.1101/2022.10.03.510643

Wang, H., Liu, B., Zhang, Y., Jiang, F., Ren, Y., Yin, L., Liu, H., Wang, S., & Fan, W. (2020). Estimation of genome size using k-mer frequencies from corrected long reads. In *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/2003.11817

Yan, F., Xi, R.-M., She, R.-X., Chen, P.-P., Yan, Y.-J., Yang, G., Dang, M., Yue, M., Pei, D., Woeste, K., & Zhao, P. (2021). Improved de novo chromosome-level genome assembly of the vulnerable walnut tree Juglans mandshurica reveals gene family evolution and possible genome basis of resistance to lesion nematode. *Molecular Ecology Resources*, *21*(6), 2063–2076.

Yu, Y., Ouyang, Y., & Yao, W. (2018). shinyCircos: an R/Shiny application for interactive creation of Circos plot. *Bioinformatics* , *34*(7), 1229–1231.

Zhou, H., Hu, Y., Ebrahimi, A., Liu, P., Woeste, K., Zhao, P., & Zhang, S. (2021). Whole genome based insights into the phylogeny and evolution of the Juglandaceae. *BMC Ecology and Evolution*, *21*(1), 191.

Zhou, H., Yan, F., Hao, F., Ye, H., Yue, M., Woeste, K., Zhao, P., & Zhang, S. (2023). Pan-genome and transcriptome analyses provide insights into genomic variation and differential gene expression profiles related to disease resistance and fatty acid biosynthesis in eastern black walnut (Juglans nigra). *Horticulture Research*, *10*(3), uhad015.

Zhu, T., Wang, L., You, F. M., Rodriguez, J. C., Deal, K. R., Chen, L., Li, J., Chakraborty, S., Balan, B., Jiang, C.-Z., Brown, P. J., Leslie, C. A., Aradhya, M. K., Dandekar, A. M., McGuire, P. E., Kluepfel, D., Dvorak, J., & Luo, M.-C. (2019). Sequencing a Juglans regia × J. microcarpa hybrid yields high-quality genome assemblies of parental species. *Horticulture Research*, *6*, 55.