A haplotype-resolved chromosome-level assembly and annotation of European hazelnut (*C. avellana* cv. Jefferson) provides insight into mechanisms of eastern filbert blight resistance

S.C. Talbot[1], K.J. Vining[1], J.W. Snelling[1], J. Clevenger[2], and S.A. Mehlenbacher[1].

[1]Department of Horticulture, Oregon State University, Corvallis, Oregon, USA; [2]Hudson Alpha Institute for Biotechnology, Huntsville, Alabama, USA.

Abstract:

European hazelnut (*Corylus avellana* L.) is an important tree nut crop. Hazelnut production in North America is currently limited in scalability due to *Anisogramma anomala,* a fungal pathogen that causes Eastern Filbert Blight (EFB) disease in hazelnut. Successful deployment of EFB resistant cultivars has been limited to the state of Oregon, where the breeding program at Oregon State University (OSU) has released cultivars with a dominant allele at a single resistance locus identified by classical breeding, linkage mapping, and molecular markers. 'Jefferson' is resistant to the predominant EFB biotype in Oregon and has been selected by the OSU breeding program as a model for hazelnut genetic and genomic research. Here, we present a near complete, haplotype-resolved chromosome-level hazelnut genome assembly for *C. avellana* 'Jefferson'. This new assembly is a significant improvement over a previously published genome draft. Analysis of genomic regions linked to EFB resistance and self-incompatibility confirmed haplotype splitting and identified new gene candidates that are essential for downstream molecular marker development, thereby facilitating breeding efforts.

Keywords: Chromosome-level, haplotype-resolved, *Corylus*, European hazelnut, genome, fungal disease resistance genes

Introduction

European hazelnut (*Corylus avellana* L.) is an important specialty tree nut crop that is grown in temperate climates for use in the in-shell and kernel markets, typically consumed raw or roasted, in confectionaries and baked goods. The estimated value of the global hazelnut industry is three billion US dollars with Turkey representing nearly 70% of global production (FAO, 2022). Hazelnut (2n = 2x = 11) is a woody perennial that is monecious, dichogamous, wind-pollinated, and self-incompatible (Hill et al., 2021). While all hazelnut species produce edible nuts, the European hazelnut (*Corylus avellana* L.) is the most widely grown because of its desirable characteristics such as a large high-quality nuts, thin shells, and desired flavor profile. Traditional cultivars are clonally propagated and originated as selections from the wild in Europe and western Asia (Mehlenbacher and Molnar, 2021).

Commercial hazelnut production in North America has been limited due to the high susceptibility of European hazelnut to *Anisogramma anomala,* a biotrophic ascomycete, and the causal agent of the eastern filbert blight (EFB) disease. *A. anomala* has co-evolved with its endemic host, the American hazelnut (*Corylus americana*), and in the wild, the disease is widely tolerated (Capik and Molnar, 2012; Revord et al., 2020). Symptoms of EFB are apparent ~18 months following initial infection, and include branch die-back, girdling of trunks, and eventual tree and orchard death. While management techniques such as pruning, scouting for cankers, and applying fungicides can slow the disease's spread, they do not eliminate it (Pscheidt and Ocamb, 2022). Thus, breeding for genetic resistance is considered the most sustainable approach to managing EFB.

Oregon State University (OSU) has been a leader in developing improved EFB resistant cultivars for the Pacific Northwest (PNW), where Oregon represents 95% of US hazelnut production.

The OSU hazelnut breeding program's primary contribution to EFB-resistant cultivar development can be traced to a 1975 discovery in southwest Washington of the obsolete pollinizer, 'Gasaway', which was completely free of EFB in a highly infected and dying orchard of 'DuChilly' (Thompson et al., 1996). To date multiple resistant pollinizers and cultivars derived from 'Gasaway' have been released (Mehlenbacher, 2021), and underlie the expansion of acreage planted in Oregon, which increased from ~11,000 ha in 2009 to greater than 25,000 ha in 2022 (USDA-NASS, 2023). Outside of Oregon, however, cultivars with 'Gasaway' resistance have been shown to be susceptible to genetically diverse *A. anomala* populations (Muehlbauer et al., 2019). Indeed, a genome assembly of the pathogen has shown that it has one of the largest Ascomycota genomes suggesting a high capacity for pathogenic variation (Cai et al., 2013). The long-term durability of Oregon's commercial hazelnut orchards and the potential for expanding hazelnut production is limited by the pathogen's variability and narrow resistance offered by 'Gasaway'.

The availability of genomic resources in *Corylus* has been increasing in recent years. The cultivar 'Jefferson' was chosen for the first *Corylus* genome assembly because it contains 'Gasaway' EFB resistance and it was selected from the reference mapping population (Mehlenbacher et al., 2006). However, the Illumina-based first draft was highly fragmented due to hazelnut's highly heterozygous nature and the limitations imparted by short-read sequencing and assembly technologies (Rowley et al., 2018). With advances in long-read sequencing, pseudo-chromosome level genome assemblies for *Corylus* have been made available for *C. avellana* cultivars 'Tombul' and 'Tonda Gentile delle Langhe' (Lucas et al., 2021; Pavese et al., 2021) and representative accessions of two *Corylus* species, *C. heterophylla* Fisch (Liu et al., 2021; Zhao et al., 2021) and *C. mandshurica* Maxim (Li et al., 2021). However, these genome assemblies are collapsed and there has been no haplotype-resolved "phased" assembly that represents both homologous chromosomes.

Distinguishing between the two chromosomes is essential for determining the parental allelic contributions to self-incompatibility, EFB resistance, and other traits.

EFB resistance derived from ‘Gasaway’ has been characterized as a dominant allele at a single locus with 1:1 segregation (Mehlenbacher et al., 1991, 2006). This source of resistance has been mapped to linkage group (LG) 6 of the genetic map using random amplified polymorphic DNA (RAPD) and simple sequence repeat (SSR) markers in a segregating population from a cross between two heterozygous clones, susceptible ‘OSU 252.146’ x resistant ‘OSU 414.062’ (Mehlenbacher et al., 2006). From this mapping population, the elite cultivar ‘Jefferson’ was identified for release and was the source of the first *Corylus* draft genome (Mehlenbacher et al., 2011; Rowley et al., 2012). Fine mapping of the ‘Gasaway’ region using bacterial artificial chromosomes (BACs) identified a span of approximately 135 kb and five candidate EFB resistance genes (Sathuvalli et al., 2017). Other sources of EFB resistance have been identified and mapped in over 30 *C. avellana* cultivars and accessions, and while the majority map to LG6 (Sathuvalli et al., 2012; Colburn et al., 2015; Komaei Koma 2020), other sources of qualitative and quantitative resistance have been mapped to LG2 (Sathuvalli et al., 2011a; Şekerli et al., 2021), LG7(Bhattarai et al., 2017; Sathuvalli et al., 2011b; Şekerli et al., 2021), LG10 and LG11 (Lombardoni et al., 2022), and more recently LG4 and LG1 (unpublished). A complete summary of resistant cultivars and their related linkage group can be found in Table 1 of Mehlenbacher et al. (2023). The development of elite EFB resistant cultivars is a major goal in hazelnut breeding; however, the lengthy field evaluations provide more robust data on phenotypic variation. The accurate identification of candidate genetic parental contributions underlying qualitative and quantitative loci for EFB resistance will significantly aid in selection across a diverse collection of *Corylus* germplasm, thereby allowing for development of cultivars with multiple resistance loci.

The largest class of characterized plant disease resistance (R) genes encode N-terminal Nucleotide Binding Site (NBS) and C-terminal Leucine-Rich-Repeat (LRR) functional domains (McHale et al., 2006). The LRR domain is highly variable within and among plant species and is typically associated with direct or indirect pathogen effector protein interactions (Prigozhin and Krasileva, 2021). R-genes are often localized into clusters within chromosomes and can have significant variations in encoded amino acid sequence motifs, even within specific categories of R-genes (Kroj et al., 2016; Bailey et al., 2018; Wang and Chai, 2020). Extensive research conducted over the past two decades has demonstrated the successful deployment of NBS-LRR R-genes in a wide range of crops (Kourelis and van der Hoorn, 2018). Investigating the complex molecular mechanisms of R-genes both within and across different plant species is an expensive and resource-intensive task. Past work has identified candidate EFB resistance genes in ‘Jefferson’, however, the functional descriptions are more than a decade old, and recent improvements in genome assembly, annotation algorithms, and curated databases of plant genomes represent an opportunity to improve the description of candidate R-genes. To better direct future research in the 'Gasaway' resistance region, it is crucial to update the annotation of *Corylus* R-gene candidates and evaluate them for protein domain similarities. This analysis will offer insights into the putative functionality of these genes and help determine if other sources of EFB resistance share similar molecular components.

Hazelnut orchard design and elite cultivar development also require an understanding of self-incompatibility. Hazelnut exhibits sporophytic self-incompatibility (SSI), whereby compatibility between cultivars is determined by the genotypes of the plants. Incompatibility is determined by a single highly polymorphic locus, with a minimum of two genes, one each for male and female identity. The best-characterized example of SSI is in *Brassica*, which consists of two genes related to pollen-stigma recognition: a female serine/threonine receptor kinase and a cysteine-rich protein that serves as the pollen's credentials for compatibility interactions (Takasaki et al., 2000; Schopfer et al., 1999).

Both proteins co-localize in clusters on the genome containing similar sequences and in the plasma membrane, and are thought to be adapted from pre-existing signaling systems related to pathogen defense (Zhang et al., 2011). To identify SI alleles in hazelnut, the current method is a time-consuming process that requires a library of tester pollens and fluorescence microscopy to visualize pollen germination (Mehlenbacher, 1997); a total of thirty-three SI alleles have been identified thus far with an nine-level dominance hierarchy (Mehlenbacher, 2014). The locus responsible for SI has been mapped to LG 5 (Mehlenbacher et al., 2006). Fine mapping of this locus revealed a region spanning 193 kb and containing 18 predicted genes that differentiate between two SI-alleles, S_1 and S_3 (Hill et al., 2021). Previous studies have shown that *Corylus* displays a unique SSI mechanism and is independent of the well-characterized SSI system in Brassica (Hou et al., 2022). Remapping the SI locus will increase the precision of molecular marker development for SI-alleles, enabling further investigation into the genic contributions from parental plants. This will also help reveal the molecular mechanisms involved in *Corylus* SSI and identify candidate genes responsible for SI specificity.

Here we present a chromosome-length haplotype-resolved genome assembly and annotation of 'Jefferson'. The assembly was produced using Pacific Biosciences HiFi reads and chromosome-scaffolded using high throughput chromosome conformation capture (Hi-C) sequence data. The practical value of this genome assembly is demonstrated by the separation of the two parents into haplotypes at the previously mapped locus for self-incompatibility alleles. Additionally, haplotype separation identified new candidate genes derived from the parent that contributed 'Gasaway' EFB resistance, providing insight into the molecular mechanisms of resistance.

Materials and methods

Plant material

The *C. avellana* cultivars 'Jefferson', and its parents, female 'OSU 252.146' and male 'OSU 414.062' were used for genome sequencing and assembly. 'OSU 252.146' is susceptible to EFB and carries the SI-alleles S_3 and S_8, whereas 'OSU 414.062' has 'Gasaway' resistance and is homozygous for the SI-allele S_1. Young leaf material was collected from field grown trees in Corvallis, Oregon, USA. Plants were dark-caged for 2-3 days prior to collection, and collected leaves were frozen in liquid nitrogen for Illumina, PacBio, and Hi-C sequencing. For same-day flow cytometry analysis, young leaf tissue was collected in the early morning of May 2020, from a field grown tree of 'Jefferson' following leaf budbreak. Flow cytometry reference material was collected the same day from young tomato leaf tissue (*Solanum lycopersicum* L. 'Stupicke') from two-week old potted plants grown in the greenhouse.

DNA extraction, library preparation, and sequencing

PacBio library prep and sequencing were done at the University of Oregon Genomics & Cell Characterization Core Facility (GC3F). High molecular weight genomic DNA was extracted from flash-frozen leaves. Two 8M SMRT cells were sequenced for 'Jefferson'. To generate HiFi reads, SMRTbell subreads were combined and post-processed with default parameters (CCS.how). Illumina library prep and sequencing of the parents, 'OSU 252.146' and 'OSU 414.062,' were done at GC3F according to then current Illumina HiSeq 4000 protocols and the iTRU library prep protocol (Glenn et al., 2019) to generate 150 bp paired-end (PE) reads. For Hi-C sequencing, tissue processing, chromatin isolation, and library preparation was performed by Dovetail Genomics (Santa Cruz, CA, USA). The parental libraries were prepared in a manner similar to that of Erez Lieberman-Aiden et al. (2009) and sequenced as 150bp PE reads using the Illumina Hiseq 4000 platform. Illumina reads were demultiplexed using the Stacks v2.0 Beta 10 process_radtags module (Rochette et al., 2019).

Demultiplexed reads were checked for quality using FASTQC (version 0.11.5) (Andrews, 2010) and then cleaned by removing adapters, trimming, and quality filtering using the BBTools software suite (Bushnell, 2016); the filterbytile.sh script was used to remove reads associated with low-quality regions of the flow cells containing bubbles, BBDuk was then implemented to trim or remove contaminating iTRU adapters, keep paired reads larger than 130bp, and quality filtering removed reads below Q20.

Flow cytometry

Flow cytometry was done on 'Jefferson' using the propidium iodide (PI) staining technique (Doležel et al., 2005). Solutions of nuclei extraction buffer and staining buffer for PI were prepared using the Cystain® PI kit according to manufacturer protocols (Sysmex, Lincolnshire, IL). Tomato (*Solanum lycopersicum* L. 'Stupicke') was used as a reference standard. The 2C DNA content of tomato has been determined to be 1.96 picograms (pg), where 1pg DNA = 0.978 x 10^9 bp (Doležel et al., 2005). Absolute genomic DNA was calculated by the following formula:

$$Sample\ 2C\ DNA\ content = \left[\frac{(sample\ G1\ peak\ mean)}{(standard\ G1peak\ mean)}\right] x\ standard\ 2C\ DNA\ content\ (pg\ DNA)$$

Briefly, sliced leaf squares of tomato and 'Jefferson' of equal size (~0.5cm^2) were placed in a petri dish together before the addition of 0.5 mL of nuclei extraction buffer. The *C. avellana* samples and tomato standard samples were co-chopped for 30 seconds using a razor blade prior to filtering through a 30 µm nylon-mesh CellTrics® into a 3.5 mL tube. Then, 2 mL of PI staining solution was added to the remaining tissue within the filter. The mixture was incubated at room temperature for 30 minutes inside a Styrofoam cooler to protect against light. Two replicated runs were conducted on different days to account for instrument variation. Stained nuclei were analyzed using a QuantaCyte Quantum P flow cytometer and CyPad software version 1.1. A minimum of 15,000 nuclei counts occurred before the manual gating of G1 sample and standard peaks for each run.

Genome sequence assembly

An initial Genome size was estimated with a *k-mer* analysis of HiFi reads using Jellyfish (version 2.3.0, RRID: SCR_005491) and the web version of GenomeScope (version 2.0, RRID: SCR_017014) with settings: *k-mer* length of 21 and read length of 15,000 bp (Marçais et al., 2011; Vurture et al., 2017). A haplotype-resolved contig assembly was generated using hifiasm trio-partition algorithm (version 0.16.1-r375, RRID: SCR_021069) (Cheng et al., 2021). First, individual *k-mer* counts of parental Illumina reads of the parents 'OSU 252.146' and 'OSU 414.062' were acquired using Yak (version 1.1) as input evidence for hifiasm trio binning. The Arima Hi-C mapping pipeline was followed to generate mapped Hi-C reads (Github.com/ArimaGenomics/mapping_pipeline). YaHs (version 1.1, RRID: SCR_022965) was run independently on both haplotype assemblies produced by hifiasm with their respective Hi-C aligned, read-name sorted bam file (Zhou et al., 2022). A Hi-C contact map was generated for each respective haplotype. Contigs were combined and gapfilled using Juicebox (version 1.11.08, RRID: SCR_021172) (Durand et al., 2017); finalized Hi-C contact maps were curated by Hudson Alpha (Huntsville, AL, USA), using an unpublished Hi-C scaffolding and alignment tool that oriented 'Jefferson' chromosomes based on the 'Tombul' genome pseudo-chromosomal scaffolds (Lucas et al., 2020). To verify haplotype assignment accuracy, parental reads were realigned to each haplotype assembly. Final assembly metrics were generated by QUAST (version 5.0.0, RRID: SCR_001228) (Mikheenko et al., 2018). Assembly completeness was assessed with BUSCO (version 5.4.6, RRID: SCR_015008) in genome mode, using the Embryophyta odb10 dataset (Manni et al., 2021). The quality of assembling repetitive genomic regions were assessed using the long terminal repeat (LTR) assembly index (LAI); this pipeline was composed of LTRharvest within GenomeTools (version 1.6.1, RRID: SCR_016120), LTR_FINDER (version 1.2, RRID: SCR_015247), and LTR_retriever (version 2.9.4, RRID: SCR_017623) using suggested default parameters to predict and combine likely full length candidate LTR-RTs (retrotransposons) (Ou et al., 2018). Calculation of the LAI index was based on the formula: LAI= (intact LTRs/total LTR length) x 100.

Structural gene annotation

Gene prediction and annotation was facilitated by Illumina transcriptome data from the following sources: 1) 'Jefferson' style, bark and leaf tissue, *C. avellana* 'Barcelona' catkins, whole seedling of 'OSU 954.076' x 'OSU 976.091' including root tissue (Rowley et al., 2012; Sathuvalli, unpublished); and 2) leaf bud tissue from *C. avellana* 'Tombul', 'Çakildak', and 'Palaz', publicly available from the National Center for Biotechnology Information (SRA: PRJNA316492) (Kavas et al., 2020). The resulting set of reads putatively representing *C. avellana* was ~423 million PE 150bp RNA-seq reads. Similarly, a protein set consisting of 61,590 annotated proteins was curated from a previous unpublished version 3 'Jefferson' genome assembly and *C. avellana* 'Tombul' (Lucas et al., 2021). Gene annotation was performed for both 'Jefferson' haplotype assemblies. To create a repeat library of transposable element families, a RepeatModeler (RRID: SCR_015027) families set was concatenated with the haplotype-resolved chromosome-level assemblies of 'Jefferson' and six other OSU *C. avellana* accessions that were trio-assembled using the same methods as 'Jefferson' but without chromosome scaffolding (unpublished). Low complexity DNA sequences and repetitive regions were soft masked prior to gene annotation using the default parameters of RepeatMasker (version 4.1.0, RRID: SCR_012954). Structural annotations of protein-coding genes were identified using the gene prediction software AUGUSTUS, GeneMark-ES/EP+, and GenomeThreader, integrated by BRAKER1 and BRAKER2 (RRID: SCR_018964) (Stanke et al., 2006a,b, 2008; Li et al., 2009; Barnett et al., 2011; Gremme, 2013; Lomsadze et al., 2014; Buchfink et al., 2015; Hoff et al., 2016, 2019; Brůna et al., 2020, 2021). First, BRAKER1 used a unique .bam file generated from the splice-aware aligner Hisat2 (Kim et al., 2019), of the previously described RNA-seq set aligned to each haplotype assembly. Second, BRAKER2 was run using the AUGUSTUS *Arabidopsis thaliana* training set and gene structures were predicted via spliced alignments with AUGUSTUS ab-initio and GenomeThreader integration on each masked haplotype genome using the combined protein dataset previously described.

Gene predictions of the respective BRAKER1 and BRAKER2 haplotype runs were assessed for quality, deduplicated, and combined using TSEBRA with default settings (Gabriel et al., 2021).

To further improve this original gene annotation set, BRAKER3 was used (Gabriel et al., 2023). A new masked genome was generated for both haplotype assemblies using EDTA (version 2.1.0, RRID: SCR_022063) (Ou et al., 2019) with parameters: --anno 1 --cds --sensitive including the respective coding sequences and gene locations generated by the BRAKER1/BRAKER2 pipeline. Finalized gene prediction sets were produced using BRAKER3 that included soft-masked genomes, a curated Viridiplantae ODB11 protein set consisting of roughly 5.3 million proteins, and the previously described RNA-seq dataset. BRAKER3 outputs were used as input for TSEBRA, with the -k parameter, to enforce and recover potential missing genes and transcripts produced by the BRAKER1/BRAKER2 pipeline.

Functional gene annotation

Both haplotype annotation sets from TSEBRA were subject to predictive functional analysis using the transcript set within OmicsBox (version 3.0); the OmicsBox pipeline included CloudBLAST using BLASTx, InterPro, GO Merge, GO Mapping, and GO Annotation plus validation (Altschul et al., 1990; Götz et al., 2008; Paysan-Lafosse et al., 2022). Completeness of the predicted annotation sets was assessed using BUSCO --protein mode, inputting translated amino-acid sequences derived from CDS of gene transcripts and the Embryophyta odb10 dataset. To assess long-range structural variation between haplotype assemblies, translocations, inversions, and copy number variation were identified using minimap2 (version 2.23-r1111, RRID: SCR_018550) (Li H., 2018), and SyRI (version 1.6.3, RRID: SCR_023008) and visualized by plotsr (Goel et al., 2019, 2022). Conservation of putative high confidence homologs between assemblies were compared using Orthofinder (version 2.5.4, RRID: SCR_017118) (Emms and Kelly, 2019).

Identification of candidate genes for EFB resistance and self-incompatibility

To identify potential disease resistance gene homologs, the amino acid sequence of annotated protein-coding genes from each assembly were queried against the Plant Resistance Gene Database (version 3.0) using DRAGO2-api (Osuna-Cruz et al., 2018). DNA alignments of previously identified RAPD and SSR marker sequence fragments, BAC-end libraries, and annotated protein-coding genes from 'Jefferson' were aligned to the new genome assemblies using minimap2 (Heng Li, 2018). Marker locations were secondarily assessed for off-target allele-size amplification and multimapping by *in silico* PCR using each marker's corresponding primer pair mapped against the Jefferson V4 haplotype 1 and 2 genomes, allowing for 1-2 mismatches per primer pair. A multiple sequence alignment of the translated candidate R-genes from each haplotype was generated with MUSCLE (version 5.1.0, RRID: SCR_011812) using default settings (Edgar, 2021). A phylogenetic tree of these sequences was created using the neighbor joining tree (BLOSUM62) calculation in JalView (Waterhouse et al., 2009). MEME software (version 5.4.1, RRID: SCR_001783) was utilized to identify conserved subdomains among the putative R-gene candidate proteins using the settings: -mod anr -nmotifs 10 -protein (Bailey et al., 2009).

In a similar approach, genes involved in self-incompatibility were remapped to both haplotype assemblies using previously identified fine-mapped markers and gene sets (Hill et al., 2021). These markers and genes served as query evidence in BLASTn/BLASTp searches of both haplotype assemblies. A multiple sequence alignment of the identified proteins of interest in each haplotype was generated using MUSCLE and visualized using the neighbor joining tree (BLOSUM62) within JalView. The complete genome assembly and annotation pipeline are summarized (Supplemental Figure S1).

Results and discussion

Genome assembly

A combined total of 3.6 million PacBio HiFi reads with an average length of 15,597 bp were generated from two 8M SMRT cells, resulting in 56.8 Gb of sequence data (~147x genome coverage) (Supplementary table S1). For the two parents, 'OSU252.146' and 'OSU414.062', 295 and 218 million PE 150 bp Illumina reads were generated, yielding 44 Gb (115x coverage) and 32 Gb (85x coverage), respectively (Supplemental table S1). These reads were used to generate hifiasm trio binned haploid genome assemblies spanning 385,825,918 bp and 372,534,284 bp, containing 663 and 229 contigs for haplotype 1 and 2, with N50s of 23.4 Mb and 22.5 Mb, respectively (Table 1).

The hifiasm haplotype assemblies were used as inputs to the chromosome scaffolding process. Hi-C sequencing of 'Jefferson' generated ~428 million PE 150 bp reads, for a total yield of ~64.6 Gb (168x coverage, Supplemental table S1). The resulting 'Jefferson V4' Hi-C scaffolded genome assemblies of each haplotype consisted of 11 pseudo-chromosomal scaffolds. The chromosome-level assemblies spanned a total length of 349,702,244 bp and 352,009,510 bp for haplotype 1 and haplotype 2, an N50 of 32.5 Mb and 32.4 Mb (Table 1, Supplemental table S2). The Hi-C interaction matrix clearly differentiated between individual chromosomes in both haplotypes (Figure 1A, 1B). Alignment of parental reads to each genome assembly haplotype showed that the majority of reads from 'OSU 252.146' aligned to haplotype 2, whereas the majority of reads from 'OSU 414.062' aligned to haplotype 1 (Supplemental table S3). BUSCO results in genome mode showed that both chromosome-level haplotype genome assemblies were of high, comparable quality and captured >97% of conserved genes in the Embryophyta dataset (Table 2).

Table 1. Summary statistics for the assembled *C. avellana* ‘Jefferson’ genomes.

Statistics	‘Jeff V4 Hap1’	‘Jeff V4 Hap2’	"Jeff V4 Hap1" Chr-resolved	"Jeff V4 Hap2" Chr-resolved
Total Scaffold number	663	229	11	11
Total assembly Length (Mb)	385.8	372.5	349.7	352
N_{50} (Mb)	23.4	22.5	32.5	32.4
Largest contig (Mb)	34.0	38.9	48.25	47.6
L50 (Mb)	7	7	5	5
Number of contigs merged	NA	NA	22	21
Number of predicted protein-coding genes	NA	NA	33,506	34,379

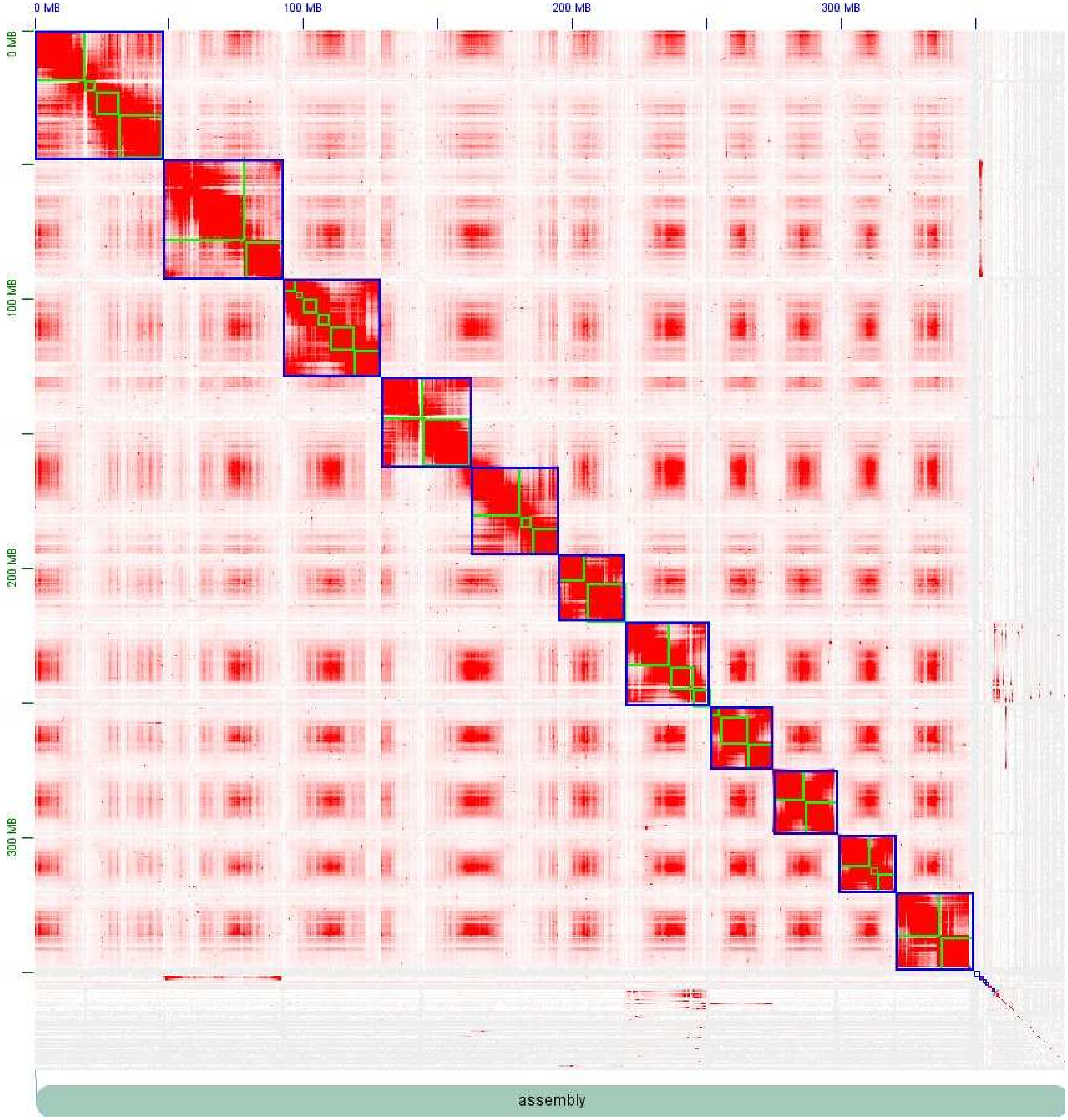


Figure 1A. Hi-C interaction matrix for the 'Jefferson' (*C.* avellana) haplotype 1 assembly. On the X and Y-axes is the distance in the genome assembly (Mb), the green squares represent contigs that are scaffolded within the blue square, which represent a chromosome. The red indicates chromatin interaction loci which are most abundant within chromosomes. The grey space in the lower right represent unaligned contigs which did not have sufficient Hi-C mapping depth to be incorporated into chromosomal scaffolds.

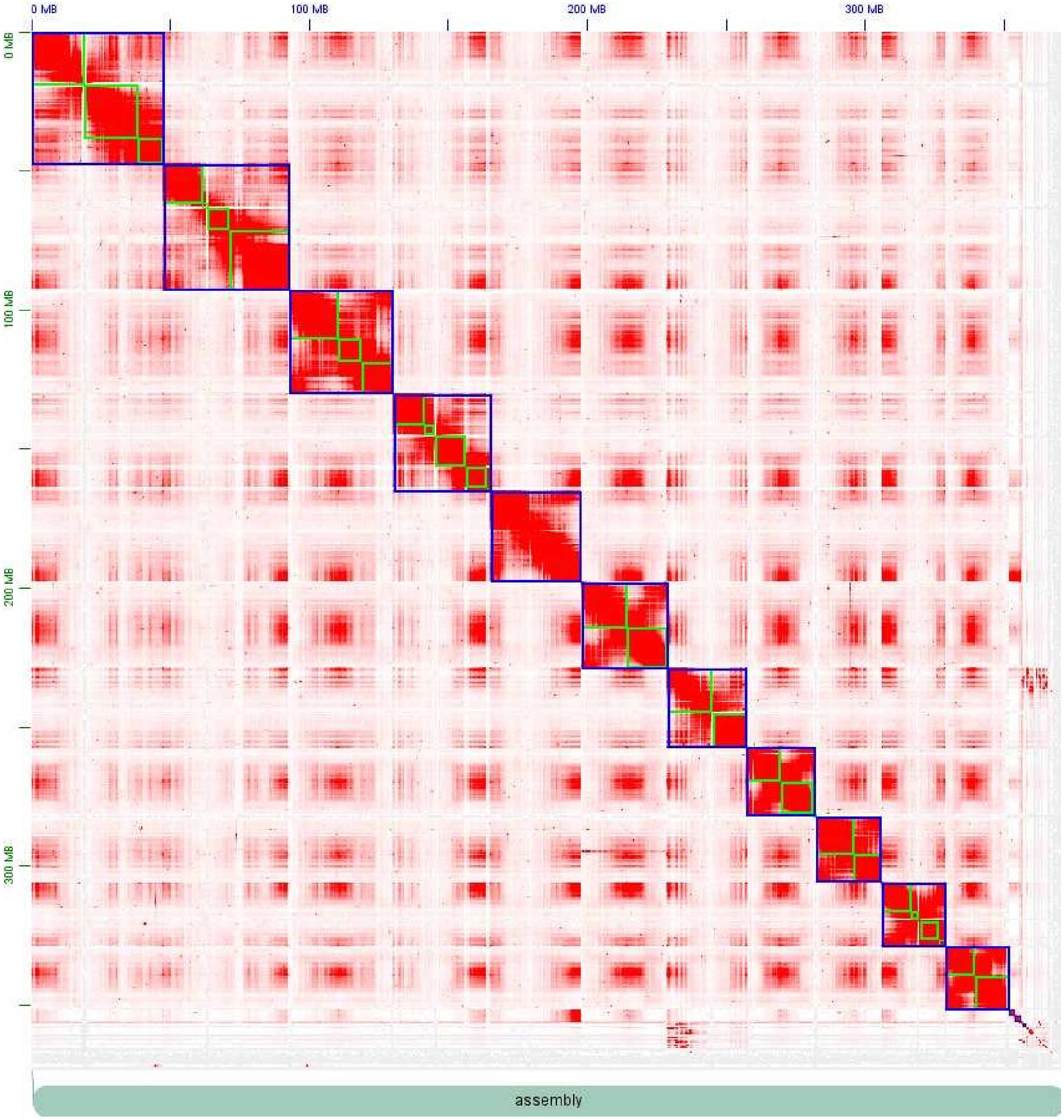


Figure 1B. Hi-C interaction matrix for 'Jefferson' (*C. avellana*) haplotype 2 assembly.

Table 2. Assessment of genome completeness in 'Jefferson' haplotypes using BUSCO.

Searching Model	Protein categories	BUSCO			
		Haplotype 1		Haplotype 2	
		Number	Percentage	Number	Percentage
Genome	Complete BUSCOs (C)	1565	97.0	1575	97.5
	Complete and single-copy BUSCOs (S)	1535	95.1	1534	95.0
	Complete and duplicated BUSCOs (D)	30	1.9	41	2.5
	Fragmented BUSCOs (F)	6	0.4	6	0.4
	Missing BUSCOs (M)	43	2.6	33	2.1
	Total BUSCO groups searched	1614	100.0	1614	100.0
Protein	Complete BUSCOs (C)	1572	97.4	1584	98.2
	Complete and single-copy BUSCOs (S)	1012	62.7	1008	62.5
	Complete and duplicated BUSCOs (D)	560	34.7	576	35.7
	Fragmented BUSCOs (F)	4	0.2	4	0.2
	Missing BUSCOs (M)	38	2.4	26	1.6
	Total BUSCO groups searched	1614	100.0	1614	100.0

Genome size estimation

Flow cytometry was used to estimate a 1C genome size of 'Jefferson' of 365.65 Mb (1C = 0.37 pg). This estimate is slightly smaller than a previously reported estimate of 'Jefferson' (370 Mb) (Rowley et al., 2018) and the reported range of other cultivars and diploid species in the subgenus *Corylus* (C*. cornuta*, *C. colurna*), which was between 1C = 0.41 - 0.43 pg (Bai et al., 2012; Vallès et al., 2014). PacBio HiFi reads of 'Jefferson' were also input to GenomeScope to provide a secondary genome size estimate and heterozygosity of 274.8 Mb and 1.54%, respectively (Supplementary Figure S2). The *k-mer* based estimate is significantly less than the flow cytometry estimate, likely due to limitations of the algorithm in accounting for long-read length and high heterozygosity. The chromosome-resolved assemblies were 4% smaller than the flow cytometry prediction.

Linkage map of 'Jefferson'

The first available *Corylus avellana* linkage map was constructed using random amplified polymorphic DNA and simple sequence repeat (SSRs)markers segregating in an F1 mapping population derived from a cross between 'OSU 252.146' and 'OSU 414.062', the same population from which 'Jefferson' was selected (Mehlenbacher et al., 2004). Since then, this linkage map has been improved by additional SSRs and data from a bacterial artificial chromosome (BAC) library (Sathuvalli et al., 2017; Mehlenbacher and Bhattarai, 2018). To assign the linkage groups to pseudo-chromosomal scaffolds, 18 RAPD, 874 microsatellite, 4,100 paired BAC-ends with proper insert size, and 15,000 biallelic SNP marker sequence fragments were aligned to both Jefferson haplotypes using minimap2 (Li H., 2018), and compared to previous linkage mapping designations (Koma Komaei, 2020). Both haplotypes were successfully assigned the same linkage group for each corresponding pseudo-chromosomal scaffold and renamed appropriately.

Synteny of 'Jefferson' haplotypes

The 'Jefferson' haplotype assemblies showed a high degree of synteny (Figure 2). Differences in length between pseudo-chromosome haplotypes ranged from ~16,000 bp (chromosome 6) to ~2.5 Mb (chromosome 5); most scaffolds representing homologous chromosomes differed in length by an average of ~892 kb. Between haplotypes there were three large scale translocations (chromosome 2, 7, and 9), two inversions (chromosome 5 and 6), and several small duplications, translocations, and gaps. The most notable of non syntenous regions were three large scale translocations on chromosomes two, seven, and nine, comprising total lengths of 14 Mb, 13 Mb, and 9.7 Mb, respectively (Supplemental table S4). Despite nearly 93% of the haplotype assemblies mapping to one another, 33% of the alignments were categorized as having high divergence (Supplemental table S5).

Past cytological work has categorized three chromosome sizes, with two homologous pairs being large, five medium, and three small (Falistocco and Marconi, 2013). Translocations have also been observed in *Corylus* (Salesses and Bonnet, 1988). Reciprocal translocations are thought to frequently confound genetic map generation for many hazelnut populations (Lunde et al., 2006; Bhattarai et al., 2017; Marioni et al., 2018), and are hypothesized to be the result of cytogenetic abnormalities, such as irregular chromosomal migration during cell division, or nondisjunction during microsporogenesis or megasporogenesis (Lagerstedt, 1977). Mono-, bi-, and multi- valent chromosome pairings have been observed frequently in *Corylus* spp. and their hybrids (Woodworth, 1929; Kasapligil, 1968); this suggests that unequal crossover events may be common, especially when diverse germplasm is used. However, it is also possible these apparent translocations are errors from orienting the ‘Jefferson’ Hi-C alignment against ‘Tombul’.

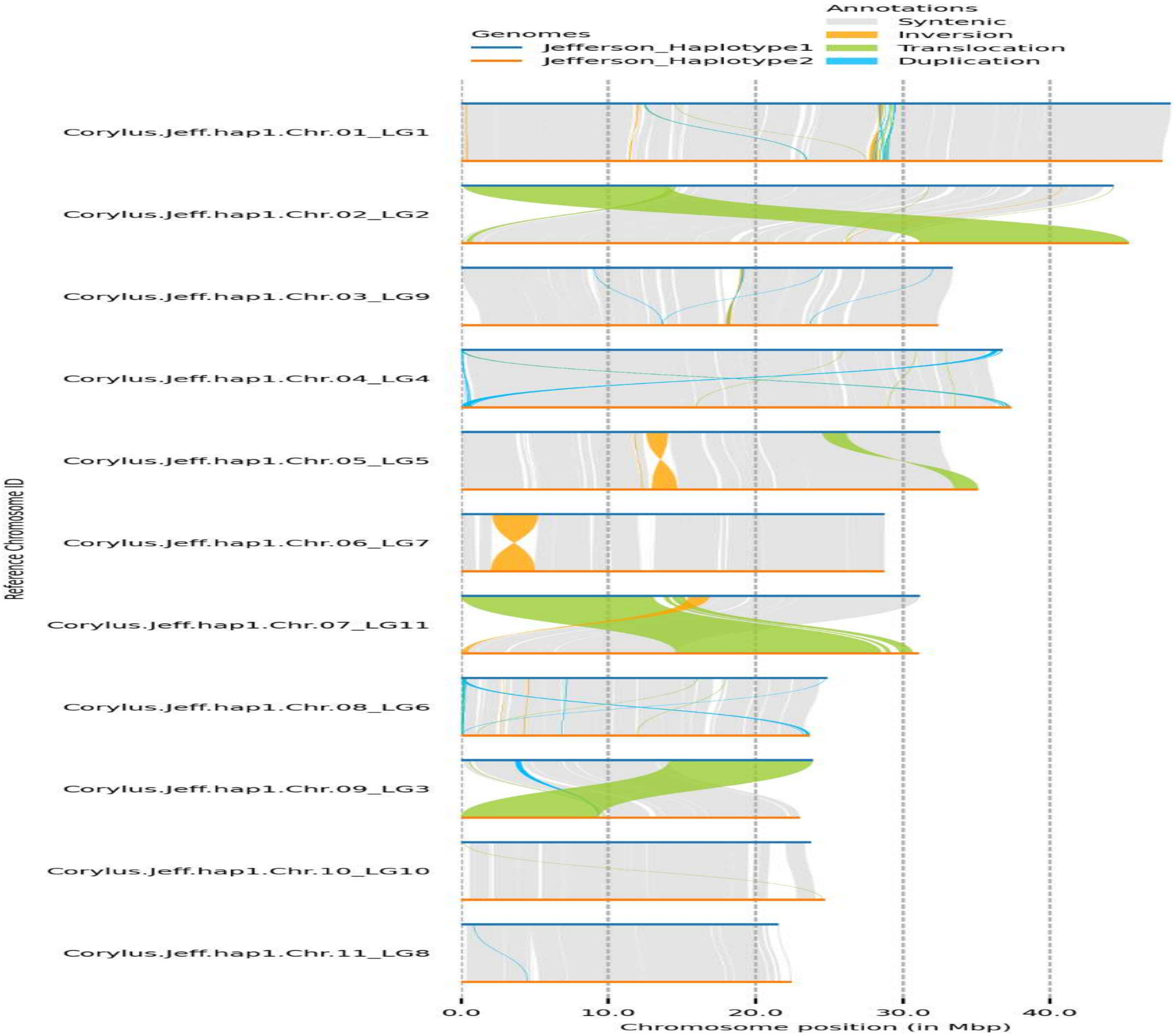


Figure 2. Synteny plot of the two 'Jefferson' chromosome-resolved haplotype assemblies. Pseudo-chromosomal scaffolds of each haplotype were aligned to each other, and labelled on the Y-axis with the chromosome ID and related linkage group. The X-axis shows the chromosome size in Mbp. Chromosomes of haplotypes 1 and 2 are displayed as blue and orange lines, respectively. Grey shading represents complete synteny between genomic positions, yellow represents an inversion, green represents a translocation, and light blue represents a duplication.

Characterization of repeats

Prior to annotating protein-coding genes, genome repeat identification and masking was performed on the chromosome-level haplotype assemblies. The proportion of repeats and unknown elements identified in the initial RepeatModeler and RepeatMasker runs for the 'Jefferson' haplotypes was higher than those reported for other *C. avellana* cultivars and *Corylus* species, with ~65% of bases being masked. The high proportion of LTRs identified suggested potentially erroneous repeat calls that were introduced by the large concatenated LTR families dataset. By rerunning the analysis using EDTA, a more stable view of LTRs was obtained, with 38.26% and 35.29% of repeats masked for haplotype 1 and 2 (Supplemental table S6, S7). Class I retroelements made up 46-54% of all repeats identified for haplotype 1 and 2, respectively. *Gypsy* superfamilies were nearly double those of *Copia*, which is opposite of what has been previously reported in *C. avellana 'Tombul'* but on par with *C. avellana 'Tonda Gentile delle Langhe'* and Silver birch (*Betula pendula*) 'SB1' (Lucas et al., 2021; Pavese et al., 2021; Salojärvi et al., 2017). Nearly 20% of the total repeat length identified in either haplotype had LTRs categorized as 'unknown.' The most significant difference observed between repeat elements of the haplotype assemblies was a doubling of the loosely-defined annotated "repeat_region", with 21 Mb and 9.5 Mb for haplotype 1 and 2, respectively. LAI analysis of haplotype 1 (LAI=16.9) and haplotype 2 (LAI=16.2), indicates that the repetitive and intergenic sequence space is of reference genome quality and a significant improvement from *'Tombul'* (LAI=8.76) (Supplementary Figure S3).

Structural and functional gene annotation

A total of 32,431 and 33,159 protein-coding genes were identified in haplotypes 1 and 2, respectively, and when considering alternative isoforms, these numbers increased to 48,832 and 50,663 coding transcripts, respectively. The protein-coding genes of both haplotype assemblies had an average length of 3,653/3,695 bp, with an average of 3.5 introns per longest isoform and median intron and exon lengths of 232 and 138 bp, respectively. For haplotypes 1 and 2, 21,201/21,354 (~64%) of genes had no alternative isoforms, 7,767/8,089 (~24%) had one alternative isoform and 3,453/3,716 (~11%) had two or more isoforms. For each haplotype's predicted gene set, >97% of *C. avellana* genes were complete BUSCOs for the ODB10 Embryophyta gene families (Table 2). Approximately 35% of highly conserved BUSCO genes were predicted as complete-duplicated, likely due to alternative transcripts.

For haplotype 1, functional annotation analyses assigned GO terms and InterPro domains to 24,369 (72.7%) of transcripts. For the remaining transcripts in haplotype 1, 3,907 (11.4%) had no blast hits, 3,605 (10.8%) had only blast hits, 1,666 (5%) were identified with GO mapping. Similarly for haplotype 2, 24,932 (72.5%) of transcripts were assigned GO terms and InterPro domains. Of the remaining transcripts in haplotype 2, 3,907 (11.4%) had no blast hits, 3,725 (10.8%) had only blast hits, and 1,815 (5.3%) of transcripts were GO mapped (Supplemental Figure S4, S5). OrthoFinder was used to further characterize and assess conservation between predicted gene sets of each haplotype assembly. Of the combined 99,495 transcripts from haplotype 1 and 2, 96,193 (96.7%) were placed in a total of 31,779 orthogroups, with only 4,618 (4.6%) of genes being categorized as unique to a haplotype. To assess the overall distribution of disease resistance genes, DRAGO2 identified 3,620 and 3,659 putative genes with resistance-like domains for haplotype 1 and haplotype 2 assemblies. The majority of these genes identified by DRAGO2 were receptor-like kinases and proteins (~25%), with a small fraction being identified as NBS-LRRs (~10%) (Supplement table S8).

Potential candidate genes for self-incompatibility

The locus for pollen-stigma incompatibility was fine-mapped by Hill et al. 2021, who identified 18 genes within a 193.5 kb region on linkage group 5 that were associated with SI alleles S_1 and S_3. To remap the SI locus, BLASTn was used to align genes from the previous assembly to both chromosome-resolved haplotype assemblies of 'Jefferson.' BLASTn searches returned twelve genes with 100% identity to the S_1 allele among the newly predicted genes in haplotype 1, chromosome 5. In chromosome 5 of haplotype 2, eleven genes with 100% identity to the S_3 allele were identified. Multiple genes that were previously identified as candidates for SI interactions in *Corylus*, PIX7 (Putative interactor of XopAC$_7$) and MIK2 (*MDIS$_1$-interacting receptor like kinase*) were also found in both Jefferson haplotypes. Haplotype 1 contained two copies of PIX7 and eight copies of MIK2, whereas haplotype 2 contained three copies of PIX7 and five copies of MIK2. The SI-locus occupied 86.6 kb in haplotype 1 and 222 kb in haplotype 2. The phasing of alleles within the chromosome 5 SI locus agrees with the previous fine mapping results showing that 'OSU 252.146' contributes S_3 to 'Jefferson', and is represented in the haplotype 2 assembly, whereas 'OSU 414.062' which contributed S_1 to 'Jefferson', is represented in the haplotype 1 assembly.

The similarity of PIX7 and MIK2 candidates was assessed using OrthoFinder, which assigned these genes to seven orthogroups. All seven PIX7 homologs were assigned to three orthogroups, whereas the majority of MIK2 homologs were assigned to a single orthogroup. This suggests that putative PIX7 and MIK2 candidate gene copies are highly conserved, but there may be some variation in protein subdomains that lead to the identification of multiple orthogroups. Indeed, of the eighteen genes identified as PIX7 or MDIS-1 homologs, all were variable in total length (Table 3). Recent studies have shown that in *Brassica*, the most well characterized SSI system, a small RNA is crucial for inducing methylation of recessive SI allele, in order to induce compatibility (Yasuda et al., 2021).

When considering the large number of SI-alleles in *Corylus* (33 to date), it is possible that unannotated sRNA(s) are acting upon different variants of PIX7 or MIK2 to establish allelic dominance. Additional genomes of other *Corylus* cultivars with confirmed SI-alleles will be needed to verify differences in SI-alleles and putative candidate genes to further elucidate the complex molecular mechanism driving SSI and allelic hierarchy in *Corylus*.

Table 3. *Corylus avellana* 'Jefferson' self-incompatibility homologs identified in the self-incompatibility region of both haplotypes of chromosome 5 (LG 5).

***Corylus avellana* gene**	**Amino acid length (bp)**	**Function**
Hap1_g18435	513	probable serine/threonine protein kinase PIX7
Hap1_g18437	695	MDIS1 interacting receptor like kinase 2 like
Hap1_g18438	328	MDIS1 interacting receptor like kinase 2 like
Hap1_g18439	937	MDIS1 interacting receptor like kinase 2 like
Hap1_g18441	357	MDIS1 interacting receptor like kinase 2 like
Hap1_g18442	767	MDIS1 interacting receptor like kinase 2 like isoform
Hap1_g18443	1,056	MDIS1 interacting receptor like kinase 2 like isoform
Hap1_g18444	112	probable serine/threonine protein kinase PIX7
Hap1_g18445	177	MDIS1 interacting receptor like kinase 2 like isoform
Hap1_g18450	787	MDIS1 interacting receptor like kinase 2 like isoform
Hap2_g19113	477	probable serine/threonine protein kinase PIX7
Hap2_g19115	417	MDIS1 interacting receptor like kinase 2-like
Hap2_g19117	937	MDIS1 interacting receptor like kinase 2-like
Hap2_g19118	182	probable serine/threonine protein kinase PIX7
Hap2_g19119	950	MDIS1 interacting receptor like kinase 2-like
Hap2_g19124	793	MDIS1 interacting receptor like kinase 2-like
Hap2_g19138	1,296	MDIS1-interacting receptor like kinase 2-like
Hap2_g19148	513	probable serine/threonine protein kinase PIX7

Potential candidate genes for EFB resistance in hazelnut

In 'Jefferson,' EFB resistance is derived from 'Gasaway' and is conferred by a dominant allele at a single locus that has been mapped between RAPD markers 152-800 and 268-580 on linkage group 6 (Mehlenbacher et al., 2006). Recent QTL (Quantitative Trail Loci) mapping in *C. americana* x *C. avellana* mapping populations associated LG6 EFB resistance in *C. avellana* cv. 'Tonda di Giffoni', with SNP 93212 (Lombardoni et al., 2022). Aligning the associated paired-end sequences from SNP 93212 to 'Jefferson' V4 haplotype 1 placed the QTL peak 20 kb upstream from the markers most closely associated with EFB resistance, and within BAC contig 43F13 in the fine-mapped region defined by Sathuvalli et al. (2017). When mapping the Sanger sequence of CC875206.1 W07-365 (365 bp), the RAPD marker originally extracted from the PCR band associated with W07 'Gasaway' resistance, the sequence is repeated 3 times in this region in both haplotypes of 'Jefferson;' however, the sequence is truncated by ~60 bp in haplotype 2 and spans an additional 100 kb in chromosomal space. Mapping the original Illumina reads from BAC 43F13 to both haplotypes revealed haplotype 1 as the source of the BAC contig and clearly defined the region coinciding with the associated BAC-end markers. The higher percentage of Illumina reads aligning to haplotype 1 from EFB-resistant parent 'OSU 414.062', provides additional support for an EFB-resistance model with R-gene contributions derived from 'Gasaway' present in haplotype 1 only.

Functional annotation of the 'Jefferson' EFB resistance region on haplotypes of chromosome 8 (LG 6) identified several probable receptor-like kinases and putative disease resistance genes. On haplotype 1, a region of approximately 125 kb contained five CNLs identified by DRAGO2 but eight genes with functional descriptions relating to "RGA" (Resistance Gene Analog). On haplotype 1, Hap1_g26572 and Hap1_g26573 were identified as having homology to RGA3 and a short 232aa RGA2-like isoform, respectively. Six other putative resistance genes were identified in haplotype 1, including a long 1,116 aa copy of disease resistance RGA2-like isoform in Hap1_g26576, three copies of RGA3 in Hap1_g26579, Hap1_g26581, and Hap1_g26582, and two copies of RGA4 in Hap1_g26580 and Hap1_g26583.

Similarly, haplotype 2 contained fourteen genes with functional descriptions related to "RGA3" and "RGA2-like isoform" (Table 4), but only eleven were identified as CNLs by DRAGO2. None of the R-genes from haplotype 1 had a 100% match to haplotype 2 R-genes. In Figure 3, the genomic location and orientation of the putative EFB R-gene candidates on chromosome 8 (LG6) are depicted for both haplotypes, showing that RGA3 homologs are closely linked to an RGA2-like isoform and an RGA4 homolog on haplotype 1, whereas R-gene candidates on haplotype 2 are identified as only RGA3 and one as RGA2-like isoforms, all ranging in distance from one another by 20-60 kb.

RGA4 has been characterized as an auto-inducer of immune response to the fungal disease rice blast caused by *Magnaporthe oryzae*, whereby RGA4 is tightly linked with RGA5, with the encoded proteins interacting as a homo and hetero dimer, such that both are required for resistance (Césari et al., 2014). Research suggests that the presence of an integrated heavy metal associated (HMA) domain within RGA5 mimics the pathogen effector target as a "decoy", and upon direct binding to the effector, a signal is transduced to RGA4, relieving RGA4 repression and initiating an immune response (Xi et al., 2022). Heavy metal-associated isoprenylated plant proteins (HIPPs) in rice (*Oryza sativa*) contain HMA domains, and have been identified as putative effector hubs (Bentham et al., 2020; Maidment et al., 2021) as HIPPs have been shown to be the target of multiple fungal effector proteins, having a greater binding affinity to *M. oryzae* AVR-Pik variants than the integrated HMA domains present in rice CC-NLR resistance genes *Pik-1* and *Pik-2* (Maidment et al., 2021). Importantly, HMA domain variants have been shown to perceive new effectors (Césari et al., 2022). On haplotype 1, the genes Hap1_g26587 and Hap1_g26589 were given the functional description "heavy metal-associated isoprenylated plant protein 47" and are located 19 kb and 43 kb upstream, respectively, of the closest RGA4 on the minus strand. Conversely, on haplotype 2, four HMA genes with the same descriptor (Hap2_g27459, Hap2_g27477, Hap2_g27482, and Hap2_g27484) were identified. These genes ranged from 20-72 kb away from the nearest putative RGA3 gene.

HIPP genes of haplotypes 1 and 2, respectively, share high identity with minimal amino acid substitutions among each other. Performing a BLASTp of these predicted proteins against the entire protein set of both haplotypes resulted in matches with other predicted HIPPs, with no homology to suggest that the nearby RGA cluster has a unique synonymous integrated HMA domain like that in rice.

In recent years it has become apparent that cysteine-rich receptor-like secreted proteins (CRRSPs) have crucial involvement in plant-fungal pathogen interactions (Zeiner et al., 2023). *Gnk2* from ginkgo (*Ginkgo biloba*) and two maize (*Zea mays*) proteins, *AFP1* and *AFP2*, bind to mannose during the defense response against fungal pathogens (Miyakawa et al., 2014; Ma et al., 2018). Mannose and its reduced sugar alcohol, mannitol, are independently important to both host plant and fungal pathogen metabolism and signaling during plant growth and pathogen invasion (Patel and Williamson, 2016). CRRSPs have also been shown to be directly involved in fungal pathogen recognition as co-receptors for pathogen effectors (Wang et al., 2023). Recently *TaCRK3,* a CRRSP in wheat, was revealed to inhibit mycelial growth *in vitro (*Guo et al., 2021). Five genes were given the functional description “cysteine-rich repeat secretory protein 38”: two in haplotype 1, Hap1_g26574 and Hap1_g26585, and three in haplotype 2, Hap2_g27475, Hap2_g27480, and Hap2_g27457. To further investigate similarity between these CRRSPs, we performed a BLASTp and used MUSCLE to generate a neighbor-joining tree in JalView (Figure 4). The haplotype 1 gene Hap1_g26574 has two transcripts, with .t1 containing a 20 bp deletion at the 5’ end; the two transcripts have an 86% and 88% similarity to the haplotype 1 gene (Hap1_g26585) and the haplotype 2 genes, respectively, whereas all haplotype 2 genes are 100% identical. These genes contained an extracellular domain composed of two DUF26 (domain of unknown function 26) motifs, but notably lacked an intracellular serine/threonine kinase domain and transmembrane domain (Figure 5).

Despite identifying candidate EFB resistance genes on haplotype 1, the overall similarity between these genes and haplotype 2 R-genes makes it challenging to determine whether one or several resistance genes are involved in the activation of ‘Gasaway’ resistance. It remains to be determined how the unique CRRSP (Hap1_g24474) is involved in processes of pathogen detection and downstream signaling response with close proximity to numerous NBS-LRRs. Thus, it appears that the uncharacterized disease resistance signaling pathway of ‘Gasaway’ involves NBS-LRR RGA homologs and CRRSP, whereby pathotype specific effector(s) might target a decoy of RGA homologs, a unique CRRSP, or possibly both, supporting the traditional R-gene guard-decoy hypothesis. Further research is needed to characterize which haplotype 1 gene(s) are truly responsible for ‘Gasaway’ EFB resistance and whether other EFB resistance sources are derived from this same hypothesized molecular mechanism, with R-gene homologs acting in congruence with unique CRRSP proteins. The hazelnut breeding program at OSU has used many different sources of EFB resistance and sequenced their genomes in an effort to expand knowledge of the allelic diversity of putative resistance gene candidates. Future work in determining EFB resistance mechanisms of other *C. avellana* cultivars should be based on comparisons between the pool of R-genes and CRRSP proteins derived from haplotype 1 of ‘Jefferson’ to prospective EFB resistance genes in order to narrow the list of putative candidate genes.

Table 4. *Corylus avellana* 'Jefferson' candidate EFB R-gene homologs identified in the 'Gasaway' resistance region locus on chromosome 8 (linkage group 6) of both haplotypes.

Corylus avellana **gene**	**Amino acid length (bp)**	**Function**
Hap1_g26572	1,213	putative disease resistance protein RGA3
Hap1_g 26573	237	disease resistance protein RGA2-like isoform X2
Hap1_g 26574	224	Cysteine-rich repeat secretory protein 38
Hap1_g 26576	1,116	disease resistance protein RGA2-like isoform X2
Hap1_g 26579	891	putative disease resistance protein RGA3
Hap1_g 26580	323	putative disease resistance protein RGA4
Hap1_g 26581	1,191	putative disease resistance protein RGA3
Hap1_g 26582	849	putative disease resistance protein RGA3
Hap1_g 26583	274	putative disease resistance protein RGA4
Hap1_g 26585	230	Cysteine-rich repeat secretory protein 38-like
Hap2_g27443	1,215	putative disease resistance protein RGA3
Hap2_g27444	1,164	putative disease resistance protein RGA3
Hap2_g27448	1,145	putative disease resistance protein RGA3
Hap2_g27452	1,159	putative disease resistance protein RGA3
Hap2_g27455	1,159	putative disease resistance protein RGA3
Hap2_g27456	1,150	disease resistance protein RGA2-like isoform X2
Hap2_g27461	1,178	putative disease resistance protein RGA3
Hap2_g27465	1,145	putative disease resistance protein RGA3
Hap2_g27466	847	putative disease resistance protein RGA3
Hap2_g27468	1,159	putative disease resistance protein RGA3
Hap2_g27470	472	putative disease resistance protein RGA3
Hap2_g27471	725	disease resistance protein RGA2-like isoform X2
Hap2_g27473	1,150	disease resistance protein RGA2-like isoform X2
Hap2_g27478	473	putative disease resistance protein RGA3

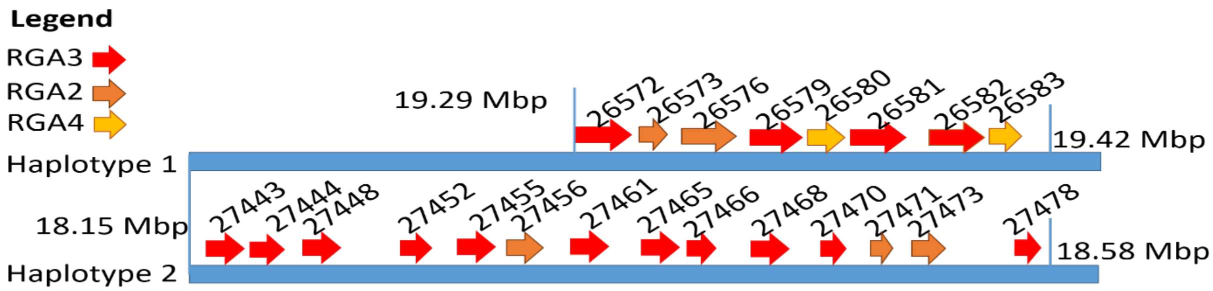


Figure 3. Putative EFB R-gene candidates (RGA-homologs) plotted on chromosome 8 of both haplotypes. Red arrows represent RGA3 homologs, orange arrows represent RGA2 isoform-X2 homologs and yellow arrows represent RGA4 homologs. The gene ID for each respective homolog is listed above the arrow where haplotype 1 represents Hap1_g and haplotype 2 represents Hap2_g. Denoted as vertical lines in Mb are the start and stop positions of the R-gene cluster.

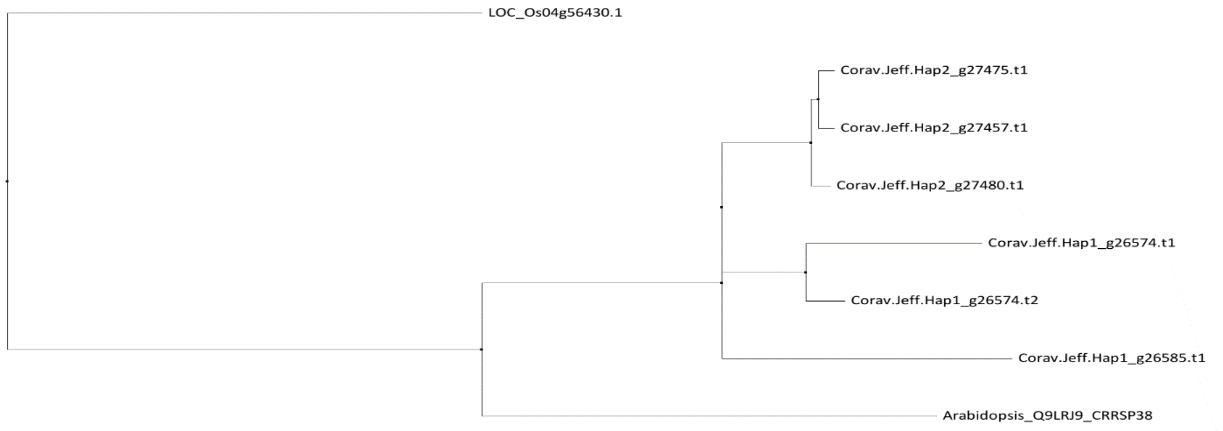
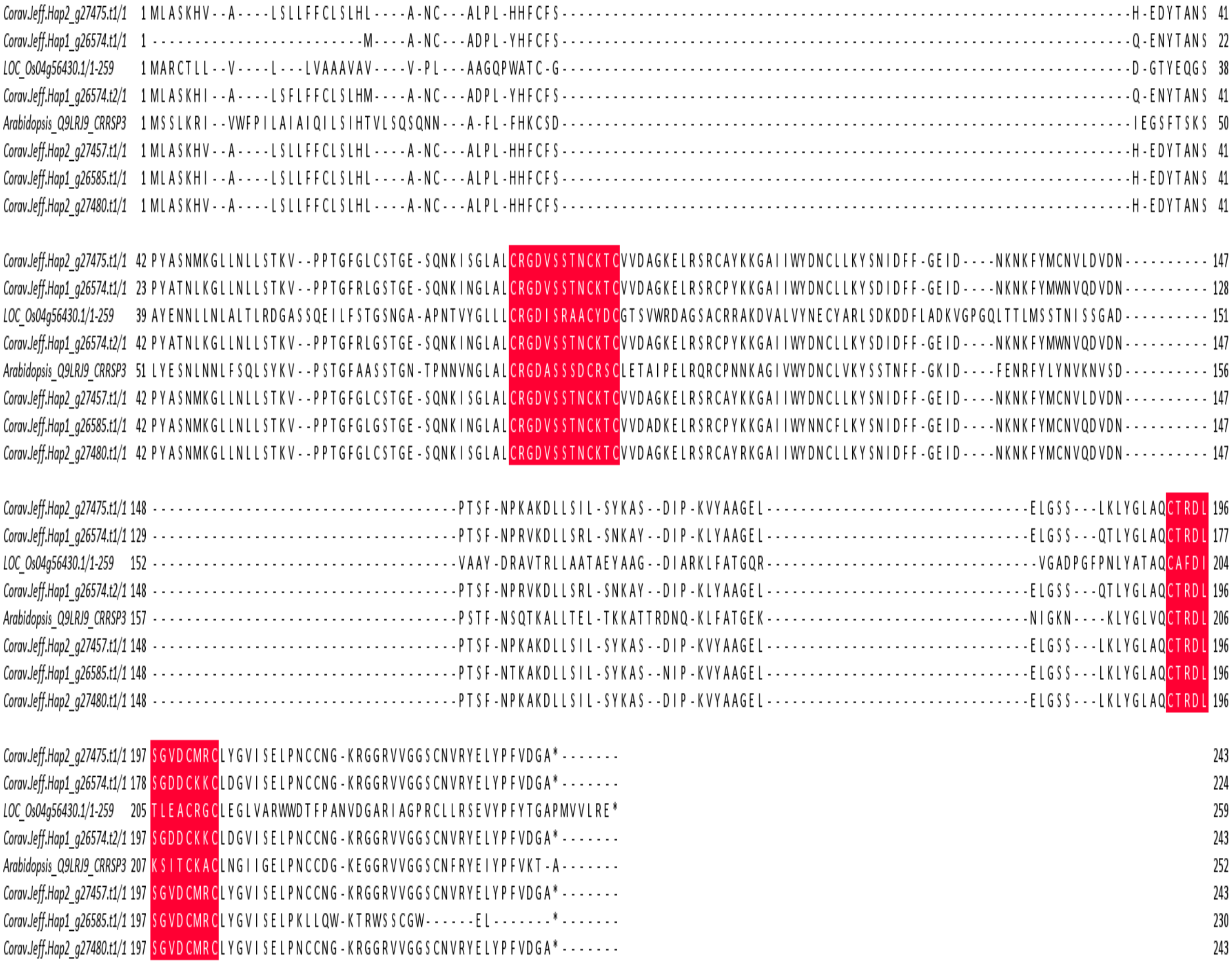

LOC_Os04g56430.1
Corav.Jeff.Hap2_g27475.t1
Corav.Jeff.Hap2_g27457.t1
Corav.Jeff.Hap2_g27480.t1
Corav.Jeff.Hap1_g26574.t1
Corav.Jeff.Hap1_g26574.t2
Corav.Jeff.Hap1_g26585.t1
Arabidopsis_Q9LRJ9_CRRSP38

Figure 4. Neighbor joining tree of seven cysteine-rich secretory proteins (CRSPs) within the EFB R-gene region of both haplotypes with *Arabidopsis* and rice (*RCR3*) homologs aligned by MUSCLE.


Corav.Jeff.Hap2_g27475.t1/1 1 MLASKHV--A----LSLLFFCLSLHL----A-NC---ALPL-HHFCFS--H-EDYTANS 41
Corav.Jeff.Hap1_g26574.t1/1 1 -------------------------M----A-NC---ADPL-YHFCFS--Q-ENYTANS 22
LOC_Os04g56430.1/1-259 1 MARCTLL--V----L---LVAAAVAV----V-PL---AAGQPWATC-G--D-GTYEQGS 38
Corav.Jeff.Hap1_g26574.t2/1 1 MLASKHI--A----LSFLFFCLSLHM----A-NC---ADPL-YHFCFS--Q-ENYTANS 41
Arabidopsis_Q9LRJ9_CRRSP3 1 MSSLKRI--VWFPILAIAIQILSIHTVLSQSQNN---A-FL-FHKCSD--IEGSFTSKS 50
Corav.Jeff.Hap2_g27457.t1/1 1 MLASKHV--A----LSLLFFCLSLHL----A-NC---ALPL-HHFCFS--H-EDYTANS 41
Corav.Jeff.Hap1_g26585.t1/1 1 MLASKHI--A----LSLLFFCLSLHL----A-NC---ALPL-HHFCFS--H-EDYTANS 41
Corav.Jeff.Hap2_g27480.t1/1 1 MLASKHV--A----LSLLFFCLSLHL----A-NC---ALPL-HHFCFS--H-EDYTANS 41
Corav.Jeff.Hap2_g27475.t1/1 42 PYASNMKGLLNLLSTKV--PPTGFGLCSTGE-SQNKISGLALCRGDVSSTNCKTCVVDAGKELRSRCAYKKGAIIWYDNCLLKYSNIDFF-GEID----NKNKFYMCNVLDVDN---------- 147
Corav.Jeff.Hap1_g26574.t1/1 23 PYATNLKGLLNLLSTKV--PPTGFRLGSTGE-SQNKINGLALCRGDVSSTNCKTCVVDAGKELRSRCPYKKGAIIWYDNCLLKYSDIDFF-GEID----NKNKFYMWNVQDVDN---------- 128
LOC_Os04g56430.1/1-259 39 AYENNLLNLALTLRDGASSQEILFSTGSNGA-APNTVYGLLLCRGDISRAACYDCGTSVWRDAGSACRRAKDVALVYNECYARLSDKDDFLADKVGPGQLTTLMSSTNISSGAD---------- 151
Corav.Jeff.Hap1_g26574.t2/1 42 PYATNLKGLLNLLSTKV--PPTGFRLGSTGE-SQNKINGLALCRGDVSSTNCKTCVVDAGKELRSRCPYKKGAIIWYDNCLLKYSDIDFF-GEID----NKNKFYMWNVQDVDN---------- 147
Arabidopsis_Q9LRJ9_CRRSP3 51 LYESNLNNLFSQLSYKV--PSTGFAASSTGN-TPNNVNGLALCRGDASSSDCRSCLETAIPELRQRCPNNKAGIVWYDNCLVKYSSTNFF-GKID----FENRFYLYNVKNVSD---------- 156
Corav.Jeff.Hap2_g27457.t1/1 42 PYASNMKGLLNLLSTKV--PPTGFGLCSTGE-SQNKISGLALCRGDVSSTNCKTCVVDAGKELRSRCAYKKGAIIWYDNCLLKYSNIDFF-GEID----NKNKFYMCNVLDVDN---------- 147
Corav.Jeff.Hap1_g26585.t1/1 42 PYASNMKGLLNLLSTKV--PPTGFGLGSTGE-SQNKINGLALCRGDVSSTNCKTCVVDADKELRSRCPYKKGAIIWYNNCFLKYSNIDFF-GEID----NKNKFYMCNVQDVDN---------- 147
Corav.Jeff.Hap2_g27480.t1/1 42 PYASNMKGLLNLLSTKV--PPTGFGLCSTGE-SQNKISGLALCRGDVSSTNCKTCVVDAGKELRSRCAYRKGAIIWYDNCLLKYSNIDFF-GEID----NKNKFYMCNVQDVDN---------- 147
Corav.Jeff.Hap2_g27475.t1/1 148 ----------------------------------PTSF-NPKAKDLLSIL-SYKAS--DIP-KVYAAGEL------------------------------ELGSS---LKLYGLAQCTRDL 196
Corav.Jeff.Hap1_g26574.t1/1 129 ----------------------------------PTSF-NPRVKDLLSRL-SNKAY--DIP-KLYAAGEL------------------------------ELGSS---QTLYGLAQCTRDL 177
LOC_Os04g56430.1/1-259 152 ----------------------------------VAAY-DRAVTRLLAATAEYAAG--DIARKLFATGQR-------------------------------VGADPGFPNLYATAQCAFDI 204
Corav.Jeff.Hap1_g26574.t2/1 148 ----------------------------------PTSF-NPRVKDLLSRL-SNKAY--DIP-KLYAAGEL------------------------------ELGSS---QTLYGLAQCTRDL 196
Arabidopsis_Q9LRJ9_CRRSP3 157 ----------------------------------PSTF-NSQTKALLTEL-TKKATTRDNQ-KLFATGEK------------------------------NIGKN----KLYGLVQCTRDL 206
Corav.Jeff.Hap2_g27457.t1/1 148 ----------------------------------PTSF-NPKAKDLLSIL-SYKAS--DIP-KVYAAGEL------------------------------ELGSS---LKLYGLAQCTRDL 196
Corav.Jeff.Hap1_g26585.t1/1 148 ----------------------------------PTSF-NTKAKDLLSIL-SYKAS--NIP-KVYAAGEL------------------------------ELGSS---LKLYGLAQCTRDL 196
Corav.Jeff.Hap2_g27480.t1/1 148 ----------------------------------PTSF-NPKAKDLLSIL-SYKAS--DIP-KVYAAGEL------------------------------ELGSS---LKLYGLAQCTRDL 196
Corav.Jeff.Hap2_g27475.t1/1 197 SGVDCMRCLYGVISELPNCCNG-KRGGRVVGGSCNVRYELYPFVDGA*------- 243
Corav.Jeff.Hap1_g26574.t1/1 178 SGDDCKKCLDGVISELPNCCNG-KRGGRVVGGSCNVRYELYPFVDGA*------- 224
LOC_Os04g56430.1/1-259 205 TLEACRGCLEGLVARWWDTFPANVDGARIAGPRCLLRSEVYPFYTGAPMVVLRE* 259
Corav.Jeff.Hap1_g26574.t2/1 197 SGDDCKKCLDGVISELPNCCNG-KRGGRVVGGSCNVRYELYPFVDGA*------- 243
Arabidopsis_Q9LRJ9_CRRSP3 207 KSITCKACLNGIIGELPNCCDG-KEGGRVVGGSCNFRYEIYPFVKT-A------- 252
Corav.Jeff.Hap2_g27457.t1/1 197 SGVDCMRCLYGVISELPNCCNG-KRGGRVVGGSCNVRYELYPFVDGA*------- 243
Corav.Jeff.Hap1_g26585.t1/1 197 SGVDCMRCLYGVISELPKLLQW-KTRWSSCGW------EL-------*------- 230
Corav.Jeff.Hap2_g27480.t1/1 197 SGVDCMRCLYGVISELPNCCNG-KRGGRVVGGSCNVRYELYPFVDGA*------- 243

562

563 **Figure 5.** Amino-acid sequence alignment of all Cysteine-rich repeat secretory protein-38 in *C. avellana* 564 *'Jefferson' and* a homolog from Arabidopsis and Rice (*RMC*). The numbers on the right side indicate the 565 positions of the residues in the corresponding protein. Red shading indicates the conserved motif of the 566 DUF26 domain C-X8-C-X2-C.

567

568

569

570

Conclusions

Here, we report the first haplotype-resolved chromosome-level genome assembly and annotation of the diploid *C. avellana* 'Jefferson'. BUSCO analysis showed that the genome assemblies and structural annotations were of high quality. The ability of haplotype-phasing to identify parental genic contributions was successfully demonstrated by the complete separation of SI-alleles to their respective parental haplotypes. Furthermore, the region associated with 'Gasaway' EFB resistance was remapped with high confidence to the resistant parental haplotype, and several new candidate resistance genes were identified. The molecular mechanism behind 'Gasaway' resistance remains to be investigated, however, the RGA cluster in congruence with a cysteine-rich secretory protein provides evidence of a guard model hypothesis. The haplotype-resolved 'Jefferson' genome assembly and annotation presented here will serve as a powerful resource for hazelnut breeders and plant scientists in the further development of molecular markers for genomics-assisted breeding and facilitate future studies of *Corylus* biology and genetics.

Data availability

The haplotype genome assemblies and annotations of *C. avellana* 'Jefferson' presented here is available at the United States Department of Energy's Joint Genomics Institute Phytozome web browser (accepted, pending release) (available . The 'Jefferson' genome assembly, annotation, and respective read tracks will also be available soon as a genome browser via JBrowse2 at Hazelnutgenomes.oregonstate.edu.

Acknowledgements

SCT performed flow cytometry, genome assembly, quality assessments, Hi-C guided assembly, synteny analysis, structural and functional annotation, and related analyses of potential candidate genes for resistance and self-incompatibility. JC assisted with Hi-C assembly. JWS conducted remapping of linkage groups and provided insight into candidate gene analysis. KJV coordinated research and provided conceptual guidance, and assisted with assembly, annotation, repeat content, and resistance gene analysis. SAM conceived the study and provided overarching guidance. SCT and KJV authored the manuscript. Oregon State University Center for Quantitative Life Sciences provided the support for computational resources used. All authors approved the final manuscript. The Oregon Hazelnut Commission supported this research.

Literature cited

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. Basic local alignment search tool. J. Mol. Biol*,* 1990; 215: 403-410. doi.org/10.1016/S0022-2836(05)80360-2.

Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bai C., Alverson W.S., Follansbee A., Waller D.M. New reports of nuclear DNA content for 407 U.S. plant species. *Annals of Botany*, 2012; 110: 1623-1629. doi.org/10.1093/aob/mcs222

Bailey T.L., Boden M., Buske F.A., Frith M., Grant C.E., Clementi L., Ren J., Li W.W., and Noble W.S. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 2009; 37: W202-W208. doi.org/10.1093/nar/gkp335

Bailey P.C., Schudoma C. Jackson W., Baggs E., Dagdas G., Haerty W., Mouscou M., and Krasileva K.V. Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. *Genome Biology*, 2018; 19(23). doi.org/10.1186/s13059-018-1392-6

Barnett D. W., Garrison E. K., Quinlan A. R., Strömberg M. P., and Marth G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 2011; 12: 1691-1692. doi.org/10.1093/bioinformatics/btr174

Bentham A.R., Concepcion J.C., Mukhi N., Zdrzałek R., Draeger M., Gorenkin D., Hughes R.K., and Banfield M.J. A molecular roadmap to the plant immune system. *Journal of Biological Chemistry*, 2020; 295(44): 14916-14935. doi.org/10.1074/jbc.REV120.010852

Bhattarai G., Mehlenbacher S.A., and Smith D.C. Eastern filbert blight disease resistance from *Corylus americana* 'Rush' and selection 'Yoder #5' maps to linkage group 7. *Tree Genetics & Genomes,* 2017; 13(45). doi.org/10.1007/s11295-017-1129-9

Brůna T., Lomsadze A., and Borodovsky M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, 2020; 2(2): lqaa026. doi.org/10.1093/nargab/lqaa026

Brůna T., Hoff K.J., Lomsadze A., Stanke M., and Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, 2021; 3(1): lqaa108. doi.org/10.1093/nargab/lqaa108

Buchfink B., Xie C., and Huson D.H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 2015; 12(1):59. doi.org/10.1038/nmeth.3176

Bushnell B. 2016. BBTools. https://jgi.doe.gov/data-and-tools/bbtools.

Cai G., Leadbetter C.W., Muehlbauer M.F., Molnar T.J., and Hillman B.I. Genome-wide microsatellite identification in the fungus *Anisogramma anomala* using Illumina sequencing and genome assembly. *PLoS One*, 2013; 8(11): e82408. doi.org/10.1371/journal.pone.0082408

Capik J.M, and Molnar T.J. Assessment of Host (*Corylus* sp.) Resistance to eastern filbert blight in New Jersey. *American Society for Horticultural Science*, 2012; 137(3): 157-172. doi.org/10.21273/JASHS.137.3.157

Césari S., Kanzaki H., Fujiwara T., Bernoux M., Chalvon V., Kawano Y., Shimamoto K., Dodds P., Terauchi R., and Kroj T. The NB-LRR proteins RGA4 and RGA5 interact functionally and physically to confer disease resistance. *EMBO Journal*, 2014; 33(17): 1941-1959. doi.org/10.15252/embj.201487923

Césari S., Xi Y., Declerck N., Chalvon V., Mammri L., Pugnière M., Henriquet C., de Guillen K., Chochois V., Padilla A., and Kroj T. New recognition specificity in a plant immune receptor by molecular engineering of its integrated domain. *Nature Communications*, 2022; 13: 1524. doi.org/10.1038/s41467-022-29196-6

Cheng H., Concepcion G.T., Feng X., Zhang H., and Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 2021; 18:170-175. doi.org/10.1038/s41592-020-01056-5

Colburn B.C., Mehlenbacher S.A., Sathuvalli V.R., and Smith D.C. Eastern filbert blight resistance in hazelnut accessions 'Cuplà', Crvenje', and OSU 495.072. *Journal of the American Society for Horticultural Science*, 2015; 140(2): 191-200. doi.org/10.21273/JASHS.140.2.191

Durand N.C., Robinson J.T., Shamim M.S., Machol I., Mesirov J.P., Lander E.S., and Aiden E.L. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell system*, 2017; 3(1):99-101. doi.org/10.1016/j.cels.2015.07.012

Edgar R.C. MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping. *bioRxiv*, 2021; 06.20.449169. doi.org/10.1101/2021.06.20.449169

Emms D.M., and Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 2019; 238(20). doi.org/10.1186/s13059-019-1832-y

Falistocco E. and Marconi G. Cytogenetic characterization by in situ hybridization techniques and molecular analysis of 5S rRNA genes of the European hazelnut (Corylus avellana). *Genome*, 2013; 56(3): 155-159. doi.org/10.1139/gen-2013-0045

Food and Agriculture Organization of the United Nations. FAOSTAT Statistical Database, 2022; [Accessed 12 August 2022]. https://www.fao.org/faostat/en/#search/hazelnut

Gabriel L., Hoff K.J., Brůna T., Borodovsky M., and Stanke M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics*, 2021; 22(566). doi.org/10.1186/s12859-021-04482-0

Gabriel L., Brůna T., Hoff K.J., Ebel M., Lomsadze A., Borodovsky M., and Stanke M. BRAKER3: fully automated genome annotation using RNA-Seq and protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *BioRxiv,* 2023. doi.org/10.1101/2023.06.10.544449

Glenn T.C., Nilsen R.A., Kieran T.J., Sanders J.G., Bayona-Vásquez N.J., Finger J.W., Pierson T.W., Bentley K.E., Hoffberg S.L., Louha S., Garcia-De Leon F.J., del Rio Portilla M.A., Reed K.D., Anderson J.L., Meece J.K., Aggrey S.E., Rekaya R., Alabady M., Belanger M., Winker K., Faircloth B.C. Adapterama I: universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext). *PeerJ*, 2019; e7755. doi.org/10.7717/peerj.7755

Goel M., Sun H., Jiao W.-B., and Schneeberger K. SyRi: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, 2019; 20(277). doi.org/10.1186/s13059-019-1911-0

Goel M., Sun H., Jiao W.-B., and Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, 2019; 20: 277. doi.org/10.1186/s13059-019-1911-0

Goel M. and Schneeberger K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics*, 2022; 38(10): 2922-2926. doi.org/10.1093/bioinformatics/btac196

Götz S., Garcia-Gomez J.M., Terol J., Williams T.D., Nagaraj S.H., Nueda M.J., Robles M., Talon M., Dopazo J., and Conesa A. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, 2008; 36(10): 3420-3435. doi.org/10.1093/nar/gkn176

Gremme, G. Computational gene structure prediction. 2013. PhD dissertation.

Guo F., Wu T., Shen F., Xu G., Qi H., and Zhang Z. The cysteine-rich receptor-like kinase TaCRK3 contributes to defense against Rhizoctonia cerealis in wheat. *Journal of Experimental Botany*, 2021; 72(20): 6904-6919. doi.org/10.1093/jxb/erab328

Hill R.J., Baldassi C., Snelling J.W., Vining K.J., and Mehlenbacher S.A. Fine mapping of the locus controlling self-incompatibility in European hazelnut. *Tree Genetics & Genomes,* 2021; 17:6. doi.org/10.1007/s11295-020-01485-5

Hoff K. J., Lange, S., Lomsadze A., Borodovsky M., and Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, 2016; 32(5): 767-769. doi.org/10.1093/bioinformatics/btv661

Hoff K. J., Lomsadze A., Borodovsky, M., and Stanke, M. Whole-genome annotation with BRAKER. In *Gene Prediction.* Humana, New York, NY. 2019; 1962: 65-95. doi.org/10.1007/978-1-4939-9173-0_5

Hou S., Zhao T., Yang Z., Liang L., Ma W., Wang G., and Ma Q. Stigmatic transcriptome analysis of self-incompatible and compatible pollination in *Corylus heterophylla* Fisch x *Corylus avellana* L. *Frontiers in Plant Science,* 2022; 13: 800768. doi.org/10.3389/fpls.2022.800768

Kasapligil B. *Corylus colurna* and its varieties. *Journal of the California Horticultural Society, 1963; 24: 95-104.*

Kavas M., Yıldırım K., Seçgin Z., and Gökdemir G. Discovery of simple seqeuence repeat (SSR) markers in hazelnut (*Corylus avellana* L.) by transcriptome sequencing and SSR-based characterization of hazelnut cultivars. *Scandinavian Journal of Forest Research*, 2020; 35(5-6). doi.org/10.1080/02827581.2020.1797155

Kim D., Paggi J.M., Park C., Bennett C. and Salzberg S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 2019; 37: 907-915. doi.org/10.1038/s41587-019-0201-4

Komaei Koma G. High-density linkage maps for European hazelnut (*Corylus avellana* L.) from single nucleotid polymorphism markers and mapping new sources of resistance to eastern filbert blight. *PhD Dissertation*, 2020.

Kourelis J., and van der Hoorn R.A.L. Defended to the nines: 25 Years of resistance gene cloning identifies nine mechanisms for R protein function. *Plant Cell,* 2018; 30(2): 285-299. doi.org/10.1105/tpc.17.00579

Kroj T., Chanclud E., Michel-Romiti C., Grand X., Morel J.B. Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. *New Phytol*, 2016; 210(2): 618-26. doi.org/10.1111/nph.13869.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., and Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics*, 2009; 25(16): 2078-2079. doi.org/10.1093/bioinformatics/btp352

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 2018; 34(18): 3094-3100. doi.org/10.1093/bioinformatics/bty191

Li Y., Sun P., Lu Z., Chen J., Wang Z., Du X., Zheng Z., Wu Y., Hu H., Yang J., Ma J., Liu J., and Yang Y. The *Corylus mandshurica* genome provides insights into the evolution of Betulaceae genomes and hazelnut breeding. *Horticulture Research*, 2021; 8: 54. doi.org/10.1038/s41438-021-00495-1

Lieberman-Aiden E., Berkum van N.L, Williams L., Imakaev M., Ragoczy T., Telling A., Amit I., Lajoie B.R., Sabo P.J., Dorschner M.O., Sandstrom R., Bernstein B., Bender M.A., Groudine M., Gnirke A., Stamatoyannopoulos J., Mirny L.A., Lander E.S., and Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 2009; 326(5950): 289-293. doi.org/10.1126/science.1181369

Liu J., Wei H., Zhang X., He H., Cheng Y., and Wang D. Chromosome-level genome assembly and HazelOmics database construction provides insights into unsaturated fatty acid synthesis and cold resistance in hazelnut (*Corylus heterophylla*). *Frontiers in Plant Science*, 2021; 12:766548. doi.org/10.3389/fpls.2021.766548

Lombardoni J.L., Honig J.A., Vaiciunas J.N., Revord R.S., and Molnar T.J. Segregation of eastern filbert blight disease response and single nucleotide polymorphism markers in three European-American interspecific hybrid hazelnut populations. *Journal of the American Society for Horticultural Science,* 2022; 147(4): 196-207. doi.org/10.21273/JASHS05112-22

Lomsadze A., Burns P.D., and Borodovsky M. Integration of mapped RNA-seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, 2014; 42(15): e119. doi.org/10.1093/nar/gku557

Lucas S.J., Kahraman K., Avşar B., Buggs R.J.A., and Bilge I. A chromosome-scale genome assembly of European hazel (*Corylus avellana* L.) reveals targets for crop improvement. *The Plant Journal*, 2021; 105: 1413-1430. doi.org/10.1111/tpj.15099

Lunde C.F., Mehlenbacher S.A., and Smith D.C. Segregation for resistance to eastern filbert blight in progeny of 'Zimmerman' hazelnut. *Journal of the American Society for Horticultural Science*, 2006; 131(6): 731-737. doi.org/

Ma L., Wang L., Trippel C., Mendoza-Mendoza A., Ullmann S., Moretti M., Carsten A., Kahnt J., Reissmann S., Zechmann B., Bange G., and Kahmann R. The Ustilago maydis repetitive effector Rsp3 blocks the antifungal activity of mannose-binding maize proteins. *Nature Communications*, 2018; 9(1): 1711. doi.org/10.1038/s41467-018-04149-0

Maidment J.H.R., Franceschetti M., Maqbool A., Saitoh H., Jantasuriyarat C., Kamoun S., Terauchi R., and Banfield M.J. Multiple variants of the fungal effector AVR-Pik bind the HMA domain of the rice protein OsHIPP19, providing a foundation to engineer plant defense. *Journal of Biological Chemistry*, 2021; 296: 100371. doi.org/10.1016/j.jbc.2021.100371

Manni M., Berkeley M.R., Seppey M., Simão F.A., and Zdobnov E.M. 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*, 2021; 38(10): 4647-4654. doi.org/10.1093/molbev/msab199

Marçais G., and Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 2011; 27(6):764-770. doi.org/10.1093/bioinformatics/btr011

Marinoni D.T., Valentini N., Portis E., Acquadro A., Beltramo C., Mehlenbacher S.A., Mockler T.C., Rowley E.R., and Botta R. High density SNP and QTL analysis for time of leaf budburst in *Corylus avellana* L. *PLOS One*, 2018. doi.org/10.1371/journal.pone.0195408

McHale L., Tan X., Koehl P., and Michelmore R.W. Plant NBS-LRR proteins: adaptable guards. *Genome Biology*, 2006; 7(212). doi.org/10.1186/gb-2006-7-4-212

Mehlenbacher S.A., Thompson M.M., and Cameron H.R. Occurrence and inheritance of resistance to eastern filbert blight in 'Gasaway' hazelnut. *HortScience*, 1991; 26:410-411. doi.org/ doi.org/10.21273/HORTSCI.26.4.442

Mehlenbacher S.A. Revised dominance hierarchy for S-alleles in *Corylus avellana* L. *Theoretical and Applied Genetics*, 1997; 94: 360-366. doi.org/10.1007/s001220050424

Mehlenbacher S.A., Brown R.N., Nouhra E.R., Gökirmak T., Bassil N.V. and Kubisiak T.L. 2006. A genetic linkage map for hazelnut (Corylus *avellana* L.) based on RAPD and SSR markers. *Genome*, 2006; 49:122-133. doi.org/10.1139/g05-091

Mehlenbacher S.A., Smith D.C., and McCluskey R.L. 2011. 'Jefferson' hazelnut. *HortScience*, 2011; 46: 662-664. doi.org/10.21273/HORTSCI.46.4.662

Mehlenbacher S.A. Geographic distribution of incompatibility alleles in cultivars and selections of European hazelnut. *Journal of American Society of Horticultural Science,* 2014; 139: 191-212. doi.org/10.21273/JASHS.139.2.191

Mehlenbacher S.A. and Bhattarai G. An updated linkage map for hazelnut with new simple sequence repeat markers. Acta Horticulture. 2018; 1226:31-38. doi.org/10.17660/ActaHortic.2018.1226.4

Mehlenbacher S.A. and Molnar T.J. Hazelnut Breeding. In Plant Breeding Reviews, I. Goldman (Ed.), 2021; chapter 2. doi.org/10.1002/9781119828235.ch2

Mehlenbacher S.A, Heilsnis B.J., Mooneyham R.T., and J.W. Snelling. Breeding hazelnuts resistant to eastern filbert blight. *Acta Horticulture ISHS*, 2023; 1362: 557-562. doi.org/10.17660/ActaHortic.2023.1362.75

Mikheenko A., Prjibelski A., Saveliev V., Antipov D., and Gurevich A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, 2018; 34(13):i142-150. doi.org/10.1093/bioinformatics/bty266

Miyakawa T., Hatano K., Miyauchi Y., Suwa Y., Sawano Y., and Tanokura M. A secreted protein with plant-specific cysteine-rich motif functions as a mannose-binding lectin that exhibits antifungal activity. *Plant Physiology*, 2014; 166(2): 766-78. doi.org/10.1104/pp.114.242636

Muehlbauer M.F., Tobia J., Honig J.A., Zhang N., Hillman B.I., Gold K.M., and Molnar T.J. Population differentiation within Anisogramma anomala in North America. *Journal of Phytopathology*, 2019; 109(6): 1074-1082. doi.org/10.1094/PHYTO-06-18-0209-R

Osuna-Cruz C.M., Paytuvi-Gallart A., Donato A.D., Sundesha V., Andolfo G., Cigliano R.A., Sanseverino W., and Ercolano M.R. PRGdb3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Research*, 2018; 46: D1197-D1201. doi.org/10.1093/nar/gkx1119

Ou S., Chen J., and Jiang N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research,* 2018; 46(21): e216. doi.org/10.1093/nar/gky730/

Ou S., Su W., Liao Y., Chougule K., Agda J.R.A., Hellinga A.J., Lugo C.S.B., Elliott T.A., Ware D., Peterson T., Jiang N., Hirsch C.N., and Hufford M.B. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, 2019; 275. doi.org/10.1186/s13059-019-1905-y

Patel T.K., and Williamson J.D. Mannitol in plants, fungi, and plant-fungal interactions. *Trends in Plant Science*, 2016; 21(6): 486-497. doi.org/10.1016/j.tplants.2016.01.006

Pavese V., Cavalet-Giorsa E., Barchi L., Acquadro A., Marinoni D.T., Portis E., Lucas S.J., and Botta R. Whole-genome assembly of *Corylus avellana* cv “Tonda Gentile delle Langhe” using linked-reads (10x Genomics). *G3-Genes Genomes Genetics*, 2021; 11(7). doi.org/10.1093/g3journal/jkab152

Paysan-Lafosse T., Blum M., Chuguransky S., Grego T., Pinto B.L., Salazar G.A., Bileschi M.L., Bork P., Bridge A., Colwell L., Gough J., Haft D.H., Letunić I., Marchler-Bauer A., Mi H., Natale D.A., Orengo C.A., Pandurangan A.P., Rivoire C., Sigrist C.J.A., Sillitoe I., Thanki N., Thomas P.D., Tosatto S.C.E., Wu C.H., and Bateman A. Interpro in 2022. *Nucleic Acids Research*, 2022; gkac993. doi.org/10.1093/nar/gkac993 .

Prigozhin D.M. and Krasileva K.V. Analysis of intraspecies diversity reveals a subset of highly variable plant immune receptors and predicts their binding sites. *The Plant Cell*, 2021; 33(4):998-1015. doi.org/10.1093/plcell/koab013 .

Pscheidt J.W. and Ocamb C.M., senior editors. Pacific Northwest Plant Disease Management Handbook. Oregon State University, 2022; [accessed 1 September 2022]. https://pnwhandbooks.org/plantdisease

Revord R.S., Lovell S.T., Brown P., Capik J., and Molnar T.J. Using genotyping-by-sequencing derived SNPs to examine the genetic structure and identify a core set of *Corylus americana* germplasm. *Tree Genetics & Genomes*, 2020; 16(65): 1-11. doi.org/10.1007/s11295-020-01462-y

Rochette N.C., Rivera-Colón A.G., and Catchen J.M. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol. Ecol.*, 2019; 28: 4737–4754. doi.org/10.1111/mec.15253

Rowley E.R., Fox S.E., Bryant D.W., Sullivan C.M., Priest H.D., Givan S.A., Mehlenbacher S.A., and Mockler T.C. Assembly and characterization of the European hazelnut ‘Jefferson’ transcriptome. *Crop Science*, 2012; 52(6): 2679-2686. doi.org/10.2135/cropsci2012.02.0065.

Rowley E.R, Vanburen R., Bryant D.W., Priest H.D., Mehlenbacher S.A., and Mockler T.C. A draft genome and high-density genetic map of European hazelnut (*Corylus avellana* L.). *Biorxiv*, 2018; 1-25. doi.org/10.1101/469015.

Salesses G. and Bonnet A. Cytogenetic study of hybrids between hazelnut varieties carrying a translocation in heterozygous state. *Cytologia*, 1988; 53: 407-413. [In French.]

Salojärvi J., Smolander O.-P., Nieminen K., Rajaraman S., Safronov O., Safdari P., Lamminmäki A., Immanen J., Lan T., Tanskanen J., Rastas P., Amiryousefi A., Jayaprakash B., Kammonen J. I., Hagqvist R., Eswaran G., Ahonen V. H., Serra J. A., Asiegbu F. O., Barajas-Lopez J. d. D., Blande D., Blokhina O., Blomster T., Broholm S., Brosché M., Cui F., Dardick C., Ehonen S. E., Elomaa P., Escamez S., Fagerstedt K. V., Fujii H., Gauthier A., Gollan P. J., Halimaa P., Heino P. I., Himanen K., Hollender C., Kangasjärvi S., Kauppinen L., Kelleher C. T., Kontunen-Soppela S., Koskinen J. P., Kovalchuk A., Kärenlampi S. O., Kärkönen A. K., Lim K.-J., Leppälä J., Macpherson L., Mikola J., Mouhu K., Mähönen A. P., Niinemets Ü., Oksanen E., Overmyer K., Palva E. T., Pazouki L., Pennanen V., Puhakainen T., Poczai P., Possen B. J. H. M., Punkkinen M., Rahikainen M. M., Rousi M., Ruonala R., van der Schoot C., Shapiguzov A., Sierla M., Sipilä T. P., Sutela S., Teeri T. H., Tervahauta A. I., Vaattovaara A., Vahala J., Vetchinnikova L., Welling A., Wrzaczek M., Xu E., Paulin L. G., Schulman A. H., Lascoux M., Albert V. A., Auvinen P., Helariutta Y., and Kangasjärvi J. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics*, 2017; 49: 904-912. doi.org/10.1038/ng.3862

Sathuvalli V.R., Mehlenbacher S.A., and Smith D.C. DNA markers linked to eastern filbert blight resistance from a hazelnut selection from the Republic of Georgia. *Journal of American Society for Horticultural Science*, 2011a; 136: 350-357. doi.org/10.21273/JASHS.136.5.350

Sathuvalli V.R., Chen H., and Mehlenbacher S.A. DNA markers linked to eastern filbert blight resistance in “Ratoli” hazelnut (*Corylus avellana* L.). *Tree Genetics & Genomes*, 2011b; 7: 337-345. doi.org/ 10.1007/s11295-010-0335-5

Sathuvalli V.R., Mehlenbacher S.A., and Smith D.C. Identification and mapping of DNA markers linked to eastern filbert blight resistance from OSU 408.040 hazelnut. *HortScience*, 2012*;* 47: 570–573.

Sathuvalli V.R., Mehlenbacher S.A., and Smith D.C. High-Resolution Genetic and Physical Mapping of the Eastern Filbert Blight Resistance Region in ‘Jefferson’ Hazelnut (*Corylus avellana* L.). *The Plant Genome*, 2017; 10(2). doi.org/10.3835/plantgenome2016.12.0123

Schopfer C.R., Nasrallah M.E., and Nasrallah J.B. The male determinant of self-incompatibility in Brassica. *Science,* 1999; 286(5445): 1697-1700. doi.org/10.1126/science.286.5445.1697

Şekerli M., Koma G.K., Snelling J.W., and Mehlenbacher S.A. New simple sequence repeat markers on linkage groups 2 and 7, and investigation of new sources of eastern filbert blight resistance in hazelnut. *Journal of the American Society for Horticultural Science,* 2021; 146(4): 252-66. doi.org/10.21273/JASHS05040-21

Stanke M., Keller O., Gunduz I., Hayes A., Waack S., and Morgenstern B. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research*, 2006a; 34(2): W435-W439. doi.org/10.1093/nar/gkl200

Stanke M., Schöffmann O., Morgenstern B., and Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, 2006b; 7(1): 62. doi.org/10.1186/1471-2105-7-62

Stanke M., Diekhans M., Baertsch R., and Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 2008; 24(5): 637-644. doi.org/10.1093/bioinformatics/btn013

Takasaki T., Hatakeyama K., Suzuki G., Watanabe M., Isogai A., and Hinata K. The S receptor kinase determines self-incompatibility in Brassica stigma. *Nature,* 2000; 403(6772): 913-916. doi.org/10.1038/35002628

Thompson M.M., Lagerstedt H.B., and Mehlenbacher S.A. New York: Wiley. In Janick J. and Moore J.N. (Eds.), Fruit breeding, 1996; Vol. 3. Nuts: 125-184.

USDA national agricultural statistics service NASS - quick stats. *USDA National Agricultural Statistics Service*, 2022; [Accessed 17 August 2023]. https://data.nal.usda.gov/dataset/nass-quick-stats

Vallès J., Bašić N., Bogunić F., Bourge M., Brown S.C., Garnatje T., Hajrudinović A., Muratović E., Pustahija F., Šolić E.M., and Siljak-Yakovev S. Contribution to plant genome size knowledge: first assessments in five genera and 30 species of angiosperms from western Balkans. *Botanica Serbica*, 2014; 38(1): 25-33.

Vurture G.W., Sedlazeck F.J., Nattestad M., Underwood C.J., Fang H., Gurtowski J., and Schatz M.C. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, 2017; 33(14, 15): 2202-2204. doi.org/10.1093/bioinformatics/btx153

Wang J. and Chai J. Structural insights into the plant immune receptors PRRs and NLRs. *Plant Physiology*, 2020; 182(4): 1566-1581. doi.org/10.1104/pp.19.01252

Wang Y., Teng Z., Li H., Wang W., Xu F., Sun K., Chu J., Qian Y., Loake G.J., Chu C., and Tang J. An activated form of NB-ARC protein RLS1 functions with cysteine-rich receptor-like protein RMC to trigger cell death in rice. *Plant Communications*, 2023; 4(2): 100459. doi.org/10.1016/j.xplc.2022.100459

Waterhouse A.M., Procter J.B., Martin D.M.A., Clamp M., and Barton G.J. Jalview version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 2009; (25): 1189-1191. doi.org/10.1093/bioinformatics/btp033

Woodworth R.H. Cytological studies in the Betulaceae. I. *Betula*. *Botanical Gazette,* 1929; 87: 383-399.

Wróblewski T., Spiridon L., Martin E.C., Petrescu A.-J., Cavanaugh K., Truco M.J., Xu H., Gozdowski D., Michelmore R.W., and Takken F.L.W. Genome-wide functional analyses of plant coiled-coil NLR-type pathogen receptors reveal essential roles of their N-terminal domain in oligomerization, networking and immunity. *PLOS Biology*, 2018. doi.org/10.1371/journal.pbio.2005821

Xi Y., Chalvon V., Padilla A., Cesari S., and Kroj T. The activity of the RGA5 sensor NLR from rice requires binding of its integrated HMA domain to effectors but not HMA domain self-interaction. *Molecular Plant Pathology*, 2022; 23(9): 1320-1330. doi.org/10.1111/mpp.13236

Yasuda S., Kobayashi R., Ito T., Wada Y., and Yakayama S. Homology-Based Interactions between Small RNAs and Their Targets Control Dominance Hierarchy of Male Determinant Alleles of Self-Incompatibility in *Arabidopsis lyrata*. *International Journal of Molecular Science*, 2021; 22(13): 6990. doi.org/10.3390/ijms22136990

Zeiner A., Colina F.J., Citterico M., and Wrzaczek M. Cysteine-rich receptor-like protein kinases: their evolution, structure, and roles in stress response and development. *Journal of Experimental Botany*, 2023; 74(17): 4910-4927. doi.org/10.1093/jxb/erad236

Zhang X., Wang L., Yuan Y., Tian D., and Yang S. Rapid copy number expansion and recent recruitment of domains in S-receptor kinase-like genes contribute to the origin of self-incompatibility. *The FEBS Journal,* 2011; 278(22): 4323-4337. doi.org/10.1111/j.1742-4658.2011.08439.x

Zhao T., Ma W., Yang Z., Liang L., Chen X., Wang G., Ma Q., and Wang L. A chromosome-level reference genome of the hazelnut, *Corylus heterophylla* Fisch. *Gigascience*, 2021; 10(4): giab027. doi.org/10.1093/gigascience/giab027

Zhou C., McCarthy S.A., and Durbin R. YaHs: yet another Hi-C scaffolding tool. *Bioinformatics, 2022;* 39(1): btac808. doi.org/10.1093/bioinformatics/btac808

Supplemental

File S1. List of Supplementary Materials

Table S1. Summary of sequencing data from Illumina, Hi-C, and PacBio platforms.

Table S2. Summary statistics for the eleven haplotype scaffolds corresponding to the ‘Jefferson’ European hazelnut (*C. avellana*) base chromosomes.

Table S3. Number and percentage of aligned Illumina 150 bp PE reads derived from ‘Jefferson’ parents to ‘Jefferson’ chromosome-level haplotype-resolved assemblies

Table S4. Structural variations by SyRI of ‘Jefferson’ haplotype 1 and haplotype 2.

Table S5. Sequence variations by SyRI of ‘Jefferson’ haplotype 1 and haplotype 2.

Table S6. ‘Jefferson’ haplotype 1 assembly EDTA output.

Table S7. ‘Jefferson’ haplotype 2 assembly EDTA output.

Table S8. Distribution of resistance-like genes identified by DRAGO2 among 11 pseudo-chromosomal scaffolds of the ‘Jefferson’ haplotype 1 and haplotype 2 assemblies.

Figure S1**.** Genome assembly and annotation workflow of *C. avellana* ‘Jefferson’.

Figure S2. GenomeScope of raw ‘Jefferson’ PacBio HiFi reads with k-mer length = 31.

Figure S3. LAI scores of *C. avellana* ‘Jefferson’ haplotypes and ‘Tombul’.

Figure S4. OmicsBox summary metrics of ‘Jefferson’ haplotype 1 functional annotation.

Figure S5. OmicsBox summary metrics of ‘Jefferson’ haplotype 2 functional annotation.

Table S1 Summary of sequencing data from Illumina, Hi-C, and PacBio platforms.

Sequencing platform	Sample	Insert length	Sequencing model	Number of reads	Total nucleotides
PacBio Sequel IIe	'Jefferson'	>20kb	2x 8M SMRT cell	3.64 M	56.8 Gb
Hiseq 4000	'OSU 252.146'	300bp	2x150	295.9 M	44.38 Gb
Hiseq 4000	'OSU 414.062'	300bp	2x150	218.2 M	32.73 Gb
Dovetail Hi-C on Hiseq 4000	'Jefferson'	300bp	2x150	428.46 M	64.69 Gb

Table S2. Summary statistics for the eleven haplotype scaffolds corresponding to the 'Jefferson' European hazelnut (*C. avellana*) base chromosomes.

Chromosomes	**Haplotype 1**			**Haplotype 2**		
	Total length (bp)	N count[1]	Gaps	Total length (bp)	N count[1]	Gaps
1	48,258,603	600	3	47,666,154	400	2
2	44,374,200	200	1	45,407,320	600	3
3	33,425,378	200	1	32,429,289	0	0
4	36,823,049	1,200	6	37,439,655	400	2
5	32,584,664	400	2	35,167,529	800	4
6	28,787,360	0	0	28,771,021	200	1
7	31,176,844	200	1	31,107,465	200	1
8	24,916,583	400	2	23,719,546	200	1
9	23,914,400	400	2	23,029,850	800	4
10	23,820,452	200	1	24,758,360	400	2
11	21,620,711	600	3	22,513,321	200	1
Total genome size	**349,702,244**	**4,400**	**22**	**352,009,510**	**4,200**	**21**

[1] Ns are inserted by YaHs at a fixed rate of 200 nucleotides for every contig merge.

Table S3. Number and percentage of aligned Illumina 150 bp PE reads derived from 'Jefferson' parents to 'Jefferson' chromosome-level haplotype-resolved assemblies.

	Number and percentage of aligned reads for each parent	
	'OSU 252.146'	'OSU 414.062'
'Jefferson' V4 Haplotype 1	267,993,778 (90.57%)	200,943,531 (92.08%)
'Jefferson' V4 Haplotype 2	272,354,026 (92.04%)	199,451,255 (91.39%)

Table S4. Structural variations by SyRI[1] of 'Jefferson' haplotype 1 and haplotype 2.

Structural variation	**Count**	**Haplotype 1 length (bp)**	**Haplotype 2 length (bp)**
Syntenic regions	253	277,416,244	276,792,668
Inversions	37	6,741,817	6,220,076
Translocations	216	41,832,328	42,723,238
Duplications (reference)	259	4,088,393	-------
Duplications (query)	353	----	2,863,543
Not aligned (reference)	593	23,471,255	----
Not aligned (query)	677	----	23,296,270

[1] Count table output from default SyRI run, derived from a minimap2 .bam alignment between both haplotypes with the parameter: --eqx.

Table S5. Sequence variations by SyRI[1] of 'Jefferson' haplotype 1 and haplotype 2.

Sequence variation	**Count**	**Haplotype 1 length (bp)**	**Haplotype 2 length (bp)**
SNPs	1,593,404	1,593,404	1,593,404
Insertions	213,321	----	2,792,929
Deletions	150,213	2,897,238	----
Copygains	91	-----	471,299
Copylosses	80	253,126	-----
Highly diverged	13,753	116,506,195	116,208,998
Tandem repeats	22	86,034	71,726

[1] Count table output from default SyRI run, derived from a minimap2 .bam alignment between both haplotypes with the parameter: --eqx.

Table S6. 'Jefferson' haplotype 1 assembly EDTA[1] output.

Class	Number of elements	Length (bp)	Percentage of genome
LTR	**103,937**	**62,391,913**	**17.84%**
Copia	23,301	15,899,955	4.55%
Gypsy	25,625	21,135,697	6.04%
unknown	55,011	25,356,261	7.25%
TIR	**158,209**	**40,419,512**	**11.55%**
CACTA	33,740	9,654,395	2.76%
Mutator	77,114	17,487,304	5.00%
PIF_Harbinger	22,873	5,671,550	1.62%
Tc1_Mariner	3,700	878,845	0.25%
hAT	20,782	6,727,418	1.92%
nonLTR	**1,065**	**361,136**	**0.10%**
LINE_element	1,031	352,462	0.10%
unknown	34	8,674	0.00%
nonTIR	--	--	--
helitron	**35,911**	**9,497,804**	**2.72%**
repeat_region	**81,361**	**21,122,389**	**6.04%**
Total Genome Masked	**380,483**	**133,792,754**	**38.26%**

[1]EDTA run with parameters: --cds --bed --sensitive 1 --analysis 1; the CDS and bed file provide is derived from the BRAKER1/BRAKER2 gene set produced for the respective haplotype.

Table S7. 'Jefferson' haplotype 2 assembly EDTA[1] output.

Class	Number of elements	Length (bp)	Percentage of genome
LTR	**126,989**	**67,657,376**	**19.22%**
Copia	27,892	17,200,312	4.89%
Gypsy	26,884	21,861,582	6.21%
unknown	72,213	28,595,482	8.12%
TIR	**145,984**	**37,854,874**	**10.75%**
CACTA	27,535	7,250,762	2.06%
Mutator	74,679	18,295,776	5.20%
PIF_Harbinger	17,714	4,238,613	1.20%
Tc1_Mariner	3,122	681,523	0.19%
hAT	22,934	7,388,200	2.10%
nonLTR	**1,100**	**442,005**	**0.12%**
LINE_element	1,047	425,994	0.12%
unknown	53	16,011	0.00%
nonTIR	--	--	--
helitron	**39,521**	**8,684,384**	**2.47%**
repeat_region	**37,689**	**9,592,939**	**2.73%**
Total Genome Masked	**380,483**	**124,231,578**	**35.29%**

[1]EDTA run with parameters: --cds --bed --sensitive 1 --analysis 1; the CDS and bed file provide is derived from the BRAKER1/BRAKER2 gene set produced for the respective haplotype.

Table S8. Distribution of resistance-like transcripts identified by DRAGO2 among 11 pseudo-chromosomal scaffolds of the 'Jefferson' haplotype 1 and haplotype 2 assemblies.

'Jefferson' Pseudo-chromosomal scaffolds	**CN[1]**	**CNL[1]**	**NL[1]**	**RLK[2]**	**RLP[2]**	**TN[3]**	**TNL[3]**	**Other[4]**	**Total**
	H1/H2	H1/H2	H1/H2	H1/H2	H1/H2	H1/H2	H1/H2	H1/H2	H1/H2
1	17/16	63/60	43/47	20/17	29/32	0/2	8/5	304/276	484/455
2	18/23	12/14	22/28	84/65	91/58	3/1	0/0	383/403	613/592
3	4/5	5/5	8/7	36/36	28/25	3/5	14/11	162/169	260/263
4	3/10	14/8	1/4	50/57	24/24	0/0	1/1	189/207	284/315
5	3/1	4/0	12/11	41/40	43/23	5/2	6/6	224/225	342/298
6	1/2	2/1	4/5	45/49	15/31	15/13	21/15	233/240	335/356
7	0/0	5/5	2/1	66/71	44/62	0/0	0/0	165/179	284/320
8	5/8	11/20	5/5	39/37	70/58	14/10	47/31	166/169	360/345
9	11/14	10/9	3/6	15/21	16/21	1/1	12/10	126/128	196/211
10	2/1	11/9	6/2	45/50	63/66	0/0	2/3	136/147	265/280
11	0/0	2/2	0/0	45/45	22/23	0/0	0/0	127/154	197/224
Total	**66/80**	**139/133**	**117/122**	**486/445**	**445/ 423**	**41/34**	**111/82**	**2,215/ 2,297**	**3,620/ 3,659**

[1] Coiled-coil nucleotide binding site [(CC-NBS (CN))]; CC-NBS-leucine rich repeat [(CC-NBS-LRR (CNL))]; NBS-LRR (NL).

[2] Receptor-like Kinase (RLK); Receptor-like Protein (RLP).

[3] TIR-NBS (TN); TIR-NBS-LRR (TNL).

[4] Includes kinases (K), NBS (N), LRRs (L), CKs, CTs, CTLs, Lysine motif containing proteins (LYK and LYP) and Lectin-like motif containing proteins (LECM).

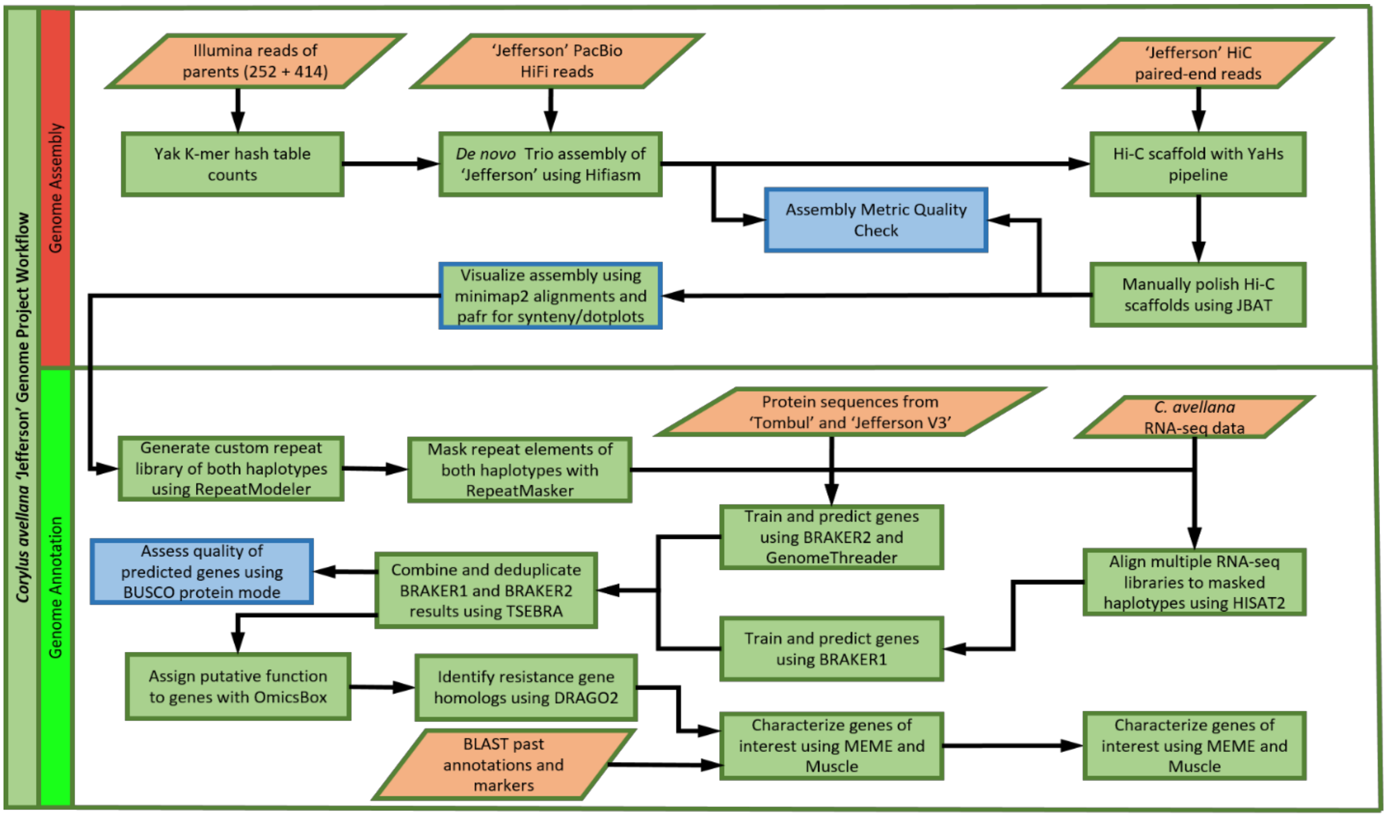

Corylus avellana 'Jefferson' Genome Project Workflow
Genome Assembly
Genome Annotation
Illumina reads of parents (252 + 414)
'Jefferson' PacBio HiFi reads
'Jefferson' HiC paired-end reads
Yak K-mer hash table counts
De novo Trio assembly of 'Jefferson' using Hifiasm
Hi-C scaffold with YaHs pipeline
Assembly Metric Quality Check
Visualize assembly using minimap2 alignments and pafr for synteny/dotplots
Manually polish Hi-C scaffolds using JBAT
Protein sequences from 'Tombul' and 'Jefferson V3'
C. avellana RNA-seq data
Generate custom repeat library of both haplotypes using RepeatModeler
Mask repeat elements of both haplotypes with RepeatMasker
Train and predict genes using BRAKER2 and GenomeThreader
Assess quality of predicted genes using BUSCO protein mode
Combine and deduplicate BRAKER1 and BRAKER2 results using TSEBRA
Align multiple RNA-seq libraries to masked haplotypes using HISAT2
Train and predict genes using BRAKER1
Assign putative function to genes with OmicsBox
Identify resistance gene homologs using DRAGO2
Characterize genes of interest using MEME and Muscle
Characterize genes of interest using MEME and Muscle
BLAST past annotations and markers

Figure S1. Genome assembly and annotation workflow of *C. avellana* 'Jefferson'. Figure shows the genome assembly and annotation pipeline with processes shown in green, extraneous data in orange and quality checks in blue.

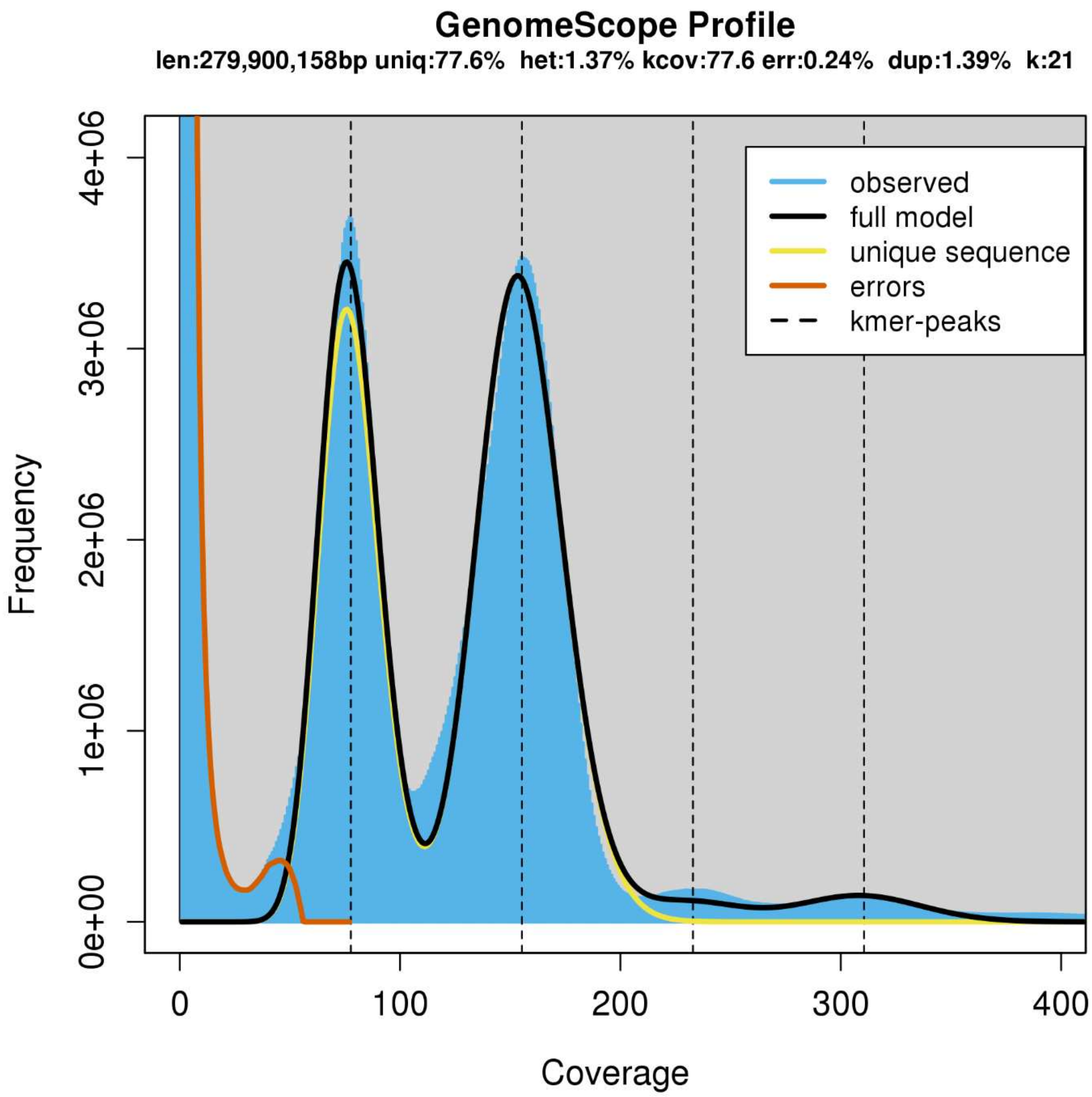


Figure S2. GenomeScope result of raw 'Jefferson' PacBio HiFi reads for k-mer length = 21. GenomeScope output derived from jellyfish count -C -m 21 -s 1000000000 and jellyfish histo.

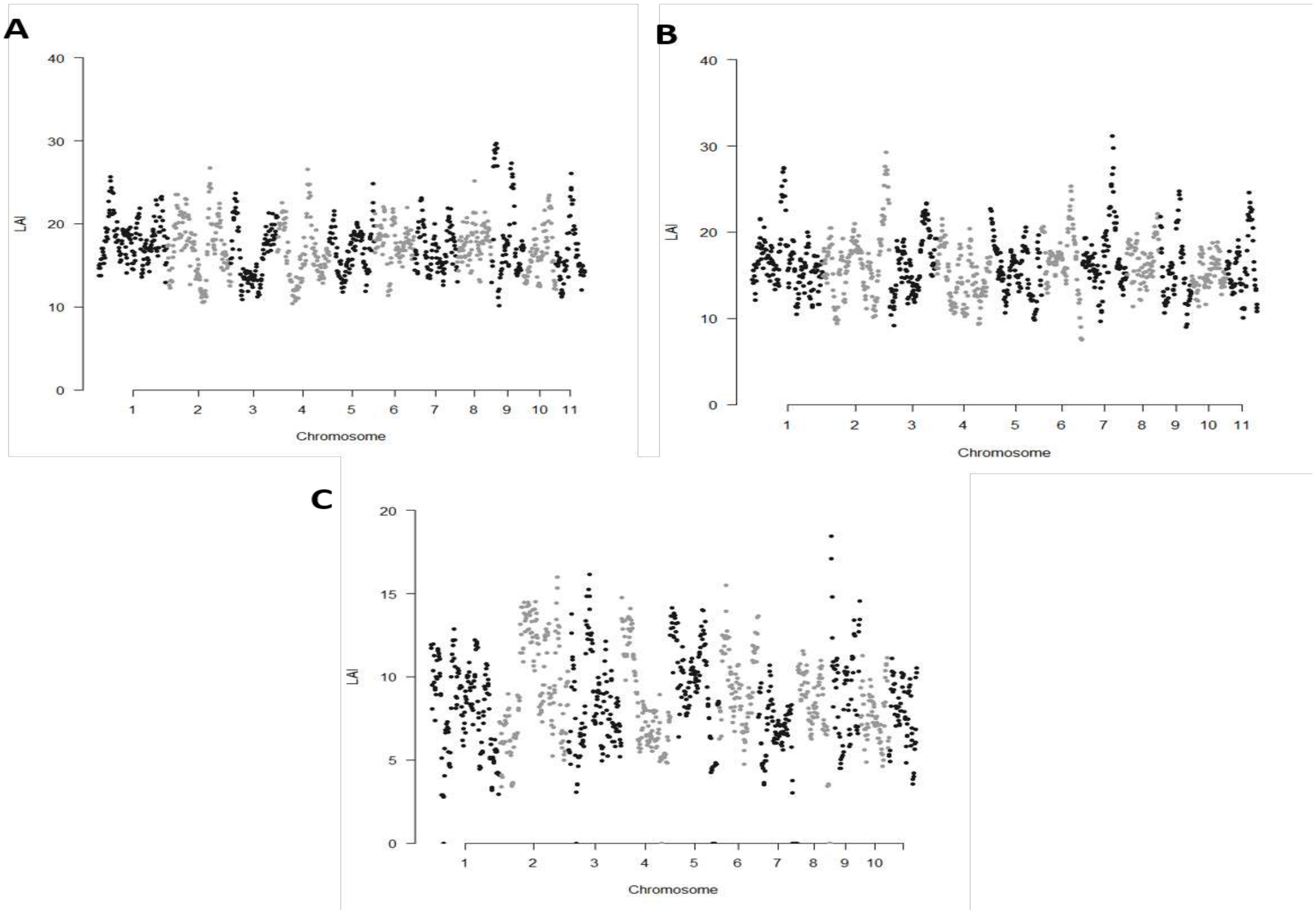


Figure S3. LAI scores of *C. avellana* 'Jefferson' haplotypes and 'Tombul'. LAI scores were obtained by LTR_Retriever from a concatenated set of LTRs derived from LTR harvest and LTR_FINDER_parallel for each respective assembly. Each dot represents LAI score of a 3 Mb-sliding window with 300-Kb increment, adjusted by the whole-genome LTR identity. (**A**) 'Jefferson' haplotype 1 genome assembly. (**B**) 'Jefferson' haplotype 2 genome assembly. (**C**) 'Tombul' genome assembly.

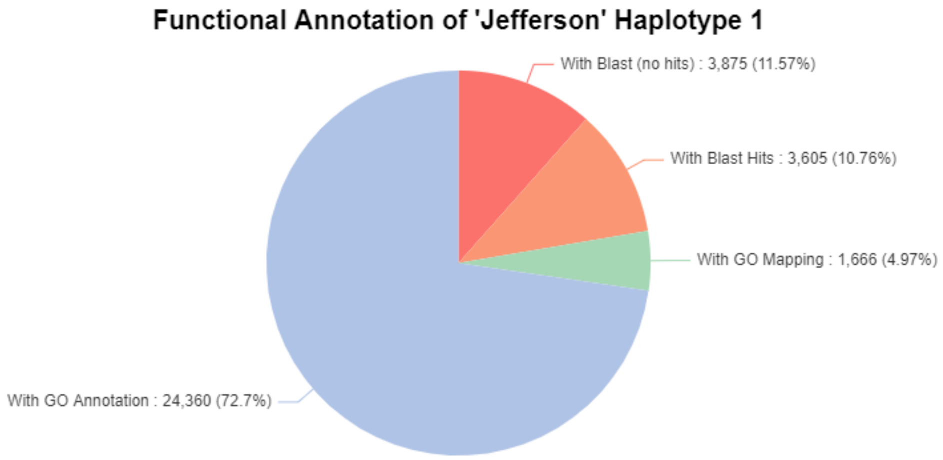


Figure S4. OmicsBox summary metrics of 'Jefferson' haplotype 1 functional annotation. Pie chart shows total distribution of OmicsBox functional annotation performed on haplotype 1 amino acid transcripts of 'Jefferson'. In red are transcripts that received no BLAST hits from the database and thus have unknown function; orange are transcripts that received only BLAST hits; green are transcripts that had GO terms associated with the initial BLAST database search; blue is transcripts that received GO annotation descriptions.

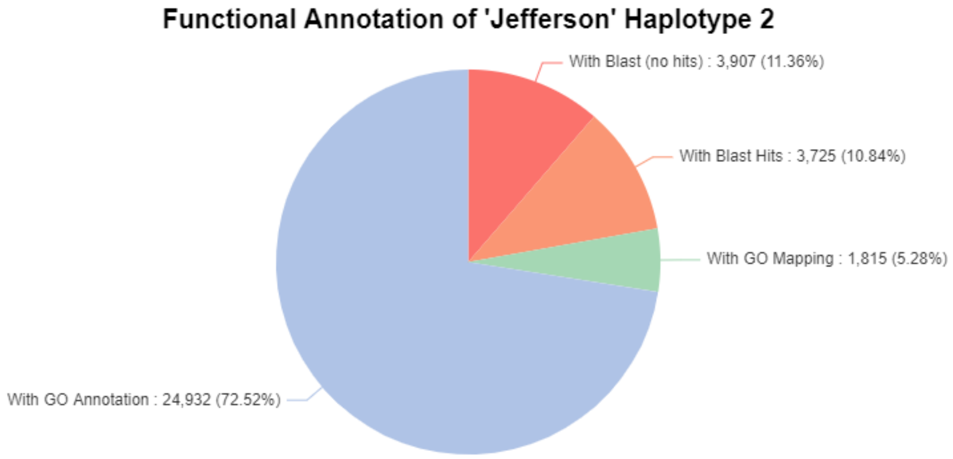


Figure S5. OmicsBox summary metrics of 'Jefferson' haplotype 2 functional annotation. Pie chart shows total distribution of OmicsBox functional annotation performed on haplotype 2 amino acid transcripts of 'Jefferson'.