

DATA NOTE

High-quality chromosome-scale assembly of the walnut (*Juglans regia* L.) reference genome

Annarita Marrano^{1,*}, Monica Britton², Paulo A. Zaini¹, Aleksey V. Zimin^{3,4}, Rachael E. Workman³, Daniela Puiu⁴, Luca Bianco⁵, Erica Adele Di Pierro⁵, Brian J. Allen¹, Sandeep Chakraborty¹, Michela Troggio⁵, Charles A. Leslie¹, Winston Timp^{3,4}, Abhaya Dandekar¹, Steven L. Salzberg^{3,4,6} and David B. Neale¹

¹Department of Plant Sciences, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA; ²Bioinformatics Core Facility, Genome Center, University of California, One Shields Avenue, Davis, CA 95616, USA; ³Department of Biomedical Engineering, Johns Hopkins University, 720 Rutland Avenue, Baltimore, MD 21205, USA; ⁴Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, 3100 Wyman Park Dr., Baltimore, MD 21211, USA; ⁵Research and Innovation Center, Fondazione Edmund Mach, Via E. Mach, 1 38010 S. Michele all'Adige (TN) 38010, Italy and ⁶Departments of Computer Science and Biostatistics, Johns Hopkins University, 3400 North Charles Street Baltimore, MD 21218, USA

*Correspondence address. Annarita Marrano, Department of Plant Sciences, University of California, Davis, CA 95616, USA. E-mail: amarrano@ucdavis.edu <http://orcid.org/0000-0001-9560-2706>

Abstract

Background: The release of the first reference genome of walnut (*Juglans regia* L.) enabled many achievements in the characterization of walnut genetic and functional variation. However, it is highly fragmented, preventing the integration of genetic, transcriptomic, and proteomic information to fully elucidate walnut biological processes. **Findings:** Here, we report the new chromosome-scale assembly of the walnut reference genome (Chandler v2.0) obtained by combining Oxford Nanopore long-read sequencing with chromosome conformation capture (Hi-C) technology. Relative to the previous reference genome, the new assembly features an 84.4-fold increase in N50 size, with the 16 chromosomal pseudomolecules assembled and representing 95% of its total length. Using full-length transcripts from single-molecule real-time sequencing, we predicted 37,554 gene models, with a mean gene length higher than the previous gene annotations. Most of the new protein-coding genes (90%) present both start and stop codons, which represents a significant improvement compared with Chandler v1.0 (only 48%). We then tested the potential impact of the new chromosome-level genome on different areas of walnut research. By studying the proteome changes occurring during male flower development, we observed that the virtual proteome obtained from Chandler v2.0 presents fewer artifacts than the previous reference genome, enabling the identification of a new potential pollen allergen in walnut. Also, the new chromosome-scale genome facilitates in-depth studies of intraspecies genetic diversity by revealing previously undetected autozygous regions in Chandler, likely resulting from inbreeding, and 195 genomic regions highly differentiated between Western and Eastern walnut cultivars. **Conclusion:** Overall, Chandler v2.0 will serve as a valuable resource to better understand and explore walnut biology.

Received: 19 October 2019; Revised: 13 March 2020; Accepted: 20 April 2020

© The Author(s) 2020. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Keywords: Nanopore; Hi-C; Iso-Seq; gene prediction; genetic diversity; proteome; allergens

Introduction

Persian walnut (*Juglans regia* L.) is among the top 3 most-consumed nuts in the world, and over the past 10 years, its global production increased by 37% [1]. Its richness in alpha-linolenic acid (ALA), proteins, minerals, and vitamins, along with documented benefits for human health, explains this increased interest in walnut consumption [2]. As suggested by its generic name *Juglans* from the Latin appellation “*Jovis glans*,” which loosely means “nut of gods,” the culinary and medical value of Persian walnut was already widely prized by ancient civilizations [3].

The origin and evolution of the Persian walnut are the results of a complex interplay between hybridization, human migration, and biogeographical forces [4]. A recent phylogenomic analysis revealed that Persian walnut (and its landrace *Juglans sigillata*) arose from an ancient hybridization that occurred between American black walnuts and Asian butternuts after a climate-driven range expansion in Eurasia during the Pliocene [5]. Evidence suggests that the mountains of Central Asia were the cradle of domestication of Persian walnut [6], whence it spread to the rest of Asia, the Balkans, Europe, and, finally, the Americas.

Today, walnut is cultivated worldwide in an area of 1,587,566 ha, mostly in China and the USA [7]. Considerable phenotypic and genetic variability can be observed in this wide distribution area, especially in the Eastern countries, where walnuts can still be found in wild fruit forests. Many studies on genetic diversity in walnut have outlined a genetic differentiation between Eastern and Western genotypes [8, 9]. Moreover, walnuts from Eastern Europe, Central Asia, and China exhibit higher genetic diversity and a higher number of rare alleles than the genotypes from Western countries [10].

The release of the first reference genome, Chandler v1.0 [11], enabled the study of walnut genetics at a genome-wide scale. For the first time, it was possible to explore the gene space of Persian walnut with the prediction of 32,498 gene models, providing the basis to untangle complex phenotypic pathways, such as those responsible for the synthesis of phenolic compounds. The availability of a reference genome marked the beginning of a genomics phase in Persian walnut, allowing whole-genome resequencing [5, 12], the development of high-density genotyping tools [9, 13], and the genetic dissection of important agronomical traits in walnut [14–17]. However, the Chandler v1.0 assembly is highly fragmented, compromising the accuracy of gene prediction and the fulfillment of advanced genomics studies necessary to resolve many, still-unanswered questions in walnut research.

The recent introduction of long-read sequencing technologies and long-range scaffolding methods has enabled chromosome-scale assembly for multiple plant species, including highly heterozygous crops such as almond (*Prunus dulcis* [18]) and kiwifruit (*Actinidia chinensis* [19]). The availability of genomes with fully assembled chromosomes provides foundations for understanding plant domestication and evolution [18, 20, 21] and the mechanisms governing important traits (e.g., flower color and scent [22]), as well as the impact of epigenetic modifications on phenotypic variability [23]. Recently, Zhu et al. [24] assembled the parental genomes of a hybrid *Juglans microcarpa* × *J. regia* (cv. Serr) at the chromosome scale using long-read Pacific Biosciences (PacBio) sequencing and optical mapping. They relied on the haplotype divergence between the 2 *Juglans* species

and demonstrated an ongoing asymmetric fractionation of the 2 subgenomes present in *Juglans* genomes.

Here we report a new chromosome-level assembly of the walnut reference genome, Chandler v2.0, which we obtained by combining Oxford Nanopore long-read sequencing [25] with chromosome conformation capture (Hi-C) technology [26]. Thanks to the increased contiguity of Chandler v2.0, we were able to substantially improve gene prediction accuracy, with new, longer gene models identified and many fewer artifacts compared to Chandler v1.0. Also, the availability of full, chromosomal sequences reveals new genetic diversity of Chandler, previously inaccessible through standard genotyping tools, and significant genetic differentiation between Western and Eastern walnuts at 195 genomic regions, including also loci involved in nut shape and harvest date. In the present research, we demonstrate the fundamental role of a chromosome-scale reference genome to integrate transcriptomics, population genetics, and proteomics, which in turn enable a better understanding of walnut biology.

Results

Genome long-read sequencing and assembly

To increase the contiguity of the Chandler genome, we first generated deep sequence coverage using Oxford Nanopore Technology (ONT), a cost-effective long-read sequencing approach that determines DNA bases by measuring the changes in electrical conductivity generated while DNA fragments pass a tiny biological pore [27]. Since the release of the first plant genome assembly generated using ONT sequencing [28], this technology has been applied to sequence and obtain chromosome-scale genomes of many other plant species [29–31]. In Persian walnut, ONT sequencing yielded 7,096,311 reads that provided 21.9 Gb of sequence, or ~35× genome coverage (assuming a genome size of 620 Mb). Read lengths averaged 3.1 kb, and the N50 read length was 6.7 kb, with the longest read being 992.2 kb (Supplementary Table S1).

One of the major limitations of long-read sequencing technologies is their high error rate, which can range between 5% and 15% for Nanopore sequencing [32]. To overcome this limitation, we adopted the hybrid assembly technique incorporated into the MaSuRCA assembler, which combines long, high-error reads with shorter but much more accurate Illumina sequencing reads to generate a robust, highly contiguous genome assembly [33]. First, using the Illumina reads, we created 3.7 million “super-reads” with a total length of 2.9 Gb. We then combined the super-reads with the ONT reads to generate 3.2 million mega-reads with a mean length of 4.7 kb, representing 24× genome coverage (Supplementary Table S2). Finally, we assembled the mega-reads to obtain the “hybrid” Illumina-ONT assembly, which comprised 1,498 scaffolds, 258 contigs, and 25,007 old scaffolds from Chandler v1.0 (Supplementary Table S3).

Even though the total number of scaffolds (>1 kb) was reduced by 80% compared to Chandler v1.0 (Table 1), the new hybrid assembly was still fragmented. To improve the assembly further and build chromosome-scale scaffolds, we applied Hi-C sequencing, which is based on proximity ligation of DNA fragments in their natural conformation within the nucleus [26]. The HiRise scaffolding pipeline processed 356 million paired-end 100-bp Illumina reads to generate the HiRise assembly (Ta-

Table 1: Comparison among the 4 assemblies of Chandler

Statistic	Chandler v1.0	Chandler v1.5	Chandler hybrid	Chandler HiRise	Chandler v2.0	JrSerr.v1.0
No. of scaffolds	27,032	4,401	3,497	2,656	2,643	73
N50 length (scaffolds) (bp)	304,423	637,984	1,640,935	32,655,472	37,114,715	35,197,335
L50 (scaffolds)	344	272	89	8	7	7
Total length of assembled scaffolds (bp)	667,299,356	650,478,320	567,378,842	567,480,142	567,796,851	534,671,929
No. of contigs	53,156	7,411	3,592	3,700	3,684	127
N50 length (contigs) (bp)	42,417	317,751	1,512,354	1,083,883	1,083,883	15,066,219
L50 (contigs)	3,630	482	97	144	144	13
Total size of assembled contigs (bp)	641,521,787	617,088,256	567,276,004	567,276,244	567,192,099	530,618,363

Scaffolds shorter than 1,000 bp are not included in these totals.

ble 1). The top 17 scaffolds from this assembly spanned >90% of the total assembly length, with a scaffold length ranging from 19.6 to 45.2 Mb (Supplementary Figs S1 and S2). As reported in Table 1, the Chandler genome contiguity increased dramatically compared with the previous assemblies. Compared with the recently published genome assembly of the walnut cultivar Serr [24], Chandler v2.0 was less contiguous at the contig level, with an N50 size of 1.1 Mb against the 15.1 Mb of JrSerr.v1.0. The higher-coverage PacBio sequencing data (57.2 Gb) used to assemble JrSerr.v1.0 may explain this discrepancy in contiguity between the 2 assemblies. Besides, our assembly presented a value of contiguity similar to that of the recently published genomes of pecan (*Carya illinoensis*; 1.1 Mb [34]), Chinese chestnut (*Castanea mollissima*; 944.4 kb [35]), and pedunculate oak (*Quercus robur*; 1.35 Mb [36]).

Validation of the HiRise assembly

To assess the quality of the HiRise assembly, we used 2 independent sources of data. First, we used the single-nucleotide polymorphism (SNP) markers mapped on the high-density genetic map of Chandler recently described by Marrano et al. [16]. Out of the 8,080 SNPs mapped into 16 linkage groups (LGs), 6,894 had probes aligning uniquely on the HiRise assembly, with 98% of identity for >95% of their length. A total of 35 scaffolds of the HiRise assembly could be anchored to a chromosomal LG by ≥ 1 SNP (Fig. 1). In particular, 13 LGs were spanned by a single HiRise scaffold, while 2–3 scaffolds each aligned the remaining 3 LGs.

Second, we anchored the HiRise assembly to the Chandler genetic map used by Luo et al. [37] to construct a walnut physical map. In total, 972 of the mapped markers (1,525 SNPs) aligned uniquely on the same 35 HiRise scaffolds anchored to the aforementioned linkage map. Overall, we observed almost perfect collinearity between the HiRise assembly and both Chandler genetic maps (Fig. 1, Supplementary Fig. S3). Therefore, we oriented, ordered, and named the HiRise scaffolds consistent with the linkage map of Luo et al. [37], generating the final 16 chromosomal pseudomolecules of *J. regia* Chandler.

These 16 contiguous chromosomal scaffolds account for 95% of the final walnut reference genome v2.0, with an N50 scaffold size of 37 Mb. We identified telomere sequences at both ends for 9 of the chromosome scaffolds, on 1 end of the other 7 chromosomes, and 1 end of 7 unanchored scaffolds. Also, all 16 chromosomes had centromeric repeats in the middle, alongside regions with low recombination rates (Fig. 2).

As compared to the previous Chandler genome assemblies (Table 1), Chandler v2.0 had a smaller genome size (573.9 Mb), much closer to the Genomescope estimate of 488.2 Mb. This

reduction in genome size represents a great improvement of Chandler v2.0 and can be related to the removal of haplotype variants, likely interpreted and annotated as different scaffolds in the previous genome versions. Compared with the Serr walnut genome (JrSerr.v1.0; 534.7 Mb) [24], Chandler v2.0 had a larger genome size, likely due to structural variation (e.g., copy number and presence/absence variants), whose central role in explaining intraspecific genomic and phenotypic diversity has been reported in different plant species [38, 39]. In addition, the higher number of unanchored scaffolds (2,631; 20.9 Mb) in Chandler v2.0 compared with JrSerr.v1.0 can represent autozygous genomic regions of Chandler, devoid of segregating markers and, therefore, difficult to anchor to linkage genetic maps [37], as also suggested by the higher fixation index (*F*) observed in Chandler (0.03) than Serr (−0.29) in previous genetic surveys [9]. The 2 walnut assemblies, however, aligned with high sequence identity (>98% for >95% of their total length) and showed high collinearity (Supplementary Fig. S4). Future comparative genomics studies will provide further insights on the functional and structural differences between the 2 genome assemblies, and their explanatory involvement in the morphological and physiological variation of these 2 walnut cultivars.

To assess the sequence accuracy of Chandler v2.0, we first compared the scaffold sequences of Chandler v2.0 with the previous version of the walnut reference genome. Approximately 578 Mb of sequence were mutual best alignments, namely, best hits of each location between Chandler v2.0 and v1.0 and vice versa, with a sequence identity of 99.6%. We also observed that 135 Mb of Chandler v1.0 (18.9%) aligned to the same locations in Chandler v2.0, suggesting the presence of redundant haplotypes in the previous version of the walnut reference genome that have been removed in our assembly. We then mapped the Illumina whole-genome shotgun data [11] against the new chromosome-scale genome. The alignment resulted in 64,950,691,681 bp mapped, of which 407,450,406 were single-base mismatches, consistent with an Illumina sequence accuracy rate of 99.5%.

Repeat annotation

More than half (58.4%) of the new Chandler v2.0 is repetitive. This estimate is higher than the previous version of the walnut reference genome (51.19%) and comparable to other *Fagales* genomes [36, 40]. As in most plant genomes, interspersed repeats were the most abundant type of repeats, with retrotransposons at 36.45% and DNA transposons at 15.86%. Gypsies (10.5%) and Copias (7.69%) were the most represented classes of long-terminal retrotransposons (LTR), and, though widely dis-

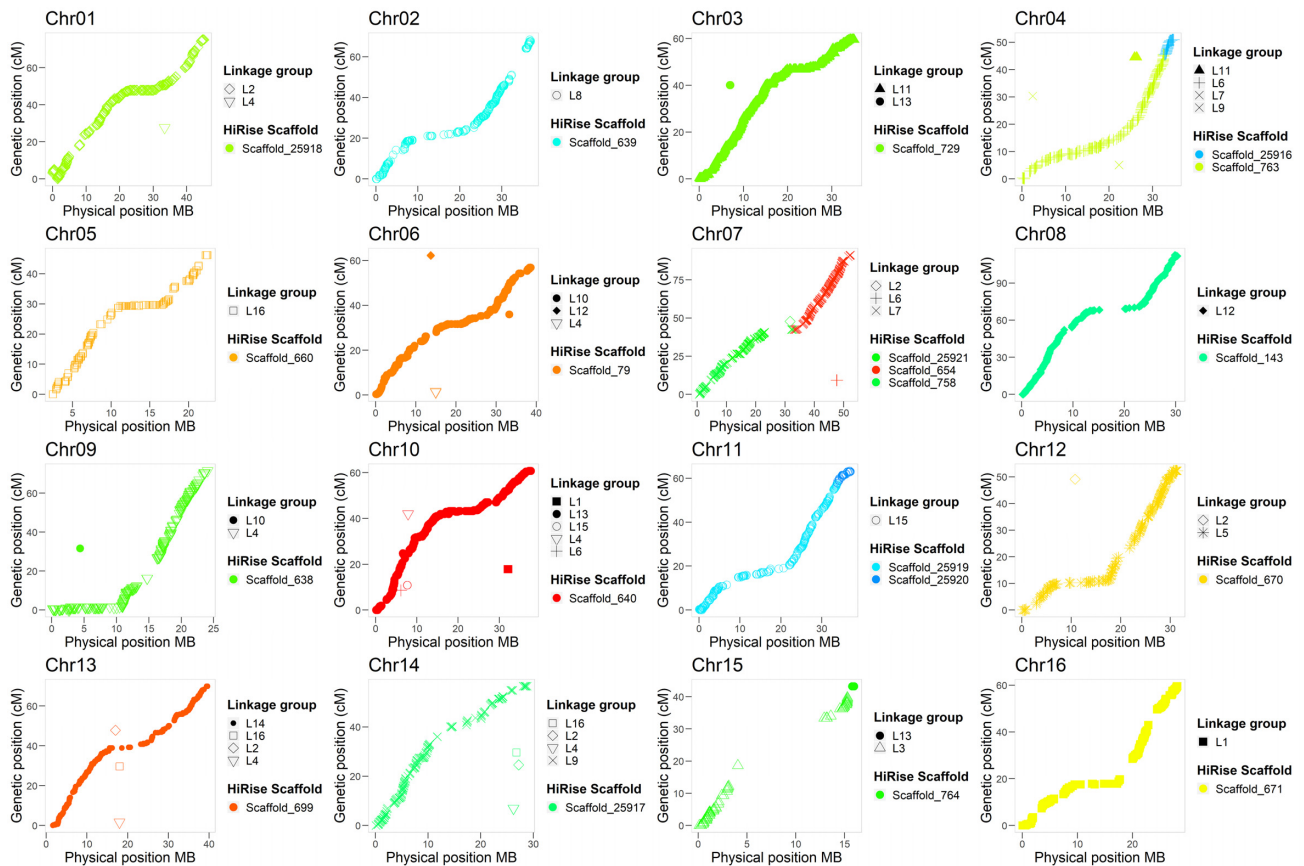


Figure 1: Collinearity between the high-density “Chandler” genetic map of Marrano et al. [16] and the 16 chromosomal pseudomolecules of Chandler v2.0.

persed throughout the genome, they were distributed differently along the 16 chromosomes (Supplementary Fig. S5): the Gypsy LTRs were more abundant alongside the centromeres, where, instead, the density of the Copia LTRs decreased, as previously observed in walnut [24]. The long-interspersed nuclear elements (L1/LINE), which possess a poly(A) tail and 2 open reading frames for autonomous retrotransposition, were the largest class of non-LTRs, at 7.14% of the genome. Simple repeats (1.91%) were also found.

PacBio Iso-Seq sequencing and gene annotation

A fragmented reference genome can severely hamper the accuracy of gene prediction because many genes will be broken across multiple small contigs (false-negative results) and because heterozygous gene variants may be annotated separately (false-positive results).

To improve the gene prediction accuracy of Chandler v2.0, we used the “Isoform Sequencing” (Iso-Seq) method, developed by PacBio, which can generate full-length transcripts up to 10 kb, allowing for accurate determination of exon-intron structure by the alignment of the transcripts to the assembly [41]. The high error rate of PacBio sequencing can be greatly reduced using circular consensus sequence (CCS), in which a transcript is circularized and then sequenced repeatedly to self-correct the errors. We applied PacBio Iso-Seq to sequence full-length transcripts from 9 tissues, chosen to cover most of the transcript diversity in walnut (Supplementary Table S4). Across the 4 SMRT cells, we obtained 26,328,087 subreads with a mean length of

1,188 bp (Supplementary Table S5) and CCSs ranging from 13,000 to 142,000 per library (Supplementary Table S6). Of the 745,730 full-length non-chimeric (FLnc) transcripts, 68,225 were classified as high quality, FL (HQ FL) consensus transcript sequences, with an average length of 1,357 bp (Supplementary Table S6). Catkin 1-inch elongated (CAT1), shoot, and root yielded the lowest number of HQ FL transcripts, while pollen and leaf had the lowest number of HQ consensus clusters obtained per CCS after polishing (Supplementary Table S6). These results can be explained by lower complementary DNA (cDNA) quality or fewer inserts of full-length transcripts from these tissues during the cDNA pooling and library preparation. Nevertheless, >99% of the HQ FL transcripts aligned onto the new chromosomal-level walnut reference genome (Supplementary Table S7).

By combining the HQ FL transcripts with available *Juglans* transcriptome sequences, we identified 37,554 gene models, which are more than those annotated in Chandler v1.0 but fewer than the predicted genes in the NCBI RefSeq *J. regia* annotation generated with the first version of the reference genome (Table 2). Thus, the new chromosome-scale genome, along with the availability of full-length transcripts, allowed us to identify genes missed in Chandler v1.0 due to genome fragmentation, as well as to remove false-positive predictions likely caused by heterologous variants of the same locus mistakenly interpreted and annotated as independent scaffolds in Chandler v1.0. Also, the mean gene length in Chandler v2.0 was higher than the previous gene annotations (Table 2), a consequence of the increased contiguity of the new chromosome-scale reference genome. The average gene density of Chandler v2.0 was 19.75 genes per 100 kb,

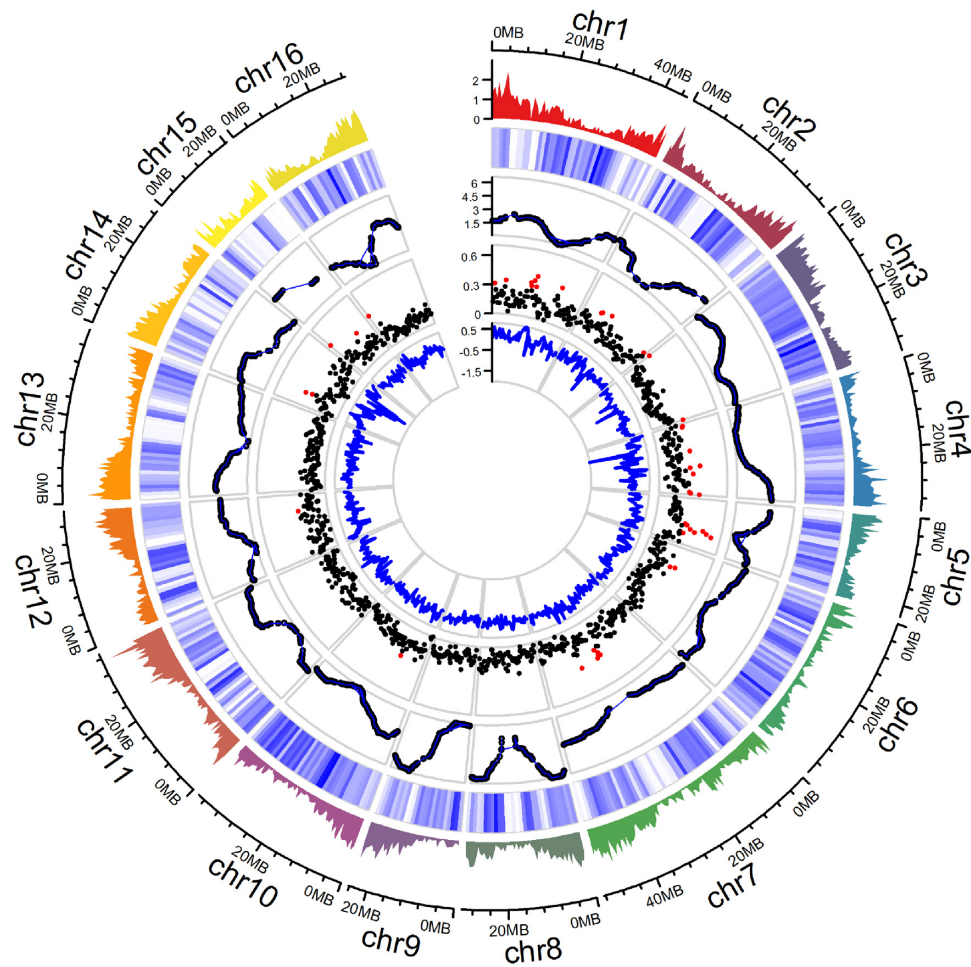


Figure 2: Summary of gene distribution and genetic diversity across the 16 chromosomes of Chandler v2.0. Tracks from outside to inside: (i) gene density of Chandler v2.0 in 1-Mb windows; (ii) Chandler heterozygosity in 1-Mb windows (white = low heterozygosity; blue = high heterozygosity); (iii) Recombination rate for sliding windows of 10 Mb (average = 2.63 cM/Mb); (iv) F_{ST} in 500-kb windows. Windows in the 95th percentiles of the F_{ST} distribution are highlighted in red; (v) ROD values for 500-kb windows.

Table 2: Statistics on the gene annotation of Chandler v2.0 compared with the previous gene annotations of the Chandler genome

Statistics	Chandler v2.0	Chandler v1.0	Chandler RefSeq v1.0
No. of genes	37,554	32,496	41,188
Mean gene length (bp)	5,319	4,358	4,641
Single-exon transcripts	6,613	6,247	6,749
Mean CDS length (bp)	1,335	1,222	1,336
No. of exons	242,208	172,273	230,261
Mean exon length (bp)	257.8	229.5	314
No. of introns	201,290	139,775	181,419
Mean intron length	853.9	730	835
Mean number of introns per gene	5.9	5.3	4.4

with higher gene content in the proximity of telomeric regions (Fig. 2), consistent with other plant genomes [21, 42]. The majority of the predicted gene models of Chandler v2.0 were supported by expression data and showed high similarity with a protein-coding transcript of other plant species (Supplementary Table S8). Also, 30,318 models were annotated with 8,243 different Gene Ontology (GO) terms (Supplementary Figs S6–S8).

Of the 40,884 transcripts identified, 84% were multi-exonic, with 5.9 exons each, on average, and longer introns than the previous gene annotations of Chandler (Table 2). The majority of intron/exon junctions were GT/AG-motif (98.2%), even though alternative splicing with non-canonical motifs was also observed (GC/AG: 0.8%; AT/AC: 0.11%). Almost 90% (36,422) of the coding sequences presented both canonical start and stop codons,

while 4,462 had either a start or a stop codon. This result represents a great improvement compared with Chandler v1.0, where only 48% of the predicted gene models presented both start and stop codons [11].

Also, we observed that 2,801 gene models had from 2 to 4 transcript isoforms each, with a mean length of 9,389 bp. This proportion of gene models with multiple transcript isoforms is smaller than in other plant species [43, 44], likely owing to the low depth of coverage of our PacBio sequencing. Of the 6,437 isoforms identified, 1,448 were covered by FL HQ transcripts in ≥ 1 tissue, while 5,689 were expressed in ≥ 1 of the 20 tissues [11], which most likely covered higher gene diversity compared with the 9 tissues used for PacBio Iso-Seq. On average, the Illumina isoforms (9,188 bp) were longer than the PacBio isoforms (6,790 bp). By running the EnTAP functional annotation pipeline with the entire NCBI RefSeq plant database [45], we observed that almost all isoforms (98%; 6,287) were annotated with a plant protein.

We also investigated possible gene family expansion and contraction among the 3 Chandler versions' gene annotations. Overall, we identified fewer gene families in Chandler v2.0 (5,163 Panther family represented) than v1.0 (5,330) and NCBI RefSeq *J. regia* annotation (5,374). However, when counting the number of members per family, we observed a gene family expansion, in general, in Chandler v2.0 compared with v1.0: 39,357 proteins were assigned to a Panther gene family in Chandler v2.0, with an average of 7.6 members per family, against the 30,639 proteins annotated with a Panther domain in v1.0 (6 members per family on average). On the contrary, we noticed an overall gene family contraction in v2.0 compared to NCBI RefSeq, where 10.4 gene members were assigned to a Panther domain on average. Both the increment of contiguity and the reduction in haplotype redundancy can explain the observed patterns of gene family expansion and contraction among the 3 Chandler versions' gene annotations, even if the different methods of gene prediction used in the 3 studies could also account for these differences.

Most of the 1,440 core genes in the embryophyte dataset from BUSCO were assembled completely (82.5% single-copy; 12.6% duplicated), similarly to other *Fagales* genome assemblies (Table 3) [34, 35, 40, 46, 47]. Also, 88% of both rosids and green sets of core gene families (coreGFs) were identified in the gene annotation, confirming the HQ and completeness of the gene space of Chandler v2.0.

Improved assessment of proteomes with the complete genome sequence

After confirming the importance of a chromosome-scale reference genome for the improvement of gene prediction accuracy, we analyzed the impact of a contiguous genome sequence using proteomic analysis. Proteomes are commonly investigated by isolating the total protein complement of a sample and fragmenting those proteins into smaller peptides that are resolved by mass and charge by means of mass spectrometry. After detection, the peptides' amino acid sequences are determined by matching their mass and charge to candidate sequences obtained from a reference proteome inferred from the reference genome (virtual proteome). A fragmented assembly of the reference genome can lead to an inaccurate prediction of a species' proteome and, then, a misidentification of the proteins expressed in specific tissues at particular stages [48].

We isolated proteins of reproductive tissues harvested from mature Chandler walnut trees, focusing on different development stages of the male flower (catkin; Supplementary Fig. S9)

and mature pollen grains. We analyzed the proteomic data generated from these samples using the virtual proteomes predicted from the gene annotation of the new chromosome-scale genome and Chandler v1.0 (NCBI RefSeq). Considering all tissues analyzed, we identified fewer unique peptides (43,083) with the new chromosome-scale walnut genome than with Chandler v1.0 (44,679). In addition, 6,966 unique proteins were detected with Chandler v2.0 against the 8,802 found using version 1 as a search database (Supplementary Table S9; Additional File 3). Most likely, the NCBI proteomic database based on the fragmented Chandler v1.0 included artifacts resulting from an overestimation of the protein-coding genes.

In the example presented below, we focused on the allergenic proteins produced during catkin and pollen development. Approximately 2% of walnut consumers have a high risk of developing allergies to nuts or pollen [49]. Initially, we clustered the samples according to their protein constituents and levels. This revealed a higher similarity between immature and mature catkins and a more distinct pattern of detected proteins between senescent catkins and pure pollen (Fig. 3).

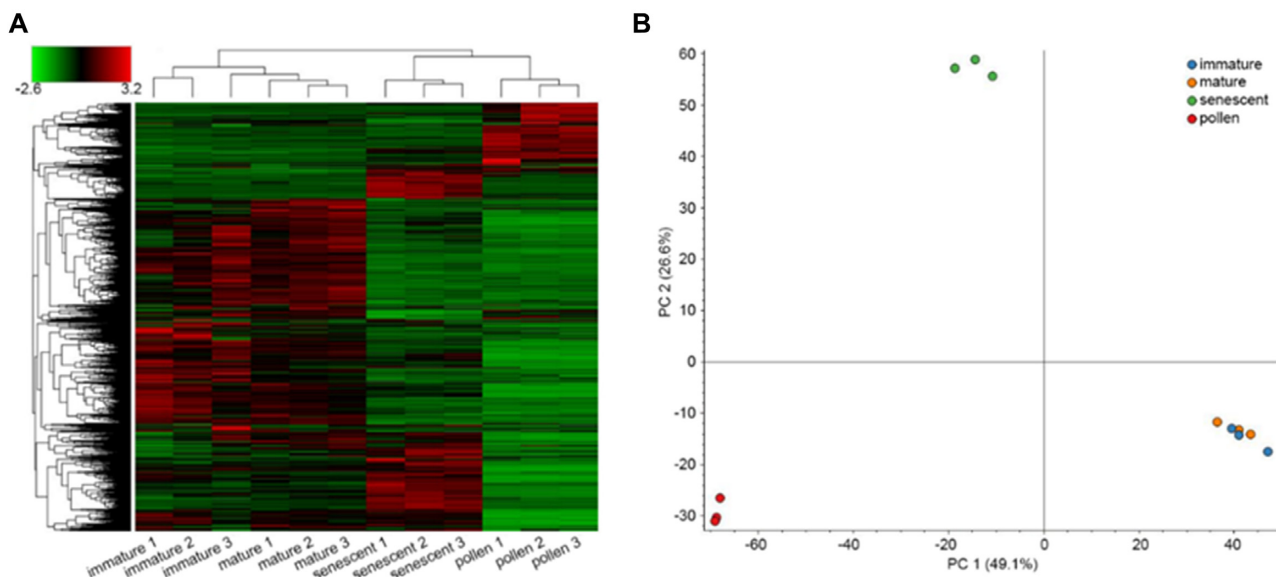
We then searched the 4 analyzed proteomes for allergenic proteins listed in the World Health Organization/International Union of Immunological Societies Allergen Database (www.allergen.org; Supplementary Table S10), as well as for additional proteins not yet registered in the allergen database but predicted in Chandler v2.0 as potential allergens given their predicted structural similarity to known allergens (Additional File 3). Four of the 8 recognized allergenic proteins were detected in ≥ 1 of the catkin developmental stages, with *Jug.r.5* (XP.018825777 | *Jr12.10750*) and *Jug.r.7* (XP.018808763 | *Jr07.28960*) present in all sample types, including pollen (Supplementary Table S10). Genes adjacent to known allergen-coding sequences, likely indicating gene duplications, encode 3 of the new potential allergens (Supplementary Table S10). Moreover, we discovered that the gene locus *Jr12.05180* encodes a non-specific lipid transfer protein (*Jug.r.9* | XP.018813928), a potential allergen highly expressed during catkin maturation and in pollen (Supplementary Tables S10 and S11). In particular, *Jug.r.9* was the most abundant protein in mature and senescent catkins, and the second most abundant in pure pollen. Another interesting allergen similar to *Jug.r.9* (same 8-cysteine configuration) is XP.018814382 | *Jr03.26970*; it decreases as the catkin matures and is absent in pollen (Supplementary Tables S10 and S11). Similarly, polyphenol oxidase (XP.018858848 | *Jr03.06780*) is high in the immature catkin and almost absent in the pollen.

The integration of these proteomic data with previously published transcriptomic data obtained from 20 walnut tissues [11] shows high reproducibility between the methods. In both datasets, allergens *Jug.r.1*, 4, and 6 were not detected in catkins, while the new putative allergen *Jug.r.9* was highly expressed in catkins (Supplementary Tables S11 and S12). Also, *Jr12.05180* transcripts were not detected in any of the 20 tissues but catkin, thus confirming the strong specificity of *Jug.r.9* for catkin and pollen tissue (Supplementary Table S12). Modeling the structure of this putative allergen reveals 4 predicted disulfide bonds, potentially conferring heat and protease resistance, and further suggesting allergenic properties (Supplementary Fig. S10). Future studies will clarify the functional role of this protein and its allergenic nature.

The detection of new potential walnut allergens confirms the positive impact of Chandler v2.0 on proteomic studies in walnut, by providing a clearer and more precise organization of the protein-coding sequences (CDSs) within a genomic region than the previous fragmented genome assembly v1.0.

Table 3: Statistics of the completeness of Chandler v2.0 assessed with BUSCO and compared with other *Fagales* genomes

Genome	BUSCO complete (%)	BUSCO duplicated (%)	BUSCO fragmented (%)	BUSCO missing (%)	Reference
<i>Juglans regia</i> cv. "Chandler" v2.0	95.1	12.6	1.3	3.6	This genome
<i>Juglans regia</i> cv. "Chandler" v1.0	94.8	13.8	1.2	4.0	[11]
<i>Juglans regia</i> cv. "Serr" v1.0	94.5	11.1	1.5	4.0	[24]
<i>Fagus sylvatica</i> v1.2	94	19	1.7	3.6	[36]
<i>Castanea mollissima</i>	96.7	7.7	1.4	1.9	[35]
<i>Carya illinoensis</i> v1	94	23	1.4	3.6	[34]
<i>Corylus avellane</i> cv. "Tombul"	96	6	1	3	[46]
<i>Quercus lobata</i> v1.0	90	52	4	6	[47]
<i>Quercus robur</i>	93	49	3	4	[36]

**Figure 3:** Clustering of the samples used in the proteomic analysis. (A) Hierarchical clustering based on Euclidian distances of normalized abundances of detected proteins. Samples are represented in columns and proteins in rows. (B) Principal component analysis of the 12 samples analyzed, clustering according to tissue type.

Chandler genomic diversity

By anchoring the HiRise assembly to the Chandler genetic map [16], we observed highly homozygous regions in Chandler, especially on Chr15, where the genetic gap spanned 14.5 cM, corresponding to a physical distance of 9.1 Mb. A large gap on Chr15 (9.23 cM; 1.5 Mb) was also observed by Luo et al. [37], which suggested inbreeding as a possible cause for the lack of segregating loci in this region in Chandler, whose parents shared Payne as an ancestor. To confirm the autozygosity of Chandler on Chr15, we used the Illumina whole-genome shotgun data of Chandler and the identified polymorphisms to study its genetic diversity across the new chromosome-scale genome. We identified 2,205,835 single heterozygous polymorphisms on the 16 chromosomal pseudomolecules, with an SNP density of 4.0 SNPs per kb (Fig. 2; Table S13). Fifty-six 1-Mb regions exhibited <377.5 SNPs (10th percentile of the genome-wide SNP number distribution), and chromosomes 15, 1, 7, and 13 were the top 4 chromosomes in the number of low-heterozygosity regions (Table S14). In particular, Chr15 presented 9 windows of 1 Mb with a significantly low number of polymorphisms, 5 of which span 4 Mb at the end of the chromosome. In these 9 low-heterozygosity regions, we found 1,536 SNPs in total (Fig. 2), of which only 25 were tiled on the Axiom *J. regia* 700 K SNPs array. The absence of these

polymorphisms segregating in Chandler in the SNP array could be related to either a failed identification during the SNP calling due to the highly fragmented reference genome v1.0 or to the SNP exclusion during the filtering process applied to build the genotyping array [9]. The low number of Chandler heterozygous SNPs in the array affected the end of Chr15 the most, causing a reduction in the genetic length of the corresponding LG (Fig. 1), as well as leaving unexplored 4 Mb of Chandler genetic variability, which is now accessible thanks to the new chromosome-scale reference genome. The failure to anchor 7 of the scaffolds with telomeric sequences can be explained by the missed detection of terminally located highly homozygous regions during genetic map constructions, due to the absence of crossing-over events with heterozygous flanking markers.

Owing to the evidence of whole-genome duplication in *Juglans* genomes [37], we searched for conserved regions of synteny between Chr15 and its homologous regions in the genome, to study their level of divergence and identify other evolutionary forces as possible causes of the localized reduction of heterozygosity on Chr15. Of the 5,739 pairs of paralogous genes (8,701 genes; Supplementary Fig. S11) identified in Chandler v2.0, 448 included genes on Chr15, and 389 of these have their respective paralogues on Chr6 (Supplementary Fig. S12), in line with what

was already reported by Luo et al. [37]. The Chr06-Chr15 pairs of paralogous genes showed average values of divergence indexes ($K_S = 0.38$; $K_A = 0.13$) similar to the ones observed genome-wide for other syntelogs ($K_S = 0.40$; $K_A = 0.09$), which are paralogous genes derived from the same ancestral genomic region. Similar values of divergence were also observed for the 178 Chr06-Chr15 syntelogs (171 genes) falling within the 9 low-heterozygosity regions on Chr15 ($K_S = 0.40$, $K_A = 0.10$), excluding different evolutionary rates for these regions. Other than paralogous genes, we found 393 singleton genes in the low-heterozygosity regions on Chr15 of Chandler. These genes are involved in different biological processes, many of which are related to signal transduction, protein phosphorylation, and response to environmental stimuli (Supplementary Table S15).

We further investigated the contribution of inbreeding to the high level of autozygosity on Chr15 by visualizing the inheritance of haplotype blocks (HB; genomic regions with little recombination) across the Chandler pedigree (Fig. 4B). We observed that Payne accounts for the entire Chandler genetic makeup (19 HBs for the total length of Chr15) inherited from Pedro (mother), where only 1 HB (2.08 Mb) shared the same allele of Conway-Mayette (maternal grandfather; Fig. 4A). Regarding the paternal genetic makeup of Chandler, 13 of 19 HBs (9.05 Mb) on Chr15 inherited Payne alleles, providing further evidence of high inbreeding on this chromosome (Fig. 4A). This is even more evident in assessing the number of alleles matching between Payne and Chandler across the genome: Chr15 (14 HBs for a total of 13.95 Mb; Supplementary Fig. S13) shares full allele identity with Payne for almost its entire length. Such allele matching between Chandler and its ancestor Payne also occurs on Chr1 (9 HBs for a total of 8.44 Mb), Chr4 (6 HBs; 7.68 Mb), Chr7 (21 HBs; 21.62 Mb), and Chr14 (7 HBs; 12.29 Mb; Fig. 5). These results suggest a high level of inbreeding in many genomic regions of Chandler (Fig. 5), even though direct and indirect selection might have caused the observed presence of extended homozygous regions in Chandler's genome.

Genomic comparison between Eastern and Western walnuts

Even though numerous surveys regarding genetic diversity within walnut germplasm collections have been reported so far [50, 51], comparative analyses at the population level and genome scans for signatures of selection are still missing in Persian walnut. The availability of a chromosome-scale reference genome enables exploration of the patterns of intraspecific variation at the genomic level, providing new insight on the extraordinary phenotypic diversity present within *J. regia*.

We used the resequencing data generated for 23 founders of the Walnut Improvement Program of the University of California, Davis (UCD-WIP; Supplementary Table S16) [12], to study the genome-wide genetic differentiation among walnut genotypes of different geographical provenance. We identified 14,988,422 SNPs, and >97% of them were distributed on the 16 chromosomal pseudomolecules, with 9.4 polymorphisms per kilobase pair. A hierarchical clustering analysis (Supplementary Fig. S14) divided the 23 founders into 2 major groups, including genotypes from Western countries (USA, France, and Bulgaria) and Asia (China, Japan, Afghanistan), respectively, as previously reported [9, 52]. High phenotypic diversity for many traits of interest in walnut, such as phenology, nut quality, and yield, has been observed within and between germplasm collections from Western and Eastern countries [53]. Walnut trees from Asia are noted for their lateral fruitfulness and precocity, rarely observed in the

USA and western Europe, so they have been used as a source of these phenotypes in different walnut breeding programs [54].

At a genomic level, we found a moderate differentiation ($F_{ST} = 0.15$) between Western and Eastern genotypes, except for 195 genomic windows (100 kb) that showed substantially high population differences ($F_{ST} \geq 0.36$; top 5% in the whole genome). In particular, chromosomes 7, 5, 1, 4, and 2 presented ~70% of the divergent sites (Fig. 2; Supplementary Fig. S15). As suggested by the mean reduction of diversity coefficient (ROD) value (0.41), in most of the genomic regions highly differentiated, the UCD-WIP founders from the USA and Europe showed lower nucleotide diversity ($\pi = 2.5 \times 10^{-4}$) than the Asian genotypes ($\pi = 5.0 \times 10^{-4}$), consistent with Bernard et al. [10] (Fig. 2; Supplementary Fig. S15). The proximity of our eastern genotypes to the supposed walnut center of domestication in central Asia can explain the high level of diversity observed in this subgroup.

More than 60% (122) of the highly differentiated windows showed a negative value of Tajima D in the EU/USA subgroup ($D_{Occ} = -1.12$), thus suggesting that selection has been likely acting on these genomic regions in the Western genotypes (Supplementary Fig. S15). Here we found 743 genes, with GO biological categories mostly related to signal transduction, embryo development, and response to stresses (Table S17). Ten candidate selective sweeps ($D_{Asia} = -0.54$) were also observed in the Eastern group (Supplementary Fig. S15), which included 57 predicted genes, related to terpenoid biosynthesis, post-embryonic development, and signal transduction (Supplementary Table S18).

Recently, many marker-trait associations have been reported for different traits of interest in walnut, such as leafing date, nut-related phenotypes, and water use efficiency [14–16]. We looked to see whether any of these trait-associated SNPs fell within regions highly differentiated between Western and Eastern genotypes. Three loci associated with shape index, nut roundness, and nut shape [14] are located in 2 genomic regions on chromosome 3 and 4 with significantly high values of F_{ST} (Supplementary Table S19). In both of these regions, Western genotypes presented lower genetic diversity and lower values of Tajima D than the Eastern walnuts. These findings may suggest that, while a selective pressure for nut shape may have occurred in the EU/USA subgroups, higher phenotypic variability can be expected for these traits in the Eastern countries. We also found that the locus AX-170770379, strongly associated with harvesting date [16], falls within a genomic region on Chr1 with an F_{ST} value equal to 0.39 and lower genetic diversity in the Western genotypes (ROD = 0.63; Supplementary Table S19). Looking at the phenotypic effect of this SNP on the harvest date of the 23 founders, we observed that most of the Western genotypes are later harvesting than the Eastern (Supplementary Fig. S16), suggesting differences in the timing of phenological events between these 2 groups as adaptation to the different climate conditions present in their countries of origin [55].

Future resequencing projects involving larger walnut collections and covering a wider area of the global walnut distribution are necessary to confirm and interpret the observed genomic differentiation between Western and Eastern walnuts, likely helping to elucidate the role of this genomic divergence in the evolutionary history of Persian walnut.

Methods

Oxford Nanopore sequencing and assembly

High molecular weight (HMW) DNA for Nanopore sequencing (Oxford Nanopore Technologies Inc., UK) was isolated through

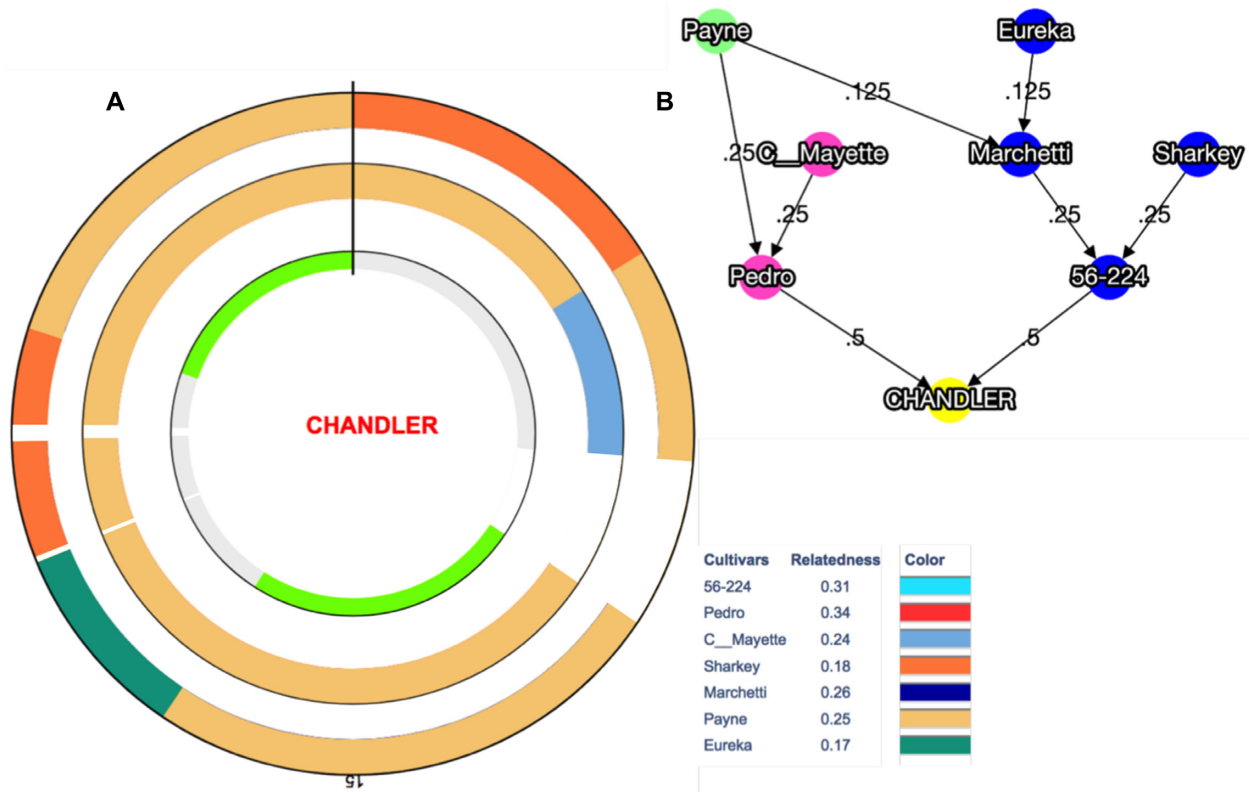


Figure 4: Graphical visualization of haplotype block (HB) inheritance on Chr15 along with the Chandler pedigree. (A) The inner circle highlights in grey 2 regions of heterozygosity (5 HB the first and 7 HB the second), and in light green 2 regions of homozygosity (3 HB the first and 4 HB the second). The circle in the middle shows maternally inherited HBs, while the HBs inherited through the paternal line are visualized in the outer circle. Payne's haplotypes are clearly present in both parental lines. White spaces represent segments of missing haplotype information. (B) Chandler pedigree, where Pedro is the maternal line and 56-224, the paternal line.

a nucleus extraction and lysis protocol. First, mature leaf tissue from the same tree used for the original *J. regia* Chandler genome [11] was homogenized with mortar and pestle in liquid nitrogen until well ground, then added to the Nuclei Isolation Buffer [56] and stirred at 4°C for 10 minutes. The cellular homogenate was filtered through 5 layers of Miracloth (Millipore-Sigma) into a 50-mL Falcon tube, then centrifuged at 4°C for 20 minutes at 3,000g. This speed of centrifugation was selected on the basis of the estimated walnut genome size of 1 Gb [57]. Extracted nuclei were then lysed for 30 minutes at 65°C in the sodium dodecyl sulfate-based lysis buffer described by Mayjonade et al. [58]. Afterwards, 0.3 volumes of 5 M potassium acetate were added to the lysate to precipitate residual polysaccharides and proteins. The sample was incubated for 5 minutes at 4°C and then centrifuged at 4°C for 10 minutes at 2,400g. After removal of the supernatant, genomic DNA (gDNA) was ethanol precipitated and then eluted in 10 mM Tris-Cl. Further purification of the gDNA was then performed using a Zymo Genomic DNA Clean and Concentrate column.

A 1-μg aliquot of the isolated gDNA was prepared for sequencing using the Ligation sequencing kit (LSK108, ONT) following the manufacturer's protocol with an optimized end repair (100 μL sample, 14 μL enzyme, 6 μL enzyme, incubated at 20°C for 20 minutes then 65°C for 20 minutes). In detail, the gDNA was end polished using the NEBNext® Ultra™ II DNA Library Prep Kit and then cleaned up with 1X Ampure XP beads (Beckman Coulter). Afterwards, the gDNA was ligated to ONT-specific adapters, followed by an additional cleanup with 0.4X Ampure XP beads. Finally, the libraries were sequenced for 48

hours on 6 flowcells of the ONT Mk1B MinION platform with the R9.4 chemistry. Raw fast5 data were base-called using Albacore version 1.25.

The ONT data and Illumina reads from Martínez-García [11] were combined to obtain the Chandler hybrid assembly using MaSuRCA v3.2.3 [59]. In detail, MaSuRCA first transformed the Illumina paired-end reads into "super-reads" using the super-reads algorithm, which uses *k*-mers from Illumina reads to extend each Illumina read uniquely in both directions. Then, each ONT read was used as a template to which super-reads can be attached, and the approximate alignments of all super-reads to each ONT read were computed. The best path of the exactly overlapping aligned super-reads on an ONT read was then defined, generating a "mega-read." The mega-reads typically have a very low error rate (<1%) because they are constructed from the super-reads, and most of them span the full length of the long reads. Finally, a customized version of the CABOG assembler [60] was used to assemble the mega-reads along with the Illumina mate pairs, which provide the linking information for the scaffolding. Gaps were closed using the gap-filling procedure implemented in MaSuRCA and described by Zimin et al. [59]. The de-duplication module implemented in MaSuRCA was then applied to remove duplicative sequences (scaffold variants due to heterozygosity).

De-duplicated scaffolds were aligned onto the previously finished *J. regia* chloroplast genome [11] using "minimap2 -x asm5," as well as to a database of 223 finished plant mitochondria (downloaded from NCBI RefSeq) using blastn with default parameters. Finally, Chandler v1.0 was aligned to the de-duplicated

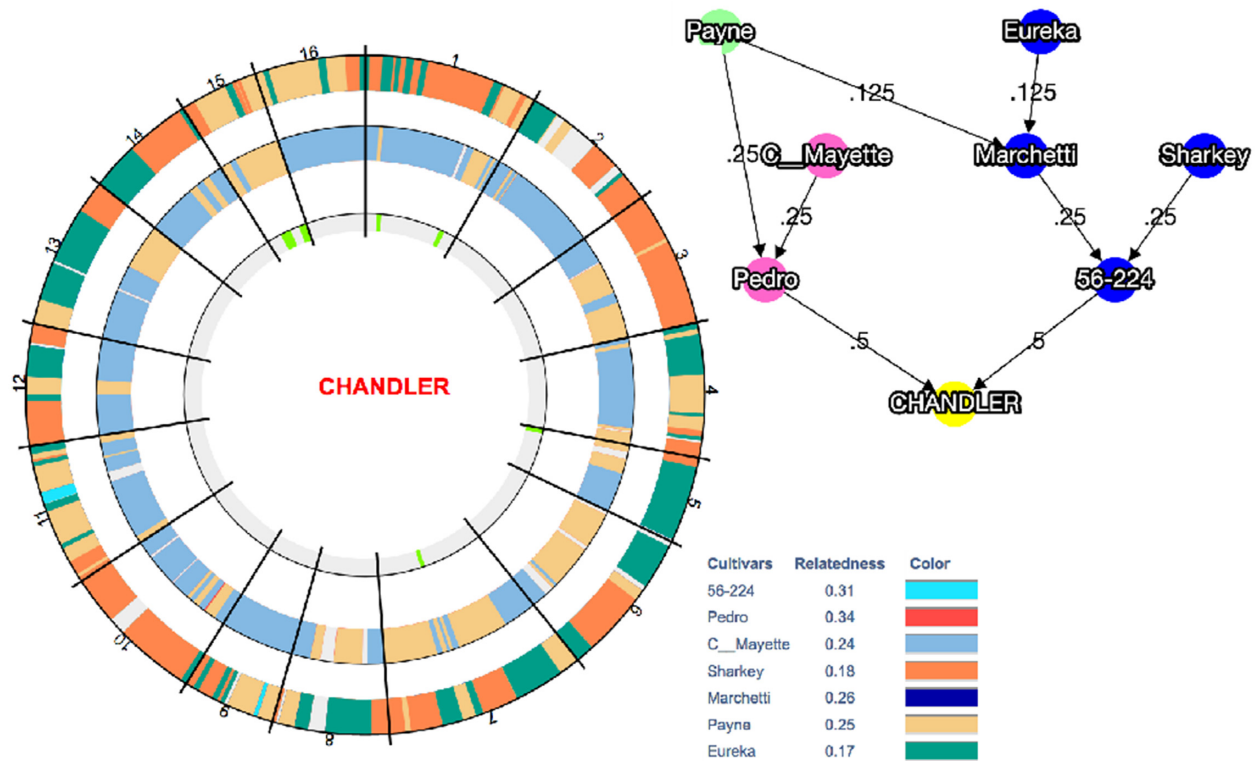


Figure 5: Graphical visualization of the haplotype block (HB) inheritance across Chandler pedigree in the 16 chromosomes. The inner circle highlights in grey the regions of heterozygosity and in light green the regions of homozygosity for each chromosome. The circle in the middle shows the maternally inherited HBs, while the HBs inherited from the paternal line are visualized in the outer circle. In both parental line circles, missing data are highlighted in grey. Payne haplotypes are inherited along both parental lines in all chromosomes but Chr5, Chr9, Chr10, Chr14, and Chr16. Chandler pedigree is represented on the side, where Pedro is the maternal line and 56-224 the paternal line.

hybrid assembly, and the unaligned regions were added to the Chandler hybrid assembly.

Hi-C sequencing

A Hi-C library was prepared by Dovetail Genomics LLC (Santa Cruz, CA, USA) as described previously [61]. Briefly, for each library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Biotin that was not internal to ligated fragments was removed from the purified DNA. Purified DNA was then sheared to ~350 bp mean fragment size. Sequencing libraries were generated using NEBNext® Ultra™ enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were then sequenced on the Illumina HiSeq4000 platform.

The hybrid ONT assembly, Illumina shotgun reads [11], and Dovetail Hi-C library reads were used as input data for the scaffolding software HiRise, which uses proximity ligation data to scaffold genome assemblies [62]. Shotgun and Dovetail Hi-C library sequences were aligned to the hybrid ONT assembly using a modified SNAP read mapper. The separations of Dovetail Hi-C read pairs mapped within the ONT scaffolds were analyzed by HiRise to produce a likelihood model for the genomic distance between read pairs, and the model was used to identify and

break putative misjoins, to score prospective joins, and make joins above a threshold. After scaffolding, Illumina shotgun sequences were used to close gaps between contigs, resulting in an improved HiRise assembly.

Validation and anchoring of the HiRise assembly to Chandler genetic maps

The HiRise assembly was first anchored to the Chandler genetic map obtained by Marrano et al. [16] from a 312-offspring F_1 population “Chandler × Idaho” genotyped with the latest Axiom *J. regia* 700 K SNP array. SNP probes (71-mers including the SNP site) from the Axiom *J. regia* 700 K SNP array were aligned onto the HiRise assembly, filtering out alignments with probe/reference identity <98%, covering <95% of the probe length, or aligning multiple times on the genome. Retained markers with a unique segregation profile were then used to anchor the HiRise scaffolds. The same procedure was also followed to anchor the HiRise assembly to the Chandler genetic map used to construct a walnut bacterial artificial chromosome clone-based physical map by Luo et al. [37]. The final ordering of scaffolds was performed by taking into consideration the marker genetic map position, and, in the final sequence, consecutive scaffolds were separated by sequences of 100,000 nucleotides.

The Tandem Repeat Finder program (TRF v4.09 [63]) was run using the recommended parameters (max mismatch delta PM PI minscore maxperiod, 2 7 7 80 10 50 500, respectively) to identify repeat elements up to 500 bp long. A histogram of repeat unit lengths was generated, and peaks at 7, 29, 33, 44, 154, and

308 bp were identified. From these data, a consensus sequence corresponding to each peak was selected. All of these repeat sequences were aligned onto the HiRise assembly using “nucmer” from the MUMmer4 package [64] with a minimum match length of 7 to capture the telomeric repeat. On the basis of the positions of these alignments along the chromosomes and contigs, we identified the 7-mer as the telomeric repeat and the 154-mer and 308-mer as centromeric repeats.

Recombination rate was estimated within sliding windows of 10 Mb with a step of 1 Mb along the chromosome sequence by using the high-density genetic map of Chandler [16] and the R/MareyMap package v1.3.4 [65]. To evaluate the Chandler v2.0 error rate, the 2 assemblies, Chandler v1.0 and 2.0, were aligned to each other using nucmer [64]. Assembly quality statistics were estimated using QUAST v5.0.2 [66], filtering for contigs with a minimum length of 1 kb. The haploid size of the walnut genome was estimated by first generating the 24-mer distribution of Illumina paired-end reads (54-fold coverage of the haploid genome) with Jellyfish v2.2.6 [67] and then uploading it to Genomescope [68]. Comparisons of Chandler v2.0 versus Jrserr.v1.0 and vice versa were performed using nucmer [64], and then the function “dnadiff” implemented in MUMmer4 was used to obtain detailed information on the differences between 2 assemblies.

RNA preparation

Five walnut tissues (leaf, catkin 1-inch elongated, catkin 3-inches elongated, pistillate flower, and pollen) were collected from Chandler trees at the UCD walnut orchards. Four additional samples (somatic embryo, callus, shoot, and roots) were taken from tissue culture material of Chandler. Several grams of each tissue were ground in liquid nitrogen and with insoluble polyvinylpyrrolidone (1% w/w). RNA was isolated using the PureLink™ Plant RNA Reagent (Invitrogen™, Carlsbad, CA) following the manufacturer's instructions, but with an additional end wash in 1 mL of 75% ethanol. For root tissue only, RNA isolation was performed using the MagMAX™ mirVana™ Total RNA Isolation Kit (Applied Biosystems™, Foster City, CA) as per protocol, except for the lysis step. A different lysis buffer was created adding 100 mg of sodium metabisulfite to 10 mL of guanidine buffer (guanidine thiocyanate 4 M, sodium acetate 0.2 M, EDTA 25 mM, PVP-40 2.5%, pH 5.0) and 1 mL of nuclease-free water. Then, 100 mg of ground root tissue was lysed in 1 mL of the new lysis buffer using a Tissue Lyser at maximum frequency for 2 minutes. The lysate was centrifuged at 4°C for 5 minutes at maximum speed. The supernatant (500 µL) was transferred to a new tube for the following steps of RNA isolation as per protocol. RNA samples were then purified, and DNase treated using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). The RNA quality was confirmed by running an aliquot of each sample on an Experion™ Automated Electrophoresis System (Bio-Rad, Hercules, CA).

PacBio Iso-Seq sequencing

Full-length cDNA Iso-Seq template libraries for PacBio Iso-Seq analysis were constructed and sequenced at the DNA Technologies & Expression Analysis Core Facility of the UCD Genome Center. FL double-stranded cDNA was generated from total RNA (2 µg per tissue) using the Lexogen Telo™ prime Full-length cDNA Kit (Lexogen, Inc., Greenland, NH). Tissue-specific cDNAs were first barcoded by PCR (16–19 cycles) using IDT barcoded primers (Integrated DNA Technologies, Inc., Coralville, IA), and then bead-size selected with AMPure PB beads (2 different size

fractions of 1X and 0.4X). The 9 cDNAs were pooled in equimolar ratios and used to prepare a SMRTbell™ library using the PacBio Template Prep Kit (PacBio, Menlo Park, CA). The SMRTbell™ library was then sequenced across 4 Sequel v2 SMRT cells with polymerase 2.1 and chemistry 2.1 (P2.1C2.1).

PacBio raw reads were processed using the Iso-Seq3 v3.0 workflow following PacBio recommendations [69]. CCSs were generated using the program “ccs.” The CCSs were demultiplexed and cleaned of cDNA primers using the program “lima.” Afterward, CCS clustering and polishing was performed using the program “Iso-Seq3,” to generate HQ FL sequences for each of the 9 tissues. FLnc and HQ clusters were aligned onto the new “Chandler” assembly v2.0 with minimap2 v2.12-r827, including the parameter “-ax splice” [70].

Repeat annotation

A genome-specific repeat database was created using the “basic” mode implemented in RepeatModeler v1.0.11 [71]. RepeatMasker v4.0.7 was then run to mask repeats in the walnut reference genome v2.0 and generate a GFF file [72].

Gene prediction and functional annotation

Juglans regia RefSeq transcripts and additional *J. regia* transcripts and protein sequences downloaded from NCBI, along with the HQ FL Iso-Seq transcripts, were used as input to the PASA pipeline v2.3.3 [73], to assemble a genome-based transcript annotation. PASA uses aligners BLAT v35 [74] and GMAP v2018-07-04 [75], along with TransDecoder v5.5.0 [76], which predicts open reading frames as genome-based GFF coordinates. The final PASA/TransDecoder GFF3 file was post-processed to name the genes and transcripts by chromosome location consistently. The chloroplast and mitochondrial genomes were annotated using the “CHLOROBOX GeSeq Annotation of Organelle Genomes” tool at [77] with default parameters [78]. NCBI accessions NC.028617.1 (*J. regia* chloroplast), KT971339.1 (*Medicago truncatula* mitochondrion), NC.029641.1 (*M. truncatula* mitochondrion), and NC.012119.1 (*Vitis vinifera* mitochondrion) were also input as custom references. The output gff3 files were then post-processed to consistently rename genes.

Functional roles were assigned to predicted peptides using Trinotate v3.1.1 [79]. In particular, similarity searches were performed against several public databases (i.e., Uniprot/Swiss-Prot, NCBI NR, *Vitis vinifera*.IGGP.12x, *J. regia* RefSeq) using BLAST v2.8.1, HMMER v3.1b2, SignalP v4.1c, and TMHMM v2.0c. Gene family analysis was performed by running Interproscan v. 5.30-69.0 [80, 81] with default parameters on each protein fasta file (v1.0 [11], NCBI RefSeq [GCF.0 014 11555.1.wgs.5d] and v2.0). The PANTHER family ID with the lowest expect value (below expect value threshold of $1.0E^{-11}$) was assigned to each protein.

The completeness and quality of both genome assembly and gene annotation of Chandler v2.0 were estimated with the BUSCO method v3 (1,440 core genes in the embryophyte dataset) [82], and the sets of coreGFs of green plants (2,928 coreGFs) and rosids (6,092 coreGFs) from PLAZA v2.5 [83]. Also, RNA-sequencing (RNA-Seq) data previously generated for 20 tissues (see [11]) were aligned to the reference genome (v1.0 and v2.0) with HISAT2 [84]. The alignments of the 20-tissue RNA-Seq data and the FL transcripts along with the new genome annotation v2.0 were then used as input to StringTie v2.0 [85] to estimate expression levels in both fragments per kilobase per million reads (FPKM) and transcripts per million (TPM) for each transcript in the v2 annotation. The percent identity and coverage of each *J.*

regia transcript compared to proteins in the NCBI plant RefSeq database was also determined by running the EnTAP pipeline v.0.9.0 [45].

Label-free shotgun proteomics

Plant tissues of immature, intermediate, and mature catkins (Additional File 1, Supplementary Fig. S9) and pure pollen from 3 individual trees of Chandler at the UCD walnut orchards were collected and frozen immediately in dry ice. Tissues were then further frozen in liquid nitrogen in the laboratory and ground with mortar and pestle. Five hundred milligrams of each sample were used for total protein extraction, following the procedure for recalcitrant plant tissues of [86], with a modification in the final buffer used to resuspend the protein pellet, consisting of 8 M urea in 50 mM triethylammonium bicarbonate (TEAB). Then, 100 μ g of total protein from each sample was used for proteomics.

Initially, 5 mM dithiothreitol (DTT) was added and incubated at 37°C for 30 minutes and 1,000 rpm shaking. Next, 15 mM iodoacetamide (IAA) was added, followed by incubation at room temperature for 30 minutes. The IAA was then neutralized with 30 mM DTT in incubation for 10 minutes. Lys-C/trypsin then was added (1:25 enzyme: total protein) followed by 4 h incubation at 37°C. Afterward, TEAB (550 μ L of 50 mM) was added to dilute the urea and activate trypsin digestion overnight. The digested peptides were desalted with Aspire RP30 Desalting Tips (Thermo Scientific), vacuum dried, and suspended in 45 μ L of 50 mM TEAB. Peptides were quantified by Pierce quantitative fluorometric assay (Thermo Scientific) and 1 μ g analyzed on a QExactive mass spectrometer (Thermo Scientific) coupled with an Easy-LC source (Thermo Scientific) and a nanospray ionization source. The peptides were loaded onto a Trap (100 μ m, C18 100 Å 5 U) and desalted online before separation using a reversed-phase (75 μ m, C18 200 Å 3 U) column. The duration of the peptide separation gradient was 60 minutes using 0.1% formic acid and 100% acetonitrile for solvents A and B, respectively. The data were acquired using a data-dependent tandem mass spectroscopy (MS/MS) method, which had a full scan range of 300–1,600 Da and a resolution of 70,000. The resolution of the MS/MS method was 17,500 and the insulation width 2 m/z with a normalized collision energy of 27. The nanospray source was operated using a spray voltage of 2.2 kV and a transfer capillary temperature heated to 250°C. Samples were analyzed at the UCD Proteomics Core.

The raw data were analyzed using X! Tandem and viewed using the Scaffold Software v.4 (Proteome Software, Inc.). Samples were searched against UniProt databases appended with the cRAP database, which recognizes common laboratory contaminants. Reverse decoy databases were also applied to the database before the X! Tandem searches. The CDSs annotated in Chandler v1.0 (NCBI accession PRJNA350852) and v2.0 were used as a reference for identification of proteins from the mass spectrometry data. The proteins identified were filtered in the Scaffold software based on the following criteria: 1.0% FDR (false-discovery rate) at protein level (following the prophet algorithm [87]), the minimum number of 2 peptides, and 0.1% FDR at the peptide level. Structure of the walnut allergen (Jug r 9) was modelled using SWISS-MODEL [88] based on the structure of a homologous allergen from lentil (PDBid:2MAL). Structures were superimposed using MUSTANG (2MAL: in red, walnut in blue) [89].

Chandler genomic diversity

Illumina whole-genome shotgun data of Chandler were aligned on Chandler v2.0 with BWA [90] with standard parameters. SNP calling was performed using SAMtools v1.9 [91] and BCFtools v2.1 [92]. SNP density for windows of 1 Mb was estimated using the command “SNPdensity” implemented in VCFtools v0.1.16 [93]. Self-collinearity analysis to detect duplicated regions in Chandler v2.0 was performed with MCScanX [94], using a simplified GFF file of the new gene annotation and a self-BLASTP as input. To improve the power of collinearity detection, tandem duplications were excluded after running the function “detect_collinear_tandem_arrays” implemented in MCScanX. Synonymous (K_s) and nonsynonymous (K_a) changes for syntenic protein-coding gene pairs were measured using the Perl script “add_ka_and_ks_to_collinearity.pl” implemented in MCScanX.

To explore the inbreeding level across the 16 chromosomal pseudomolecules of Chandler, haplotypes were built for 55 individuals of the UCD-WIP, including 25 founders and several commercially relevant walnut cultivars (e.g., Chandler, Howard, Tulara, Vina, Franquette) along with their parents and progenitors. All individuals were genotyped using the latest Axiom™ J. regia 700 K SNP array as described in [9]. To define SNP HBs, 26,544 unique and robust SNPs were selected and ordered according to the Chandler genome v2.0 physical map. Subsequently, for each SNP marker and individual, phasing and identification of closely linked groups of SNPs, without recombination in most of the pedigree, was performed using the software FlexQTL™ [95] and PediHaplotyper [96] following the approach described in [96] and [97]. In particular, HBs were defined by recombination sites detected in ancestral generation of Chandler.

Genomic comparison between Eastern and Western walnuts

The resequencing data of 23 founders of the UCD-WIP (Table S16) [12] were mapped onto the Chandler v2.0 with BWA, and SNPs were called following the same procedure described above for Chandler. SNPs with no missing data and minor-allele frequency > 10% were retained for the following genetic analyses (7,269,224 SNPs out of the 14,988,422 identified). Hierarchical cluster analysis on a dissimilarity matrix of the 23 UCD-WIP founders was performed using R/SNPRelate v.1.18.0 [98]. Fixation index (F_{ST}) was measured between genotypes from EU/USA and Asia with VCFtools v0.1.16, setting windows of 100 and 500 kb. Genomic windows with the top 5% of F_{ST} values were selected as candidate regions for further analysis. The empirical cut-off with a low FDR (5%) was verified by performing whole-genome permutation test (1,000) with a custom Python script. Nucleotide diversity (π) and Tajima D [99] were also computed along the whole genome in 100- and 500-kb windows using VCFtools. Reduction of diversity coefficient was estimated as $1 - (\pi_{Occ}/\pi_{Asia})$. The new walnut gene annotation v2.0 was used to identify predicted genes in the candidate regions under selection. The distribution of the identified genes into different biological processes was evaluated using the weight01 method provided by R/topGO [100]. The Kolmogorov-Smirnov-like test was performed to assess the significance of over-representation of GO categories compared with all genes in the walnut gene prediction. Plots were obtained using the R/circlize v.0.4.6 and R/ggplot2 v.3.5.3 packages.

Availability of Supporting Data and Materials

All raw and processed sequencing data generated in this study have been submitted to the NCBI BioProject database [101] under accession number PRJNA291087. All SNP data have been submitted to Hardwood Genomics [102]. Data further supporting this work are openly available in the GigaScience repository, GigaDB [103].

Additional Files

Supplementary Figure S1. Contiguity of the Chandler ON assembly and the final HiRise scaffolds.

Supplementary Figure S2. Mapping positions of the first and second read in the read pair respectively, grouped into bins.

Supplementary Figure S3. Collinearity between the 'Chandler' genetic map of [35] and the 16 chromosomal pseudomolecules of Chandler v2.0.

Supplementary Figure S4. Whole-genome comparison between Chandler v2.0 and JrSerr.v1.0 [22].

Supplementary Figure S5. Retrotransposons distribution across the 16 chromosomes of Chandler v2.0.

Supplementary Figure S6. Top 20 biological process GO terms.

Supplementary Figure S7. Top 20 molecular function GO terms.

Supplementary Figure S8. Top 20 cellular component GO terms.

Supplementary Figure S9. Catkin development stages sampled for the label-shotgun proteomic study (pollen is not shown).

Supplementary Figure S10. Modeled structure of the putative new allergen encoded by *Jr12.05180*.

Supplementary Figure S11. Collinear blocks among the 16 chromosomal pseudomolecules of Chandler v2.0.

Supplementary Figure S12. Dual synteny plot between Chr06 a-Chr15 of Chandler v2.0.

Supplementary Figure S13. Graphical visualization of allele identity between Chandler and its ancestor Payne for all 16 chromosomes of Chandler.

Supplementary Figure S14. Hierarchical clustering analysis among the 23 re-sequenced founders of the UCD-WIP.

Supplementary Figure S15. Genome scan for selective sweeps between walnuts from EU/USA and Asia.

Supplementary Figure S16. Phenotypic differences of harvesting date observed among the three genotypic classes of the marker AX-170770379 significantly associated to harvest date (Marrano et al., 2019).

Supplementary Table 1. Statistics on the six Nanopore flowcells used to sequence Chandler's genome.

Supplementary Table 2. Statistics on k-unitigs, super-reads and mega-reads obtained with the MaSuRCA assembler on ONT and Illumina reads.

Supplementary Table 3. Characteristics of the Chandler ONT assembly.

Supplementary Table 4. List of tissues used for PacBio IsoSeq.

Supplementary Table 5. Statistics on the PacBio IsoSeq sequencing per flow-cell.

Supplementary Table 6. Statistics on CCSs, FLnc and HQ FL transcripts obtained per tissue with PacBio IsoSeq.

Supplementary Table 7. Percentage of FLnc and HQ FL transcripts aligned on the new assembly per tissue.

Supplementary Table 8. Validation of gene annotation with expression data and public plant protein databases.

Supplementary Table 9. Summary of proteome results obtained using genome annotations v1 and v2.

Supplementary Table 10. Allergen expression data obtained from proteome data.

Supplementary Table 11. Top 10 abundant proteins detected in each tissue.

Supplementary Table 12. Expression of genes encoding allergenic proteins in different *Juglans regia* cv. Chandler tissues.

Supplementary Table 13. Number of SNPs and SNP density in 100-kb windows per chromosome in Chandler v2.0.

Supplementary Table 14. Regions (1 Mb) with less than 377.5 SNPs (10th percentile of the SNP number distribution) in Chandler.

Supplementary Table 15. Top 50 biological process GO terms for the 393 singletons genes in the low heterozygous regions on Chr15 of Chandler.

Supplementary Table 16. List of 23 UCD-WIP founders used for the selective sweep analysis.

Supplementary Table 17. Top 50 biological process GO terms for the 122 windows (100 kb) with negative value of Tajima's D in the Western genotypes.

Supplementary Table 18. Top 50 biological process GO terms for the 122 windows (100 kb) with negative value of Tajima's D in the Eastern genotypes.

Supplementary Table 19. Marker-trait associations identified within genomic regions highly differentiated between Western and Eastern walnuts.

Additional File 3. Mass-spectrometry proteome data of catkins and pollen tissues. Three samples of each tissue type (immature catkin, mature catkin, senescent catkin, and pure pollen) were analyzed using v1.0 and v2.0 reference walnut genome assemblies. Total intensity of matching peptides, number of spectra, and percentage of protein covered by the identified peptides are reported.

Abbreviations

BLAST: Basic Local Alignment Search Tool; BLAT: BLAST-Like Alignment Tool; bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; BWA: Burrows-Wheeler Aligner; CABOG: Celera Assembler with the Best Overlap Graph; CCS: circular consensus sequence; cDNA: complementary DNA; CDS: protein-coding sequence; coreGF: core gene family; DTT: dithiothreitol; FDR: false-discovery rate; FLnc: full-length non-chimeric; Gb: gigabase pairs; gDNA: genomic DNA; GO: Gene Ontology; HB: haplotype block; Hi-C: chromosome conformation capture; HQ: high quality; HMW: high molecular weight; IAA: iodoacetamide; kb: kilobase pairs; LG: linkage group; LTR: long-terminal retrotransposon; MaSuRCA: Maryland Super-Read Celera Assembler; Mb: megabase pairs; MS/MS: tandem mass spectroscopy; NCBI: National Center for Biotechnology Information; ONT: Oxford Nanopore Technology; PacBio: Pacific Biosciences; PASA: Program to Assemble Spliced Alignments; RNA-Seq: RNA-sequencing; ROD: reduction of diversity coefficient; SNP: single-nucleotide polymorphism; UCD: University of California Davis; TEAB: triethylammonium bicarbonate; WIP: Walnut Improvement Program.

Competing Interests

The authors declare that they have no competing interests.

Funding

This project has been funded by the California Walnut Board.

Authors' Contributions

D.B.N. and A.M. conceived and coordinated the research. R.E.W. and W.T. performed the HMW DNA extraction and Nanopore sequencing. A.V.Z., D.P., and S.L.S. assembled the hybrid Illumina-ONT assembly. L.B., M.T., D.P., and S.L.S. validated and anchored the HiRise assembly to the genetic maps. A.M. and B.J.A. collected and extracted all RNA samples. M.B. analyzed the PacBio Iso-Seq results and performed the repeat and gene annotation. A.D. conceived the design of the proteomic analyses; P.A.Z. and S.C. generated and analyzed the proteomic data. L.B. called the SNPs in Chandler and the 23 UCD-WIP founders, while A.M. carried out the analyses on walnut genomic diversity. E.A.D., L.B., and M.T. built and analyzed the SNP haplotypes. C.A.L. provided all the plant material. A.M. wrote the manuscript, which has been revised by all coauthors.

Acknowledgments

We are grateful to Sriema Walawage for assistance with RNA extraction, and Brett Phinney for preparing the raw proteome data.

References

1. International Nut and Dried Fruit (INC) Statistical Yearbook 2018/2019 (<https://www.nutfruit.org/industry/news/detail/statistical-yearbook>). Accessed 11 May 2020.
2. Martínez ML, Labuckas DO, Lamarque AL, et al. Walnut (*Juglans regia* L.): Genetic resources, chemistry, by-products. *J Sci Food Agric* 2010;**90**:1959–67.
3. McGranahan G, Leslie C. Walnut. In: Badenes ML, Byrne DH, eds. *Fruit Breed*. Springer; 2012:827–46.
4. Pollegioni P, Woeste K, Chiocchini F, et al. Rethinking the history of common walnut (*Juglans regia* L.) in Europe: Its origins and human interactions. *PLoS One* 2017;**12**:1–24.
5. Zhang B, Xu L, Li N, et al. Phylogenomics reveals an ancient hybrid origin of the Persian walnut. *Mol Biol Evol* 2019, doi:10.1093/molbev/msz112.
6. Zeven A, Zhukovskii PM. *Dictionary of Cultivated Plants and their Centres of Diversity, Excluding Ornamentals, Forest Trees, and Lower Plants*. Wageningen, The Netherlands: Centre for Agricultural Publishing and Documentation; 1975. <https://core.ac.uk/download/pdf/29387092.pdf>.
7. FAOSTAT statistics, 2017 <http://www.fao.org/statistics/en/>. Accessed 1 Sept 2019.
8. Ebrahimi A, Zarei A, Lawson S, et al. Genetic diversity and genetic structure of Persian walnut (*Juglans regia*) accessions from 14 European, African, and Asian countries using SSR markers. *Tree Genet Genomes* 2016;**12**:114.
9. Marrano A, Martínez-García PJ, Bianco L, et al. A new genomic tool for walnut (*Juglans regia* L.): development and validation of the high-density Axiom™ J. regia 700 K SNP genotyping array. *Plant Biotechnol J* 2018;**17**(6):1027–36.
10. Bernard A, Barreneche T, Lheureux F, et al. Analysis of genetic diversity and structure in a worldwide walnut (*Juglans regia* L.) germplasm using SSR markers. *PLoS One* 2018;**13**:e0208021.
11. Martínez-García PJ, Crepeau MW, Puiu D, et al. The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *Plant J* 2016;**87**:507–32.
12. Stevens KA, Woeste K, Chakraborty S, et al. Genomic variation among and within six *Juglans* species. *G3 (Bethesda)* 2018;**8**(7):2153–65.
13. Kefayati S, Ikhsan AS, Sutyemez M, et al. First simple sequence repeat-based genetic linkage map reveals a major QTL for leafing time in walnut (*Juglans regia* L.). *Tree Genet Genomes* 2018;**15**:13.
14. Arab MM, Marrano A, Abdollahi-Arpanahi R, et al. Genome-wide patterns of population structure and association mapping of nut-related traits in Persian walnut populations from Iran using the Axiom J. regia 700 K SNP array. *Sci Rep* 2019;**9**:6376.
15. Famula RA, Richards JH, Famula TR, et al. Association genetics of carbon isotope discrimination in the founding individuals of a breeding population of *Juglans regia* L. *Tree Genet Genomes* 2019;**15**:6.
16. Marrano A, Sideli GM, Leslie CA, et al. Deciphering of the genetic control of phenology, yield and pellicle color in Persian walnut (*Juglans regia* L.). *Front Plant Sci* 2019;**10**:1140.
17. Bernard A, Marrano A, Donkpegan A, et al. Association and linkage mapping to unravel genetic architecture of phenological traits and lateral bearing in Persian walnut (*Juglans regia* L.). *BMC Genomics* 2020;**21**:203.
18. Sánchez-Pérez R, Pavan S, Mazzeo R, et al. Mutation of a bHLH transcription factor allowed almond domestication. *Science* 2019;**364**:1095–8.
19. Tang W, Sun X, Yue J, et al. Chromosome-scale genome assembly of kiwifruit *Actinidia eriantha* with single-molecule sequencing and chromatin interaction mapping. *Giga-science* 2019;**8**(4), doi:10.1093/gigascience/giz027.
20. Jarvis DE, Ho YS, Lightfoot DJ, et al. The genome of *Chenopodium quinoa*. *Nature* 2017;**542**:307–12.
21. Maccaferri M, Harris NS, Twardziok SO, et al. Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat Genet* 2019;**51**:885–95.
22. Raymond O, Gouzy J, Just J, et al. The Rosa genome provides new insights into the domestication of modern roses. *Nat Genet* 2018;**50**:772–7.
23. Daccord N, Celton JM, Linsmith G, et al. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat Genet* 2017;**49**:1099–106.
24. Zhu T, Wang L, You FM, et al. Sequencing a *Juglans regia* × *J. microcarpa* hybrid yields high-quality genome assemblies of parental species. *Hortic Res* 2019;**6**:55.
25. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinform* 2016;**14**:265–79.
26. Belton JM, McCord RP, Gibcus JH, et al. Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods* 2012;**58**:268–76.
27. Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. *J Exp Bot* 2017;**68**:5419–29.
28. Schmidt MH-W, Vogel A, Denton AK, et al. De novo assembly of a new *Solanum pennellii* accession using nanopore sequencing. *Plant Cell* 2017;**29**:2336–48.
29. Belser C, Istace B, Denis E, et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat Plants* 2018;**4**:879–87.
30. Yasodha R, Vasudeva R, Balakrishnan S, et al. Draft genome of a high value tropical timber tree, teak (*Tectona grandis* L. f): insights into SSR diversity, phylogeny and conservation. *DNA Res* 2018;**25**:409–19.

31. Deschamps S, Zhang Y, Llaca V, et al. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat Commun* 2018;9:4844.
32. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biol* 2018;19:90.
33. Zimin AV, Luo M, Marçais G, et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res* 2017;27:787–92.
34. Huang Y, Xiao L, Zhang Z, et al. The genomes of pecan and Chinese hickory provide insights into *Carya* evolution and nut nutrition. *Gigascience* 2019;8(5), doi:10.1093/gigascience/giz036.
35. Xing Y, Liu Y, Zhang Q, et al. Hybrid de novo genome assembly of Chinese chestnut (*Castanea mollissima*). *Gigascience* 2019;8(9), doi:10.1093/gigascience/giz112.
36. Plomion C, Aury JM, Amselem J, et al. Oak genome reveals facets of long lifespan. *Nat Plants* 2018;4:440–52.
37. Luo M-C, You FM, Li P, et al. Synteny analysis in Rosids with a walnut physical map reveals slow genome evolution in long-lived woody perennials. *BMC Genomics* 2015;16:707.
38. Springer NM, Ying K, Fu Y, et al. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet* 2009;5:e1000734.
39. Marroni F, Pinosio S, Morgante M. Structural variation and genome complexity : Is dispensable really dispensable ? *Curr Opin Plant Biol* 2014;18:31–6.
40. Mishra B, Gupta DK, Pfenninger M, et al. A reference genome of the European beech (*Fagus sylvatica* L.). *Gigascience* 2018;7(6), doi: 10.1093/gigascience/giy063.
41. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics Proteomics Bioinform* 2015;13:278–89.
42. Linsmith G, Rombauts S, Montanari S, et al. Pseudochromosome-length genome assembly of a double haploid “Bartlett” pear (*Pyrus communis* L.). *Gigascience* 2019;8(12), doi:10.1093/gigascience/giz138.
43. Vanechoutte D, Estrada AR, Lin YC, et al. Genome-wide characterization of differential transcript usage in *Arabidopsis thaliana*. *Plant J* 2017;92:1218–31.
44. Clark S, Yu F, Gu L, et al. Expanding alternative splicing identification by integrating multiple sources of transcription data in tomato. *Front Plant Sci* 2019;10:689.
45. Hart AJ, Ginzburg S, Xu MS, et al. EnTAP: bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Mol Ecol Resour* 2020;20(2):591–604.
46. Lucas SJ, Kahraman K, Avşar B, et al. A chromosome-scale genome assembly of European hazel (*Corylus avellana* L.) reveals targets for crop improvement. *bioRxiv* 2019,doi:10.1101/817577;.
47. Sork VL, Fitz-Gibbon ST, Puiu D, et al. First draft assembly and annotation of the genome of a California endemic oak *Quercus lobata* Née (Fagaceae). *G3 (Bethesda)* 2016;6: 3485–95.
48. Jamet E, Santoni V. Editorial for Special Issue: 2017 Plant Proteomics. *Proteomes* 2018;6:28.
49. Costa J, Carrapatoso I, Oliveira MBPP, et al. Walnut allergens: Molecular characterization, detection and clinical relevance. *Clin Exp Allergy* 2014;44:319–41.
50. Aradhya M, Woeste K, Velasco D. Genetic diversity, structure and differentiation in cultivated walnut (*Juglans regia* L.). *Acta Hort* 2010;861:127–32.
51. Ruiz-Garcia L, Lopez-Ortega G, Fuentes Denia A, et al. Identification of a walnut (*Juglans regia* L.) germplasm collection and evaluation of their genetic variability by microsatellite markers. *Span J Agric Res* 2011;9:179–92.
52. Dangi GS, Woeste K, Aradhya MK, et al. Characterization of 14 microsatellite markers for genetic analysis and cultivar identification of walnut. *J Am Soc Hortic Sci* 2005;130:348–54.
53. McGranahan GH, Leslie CA. Walnuts. *Acta Hort* 1991;290:907–18.
54. Bernard A, Lheureux F, Dirlwanger E. Walnut: Past and future of genetic improvement. *Tree Genet Genomes* 2018;14:1741.
55. Gauthier MM, Jacobs DF. Walnut (*Juglans* spp.) ecophysiology in response to environmental stresses and potential acclimation to climate change. *Ann For Sci* 2011;68:1277–90.
56. Workman R, Timp W, Fedak R, et al. High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing. *Protoc Exch* 2018, doi:10.1038/protex.2018.059.
57. Zhang M, Zhang Y, Scheuring CF, et al. Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. *Nat Protoc* 2012;7:467–78.
58. Mayjonade B, Gouzy J, Donnadieu C, et al. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques* 2017;61:203–5.
59. Zimin AV, Marçais G, Puiu D. The MaSuRCA genome assembler. *Bioinformatics* 2013;29:2669–77.
60. Miller JR, Delcher AL, Koren S, et al. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* 2008;24:2818.
61. Lieberman-Aiden E, Van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;326:289.
62. Putnam NH, O’Connell BL, Stites JC, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* 2016;26:342–50.
63. Benson G. Tandem Repeats Finder: A program to analyze DNA sequences. *Nucleic Acids Res* 1999;27:573–80.
64. Marçais G, Delcher AL, Phillippy AM, et al. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* 2018;14:e1005944.
65. Rezvoy C, Charif D, Guéguen L, et al. MareyMap: An R-based tool with graphical interface for estimating recombination rates. *Bioinformatics* 2007;23:2188–9.
66. Gurevich A, Saveliev V, Vyahhi N, et al. QUAST : Quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–5.
67. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011;27:764.
68. Vurture GW, Sedlazeck FJ, Nattestad M, et al. GenomeScope : Fast reference-free genome profiling from short reads. *Bioinformatics* 2017;33:2202–4.
69. PacBio IsoSeq GitHub Repository <https://github.com/PacificBiosciences/IsoSeq>. Accessed 11 May 2020
70. Li H. Minimap2 : Pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–100.
71. Smit A, Hubley R. RepeatModeler Open-1.0. 2008. <http://www.repeatmasker.org>.
72. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013. <http://www.repeatmasker.org>.

73. Haas BJ, Delcher AL, Mount SM, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 2003;**31**:5654–66.
74. Kent WJ. BLAT—The BLAST-Like Alignment Tool. *Genome Res* 2002;**12**:656–64.
75. Wu TD, Watanabe CK. GMAP : A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 2005;**21**:1859–75.
76. Haas BJ, Papanicolaou A, Yassour M, et al. De novo transcript sequence reconstruction from RNA-Seq: Reference generation and analysis with Trinity. *Nat Protoc* 2013;**8**, doi:10.1038/nprot.2013.084.
77. GeSeq - Annotation of Organellar Genomes. <https://chlo-robox.mpimp-golm.mpg.de/geseq.html>. Accessed 11 May 2020
78. Tillich M, Lehwark P, Pellizzer T, et al. GeSeq - Versatile and accurate annotation of organelle genomes. *Nucleic Acids Res* 2017;**45**:W6.
79. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;**29**:644.
80. Quevillon E, Silventoinen V, Pillai S, et al. InterProScan: Protein domains identifier. *Nucleic Acids Res* 2005;**33**:W116.
81. Jones P, Binns D, Chang HY, et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 2014;**30**:1236.
82. Simão FA, Waterhouse RM, Ioannidis P, et al. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;**31**:3210–2.
83. Veeckman E, Ruttink T, Vandepoele K. Are we there yet? Reliably estimating the completeness of plant genome sequences. *Plant Cell* 2016;**28**:1759–68.
84. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**:357.
85. Pertea M, Kim D, Pertea GM, et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* 2016;**11**:1650.
86. Faurobert M, Pelpoir E, Chaïb J. Phenol extraction of proteins for proteomic studies of recalcitrant plant tissues. 2007, *Plant Proteomics*, Thiellement H, Zivy M, Damerval C, Méchin V, Humana Press Inc., Totowa, New Jersey, 9–14.
87. ProteinProphet. <http://proteinprophet.sourceforge.net/>. Accessed 11 May 2020
88. Arnold K, Bordoli L, Kopp J, et al. The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. *Bioinformatics* 2006;**22**:195.
89. Konagurthu AS, Whisstock JC, Stuckey PJ, et al. MUSTANG: A multiple structural alignment algorithm. *Proteins Struct Funct Genet* 2006;**64**:559.
90. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754–60.
91. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.
92. Narasimhan V, Danecek P, Scally A, et al. BCFtools/RoH : a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* 2016;**32**:1749–51.
93. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics* 2011;**27**:2156–8.
94. Wang Y, Tang H, Debarry JD, et al. MScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 2012;**40**:1–14.
95. Bink MCAM, Jansen J, Madduri M, et al. Bayesian QTL analyses using pedigreed families of an outcrossing species, with application to fruit firmness in apple. *Theor Appl Genet* 2014;**127**:1073–90.
96. Voorrips RE, Bink MCAM, Kruisselbrink JW, et al. PediHaplotyper: software for consistent assignment of marker haplotypes in pedigrees. *Mol Breed* 2016;**36**:119.
97. Vanderzande S, Howard NP, Cai L, et al. High-quality, genome-wide SNP genotypic data for pedigreed germplasm of the diploid outbreeding species apple, peach, and sweet cherry through a common workflow. *PLoS One* 2019;**14**:e0210928.
98. Zheng X, Levine D, Shen J, et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 2012;**28**:3326–8.
99. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989;**123**:585–95.
100. Alexa A, Rahnenfuhrer J. Gene set enrichment analysis with topGO. <https://bioconductor.riken.jp/packages/3.7/bioc/vignettes/topGO/inst/doc/topGO.pdf>. Accessed 11 May 2020.
101. BioProject. <https://www.ncbi.nlm.nih.gov/bioproject/>. Accessed 11 May 2020
102. *Juglans regia* - Reference Genome V 2.0. <https://hardwoodgenomics.org/Genome-assembly/2539069>. Accessed 11 May 2020
103. Marrano A, Britton M, Zaini PA, et al. Supporting data for “High-quality chromosome-scale assembly of the walnut (*Juglans regia* L.) reference genome.” *GigaScience Database* 2020. <http://dx.doi.org/10.5524/100735>.

High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome

--Manuscript Draft--

Manuscript Number:	
Full Title:	High-quality chromosome-scale assembly of the walnut (<i>Juglans regia</i> L) reference genome
Article Type:	Data Note
Funding Information:	
Abstract:	Background The release of the first reference genome of walnut (<i>Juglans regia</i> L.) enabled many achievements in the characterization of walnut genetic and functional variation. However, it is highly fragmented, preventing the integration of genetic, transcriptomic, and proteomic information to fully elucidate walnut biological processes. Findings Here, we report the new chromosome-scale assembly of the walnut reference genome (Chandler v2.0) obtained by combining Oxford Nanopore long-read sequencing with chromosome conformation capture (Hi-C) technology. Relative to the previous reference genome, the new assembly features an 84.4-fold increase in N50 size and the full sequence of all 16 chromosomal pseudomolecules. Using full-length transcripts from single-molecule real-time sequencing, we predicted 40,491 gene models, with a mean gene length higher than the previous gene annotations. Most of the new protein-coding genes (90%) are full-length, which represents a significant improvement compared to Chandler v1.0 (only 48%). We then tested the potential impact of the new chromosome-level genome on different areas of walnut research. By studying the proteome changes occurring during catkin development, we observed that the virtual proteome obtained from Chandler v2.0 presents fewer artifacts than the previous reference genome, enabling the identification of a new potential pollen allergen in walnut. Also, the new chromosome-scale genome facilitates in-depth studies of intraspecies genetic diversity by revealing previously undetected autozygous regions in Chandler, likely resulting from inbreeding, and 195 genomic regions highly differentiated between Western and Eastern walnut cultivars. Conclusion Overall, Chandler v2.0 will serve as a valuable resource to understand and explore walnut biology better.
Corresponding Author:	Annarita Marrano, Ph. D. Davis, CA UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Annarita Marrano, Ph. D.
First Author Secondary Information:	
Order of Authors:	Annarita Marrano, Ph. D. Monica Britton Paulo Adriano Zaini Aleksey V. Zimin Rachael E. Workman Daniela Puiu Luca Bianco

	Erica Adele Di Pierro
	Brian J. Allen
	Sandeep Chakraborty
	Michela Troggio
	Charles A. Leslie
	Winston Timp
	Abhaya Dandekar
	Steven L. Salzberg
	David B. Neale
Order of Authors Secondary Information:	
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist . Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist ?	Yes

Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes
---	------------

High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome

Annarita Marrano¹, amarrano@ucdavis.edu; *corresponding author*

Monica Britton², mtbritton@ucdavis.edu

Paulo A. Zaini¹, pazaini@ucdavis.edu

Aleksey V. Zimin³, alekseyz@jhu.edu

Rachael E. Workman³, rachael.e.workman@gmail.com

Daniela Puiu⁴, dpuiu@jhu.edu

Luca Bianco⁵, luca.bianco@fmach.edu

Erica Adele Di Pierro⁵, erica.dipierro@fmach.it

Brian J. Allen¹, brallen@ucdavis.edu

Sandeep Chakraborty¹, sanchak@gmail.com

Michela Troglio⁵, michela.troggio@fmach.edu

Charles A. Leslie¹, caleslie@ucdavis.edu

Winston Timp³, wtimp@jhu.edu

Abhaya Dandekar¹, amdandekar@ucdavis.edu

Steven L. Salzberg^{3,4,6}, salzberg@jhu.edu

David B. Neale¹, dbneale@ucdavis.edu

¹ Department of Plant Sciences, University of California, Davis, CA 95616, USA

² Bioinformatics Core Facility, Genome Center, University of California Davis, CA 95616, USA

³ Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205, USA

⁴ Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21205, USA

⁵ Research and Innovation Center, Department of Genomics and Biology of Fruit Crops, Fondazione E Mach, San Michele all' Adige (TN) 38010, Italy

⁶ Departments of Computer Science and Biostatistics, Johns Hopkins University, Baltimore, MD 21218

Abstract

Background: The release of the first reference genome of walnut (*Juglans regia* L.) enabled many achievements in the characterization of walnut genetic and functional variation. However, it is highly fragmented, preventing the integration of genetic, transcriptomic, and proteomic information to fully elucidate walnut biological processes. **Findings:** Here, we report the new chromosome-scale assembly of the walnut reference genome (Chandler v2.0) obtained by combining Oxford Nanopore long-read sequencing with chromosome conformation capture (Hi-C) technology. Relative to the previous reference genome, the new assembly features an 84.4-fold increase in N50 size and the full sequence of all 16 chromosomal pseudomolecules. Using full-length transcripts from single-molecule real-time sequencing, we predicted 40,491 gene models, with a mean gene length higher than the previous gene annotations. Most of the new protein-coding genes (90%) are full-length, which represents a significant improvement compared to Chandler v1.0 (only 48%). We then tested the potential impact of the new chromosome-level genome on different areas of walnut research. By studying the proteome changes occurring during catkin development, we observed that the virtual proteome obtained from Chandler v2.0 presents fewer artifacts than the previous reference genome, enabling the identification of a new potential pollen allergen in walnut. Also, the new chromosome-scale genome facilitates in-depth studies of intraspecies genetic diversity by revealing previously undetected autozygous regions in Chandler, likely resulting from inbreeding, and 195 genomic regions highly differentiated between Western and Eastern walnut cultivars. **Conclusion:** Overall, Chandler v2.0 will serve as a valuable resource to understand and explore walnut biology better.

Keywords: Nanopore, Hi-C, IsoSeq, gene prediction, genetic diversity, proteome, allergens.

53

54 **Introduction**

55 Persian walnut (*Juglans regia* L.) is among the top three most-consumed nuts in the world, and
56 over the last ten years, its global production increased by 37% (International Nut and Dried Fruit
57 Council, 2019). Its richness in alpha-linolenic acid (ALA), proteins, minerals, and vitamins, along
58 with documented benefits for human health, explains this increased interest in walnut consumption
59 [1]. As suggested by its generic name *Juglans* from the Latin appellation ‘*Jovis glans*’, which
60 loosely means ‘nut of gods’, the culinary and medical value of Persian walnut was already widely
61 prized by ancient civilizations [2].

62 The origin and evolution of the Persian walnut are the results of a complex interplay between
63 hybridization, human migration, and biogeographical forces [3]. A recent phylogenomic analysis
64 revealed that Persian walnut (and its landrace *J. sigillata*) arose from an ancient hybridization
65 between American black walnuts and Asian butternuts during the late Pliocene (3.45 Mya) [4].
66 Evidence suggests that the mountains of Central Asia were the cradle of domestication of Persian
67 walnut [5], from where it spread to the rest of Asia, the Balkans, Europe, and, finally, the Americas.

68 Today, walnut is cultivated worldwide in an area of 1,587,566 ha, mostly in China and the USA
69 (FAOSTAT statistics, 2017). Considerable phenotypic and genetic variability can be observed in
70 this wide distribution area, especially in the Eastern countries, where walnuts can still be found in
71 wild fruit forests. Many studies on genetic diversity in walnut have outlined a genetic
72 differentiation between Eastern and Western genotypes [6,7]. Moreover, walnuts from Eastern
73 Europe, Central Asia, and China exhibit higher genetic diversity and a higher number of rare alleles
74 than the genotypes from Western countries [8].

The release of the first reference genome, Chandler v1.0 [9], enabled the study of walnut genetics at a genome-wide scale. For the first time, it was possible to explore the gene space of Persian walnut with the prediction of 32,498 gene models, providing the basis to untangle complex phenotypic pathways, such as those responsible for the synthesis of phenolic compounds. The availability of a reference genome marked the beginning of a genomics phase in Persian walnut, allowing whole-genome resequencing [4,10], the development of high-density genotyping tools [7,11], and the genetic dissection of important agronomical traits in walnut [12–14]. However, the Chandler v1.0 assembly is highly fragmented, compromising the accuracy of gene prediction and the fulfillment of advanced genomics studies necessary to resolve many, still unanswered questions in walnut research.

The recent introduction of long-read sequencing technologies and long-range scaffolding methods has enabled chromosome-scale assembly for multiple plant species, including highly heterozygous tree crops such as almond (*Prunus dulcis*; [15] and kiwifruit (*Actinidia eriantha*; [16]. The availability of genomes with fully assembled chromosomes provides foundations for understanding plant domestication and evolution [15,17,18], the mechanisms governing important traits (e.g., flower color and scent; [19], as well as the impact of epigenetic modifications on phenotypic variability [20]. Recently, Zhu et al., (2019) assembled the parental genomes of a hybrid *J. microcarpa* × *J. regia* (cv. Serr) at the chromosome-scale using long-read PacBio sequencing and optical mapping. They relied on the haplotype divergence between the two *Juglans* species and demonstrated an ongoing asymmetric fractionation of the two subgenomes present in *Juglans* genomes.

Here we report a new chromosome-level assembly of the walnut reference genome with unprecedented contiguity, Chandler v2.0, which we obtained by combining Oxford Nanopore

long-read sequencing [22] with chromosome conformation capture (Hi-C) technology [23]. Thanks to the increased contiguity of Chandler v2.0, we were able to substantially improve gene prediction accuracy, with new, longer gene models identified and many fewer artifacts compared to Chandler v1.0. Also, the availability of full, chromosomal sequences reveals new genetic diversity of Chandler, previously inaccessible through standard genotyping tools, and significant genetic differentiation between Western and Eastern walnuts at 195 genomic regions, including also loci involved in nut shape and harvest date. In the present research, we demonstrate the fundamental role of a chromosome-scale reference genome to integrate transcriptomics, population genetics, and proteomics, which in turn enable a better understanding of walnut biology.

Genome long-read sequencing and assembly

To increase the contiguity of the Chandler genome, we first generated deep sequence coverage using Oxford Nanopore Technology (ONT), a cost-effective long-read sequencing approach that determines DNA bases by measuring the changes in electrical conductivity generated while DNA fragments pass a tiny biological pore [24]. Since the release of the first plant genome assembly generated using ONT sequencing [25], this technology has been applied to sequence and obtain chromosome-scale genomes of many other plant species [26–28]. In Persian walnut, ONT sequencing yielded 7,096,311 reads that provided 21.9 Gbp of sequence, or ~35X genome coverage (assuming a genome size of 620 Mb). Read lengths averaged 3.1 kb, and the N50 read length was 6.7 kb, with the longest read being 992.2 kb.

One of the major limitations of long-read sequencing technologies is their high error rate, which can range between 5% and 15% for Nanopore sequencing [29]. To overcome this limitation, we adopted the hybrid assembly technique incorporated into the MaSuRCA assembler, which

combines long, high-error reads with shorter but much more accurate Illumina sequencing reads to generate a robust, highly contiguous genome assembly [30]. First, using the Illumina reads, we created 3.7 million 'super-reads' with a total length of 2.9 Gb. We then combined the super-reads with the ONT reads to generate 3.2 million mega-reads with a mean length of 4.7 kb, representing 24X genome coverage (**Additional file 1**). Finally, we assembled the mega-reads to obtain the 'hybrid' Illumina-ONT assembly, which comprised 1,498 scaffolds, 258 contigs, and 25,007 old scaffolds from Chandler v1.0 (**Table 1; Additional file 2**).

Even though the total number of scaffolds (> 1 Kb) was reduced by 80% compared to Chandler v1.0 (**Table 1**), the new hybrid assembly was still fragmented. To improve the assembly further and build chromosome-scale scaffolds, we applied Hi-C sequencing, which is based on proximity ligation of DNA fragments in their natural conformation within the nucleus [23]. The HiRise scaffolding pipeline processed 356 million paired-end 100-bp Illumina reads to generate the HiRise assembly, which contained 2,656 scaffolds longer than 1 kb (**Table 1**). The top 17 scaffolds from this assembly spanned more than 90% of the total assembly length, with a scaffold length ranging from 19.6 to 45.2 Mb (**Additional files 3-4**). As compared to the previous (1.0) assembly, the Chandler genome contiguity increased dramatically, with an N50 size 98% higher than Chandler v1.0 and only 0.04% of the genome in gaps.

Validation of the HiRise assembly

To assess the quality of the HiRise assembly, we used two independent sources of data. First, we used the single nucleotide polymorphism (SNP) markers mapped on the high-density genetic map of Chandler recently described by [14]. Out of the 8,080 SNPs mapped into 16 linkage groups (LGs), 6,894 had probes aligning uniquely on the HiRise assembly with 98% of identity for more than 95% of their length. A total of 35 scaffolds of the HiRise assembly could be anchored to a

chromosomal linkage group by at least one SNP (**Figure 1**). In particular, 13 LGs were spanned by a single HiRise scaffold, while two to three scaffolds each aligned the remaining three LGs.

Second, we anchored the HiRise assembly to the Chandler genetic map used by [31] to construct a walnut physical map. In total, 972 of the mapped markers (1,525 SNPs) aligned uniquely on the same 35 HiRise scaffolds anchored to the linkage map mentioned above. Overall, we observed almost perfect collinearity between the HiRise assembly and both Chandler genetic maps (**Figure 1, Additional file 5**). Therefore, we oriented, ordered, and named the HiRise scaffolds consistent with the linkage map of [31], generating the final 16 chromosomal pseudomolecules of *J. regia* Chandler.

These 16 contiguous chromosomal scaffolds account for 95% of the final walnut reference genome v2.0, with an N50 size of 35 Mb. Chandler v2.0 has a total length of 576,258,700 bps, of which only 20.9 Mb are fragmented in 2,631 small unanchored scaffolds (> 1 kb; **Table 1**). The larger genome size of Chandler v2.0 compared to the recently published genome assembly of the cv. Serr (JrSerr_v1.0; 534.7 Mb) [21] can be explained by structural variation (e.g., copy number and presence/absence variants), whose central role in explaining intraspecific genomic and phenotypic diversity has been reported in different plant species [32,33]. In addition, the higher number of unanchored scaffolds in Chandler v2.0 compared to JrSerr_v1.0 can represent autozygous genomic regions of Chandler, devoid of segregating markers and, therefore, difficult to anchor to linkage genetic maps [31], as also suggested by the higher fixation index (F) observed in Chandler (0.03) than Serr (-0.29) in previous genetic surveys [7].

We identified telomere sequences at both ends for nine of the chromosome scaffolds, on one end of the other seven chromosomes and one end of seven unanchored scaffolds. Also, all 16

chromosomes had centromeric repeats in the middle, alongside regions with low recombination rates (**Figure 2**).

To assess the sequence accuracy of Chandler v2.0, we first compared the scaffold sequences of Chandler v2.0 with the previous version of the walnut reference genome, generating 838,173 alignments with sequence identity averaging 94.11%. We then mapped the Illumina whole-genome shotgun data (Martínez-García et al., 2016) against the new chromosome-scale genome. The alignment resulted in 64,950,691,681 bps mapped, of which 407,450,406 were single-base mismatches, consistent with an Illumina sequence accuracy rate of 99.5%.

Repeat annotation

Almost half (49.68%) of the new Chandler v2.0 is repetitive, similarly to the previous version of the walnut reference genome (51.19%). As in most plant genomes, interspersed repeats (45.13%) were the most abundant type of repeats, with retrotransposons at 18.55% and DNA transposons at 3.05%. *Gypsies* (6.58%) and *Copias* (4.1%) were the most represented classes of long-terminal retrotransposons (LTR), and, though widely dispersed throughout the genome, they were distributed differently along the 16 chromosomes (**Additional file 6**): the *Gypsies* LTRs were more abundant alongside the centromeres, where, instead, the density of the *Copia* LTRs decreased, consistent with [21]. L1/LINE (long-interspersed nuclear elements), which possess a poly(A) tail and two open reading frames (ORFs) for autonomous retrotransposition, was the largest class of non-LTRs at 6.98% of the genome. Simple repeats (4.29%) and low-complexity regions (0.26%) were also found.

PacBio IsoSeq sequencing and gene annotation

A fragmented reference genome can severely hamper the accuracy of gene prediction, because many genes will be broken across multiple small contigs (false negatives), and because multiple fragments of the same gene may be annotated separately (false positives). Also, transcriptome assemblies generated using second-generation (Illumina) sequencing data are likely to miss many transcripts due to the very short read lengths [34].

To improve the gene prediction accuracy of Chandler v2.0, we used the “Isoform Sequencing” (Iso-Seq) method, developed by Pacific Biosciences (PacBio), which can generate full-length transcripts up to 10 kb, allowing for accurate determination of exon-intron structure by the alignment of the transcripts to the assembly [35]. The high error rate of PacBio sequencing can be greatly reduced using circular consensus sequence (CCS), in which a transcript is circularized and then sequenced repeatedly to self-correct the errors. We applied PacBio IsoSeq to sequence full-length transcripts from nine tissues, chosen to cover most of the transcript diversity in walnut (**Additional file 7**). Across the four SMRT cells, we obtained 26,328,087 subreads with a mean length of 1,188 bp (**Additional file 8**) and CCSs ranging from 13K to 142K per library (**Additional file 9**). Out of the 745,730 full-length non-chimeric (FLnc) transcripts, 68,225 were classified as high quality, FL (HQ FL) consensus transcript sequences, with an average length of 1,357 bp (**Additional file 9**). Catkin 1-inch elongated (CAT1), shoot, and root yielded the lowest number of HQ FL transcripts, while pollen and leaf had the lowest number of HQ consensus clusters obtained per CCS after polishing (**Additional file 9**). These results can be explained by lower cDNA quality or fewer inserts of full-length transcripts from these tissues during the cDNA pooling and library preparation. Nevertheless, more than 99% of the HQ FL transcripts aligned onto the new chromosomal-level walnut reference genome (**Additional file 10**).

209 By combining the HQ FL transcripts with available *Juglans* transcriptome sequences, we identified
210 40,491 gene models, which are more than those annotated in Chandler v1.0 but fewer than the
211 predicted genes in the NCBI RefSeq *J. regia* annotation generated with the first version of the
212 reference genome (**Table 2**). This result suggests that the new chromosome-scale genome, along
213 with the availability of full-length transcripts, allowed us to identify genes missed during the
214 annotation of Chandler v1.0, as well as to remove false-positive predictions. Also, the mean gene
215 length in Chandler v2.0 was higher than the previous gene annotations (**Table 2**), a consequence
216 of the increased contiguity of the new chromosome-scale reference genome.

217 The average gene density of Chandler v2.0 was 19.28 genes per 100 kb, with higher gene content
218 in the proximity of telomeric regions (**Figure 2**), consistent with other plant genomes [18,36]. In
219 addition, 92% (37,102) of the predicted gene models of Chandler v2.0 was supported by expression
220 data, and 97% showed high similarity with a protein-coding transcript from either the *J. regia*
221 RefSeq v1 gene set or a protein from the wider NCBI RefSeq plant database (**Additional file 11**),
222 highlighting the accuracy and robustness of the Chandler v2.0 genome annotation. Overall,
223 Chandler v2.0 contained 339 newly predicted gene models, with a mean length of 1,533 bp. Of
224 these new predicted gene models, 150 (44%) and 329 (97%) were supported by PacBio IsoSeq
225 and Illumina RNA-seq data (Martínez-García et al., 2016), respectively. Thus, the failure to
226 identify these genes in Chandler v1.0 was most likely related to its low contiguity than to the lack
227 of the gene transcripts in the RNA-seq data.

228 Out of the 41,081 transcripts identified, 84% were multi-exonic, with, on average, 5.94 exons each
229 and a mean exon length of 257.2 bp (**Table 2**). The mean number of introns per gene was 5.9, with
230 a length ranging from 20 bp to 493 kb. These values are similar to those observed in the previous
231 gene annotations of Chandler, except for the number of exons and introns which was higher in

Chandler v2.0. Also, introns were longer, on average, contributing to the higher mean gene length observed in Chandler v2.0. The majority of intron/exon junctions were GT/AG-motif (98.2%), even though alternative splicing with non-canonical motifs was also observed (GC/AG – 0.8%; AT/AC – 0.11%). Almost 90% (36,438) of the coding sequences were full-length with canonical start and stop codons, while 4,525 presented either a start or a stop codon. This result represents a great improvement compared to Chandler v1.0, where only 48% of the predicted gene models were complete [9].

Also, we observed that 568 gene models had from two to four transcript isoforms each, with a mean length of 7,080 bp. Out of the 1,158 isoforms identified, 339 were covered by FL HQ transcripts in at least one tissue, while 835 were expressed in at least one of the 20 tissues (Martínez-García et al., 2016), which most likely covered higher gene diversity compared to the nine tissues used for PacBio IsoSeq. On average, the Illumina isoforms (7,220 bp) were longer than the PacBio isoforms (6,044 bp). By running the EnTAP functional annotation pipeline with the entire NCBI RefSeq plant database [37], we observed that 672 isoforms were annotated with a plant protein, while the remaining 486 transcript isoforms were not identified in the previous walnut gene annotation.

Of the 41,103 gene models, 83% were annotated with a plant protein, and 84% had a known Pfam domain. Also, 33,034 models were annotated with 8,244 different Gene Ontology (GO) terms. The three most common biological processes were regulation of transcription (2%), defense response (1.43%), and DNA recombination (1.2%; **Additional file 12**). ATP binding (7.7%), metal ion binding (7.1%) and DNA-binding transcription factor activity (3.7%; **Additional file 13**) were the most abundant molecular functions, while nucleus (13%), integral component of membrane (10.3%) and plasma membrane (8%) were the top three cellular components (**Additional file 14**).

The majority (95%) of the 1,440 core genes in the embryophyte dataset from Benchmarking Universal Single-Copy Orthologs (BUSCO) were identified in both the new Chandler genome assembly and gene space v2.0. Also, 88% of both rosids and green sets of core gene families (coreGFs) were identified in the gene annotation, confirming the high-quality and completeness of the gene space of Chandler v2.0.

Improved assessment of proteomes with the complete genome sequence

After confirming the importance of a chromosome-scale reference genome for the improvement of gene prediction accuracy, we studied the impact of a contiguous genome on proteomic analysis. A virtual proteome, which includes all protein sequences predicted from a reference genome, is generally used to map and assign the peptides detected in mass spectrometry (MS) to specific protein-coding genes. Therefore, a fragmented assembly of the reference genome can lead to an inaccurate prediction of a species' proteome and, then, a miss-identification of the proteins expressed in specific tissues at particular stages [38].

We analyzed the proteomic data generated from samples encompassing different developing stages of the male walnut flower (catkin) and pure pollen, by using the virtual proteomes predicted from the gene annotation of the new chromosome-scale genome and Chandler v1.0 (NCBI RefSeq). Considering all tissues analyzed, we identified fewer unique peptides (43,083) with the new chromosome-scale walnut genome than with Chandler v1.0 (44,679). Also, 6,966 unique proteins were detected with Chandler v2.0 against the 8,802 found using version 1 as a search database (**Additional file 15-16**). Most likely, the NCBI proteomic database based on the fragmented Chandler v1.0 included artifacts resulting from an overestimation of the protein-coding genes. Therefore, the new chromosome-scale genome allows accurate estimation of the proteomic

277 changes occurring in the different vegetative and reproductive stages of walnut, which is
 278 fundamental to fully understand the molecular bases of the observed phenotypic traits.

279 Sample clustering according to their protein constituents and levels showed greater similarity
 280 between immature and mature catkins and a more distinct profile between senescent catkins and
 281 pure pollen (**Figure 3**).

282 Given that ~2% of walnut consumers have high risk of developing allergies to nuts or pollen [39],
 283 we searched the four developed proteomes for allergenic proteins listed in the WHO/IUIS Allergen
 284 Database (www.allergen.org; **Additional file 17**), and additional proteins not yet registered in the
 285 allergen database but predicted in Chandler v2 as potential allergens (**Additional file 16**). Four of
 286 the eight recognized allergenic proteins were detected in at least one of the catkin developmental
 287 stages, with Jug_r_5 (XP_018825777 | *Jr12_10750*) and Jug_r_7 (XP_018808763 | *Jr07_28960*)
 288 present in all sample types, including pollen (**Additional file 17**). Three of the new potential
 289 allergens (**Additional file 17**) are encoded by genes adjacent to known allergen-coding sequences,
 290 likely indicating gene duplications. Also, we discovered that the gene locus *Jr12_05180* encodes
 291 a non-specific lipid transfer protein (nsLTP; Jug_r_9 | XP_018813928), a potential allergen highly
 292 expressed during catkin maturation and in pollen (**Additional files 17-18**). In particular, Jug_r_9
 293 was the most abundant protein in mature and senescent catkins, and the second-most abundant in
 294 pure pollen (**Additional files 17-18**). Another interesting allergen similar to Jug_r_9 (same eight
 295 cysteine configuration) is XP_018814382 | *Jr03_26970*; it decreases as the catkin matures, and is
 296 entirely absent in pollen (**Additional files 17-18**). Similarly, polyphenol oxidase (PPO,
 297 XP_018858848 | *Jr03_06780*) is high in the immature catkin and almost absent in the pollen. The
 298 integration of this proteomic data with previously published transcriptomic data obtained from 20
 299 walnut tissues [9] shows high reproducibility between the methods. In both datasets, allergens

Jug_r_1, 4, and 6 were not detected in catkins, while the new putative allergen Jug_r_9 was highly expressed in catkins (**Additional files 18-19**). Also, *Jr12_05180* transcripts were not detected in any of the 20 tissues but catkin, thus confirming the strong specificity of Jug_r_9 for catkin and pollen tissue (**Additional files 19**). Modeling the structure of this putative allergen reveals four predicted disulfide bonds, potentially conferring heat and protease-resistance, and further suggesting allergenic properties (**Figure 4**). Future studies will clarify the functional role of this protein and its allergenic nature.

The detection of new potential walnut allergens confirms the positive impact of Chandler v2.0 on proteomic studies in walnut, by providing a clearer and more precise organization of the CDSs within a genomic vicinity than the previous fragmented genome assembly v1.0.

Chandler genomic diversity

By anchoring the HiRise assembly to the Chandler genetic map [14], we observed highly homozygous regions in Chandler, especially on Chr15, where the genetic gap spanned 14.5 cM, corresponding to a physical distance of 9.1 Mb. A large gap on Chr15 (9.23 cM – 1.5 Mb) was also observed by [31], which suggested inbreeding as a possible cause for the lack of segregating loci in this region in Chandler, whose parents shared Payne as an ancestor. To confirm the autozygosity of Chandler on Chr15, we used the Illumina whole-genome shotgun data of Chandler and the identified polymorphisms to study its genetic diversity across the new chromosome-scale genome. We identified 2,205,835 single heterozygous polymorphisms on the 16 chromosomal pseudomolecules, with an SNP density of 4.0 SNPs per kb (**Figure 2; Additional file 20**). Fifty-six 1-Mb-regions exhibited less than 377.5 SNPs (10th percentile of the genome-wide SNP number distribution), and chromosomes 15, 1, 7, and 13 were the top four chromosomes in the number of low heterozygous regions (**Additional file 21**). In particular, Chr15 presented nine 1-Mb windows

with a significantly low number of polymorphisms, five of which span 4 Mb at the end of the chromosome. In these nine low heterozygous regions, we found 1,536 SNPs in total (**Figure 2**), of which only 25 were tiled on the Axiom *J. regia* 700K SNPs array. The absence of these polymorphisms segregating in Chandler in the SNP array could be related to either a failed identification during the SNP calling due to the highly fragmented reference genome v1.0 or with the SNP exclusion during the filtering process applied to build the genotyping array [7]. The low number of Chandler heterozygous SNPs in the array affected the end of Chr15 the most, causing a reduction in the genetic length of the corresponding linkage group (**Figure 1**), as well as leaving unexplored 4 Mb of Chandler genetic variability, which is now accessible thanks to the new chromosome-scale reference genome. The failure to anchor seven of the scaffolds with telomeric sequences can be explained by the missed detection of terminally located highly homozygous regions during genetic map constructions, due to the absence of crossing-over events with heterozygous flanking markers.

Due to the evidence of whole-genome duplication in *Juglans* genomes [31], we searched for conserved regions of synteny between Chr15 and its homologous regions in the genome, to study their level of divergence and identify other evolutionary forces as possible causes of the localized reduction of heterozygosity on Chr15. Of the 5,739 pairs of paralogous genes (8,701 genes; **Additional file 22**) identified in Chandler v2.0, 448 included genes on Chr15, and 389 of these have their respective paralogues on Chr6 (**Additional file 23**), in line with what was already reported by [31]. The Chr06-Chr15 pairs of paralogous genes showed average values of divergence indexes ($K_S = 0.38$; $K_A = 0.13$) similar to the ones observed genome-wide for other syntelogs ($K_S = 0.4$; $K_A = 0.09$). Similar values of divergence were also observed for the 178 Chr06-Chr15 syntelogs (171 genes) falling within the nine low heterozygous regions on Chr15 ($K_S = 0.4$, $K_A =$

0.1), excluding different evolutionary rates or positive selection for these regions, and leaving inbreeding as the most reasonable explanation. Other than paralogous genes, we found 393 singletons genes in the low heterozygous regions on Chr15 of Chandler. These genes are involved in different biological processes, many of which related to signal transduction, protein phosphorylation, and response to environmental stimuli (**Additional file 24**).

We further investigated the contribution of inbreeding to the high level of autozygosity on Chr 15 by visualizing the inheritance of haplotype-blocks (HB; genomic regions with little recombination) across the Chandler pedigree (**Figure 5B, Additional file 25**). We observed that Payne accounts for the entire Chandler genetic makeup (19 HBs for the total length of Chr15) inherited from Pedro (mother), where only one HB (2,08 Mb) shared the same allele of Conway-Mayette (maternal-grandfather; **Figure 5A**). Regarding the paternal genetic makeup of Chandler, 13 out of 19 HBs (9,05 Mb) on Chr15 inherited Payne alleles, providing further evidence of high inbreeding on this chromosome (**Figure 5A**). This is even more evident in assessing the number of alleles matching between Payne and Chandler across the genome: Chr15 (14 HBs for a total of 13,95 Mb; **Figure 6**) shares full allele identity with Payne for almost its entire length. Such allele matching between Chandler and its ancestor Payne also occurs on Chr1 (9 HBs for a total of 8,44 Mb), Chr4 (6 HBs - 7,68 Mb), Chr7 (21 HBs - 21,62Mb) and Chr14 (7 HBs – 12,29 Mb). These results confirm a high level of inbreeding in many genomic regions of Chandler (**Additional file 25**) and support the crucial role of Chandler v2.0 for understanding trait genetic inheritance in walnut.

Genomic comparison between Eastern and Western walnuts

Even though numerous surveys regarding genetic diversity within walnut germplasm collections have been reported so far [40,41], comparative analyses at the population level and genome scans for signatures of selection are still missing in Persian walnut. The availability of a chromosome-

scale reference genome enables exploration of the patterns of intraspecific variation at the genomic level, providing new insight on the extraordinary phenotypic diversity present within *J. regia*.

We used the resequencing data generated for 23 founders of the Walnut Improvement Program of the University of California, Davis (UCD-WIP; **Additional file 26**) [10] to study the genome-wide genetic differentiation among walnut genotypes of different geographical provenance. We identified 14,988,422 SNPs, and over 97% of them were distributed on the 16 chromosomal pseudomolecules, with 9.4 polymorphisms per kb. A hierarchical clustering analysis (**Additional file 27**) divided the 23 founders into two major groups, including genotypes from western countries (USA, France, and Bulgaria) and Asia (China, Japan, Afghanistan), respectively, as previously reported [7,42]. High phenotypic diversity for many traits of interest in walnut, such as phenology, nut quality, and yield, has been observed within and between germplasm collections from Western and Eastern countries [43]. Walnut trees from Asia are noted for their lateral fruitfulness and precocity, rarely observed in the USA and western Europe, so that they have been used as a source of these phenotypes in different walnut breeding programs [44].

At a genomic level, we found a moderate differentiation ($F_{ST} = 0.15$) between Western and Eastern genotypes, except for 195 genomic windows (100 kb) that showed substantially high population differences ($F_{ST} \geq 0.36$; top 5% in the whole genome). In particular, chromosomes 7, 5, 1, 4, and 2 presented about 70% of the divergent sites (**Figure 2; Additional file 28**). As suggested by the mean reduction of diversity coefficient (ROD) value (0.41), in most of the genomic regions highly differentiated, the UCD-WIP founders from the USA and Europe showed lower nucleotide diversity ($\pi = 2.5 \times 10^{-4}$) than the Asian genotypes ($\pi = 5.0 \times 10^{-4}$), consistent with [8] (**Figure 2; Additional file 28**). The proximity of our eastern genotypes to the supposed walnut center of domestication in Central Asia can explain the high level of diversity observed in this subgroup.

392 More than 60% (122) of the highly differentiated windows showed a negative value of Tajima's
 393 D in the EU/USA subgroup ($D_{Occ} = -1.12$), thus, suggesting that selection has been likely acting
 394 on these genomic regions in the Western genotypes (**Additional file 28**). Here we found 743 genes,
 395 with GO biological categories mostly related to signal transduction, embryo development, and
 396 response to stresses (**Additional file 29**). Ten candidate selective sweeps ($D_{Asia} = -0.54$) were also
 397 observed in the Eastern group (**Additional file 28**), which included 57 predicted genes, related to
 398 terpenoid biosynthesis, post-embryonic development, and signal transduction (**Additional file 30**).

399 Recently, many marker-trait associations have been reported for different traits of interest in
 400 walnut, such as leafing date, nut-related phenotypes, and water use efficiency [12–14]. We looked
 401 to see if any of these trait-associated SNPs fell within regions highly differentiated between
 402 Western and Eastern genotypes. Three loci associated with shape index, nut roundness, and nut
 403 shape [12] are located in two genomic regions on chromosome 3 and 4 with significantly high
 404 values of F_{ST} (**Additional file 31**). In both of these regions, Western genotypes presented lower
 405 genetic diversity and lower values of Tajima's D than the Eastern walnuts. These findings may
 406 suggest that, while a selective pressure for nut shape may have occurred in the EU/USA subgroups,
 407 higher phenotypic variability can be expected for these traits in the Eastern countries. We also
 408 found that the locus AX-170770379, strongly associated with harvesting date [14], falls within a
 409 genomic region on Chr1 with an F_{ST} value equal to 0.39 and lower genetic diversity in the western
 410 genotypes ($ROD = 0.63$; **Additional file 31**). Looking at the phenotypic effect of this SNP on the
 411 harvest date of the 23 founders, we observed that most of the western genotypes are later harvesting
 412 than the eastern (**Additional file 32**), suggesting differences in the timing of phenological events
 413 between these two groups as adaptation to the different climate conditions present in their countries
 414 of origin [45].

These results confirm the central role of a chromosome-scale genome assembly for population genetics studies, which are fundamental to study how the environment and human selection impacted walnut biology. Future resequencing projects involving larger walnut collections and covering a wider area of the global walnut distribution are necessary to confirm and interpret the observed genomic differentiation between Western and Eastern walnuts, likely helping to understand the role of this genomic divergence in the evolutionary history of Persian walnut.

Methods

Oxford Nanopore sequencing and assembly

High molecular weight (HMW) DNA for Nanopore sequencing (Oxford Nanopore Technologies Inc., UK) was isolated through a nuclei extraction and lysis protocol. First, mature leaf tissue from the same tree used for the original *J. regia* genome [9] was homogenized with mortar and pestle in liquid nitrogen until well ground, then added to the Nuclei Isolation Buffer [46], and stirred at 4°C for 10 minutes. The cellular homogenate was filtered through 5 layers of Miracloth (Millipore-Sigma) into a 50 mL Falcon tube, then centrifuged at 4°C for 20 minutes at 3000 x g. This speed of centrifugation was selected based on the estimated walnut genome size of 1 Gb [47]. Extracted nuclei were then lysed for 30 minutes at 65°C in the SDS-based lysis buffer described by [48]. Potassium acetate was added to the lysate to precipitate residual polysaccharides and proteins. The sample was incubated for 5 minutes at 4°C and then centrifuged at 4°C for 10 minutes at 2400 x g. After removing the supernatant, genomic DNA (gDNA) was ethanol precipitated, and then eluted in 10 mM Tris-Cl. Further purification of the gDNA was then performed using a Zymo Genomic DNA Clean and Concentrate column.

One µg of the isolated gDNA was prepared for sequencing using the Ligation sequencing kit (LSK108, Oxford Nanopore) following manufacturer's protocol with an optimized end repair (100 µl sample, 14 µl enzyme, 6 µl enzyme, incubated at 20°C for 20 minutes then 65°C for 20 minutes). Libraries were sequenced for 48 hours on the Oxford Nanopore Mk1B MinION platform with the R9.4 chemistry on eight flowcells. Raw fast5 data was base-called using Albacore version 1.25.

The ONT data and Illumina reads from [9] were combined using the assembly algorithm implemented in MaSuRCA v3.2.2 [49]. Super-reads were constructed using a k-mer size of 41 bp. De-duplicated scaffolds were aligned onto the previously finished *J. regia* chloroplast genome [9] using "minimap2 -x asm5", as well as to a database of 223 finished plant mitochondria (downloaded from NCBI RefSeq) using blastn with default parameters.

Hi-C sequencing

A Hi-C library was prepared by Dovetail Genomics LLC (Santa Cruz, CA, USA) as described previously [50]. Briefly, for each library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Biotin that was not internal to ligated fragments was removed from the purified DNA. Purified DNA was then sheared to ~350 bp mean fragment size. Sequencing libraries were generated using NEBNext® Ultra™ enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were then sequenced on the Illumina HiSeq4000 platform. The hybrid ONT assembly, Illumina shotgun reads [9], and Dovetail Hi-C library reads were used as input data for the scaffolding software HiRise, which uses proximity ligation data to scaffold

genome assemblies [51]. Shotgun and Dovetail Hi-C library sequences were aligned to the hybrid ONT assembly using a modified SNAP read mapper. The separations of Dovetail Hi-C read pairs mapped within the ONT scaffolds were analyzed by HiRise to produce a likelihood model for the genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After scaffolding, Illumina shotgun sequences were used to close gaps between contigs, resulting in an improved HiRise assembly.

Validation and anchoring of the HiRise assembly to Chandler genetic maps

The HiRise assembly was first anchored to the Chandler genetic map obtained by [14] from a 312 offspring F₁ population ‘Chandler x Idaho’ genotyped with the latest Axiom *J. regia* 700K SNP array. SNP probes (71-mers including the SNP site) from the Axiom *J. regia* 700K SNP array were aligned onto the HiRise assembly filtering out alignments with probe/reference identity lower than 98%, covering less than 95% of the probe length or aligning multiple times on the genome. Retained markers with a unique segregation profile were then used to anchor the HiRise scaffolds. The same procedure was also followed to anchor the HiRise assembly to the Chandler genetic map used to construct a walnut bacterial artificial chromosome (BAC) clone-based physical map by [31]. The final ordering of scaffolds was performed by taking into consideration the marker genetic map position, and, in the final sequence, consecutive scaffolds were separated by sequences of 100,000 Ns.

The tandem repeat finder program (trf v4.09; [52]) was run using the recommended parameters (max mismatch delta PM PI minscore maxperiod, 2 7 7 80 10 50 500 resp.) to identify repeat elements up to 500 bp long. A histogram of repeat unit lengths was generated, and peaks at 7, 29, 33, 44, 154, and 308 bp were identified. From this data, a consensus sequence corresponding to

each peak was selected. All of these repeat sequences were aligned onto the HiRise assembly using ‘nucmer’ from the MUMmer4 package [53] with a minimum match length of 7 to capture the telomeric repeat. Based on the positions of these alignments along the chromosomes and contigs, we identified the 7-mer as the telomeric repeat and the 154-mer and 308-mer as centromeric repeats.

Recombination rate was estimated within sliding windows of 10 Mb with a step of 1 Mb along the chromosome sequence by using the high-density genetic map of Chandler [14] and the R/MareyMap package v 1.3.4 [54]. To evaluate Chandler v2.0 error rate, the two assemblies, Chandler v1.0 and 2.0, were aligned to each other using the ‘nucmer’ program (Marçais et al., 2018).

RNA preparation

Five walnut tissues (leaf, catkin 1-inch elongated; catkin 3-inches elongated, pistillate flower, and pollen) were collected from ‘Chandler’ trees at the UCD walnut orchards. Four additional samples (somatic embryo, callus, shoot, and roots) were taken from tissue culture material of ‘Chandler’. Several grams of each tissue were ground in liquid nitrogen and with insoluble polyvinylpyrrolidone (PVPP; 1% w/w). RNA was isolated using the PureLink™ Plant RNA Reagent (Invitrogen™, Carlsbad, CA) following the manufacturer’s instructions, but with an additional end wash in 1 mL of 75% Ethanol. For root tissue only, RNA isolation was performed using the MagMAX™ mirVana™ Total RNA Isolation Kit (Applied Biosystems™, Foster City, CA) as per protocol, except for the lysis step. A different lysis buffer was created adding 100 mg of sodium metabisulfite to 10 mL of guanidine buffer (guanidine thiocyanate 4M, sodium acetate 0.2M, EDTA 25 mM, PVP-40 2.5%, pH 5.0) and 1 mL of nuclease-free water. Then, 100 mg of ground root tissue were lysed in 1 mL of the new lysis buffer using a Tissue Lyser at max frequency

for 2 min. The lysate was centrifuged at 4° C for 5 min at max speed. The supernatant (500 µL) was transferred to a new tube for the following steps of RNA isolation as per protocol. RNA samples were then purified, and DNase treated using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). The RNA quality was confirmed by running an aliquot of each sample on an Experion™ Automated Electrophoresis System (Bio-Rad, Hercules, CA).

PacBio IsoSeq sequencing

Full-length cDNA Iso-Seq template libraries for PacBio IsoSeq analysis were constructed and sequenced at the DNA Technologies & Expression Analysis Core Facility of the UC Davis Genome Center. FL double-stranded cDNA was generated from total RNA (2 µg per tissue) using the Lexogen Telo™ prime Full-length cDNA Kit (Lexogen, Inc., Greenland, NH, USA). Tissue-specific cDNAs were first barcoded by PCR (16-19 cycles) using IDT barcoded primers (Integrated DNA Technologies, Inc., Coralville, Iowa), and then bead-size selected with AMPure PB beads (two different size fractions of 1X and 0.4X). The nine cDNAs were pooled in equimolar ratios and used to prepare a SMRTbell™ library using the PacBio Template Prep Kit (PacBio, Menlo Park, CA). The SMRTbell™ library was then sequenced across four Sequel v2 SMRT cells with polymerase 2.1 and chemistry 2.1 (P2.1C2.1).

PacBio raw reads were processed using the Isoseq3 v.3.0 workflow following PacBio recommendations. Circular consensus sequences (CCSs) were generated using the program ‘ccs’. The CCSs were demultiplexed and cleaned of cDNA primers using the program ‘lima’. Afterward, CCS clustering and polishing was performed using the program ‘isoseq3’, to generate HQ FL sequences for each of the nine tissues. FLnc and HQ clusters were aligned onto the new ‘Chandler’ assembly v2.0 with minimap2 v.2.12-r827, including the parameter ‘-ax splice’ [55].

Repeat annotation

A genome-specific repeat database was created using the ‘basic’ mode implemented in RepeatModeler v.1.0.11 [56]. RepeatMasker v.4.0.7 was then run to mask repeats in the walnut reference genome v.2.0 and generate a GFF file [57].

Gene prediction and functional annotation

J. regia RefSeq transcripts and additional *J.regia* transcripts and protein sequences downloaded from NCBI, along with the HQ FL IsoSeq transcripts, were used as input to the PASA pipeline v.2.3.3 [58], to assemble a genome-based transcript annotation. PASA utilizes the aligners BLAT v.35 [59] and GMAP v.2018-07-04 [60], along with TransDecoder v.5.5.0 [61], which predicts open reading frames (ORFs) as genome-based GFF coordinates. The final PASA/TransDecoder GFF3 file was post-processed to name the genes and transcripts by chromosome location consistently. Functional roles were assigned to predicted peptides using Trinotate v.3.1.1 [62]. In particular, similarity searches were performed against several public databases (i.e., Uniprot/Swiss-Prot, NCBI NR, *Vitis_vinifera.IGGP_12x*, *J. regia* RefSeq) using BLAST v.2.8.1, HMMER v.3.1b2, SignalP v.4.1c, and TMHMM v.2.0c.

The completeness and quality of both genome assembly and gene annotation of Chandler v.2.0 were estimated with the BUSCO method v.3 (1,440 core genes in the embryophyte dataset) [63], and the sets of coreGFs of green plants (2,928 coreGFs) and rosids (6,092 coreGFs) from PLAZA v.2.5 [64]. Also, RNA-Seq data generated for 20 tissues (see Martínez-García et al., 2016) were aligned to the reference genome (v1 and v2) with HISAT2 [65]. The alignments of the 20 RNA-seq data and the FL transcripts along with the new genome annotation v2.0 were then used as input

to StringTie v.2.0 [66] to estimate expression levels in both fragments per kilobase per million reads (FPKM) and transcripts per million (TPM) for each transcript in the v2 annotation.

The percent identity and coverage of each *J. regia* transcript compared to proteins in the NCBI plant RefSeq database was also determined by running the EnTAP pipeline v.0.9.0 [37].

Label-free shotgun proteomics

Plant tissues of immature, intermediate, mature catkins and pure pollen from three individual trees of Chandler at the UCD walnut orchards were collected and frozen immediately in dry ice. Tissues were then further frozen in liquid nitrogen in the laboratory and ground with mortar and pestle. Five hundred milligrams of each sample were used for total protein extraction, following the procedure for recalcitrant plant tissues of [67], with a modification in the final buffer used to resuspend the protein pellet, consisting of 8M urea in 50mM triethylammonium bicarbonate (TEAB). One hundred micrograms of total protein from each sample were then used for proteomics.

Initially, 5 mM dithiothreitol (DTT) was added and incubated at 37°C for 30 min and 1,000 rpm shaking. Next, 15 mM iodoacetamide (IAA) was added, followed by incubation at room temperature for 30 min. The IAA was then neutralized with 30 mM DTT in incubation for 10 min. Lys-C/trypsin then was added (1:25 enzyme: total protein) followed by 4 h incubation at 37°C. After, TEAB (550 µl of 50 mM) was added to dilute the urea and activate trypsin digestion overnight. The digested peptides were desalted with Aspire RP30 Desalting Tips (Thermo Scientific), vacuum dried, and suspended in 45 µl of 50 mM TEAB. Peptides were quantified by Pierce quantitative fluorometric assay (Thermo Scientific) and 1 µg analyzed on a QExactive mass spectrometer (Thermo Scientific) coupled with an Easy-LC source (Thermo Scientific) and a

nanospray ionization source. The peptides were loaded onto a Trap (100 microns, C18 100 Å 5U) and desalted online before separation using a reversed-phase (75 microns, C18 200 Å 3U) column. The duration of the peptide separation gradient was 60 min using 0.1% formic acid and 100% acetonitrile (ACN) for solvents A and B, respectively. The data were acquired using a data-dependent MS/MS method, which had a full scan range of 300-1,600 Da and a resolution of 70,000. The resolution of the MS/MS method was 17,500 and the insulation width 2 m/z with a normalized collision energy of 27. The nanospray source was operated using a spray voltage of 2.2 KV and a transfer capillary temperature heated to 250°C. Samples were analyzed at the UC Davis Proteome Core.

The raw data were analyzed using X! Tandem and viewed using the Scaffold Software v.4. (Proteome Software, Inc.). Samples were searched against UniProt databases appended with the cRAP database, which recognizes common laboratory contaminants. Reverse decoy databases were also applied to the database before the X! Tandem searches. The protein-coding sequences (CDS) annotated in Chandler v1.0 (NCBI accession PRJNA350852) and v2.0 were used as a reference for identification of proteins from the mass spectrometry data. The proteins identified were filtered in the Scaffold software based on the following criteria: 1.0% FDR (false discovery rate) at protein level (following the prophet algorithm: <http://proteinprophet.sourceforge.net/>), the minimum number of 2 peptides and 0.1% FDR at the peptide level. Structure of the walnut allergen (Jug r 9) was modelled using SWISS-MODEL [68] based on the structure of a homologous allergen from lentil (PDBid:2MAL). Structures were superimposed using MUSTANG (2MAL:in red, walnut in blue) [69].

Chandler genomic diversity

Illumina whole-genome shotgun data of Chandler were aligned on the Chandler v2.0 with BWA [70] with standard parameters. SNP calling was performed using SAMtools v1.9 [71] and BCFtools v.2.1 [72]. SNP density for windows of 1 Mb was estimated using the command ‘SNPdensity’ implemented in VCFtools v0.1.16 [73]. Self-collinearity analysis to detect duplicated regions in Chandler v2.0 was performed with MCScanX [74], using a simplified GFF file of the new gene annotation and a self-BLASTP as input. To improve the power of collinearity detection, tandem duplications were excluded after running the function ‘detect_collinear_tandem_arrays’ implemented in MCScanX. Synonymous (K_S) and nonsynonymous (K_A) changes for syntenic protein-coding gene pairs were measured using the Perl script “add_ka_and_ks_to_collinearity.pl” implemented in MCScanX.

To explore the inbreeding level across the 16 chromosomal pseudomolecules of Chandler, haplotypes were built for 55 individuals of the UCD-WIP, including 25 founders and several commercially relevant walnut cultivars (e.g., Chandler, Howard, Tulare, Vina, Franquette) along with their parents and progenitors. All individuals were genotyped using the latest Axiom™ J. regia 700K SNP array as described in [7]. To define SNP HBs, 26,544 unique and robust SNPs were selected and ordered according to the Chandler genome v2.0 physical map. Subsequently, for each SNP markers and individual, phasing and identification of closely linked groups of SNPs, without recombination in most of the pedigree, was performed using the software FlexQTL™ [75] and PediHaplotyper [76] following the approach described in [77] and [76]. In particular, HB were defined by recombination sites detected in ancestral generation of Chandler.

Genomic comparison between Eastern and Western walnuts

The resequencing data of 23 founders of the UCD-WIP (**Additional file 26**)[10] were mapped onto the Chandler v2.0 with BWA, and SNPs were called following the same procedure described

above for Chandler. SNPs with no missing data and minor allele frequency (MAF) higher than 10% were retained for the following genetic analyses (7,269,224 SNPs out of the 14,988,422 identified). Hierarchical cluster analysis on a dissimilarity matrix of the 23 UCD-WIP founders was performed using R/SNPRelate v.1.18.0 [78]. Fixation index (F_{ST}) was measured between genotypes from EU/USA and Asia with VCFtools v0.1.16, setting windows of 100kb and 500kb. Genomic windows with the top 5% of F_{ST} values were selected as candidate regions for further analysis. The empirical cutoff with a low false discovery rate (5%) was verified by performing whole-genome permutation test (1000) with a custom Python script. Nucleotide diversity (π) and Tajima's D [79] were also computed along the whole genome in 100-kb and 500-kb windows using VCFtools. Reduction of diversity coefficient (ROD) was estimated as $1 - (\pi_{Occ} / \pi_{Asia})$. The new walnut gene annotation v.2.0 was used to identify predicted genes in the candidate regions under selection. The distribution of the identified genes into different biological processes was evaluated using the weight01 method provided by the R/topGO [80]. The Kolmogorov–Smirnov-like test was performed to assess the significance of over-representation of GO categories compared with all genes in the walnut gene prediction. Plots were obtained using the R/circlize v.0.4.6 and R/ggplot2 v.3.5.3 packages.

Availability of supporting data

All raw and processed sequencing data generated in this study have been submitted to the NCBI BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject/>) under accession number PRJNA291087. All SNP data have been submitted to Hardwood Genomics (<https://hardwoodgenomics.org/Genome-assembly/2539069>).

636 **Additional files**

637 **Additional file 1.** Statistics on k-unitgs, super-reads and mega-reads obtained with the MaSuRCA
638 assembler on ONT and Illumina reads.

639 **Additional file 2.** Characteristics of the Chandler ON assembly.

640 **Additional file 3.** Contiguity of the Chandler ON assembly and the final HiRise scaffolds. Each
641 curve shows the fraction of the total length of the assembly present in scaffolds of a given length
642 or smaller. The fraction of the assembly is indicated on the Y-axis and the scaffold length in base
643 pairs is given on the X-axis. The two dashed lines mark the N50 and N90 lengths of each assembly.
644 Scaffolds less than 1 kb are excluded.

645 **Additional file 4.** Mapping positions of the first and second read in the read pair respectively,
646 grouped into bins. The color of each square gives the number of read pairs within that bin. White
647 vertical and black horizontal lines show the borders between scaffolds. Scaffolds less than 1 Mb
648 are excluded.

649 **Additional file 5.** Collinearity between the ‘Chandler’ genetic map of [31] and the 16
650 chromosomal pseudomolecules of Chandler v2.0.

651 **Additional file 6.** Retrotransposons distribution across the 16 chromosomes of Chandler v2.0.

652 **Additional file 7.** List of tissues used for PacBio IsoSeq.

653 **Additional file 8.** Statistics on the PacBio IsoSeq sequencing per flow-cell.

654 **Additional file 9.** Statistics on CCSs, FLnc and HQ FL transcripts obtained per tissue with PacBio
655 IsoSeq.

656 **Additional file 10.** Percentage of FLnc and HQ FL transcripts aligned on the new assembly per
657 tissue.

658 **Additional file 11.** Validation of gene annotation with expression data and public plant protein
659 databases.

660 **Additional file 12.** Top 20 biological process GO terms.

661 **Additional file 13.** Top 20 molecular function GO terms.

662 **Additional file 14.** Top 20 cellular component GO terms.

663 **Additional file 15.** Summary of proteome results obtained using genome annotations v1 and v2.

664 **Additional file 16.** Mass-spectrometry proteome data of catkins and pollen tissues. Three samples
665 of each tissue type (immature catkin, mature catkin, senescent catkin, and pure pollen) were
666 analyzed using v1 and v2 reference walnut genome assemblies. Total intensity of matching
667 peptides, number of spectra and percentage of protein covered by the identified peptides are
668 reported.

669 **Additional file 17.** Allergen expression data obtained from proteome data.

670 **Additional file 18.** Top 10 abundant proteins detected in each tissue.

671 **Additional file 19.** Expression of genes encoding allergenic proteins in different tissues of cv.
672 Chandler.

673 **Additional file 20.** Number of SNPs and SNP density in 100-kb windows per chromosome in
674 Chandler v2.0.

675 **Additional file 21.** Regions (1 Mb) with less than 377.5 SNPs (10th percentile of the SNP number
676 distribution) in Chandler.

677 **Additional file 22.** Collinear blocks among the 16 chromosomal pseudomolecules of Chandler
678 v2.0. Different colors of dots represent different pairs of paralogous genes.

679 **Additional file 23.** Dual synteny plot between Chr06 a- Chr15 of Chandler v2.0.

680 **Additional file 24.** Top 50 biological process GO terms for the 393 singletons genes in the low
681 heterozygous regions on Chr15 of Chandler.

682 **Additional file 25.** Graphical visualization of the haplotype-blocks (HB) inheritance across
683 Chandler pedigree in the 16 chromosomes. (A) The inner circle highlights in grey the regions of
684 heterozygosity and in light green the regions of homozygosity for each chromosome. The circle in
685 the middle shows the maternally inherited HBs, while the HBs inherited from the paternal line are
686 visualized in outer circle. In both parental line circles, missing data are highlighted in grey. Payne
687 haplotypes are inherited along both parental lines in all chromosomes, but Chr5, Chr9, Chr10,
688 Chr14 and Chr16. (B) Chandler pedigree, where Pedro is the maternal line and 56-224 the paternal
689 line.

690 **Additional file 26.** List of 23 UCD-WIP founders used for the selective sweep analysis [10].

691 **Additional file 27.** Hierarchical clustering analysis among the 23 re-sequenced founders of the
692 UCD-WIP.

693 **Additional file 28.** Genome scan for selective sweeps between walnuts from EU/USA and Asia.
694 Tracks from outside to inside: (i) FST in 500-kb windows. Windows in the 95 percentiles of the
695 FST distribution are highlighted in red; (ii) ROD values for 500-kb windows; (iii) Tajima's D in

500-kb windows for genotypes from Europe a- USA; (iv) Tajima's D in 500-kb windows for genotypes from Asia.

Additional file 29. Top 50 biological process GO terms for the 122 windows (100 kb) with negative value of Tajima's D in the Western genotypes.

Additional file 30. Top 50 biological process GO terms for the 122 windows (100 kb) with negative value of Tajima's D in the Eastern genotypes.

Additional file 31. Marker-trait associations identified within genomic regions highly differentiated between Western and Eastern walnuts.

Additional file 32. Phenotypic differences of harvesting date observed among the three genotypic classes of the marker AX-170770379 significantly associated to harvest date [14].

Competing interests

The authors declare no conflict of interest.

Funding:

This project has been funded by the Californian Walnut Board.

Author's contribution

DBN and AM conceived and coordinated the research. REW and WT performed the HMW DNA extraction and Nanopore sequencing. AVZ, DP and SLS assembled the hybrid Illumina-ONT assembly. LB, MT, DP and SLS validated and anchored the HiRise assembly to the genetic maps. AM and BJA collected and extracted all RNA samples. MB analyzed the PacBio IsoSeq results

and performed the repeat and gene annotation. AD conceived the design of the proteomic analyses; PAZ and SC generated and analyzed the proteomic data. LB called the SNPs in Chandler and the 23 UCD WIP founders, while AM carried out the analyses on walnut genomic diversity. EAD, LB and MT built and analyzed the SNP haplotypes. CAL provided all the plant material. AM wrote the manuscript, which has been revised by all coauthors.

Acknowledgments

We are grateful to Sriema Walawage for assistance with RNA extraction, and Brett Phinney for preparing the raw proteome data.

References

1. Martínez ML, Labuckas DO, Lamarque AL, Maestri DM. Walnut (*Juglans regia* L.): Genetic resources, chemistry, by-products. J Sci Food Agric. 2010;90:1959–67.
2. McGranahan G, Leslie C. Walnut. In: Badenes ML, Byrne DH, editors. Fruit Breed. Springer Science+Business Media, LLC; 2012. p. 827–46.
3. Pollegioni P, Woeste K, Chiocchini F, Del Lungo S, Ciolfi M, Olimpieri I, et al. Rethinking the history of common walnut (*Juglans regia* L.) in Europe: Its origins and human interactions. PLoS One. 2017;12:1–24.
4. Zhang B, Xu L, Li N, Yan P, Jiang X, Woeste KE, et al. Phylogenomics Reveals an Ancient Hybrid Origin of the Persian Walnut. Mol Biol Evol. 2019;1–11.
5. Zeven A, Zhukovskii PM. Dictionary of cultivated plants and their centres of diversity, excluding ornamentals, forest trees, and lower plants [Internet]. Cent. Agric. Publ. Doc.

Wageningen. 1975. Available from: <https://core.ac.uk/download/pdf/29387092.pdf>

6. Ebrahimi A, Zarei A, Lawson S, Woeste KE, Smulders MJM. Genetic diversity and genetic structure of Persian walnut (*Juglans regia*) accessions from 14 European, African, and Asian countries using SSR markers. *Tree Genet Genomes* [Internet]. *Tree Genetics & Genomes*; 2016;12:114. Available from: <http://link.springer.com/10.1007/s11295-016-1075-y>

7. Marrano A, Martínez-García PJ, Bianco L, Sideli GM, Di Pierro EA, Leslie CA, et al. A new genomic tool for walnut (*Juglans regia* L.): development and validation of the high-density Axiom™ *J. regia* 700K SNP genotyping array. *Plant Biotechnol J* [Internet]. 2018;1–10. Available from: <http://doi.wiley.com/10.1111/pbi.13034>

8. Bernard A, Barreneche T, Lheureux F, Dirlewanger E. Analysis of genetic diversity and structure in a worldwide walnut (*Juglans regia* L.) germplasm using SSR markers. *PLoS One*. 2018;13:1–19.

9. Martínez-García PJ, Crepeau MW, Puiu D, Gonzalez-Ibeas D, Whalen J, Stevens KA, et al. The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *Plant J*. 2016;87:507–32.

10. Stevens KA, Woeste K, Chakraborty S, Crepeau MW, Leslie CA, Martínez-García PJ, et al. Genomic Variation Among and Within Six *Juglans* Species. *G3 Genes|Genomes|Genetics* [Internet]. 2018;8:1–37. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/29792315> <http://g3journal.org/lookup/doi/10.1534/g3.118.200030>

11. Kefayati S, Ikhsan AS, Sutyemez M, Paizila A, Topcu H, Bukubu SB, et al. First simple sequence repeat-based genetic linkage map reveals a major QTL for leafing time in walnut

760 (*Juglans regia* L.). Tree Genet Genomes. 2018;15:13.

761 12. Arab MM, Marrano A, Abdollahi-Arpanahi R, Leslie CA, Askari H, Neale DB, et al.

762 Genome-wide patterns of population structure and association mapping of nut-related traits in

763 Persian walnut populations from Iran using the Axiom *J. regia* 700K SNP array. Sci Rep

764 [Internet]. Springer US; 2019;9:6376. Available from: [http://www.nature.com/articles/s41598-](http://www.nature.com/articles/s41598-019-42940-1)

765 019-42940-1

766 13. Famula RA, Richards JH, Famula TR, Neale DB. Association Genetics of Carbon Isotope

767 Discrimination in the Founding Individuals of a Breeding Population of *Juglans regia* L. Tree

768 Genet Genomes [Internet]. Tree Genetics & Genomes; 2019;15:6. Available from:

769 <https://doi.org/10.1007/s11295-018-1307-4>

770 14. Marrano A, Sideli GM, Leslie CA, Cheng H, Neale DB. Deciphering of the genetic control

771 of phenology, yield and pellicle color in Persian walnut (*Juglans regia* L.). Front Plant Sci.

772 2019;10:1–14.

773 15. Sánchez-Pérez R, Pavan S, Mazzeo R, Moldovan C, Aiese Cigliano R, Del Cueto J, et al.

774 Mutation of a bHLH transcription factor allowed almond domestication. Science (80-)

775 [Internet]. 2019;364:1095–8. Available from:

776 <http://www.sciencemag.org/lookup/doi/10.1126/science.aav8197>

777 16. Tang W, Sun X, Yue J, Tang X, Jiao C, Yang Y, et al. Chromosome-scale genome assembly

778 of kiwifruit *Actinidia eriantha* with single-molecule sequencing and chromatin interaction

779 mapping. Gigascience. Oxford University Press; 2019;8:1–10.

780 17. Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, et al. The genome of

781 *Chenopodium quinoa*. Nature. 2017;542:307–12.

- 782 18. Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, et al. Durum
783 wheat genome highlights past domestication signatures and future improvement targets. Nat
784 Genet. 2019;51:885–95.
- 785 19. Raymond O, Gouzy J, Just J, Badouin H, Verdenaud M, Lemainque A, et al. The Rosa
786 genome provides new insights into the domestication of modern roses. Nat Genet. 2018;50:772–
787 7.
- 788 20. Daccord N, Celton JM, Linsmith G, Becker C, Choisne N, Schijlen E, et al. High-quality de
789 novo assembly of the apple genome and methylome dynamics of early fruit development. Nat
790 Genet [Internet]. Nature Publishing Group; 2017;49:1099–106. Available from:
791 <http://dx.doi.org/10.1038/ng.3886>
- 792 21. Zhu T, Wang L, You FM, Rodriguez JC, Deal KR, Chen L, et al. Sequencing a *Juglans regia*
793 $\times$ *J. microcarpa* hybrid yields high-quality genome assemblies of parental species. Hortic Res
794 [Internet]. Springer US; 2019;1–16. Available from: [http://dx.doi.org/10.1038/s41438-019-0139-](http://dx.doi.org/10.1038/s41438-019-0139-1)
795 1
- 796 22. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly.
797 Genomics, Proteomics Bioinforma [Internet]. Beijing Institute of Genomics, Chinese Academy
798 of Sciences and Genetics Society of China; 2016;14:265–79. Available from:
799 <http://dx.doi.org/10.1016/j.gpb.2016.05.004>
- 800 23. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: A comprehensive
801 technique to capture the conformation of genomes. Methods [Internet]. Elsevier Inc.;
802 2012;58:268–76. Available from: <http://dx.doi.org/10.1016/j.ymeth.2012.05.001>
- 803 24. Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. J Ex.

804 2017;68:5419–29.

805 25. Schmidt MH-W, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, et al. De Novo
806 Assembly of a New *Solanum pennellii* Accession Using Nanopore Sequencing . Plant Cell.
807 2017;29:2336–48.

808 26. Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, et al. Chromosome-scale
809 assemblies of plant genomes using nanopore long reads and optical maps. Nat Plants [Internet].
810 Springer US; 2018;4:879–87. Available from: <http://dx.doi.org/10.1038/s41477-018-0289-4>

811 27. Yasodha R, Vasudeva R, Balakrishnan S, Sakthi AR, Abel N, Binai N, et al. Draft genome of
812 a high value tropical timber tree, Teak (*Tectona grandis* L.): insights into SSR diversity,
813 phylogeny and conservation. DNA Res. 2018;25:409–19.

814 28. Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, et al. A chromosome-scale
815 assembly of the sorghum genome using nanopore sequencing and optical mapping. Nat
816 Commun. 2018;9:4844.

817 29. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: Computational
818 approaches for improving nanopore sequencing read accuracy. Genome Biol. Genome Biology;
819 2018;19:1–11.

820 30. Zimin A V., Luo M, Marçais G, Salzberg SL, Yorke JA, Puiu D, et al. Hybrid assembly of
821 the large and highly repetitive genome of *Aegilops tauschii* , a progenitor of bread wheat, with
822 the MaSuRCA mega-reads algorithm. Genome Res [Internet]. 2017;27:787–92. Available from:
823 <http://www.ncbi.nlm.nih.gov/pubmed/28130360>
824 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5411773>

825 31. Luo M-C, You FM, Li P, Wang J-R, Zhu T, Dandekar AM, et al. Synteny analysis in Rosids
826 with a walnut physical map reveals slow genome evolution in long-lived woody perennials.
827 BMC Genomics [Internet]. BMC Genomics; 2015;16:1–17. Available from:
828 <http://www.biomedcentral.com/1471-2164/16/707>

829 32. Springer NM, Ying K, Fu Y, Ji T, Yeh C, Jia Y, et al. Maize Inbreds Exhibit High Levels of
830 Copy Number Variation (CNV) and Presence / Absence Variation (PAV) in Genome Content.
831 PLoS Genet. 2009;5.

832 33. Marroni F, Pinosio S, Morgante M. Structural variation and genome complexity : is
833 dispensable really dispensable ? Curr Opin Plant Biol [Internet]. Elsevier Ltd; 2014;18:31–6.
834 Available from: <http://dx.doi.org/10.1016/j.pbi.2014.01.003>

835 34. Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver P, Montfort M, et al. Exploiting
836 single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol [Internet].
837 Genome Biology; 2015;16:1–13. Available from: <http://dx.doi.org/10.1186/s13059-015-0729-7>

838 35. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics, Proteomics
839 Bioinforma [Internet]. Beijing Institute of Genomics, Chinese Academy of Sciences and
840 Genetics Society of China; 2015;13:278–89. Available from:
841 <http://dx.doi.org/10.1016/j.gpb.2015.08.002>

842 36. Linsmith G, Rombauts S, Montanari S, Deng CH, Guérif P, Liu C, et al. Pseudo-
843 chromosome length genome assembly of a double haploid ‘ Bartlett ’ pear (*Pyrus communis* L .).
844 bioRxiv. 2019;

845 37. Hart AJ, Ginzburg S, Xu M (Sam), Fisher CR, Rahmatpour N, Mitton JB, et al. EnTAP:
846 bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes.

847 bioRxiv [Internet]. 2018;307868. Available from:
 848 <https://www.biorxiv.org/content/biorxiv/early/2018/04/28/307868.full.pdf>
 849 <https://www.biorxiv.org/content/early/2018/04/24/307868>
 850 307868

851 38. Jamet E, Santoni V. Editorial for Special Issue: 2017 Plant Proteomics. *proteomes*.
 852 2018;6:28.

853 39. Costa J, Carrapatoso I, Oliveira MBPP, Mafra I. Walnut allergens: Molecular
 854 characterization, detection and clinical relevance. *Clin Exp Allergy*. 2014;44:319–41.

855 40. Aradhya M, Woeste K, Velasco D. Genetic diversity, structure and differentiation in
 856 cultivated walnut (*Juglans regia* L.). *Acta Hortic*. 2010;861:127–32.

857 41. Ruiz-Garcia L, Lopez-Ortega G, Fuentes Denia a., Frutos Tomas D. Identification of a
 858 walnut (*Juglans regia* L.) germplasm collection and evaluation of their genetic variability by
 859 microsatellite markers. *Spanish J Agric Res*. 2011;9:179–92.

860 42. Dangl GS, Woeste K, Aradhya MK, Koehmstedt A, Simon C, Potter D, et al.
 861 Characterization of 14 Microsatellite Markers for Genetic Analysis and Cultivar Identification of
 862 Walnut. *J Am Soc Hortic Sci* [Internet]. 2005;130:348–54. Available from:
 863 <http://journal.ashspublications.org/content/130/3/348>
 864 <http://journal.ashspublications.org/content/130/3/348.full.pdf>

865 43. McGranahan GH, Leslie CA. Walnuts. In: Moore JN, Ballington JRJ, editors. *Genet Resour*
 866 *Temp Fruit Nut Crop*. International Society for Horticultural Science; 1991. p. 907–18.

867 44. Bernard A, Lheureux F, Dirlewanger E. Walnut: past and future of genetic improvement.

868 Tree Genet Genomes. Tree Genetics & Genomes; 2018;14:1–28.

869 45. Gauthier MM, Jacobs DF. Walnut (*Juglans* spp.) ecophysiology in response to environmental
870 stresses and potential acclimation to climate change. Ann For Sci. 2011;68:1277–90.

871 46. Workman R, Fedak R, Kilburn D, Hao S, Liu K, Timp W. High Molecular Weight DNA
872 Extraction from Recalcitrant Plant Species for Third Generation Sequencing. Protoc Exch.
873 2018;1–12.

874 47. Zhang M, Zhang Y, Scheuring CF, Wu CC, Dong JJ, Zhang H Bin. Preparation of megabase-
875 sized DNA from a variety of organisms using the nuclei method for advanced genomics
876 research. Nat Protoc [Internet]. Nature Publishing Group; 2012;7:467–78. Available from:
877 <http://dx.doi.org/10.1038/nprot.2011.455>

878 48. Mayjonade B, Gouzy J, Donnadieu C, Pouilly N, Marande W, Callot C, et al. Extraction of
879 high-molecular-weight genomic DNA for long-read sequencing of single molecules.
880 Biotechniques. 2017;62:xv.

881 49. Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome
882 assembler. Bioinformatics. 2013;29:2669–77.

883 50. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al.
884 Comprehensive mapping of long-range interactions reveals folding principles of the human
885 genome. Science (80). 2009;

886 51. Putnam NH, O’Connell, Brendan L. Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, et
887 al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage.
888 Genome Res. 2016;26:342–50.

889 52. Benson G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids*
890 *Res.* 1999;27:573–80.

891 53. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast
892 and versatile genome alignment system. *PLoS Comput Biol.* 2018;14:1–14.

893 54. Rezvoy C, Charif D, Guéguen L, Marais GAB. MareyMap: An R-based tool with graphical
894 interface for estimating recombination rates. *Bioinformatics.* 2007;23:2188–9.

895 55. Li H. Minimap2 : pairwise alignment for nucleotide sequences. *Bioinformatics.*
896 2018;34:3094–100.

897 56. Smit A, Hubley R. RepeatModeler Open-1.0. [Internet]. 2008. Available from:
898 <http://www.repeatmasker.org>

899 57. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. [Internet]. 2013. Available from:
900 <http://www.repeatmasker.org>

901 58. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, et al. Improving
902 the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic*
903 *Acids Res.* 2003;31:5654–66.

904 59. Kent WJ. BLAT — The BLAST -Like Alignment Tool. *Genome Res.* 2002;12:656–64.

905 60. Wu TD, Watanabe CK. GMAP : a genomic mapping and alignment program for mRNA and
906 EST sequences. *Bioinformatics.* 2005;21:1859–75.

907 61. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo
908 transcript sequence reconstruction from RNA-Seq: reference generation and analysis with
909 Trinity. *Nat Protoc.* 2013;8:1–43.

910 62. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length
911 transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011;
912 63. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO:
913 Assessing genome assembly and annotation completeness with single-copy orthologs.
914 *Bioinformatics.* 2015;31:3210–2.

915 64. Veeckman E, Ruttink T, Vandepoele K. Are We There Yet? Reliably Estimating the
916 Completeness of Plant Genome Sequences. *Plant Cell.* 2016;28:1759–68.

917 65. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory
918 requirements. *Nat Methods.* 2015;

919 66. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of
920 RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc.* 2016;

921 67. Valerie M, Catherine D, Michel Z, Hervé T, Faurobert M, Pelpoir E, et al. Phenol Extraction
922 of Proteins for Proteomic Studies of Recalcitrant Plant Tissues. *Plant Proteomics.* 2006.

923 68. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: A web-based
924 environment for protein structure homology modelling. *Bioinformatics.* 2006;

925 69. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. MUSTANG: A multiple structural
926 alignment algorithm. *Proteins Struct Funct Genet.* 2006;

927 70. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
928 *Bioinformatics.* 2009;25:1754–60.

929 71. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
930 Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25:2078–9.

931 72. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-smith C, Durbin R. BCFtools/RoH : a
932 hidden Markov model approach for detecting autozygosity from next-generation sequencing
933 data. *Bioinformatics*. 2016;32:1749–51.

934 73. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call
935 format and VCFtools. *Bioinformatics*. 2011;27:2156–8.

936 74. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: A toolkit for detection
937 and evolutionary analysis of gene synten and collinearity. *Nucleic Acids Res*. 2012;40:1–14.

938 75. Bink MCAM, Jansen J, Madduri M, Voorrips RE, Durel CE, Kouassi AB, et al. Bayesian
939 QTL analyses using pedigreed families of an outcrossing species, with application to fruit
940 firmness in apple. *Theor Appl Genet*. 2014;127:1073–90.

941 76. Voorrips RE, Bink MCAM, Kruisselbrink JW, Koehorst-van Putten HJJ, van de Weg WE.
942 PediHaplotyper: software for consistent assignment of marker haplotypes in pedigrees. *Mol*
943 *Breed*. Springer Netherlands; 2016;36.

944 77. Vanderzande S, Howard NP, Cai L, Da Silva Linge C, Antanaviciute L, Bink MCAM, et al.
945 High-quality, genome-wide SNP genotypic data for pedigreed germplasm of the diploid
946 outbreeding species apple, peach, and sweet cherry through a common workflow. *PLoS One*.
947 2019;

948 78. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance
949 computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*.
950 2012;28:3326–8.

951 79. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA

polymorphism. Genetics. 1989;123:585–95.

80. Alexa A. Gene set enrichment analysis with topGO. 2015;47–53.

Tables

Table 1. Comparison among the four assemblies of Chandler. Scaffolds shorter than 1,000 bp are not included in these totals.

Assembly	Genome size (bp)	Number of Scaffolds	N50 (bp) G=620M
Chandler v1.0	712,759,961	27,032	415,376
Chandler v1.5	651,682,552	4,402	687,445
Chandler hybrid	573,816,693	3,497	1,361,770
Chandler HiRise	573,917,993	2,656	31,492,331
Chandler v2.0 (chrs)	547,778,456	16	35,064,427

Table 2. Statistics on the gene annotation of Chandler v2.0 compared to the previous gene annotations of the Chandler genome.

	Chandler v2.0	Chandler v1.0	Chandler RefSeq v1.0
Number of genes	40,491	32,496	41,188
Average gene length (bp)	5,776	4,358	4,641
Single-exon transcripts	6,616	6,247	6,749
Average CDS length (bp)	1,335	1,222	1,336
Number of exons	244,238	172,273	230,261
Average exon length (bp)	257.1	229.5	314
Number of Introns	203,157	139,775	181,419
Average intron length	856.9	730	835

Figure legends

Figure 1. Collinearity between the high-density ‘Chandler’ genetic map of [14] and the 16 chromosomal pseudomolecules of Chandler v2.0.

Figure 2. Summary of gene distribution and genetic diversity across the 16 chromosomes of Chandler v2.0. Tracks from outside to inside: **(i)** gene density of Chandler v2.0 in 1-Mb windows; **(ii)** Chandler heterozygosity in 1-Mb windows (white = low heterozygosity; blue = high heterozygosity); **(iii)** Recombination rate for sliding windows of 10 Mb (average = 2.63 cM/Mb); **(iv)** F_{ST} in 500-kb windows. Windows in the 95 percentiles of the F_{ST} distribution are highlighted in red; **(v)** ROD values for 500-kb windows.

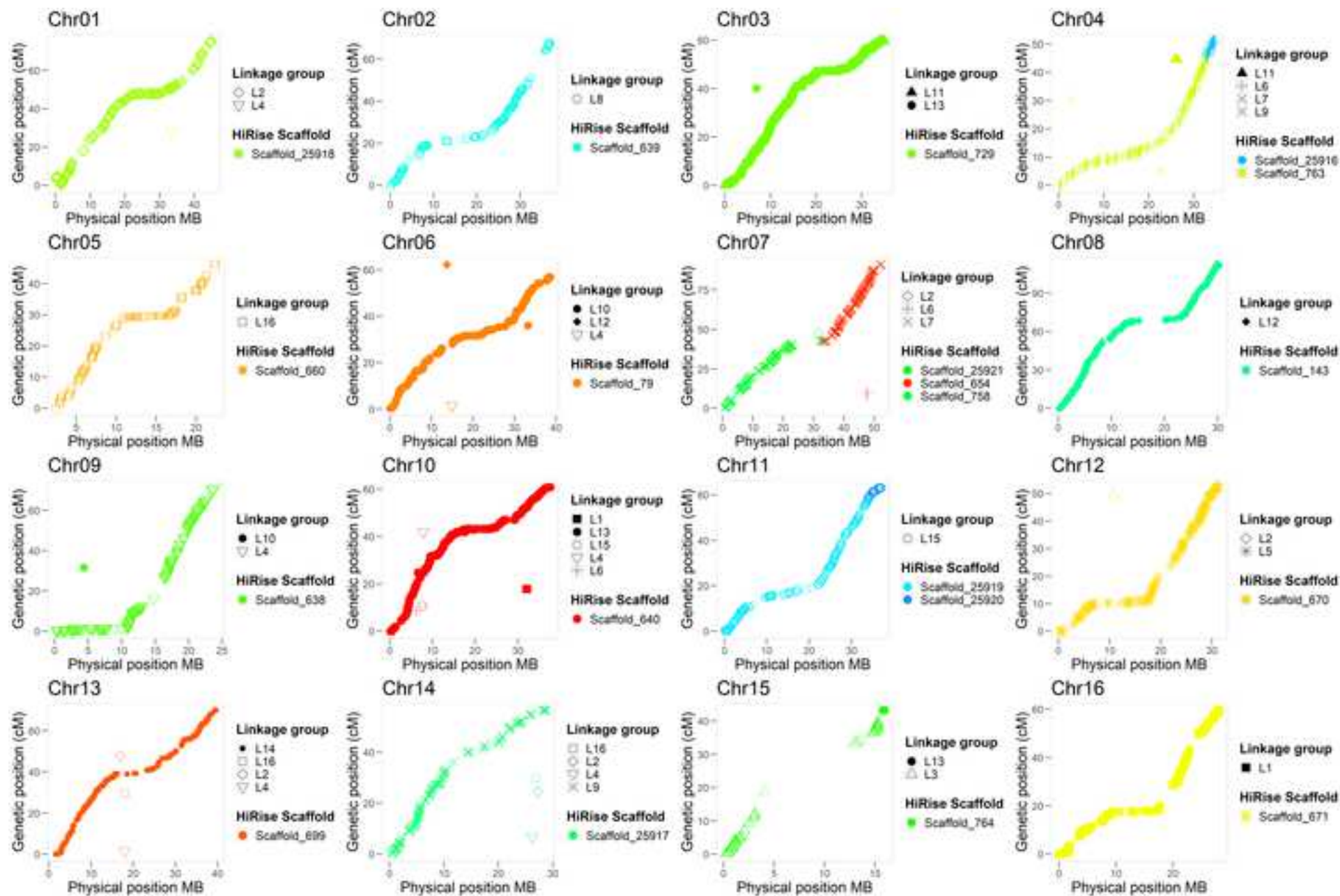
Figure 3. Clustering of the samples used in the proteomic analysis. **(A)** Hierarchical clustering based on Euclidian distances of normalized abundances of detected proteins. Samples are represented in columns and proteins in rows. **(B)** Principal component analysis of the 12 samples analyzed, clustering according to tissue type.

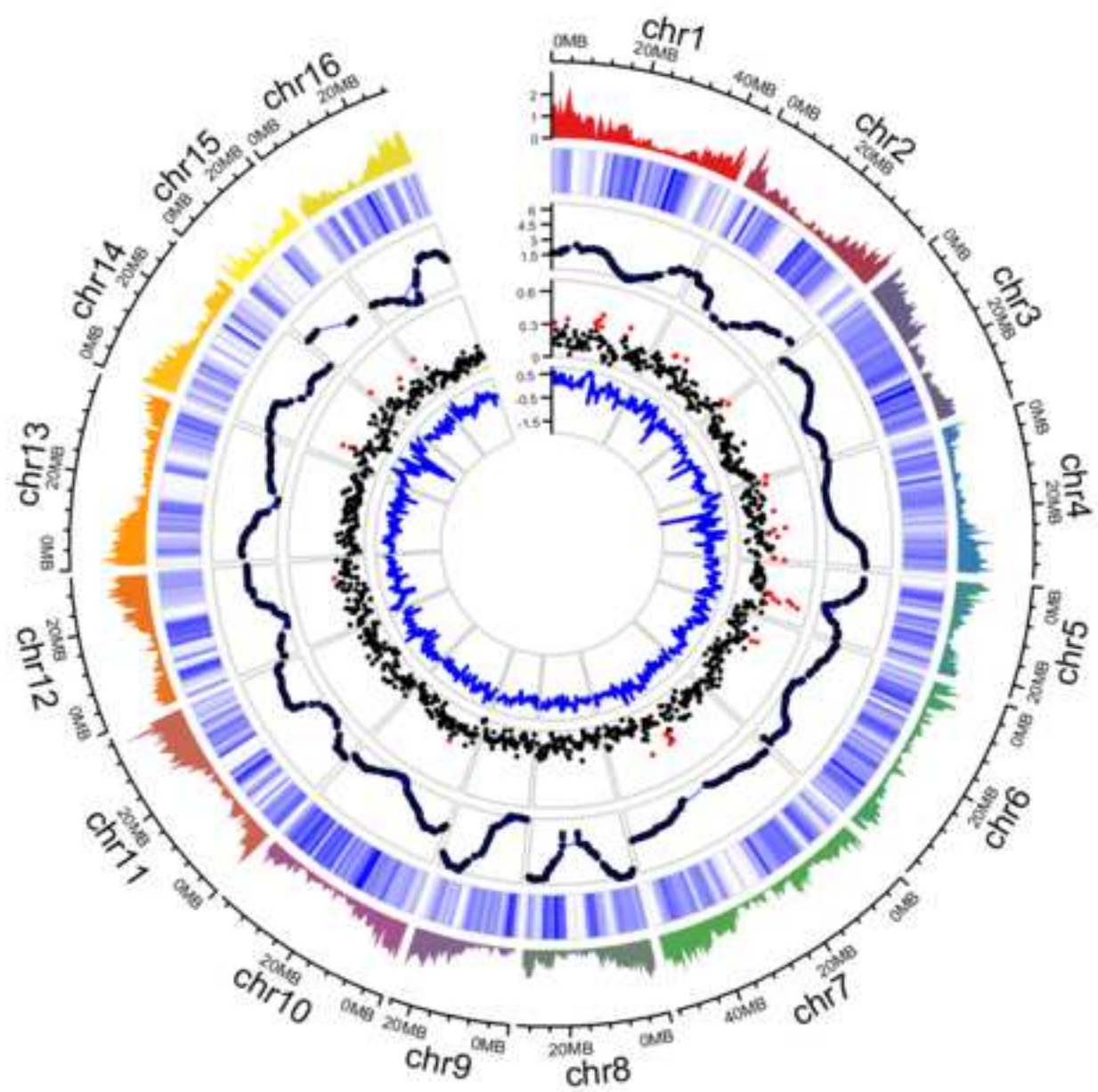
Figure 4. Modeled structure of the putative new allergen encoded by *Jr12_05180*. The compact structure is stabilized by four disulfide bonds, common in other allergenic proteins. The model in blue is superimposed with a homologous allergen from lentil (PDBid:2MAL) represented in red. Structure rendered with Pymol 2.3.

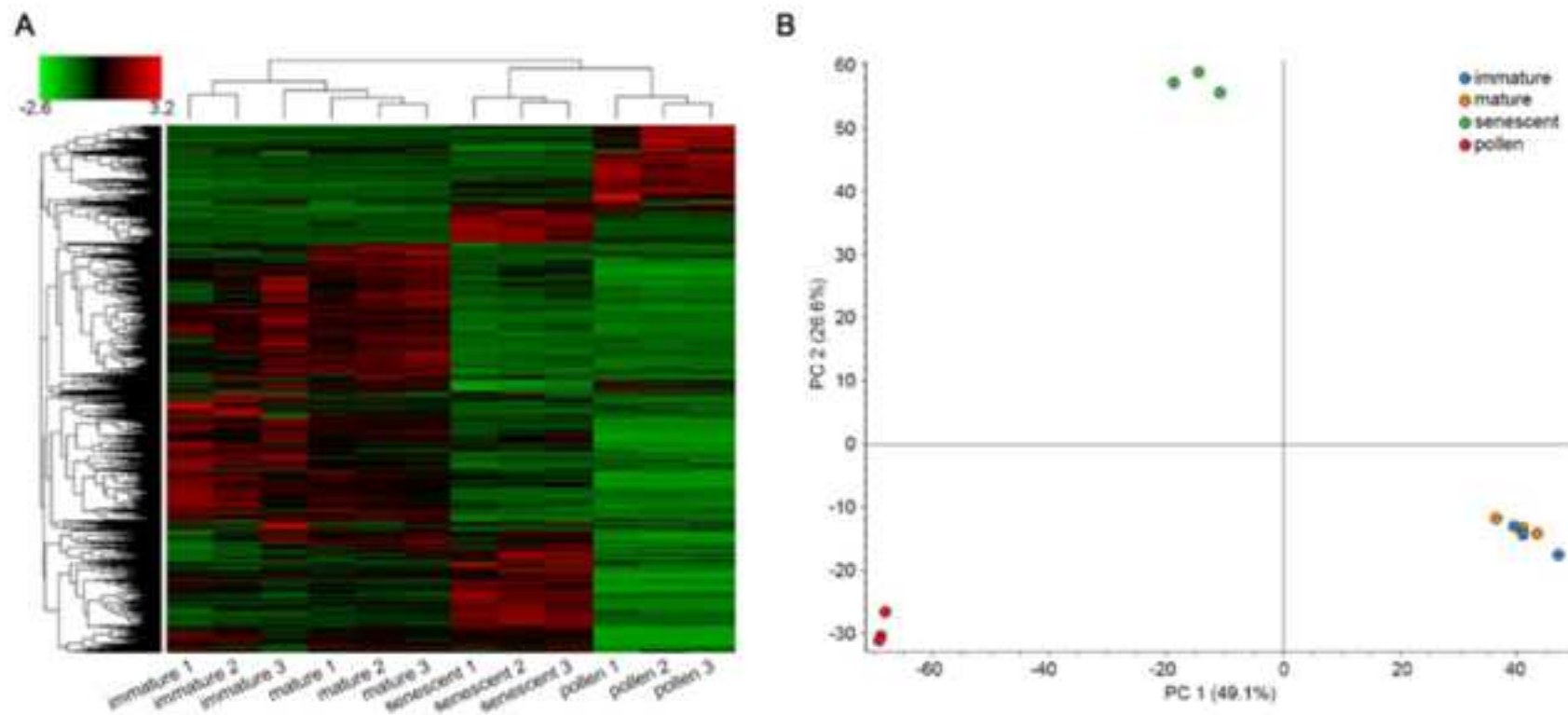
Figure 5. Graphical visualization of haplotype-blocks (HB) inheritance on Chr15 along with the Chandler pedigree. **(A)** The inner-circle highlights in grey two regions of heterozygosity (5 HB the first and 7 HB the second), and in light green two regions of homozygosity (3 HB the first and 4 HB the second). The circle in the middle shows maternally inherited HBs, while the HBs inherited through the paternal line are visualized in the outer circle. Payne’s haplotypes are clearly

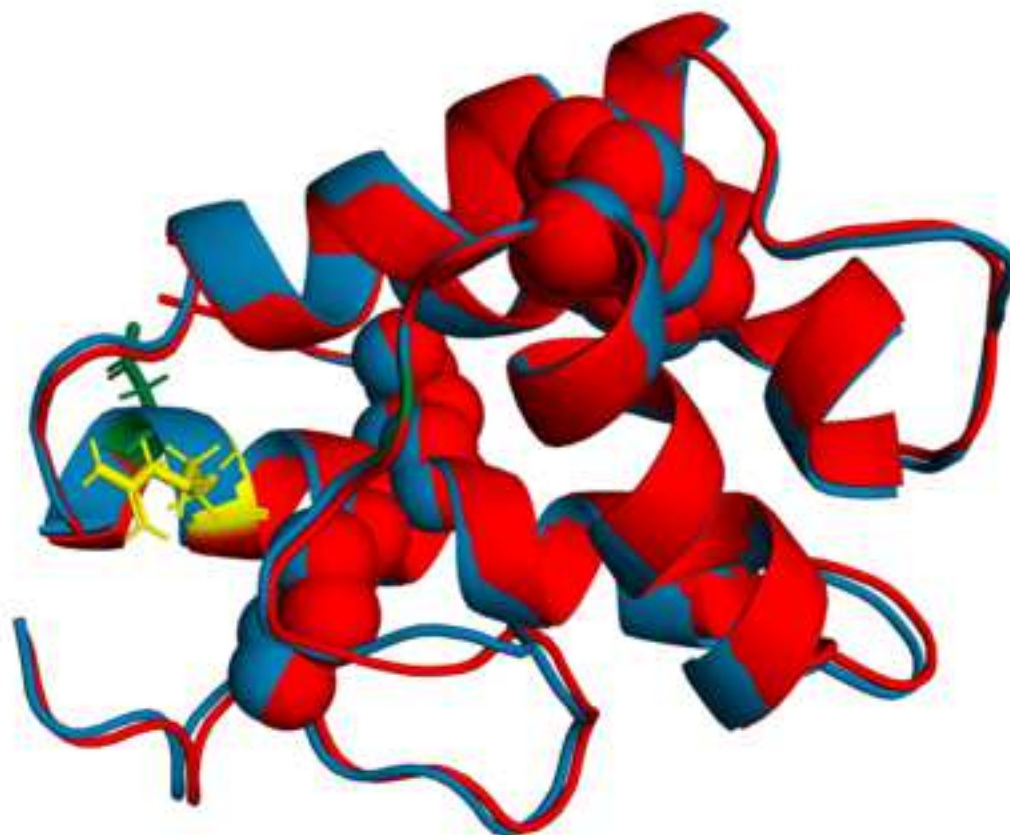
985 present in both parental lines. **(B)** Chandler pedigree, where Pedro is the maternal line and 56-224,
986 the paternal line.

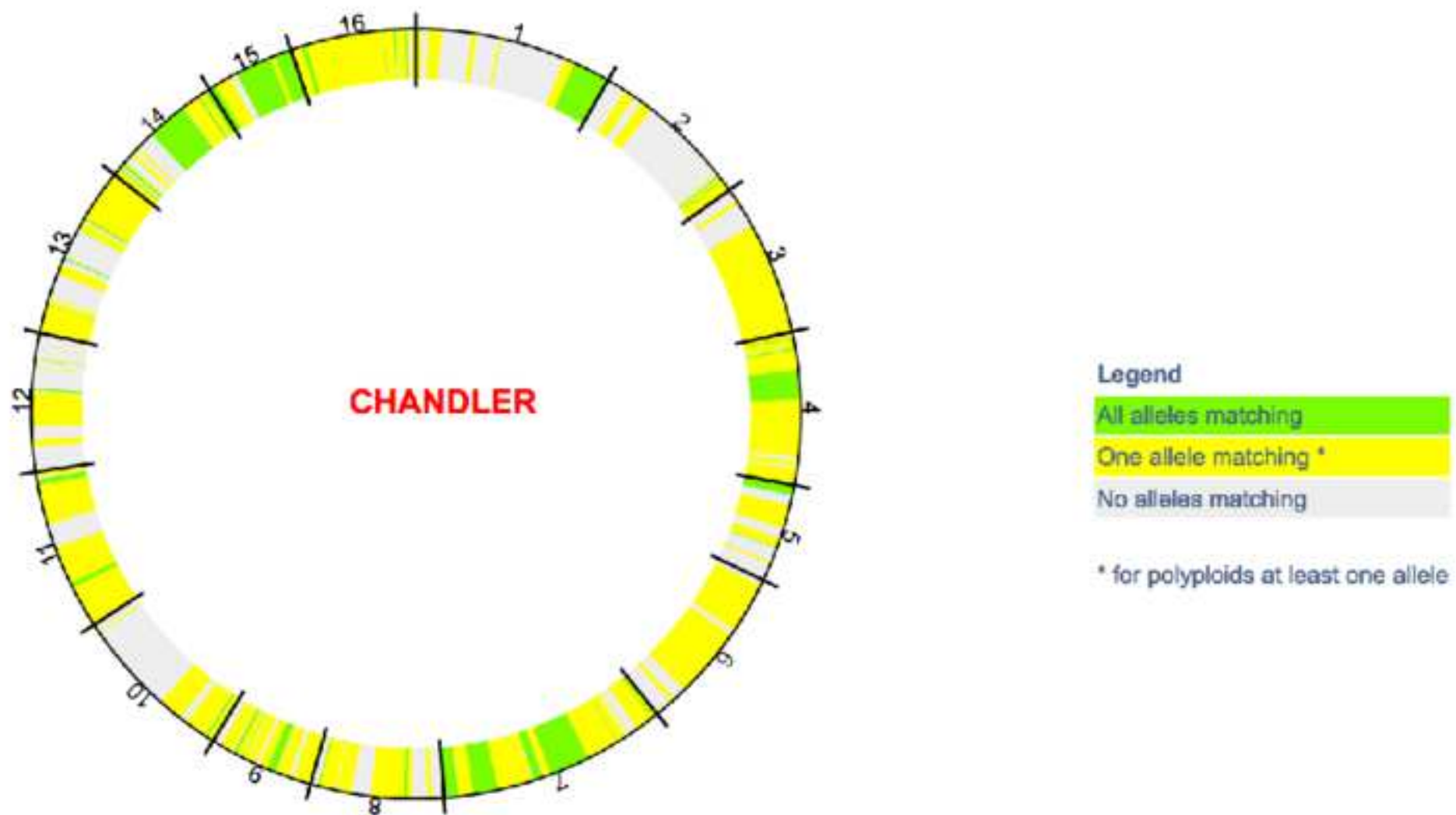
987 **Figure 6.** Graphical visualization of allele identity between Chandler and its ancestor Payne for
988 all 16 chromosomes of Chandler.

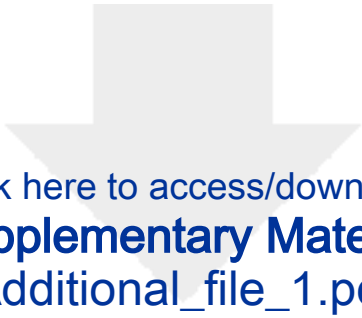




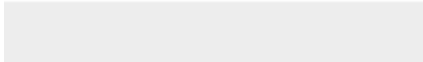


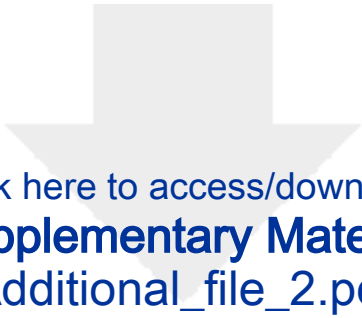




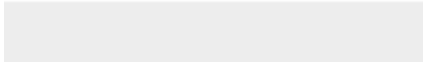


Click here to access/download
Supplementary Material
Additional_file_1.pdf



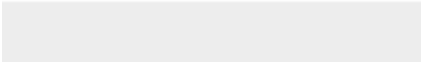


Click here to access/download
Supplementary Material
Additional_file_2.pdf



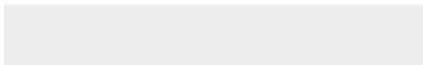


Click here to access/download
Supplementary Material
Additional_file_3.tif



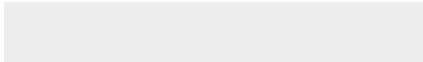


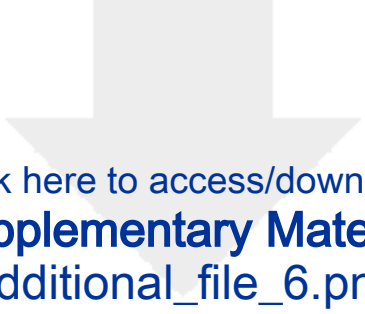
Click here to access/download
Supplementary Material
Additional_file_4.tif



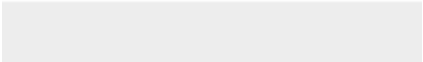


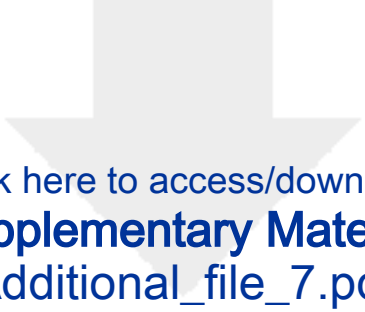
Click here to access/download
Supplementary Material
Additional_file_5.tiff



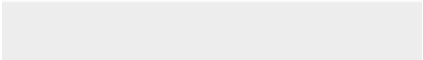


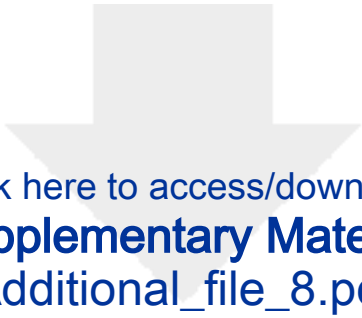
Click here to access/download
Supplementary Material
Additional_file_6.png



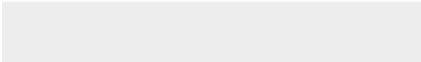


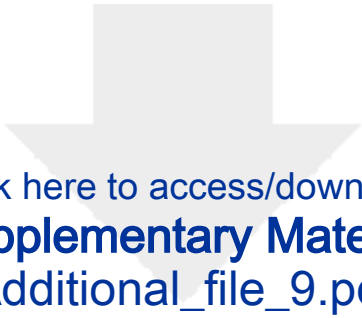
Click here to access/download
Supplementary Material
Additional_file_7.pdf



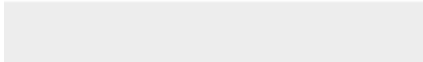


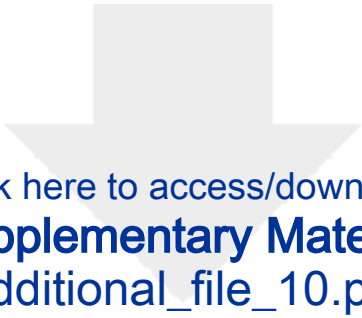
Click here to access/download
Supplementary Material
Additional_file_8.pdf



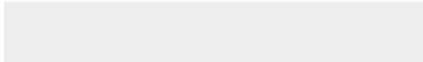


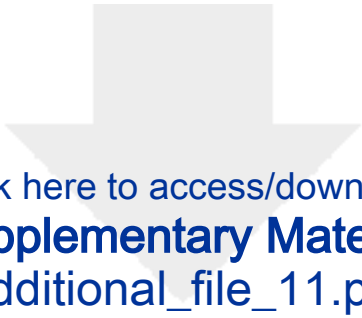
Click here to access/download
Supplementary Material
Additional_file_9.pdf



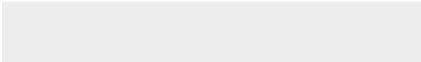


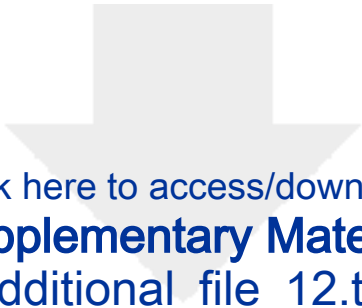
Click here to access/download
Supplementary Material
Additional_file_10.pdf



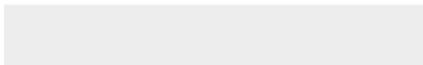
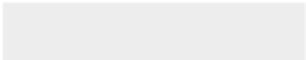


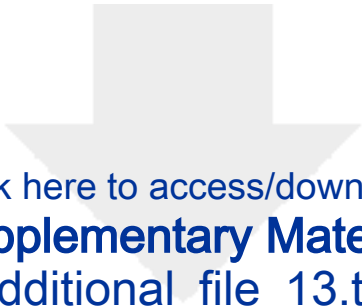
Click here to access/download
Supplementary Material
Additional_file_11.pdf



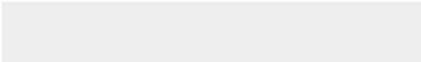


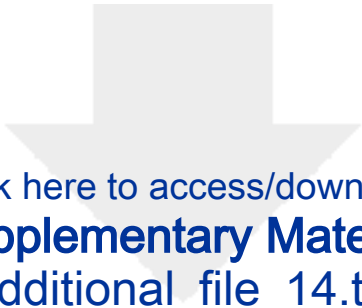
Click here to access/download
Supplementary Material
Additional_file_12.tiff



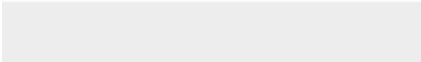
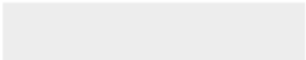


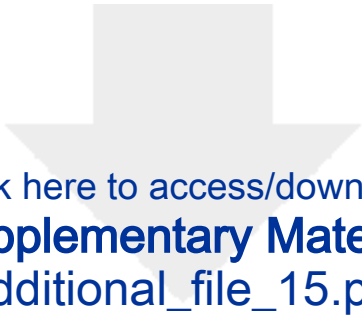
Click here to access/download
Supplementary Material
Additional_file_13.tiff



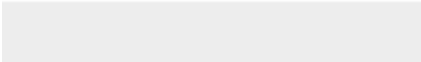


Click here to access/download
Supplementary Material
Additional_file_14.tiff





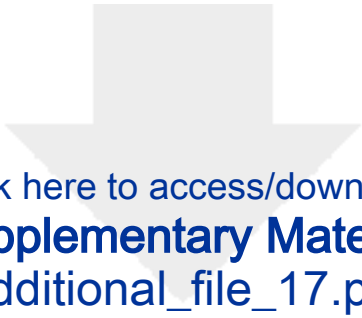
Click here to access/download
Supplementary Material
Additional_file_15.pdf



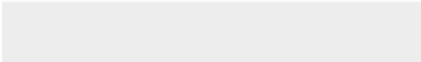


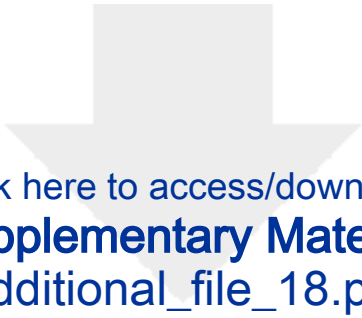
Click here to access/download
Supplementary Material
Additional_file_16.xlsx



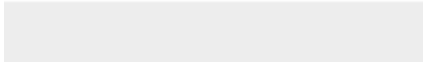


Click here to access/download
Supplementary Material
Additional_file_17.pdf



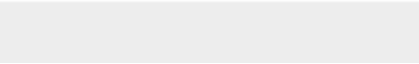


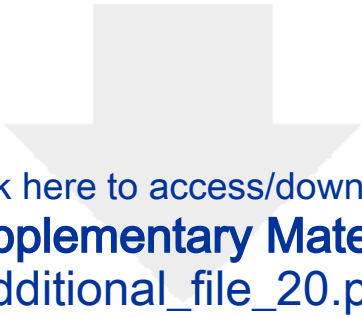
Click here to access/download
Supplementary Material
Additional_file_18.pdf



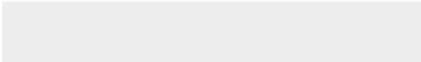


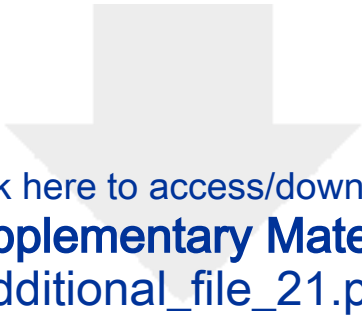
Click here to access/download
Supplementary Material
Additional_file_19.xlsx



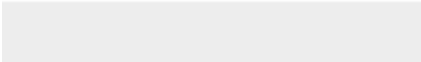


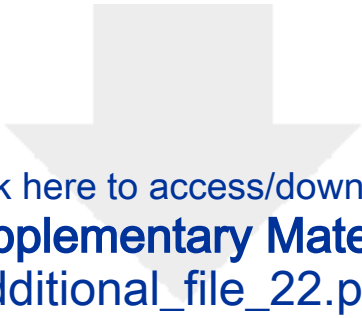
Click here to access/download
Supplementary Material
Additional_file_20.pdf



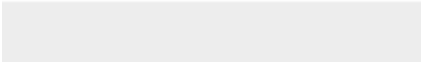


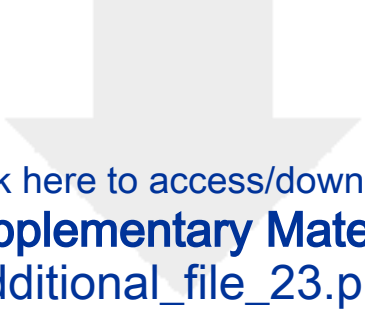
Click here to access/download
Supplementary Material
Additional_file_21.pdf



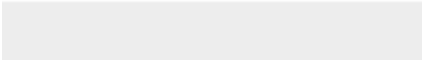


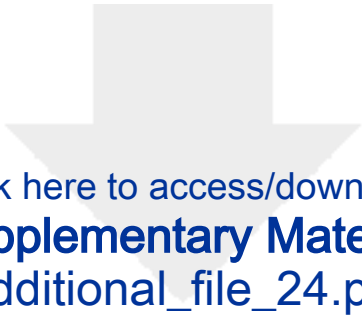
Click here to access/download
Supplementary Material
Additional_file_22.png



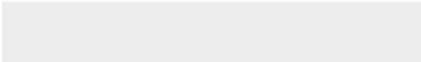


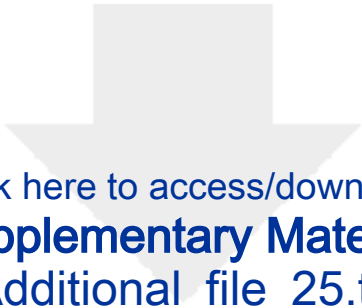
Click here to access/download
Supplementary Material
Additional_file_23.png



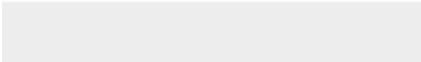


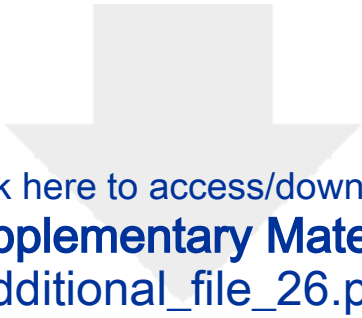
Click here to access/download
Supplementary Material
Additional_file_24.pdf



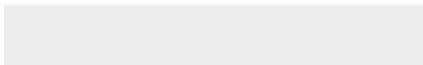


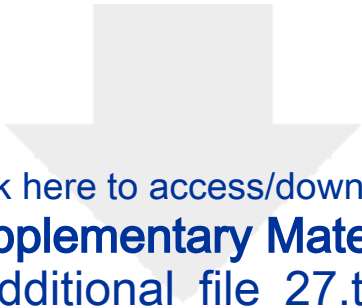
Click here to access/download
Supplementary Material
Additional_file_25.tif



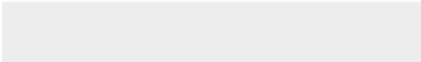
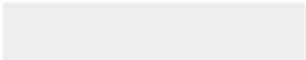


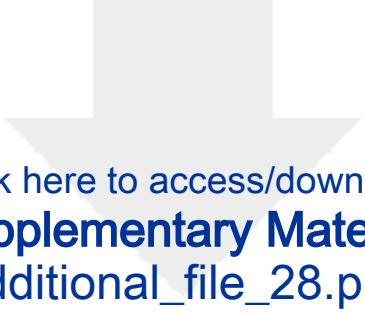
Click here to access/download
Supplementary Material
Additional_file_26.pdf



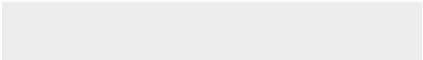


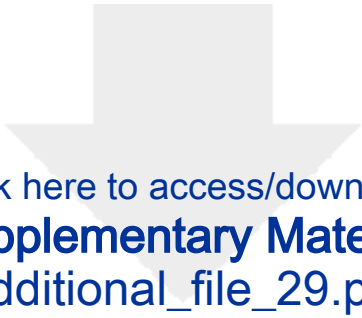
Click here to access/download
Supplementary Material
Additional_file_27.tiff



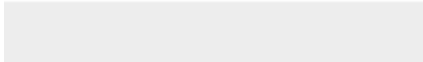


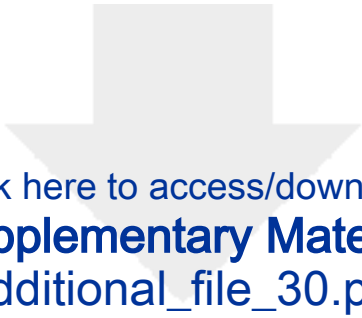
Click here to access/download
Supplementary Material
Additional_file_28.png



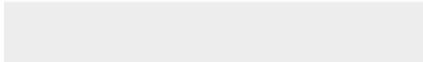


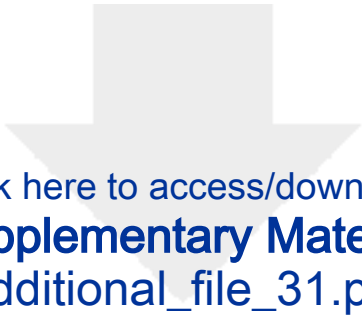
Click here to access/download
Supplementary Material
Additional_file_29.pdf



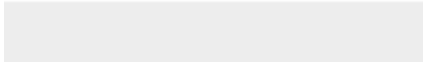


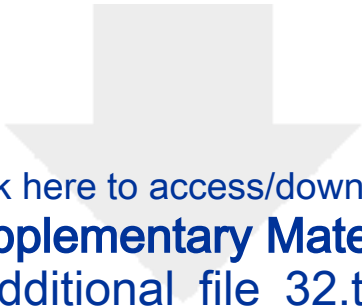
Click here to access/download
Supplementary Material
Additional_file_30.pdf



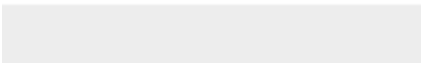
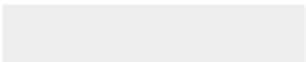


Click here to access/download
Supplementary Material
Additional_file_31.pdf





Click here to access/download
Supplementary Material
Additional_file_32.tiff



UNIVERSITY OF CALIFORNIA, DAVIS

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO



SANTA BARBARA • SANTA CRUZ

DEPARTMENT OF PLANT SCIENCES
MAIL STOP 6
UNIVERSITY OF CALIFORNIA
ONE SHIELDS AVE
DAVIS, CALIFORNIA 95616-8780
TELEPHONE: 530-752-1703
FAX: 530-752-1819

COLLEGE OF AGRICULTURAL AND
ENVIRONMENTAL SCIENCES
AGRICULTURAL EXPERIMENT STATION
COOPERATIVE EXTENSION

October 19th, 2018

Dear Editor,

The enclosed manuscript entitled “High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome” is being submitted to *GigaScience* for publication as a Data Note.

By submitting this manuscript, we declare it represents original work that has not been published elsewhere, though a preprint of this manuscript is available on bioRxiv (<http://biorxiv.org/cgi/content/short/809798v1>). All authors have approved the manuscript for submission. We declare no conflict of interest, and that all previously published work has been thoroughly acknowledged in our manuscript.

In the present research, we describe the development of the new chromosome-scale assembly of the walnut reference genome (Chandler v2.0), and its crucial role for the fulfillment of advanced research studies in walnut. By applying Oxford Nanopore long-read sequencing and chromosome conformation capture (Hi-C) technology, we fully assembled the 16 chromosomal pseudomolecules of *J. regia* with unprecedented contiguity. The new chromosome-scale assembly considerably improved gene prediction accuracy, with longer and full-length gene models, and fewer artifacts than the previous walnut gene annotations.

We then tested the utility of the new chromosome-level reference genome in different areas of walnut research. For instance, we confirmed how the new, more contiguous reference genome allows accurate estimation of the proteomic changes occurring in the different vegetative and reproductive stages of walnut so that it enables the identification of a new potential pollen allergen in walnut. Also, by studying the genome-wide genetic diversity of Chandler, we discovered highly homozygous regions, likely arising from inbreeding due to a shared ancestor in Chandler pedigree. Lastly, we proved the crucial role of the new chromosome-scale genome for comparative analyses

at the population level and genome scans for signatures of selection in walnut, by studying the genomic differentiation among walnut genotypes from Western and Eastern countries. We believe that Chandler v2.0 is a fundamental genomic resource for the research community, and it will be of high interest for the audience of *GigaScience*.

All genomic and genetic data generated in the present research have been submitted to public databases (NCBI, Hardwood Genomics), and all gene/protein names and symbols used in the paper adhere to approved nomenclature guidelines for *J. regia*.

We kindly request to exclude the following referees:

- Ming-Cheng Luo, University of California, Davis;
- Dong Pei, Chinese Academy of Forestry;
- Zhang Junpei, Chinese Academy of Forestry.

We suggest the following referees:

- Pere Arus, CRAG, Barcelona; pere.arus@cragenomica.es
- Elisabeth Dirlewanger, INRA, Bordeaux; elisabeth.dirlewan@inra.fr

Sincerely,

Annarita Marrano



Ph.D.

Dept. Plant Sciences, University of California, Davis

High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome

--Manuscript Draft--

Manuscript Number:	GIGA-D-19-00363R1
Full Title:	High-quality chromosome-scale assembly of the walnut (<i>Juglans regia</i> L) reference genome
Article Type:	Data Note
Funding Information:	
Abstract:	Background: The release of the first reference genome of walnut (<i>Juglans regia</i> L.) enabled many achievements in the characterization of walnut genetic and functional variation. However, it is highly fragmented, preventing the integration of genetic, transcriptomic, and proteomic information to fully elucidate walnut biological processes. Findings: Here, we report the new chromosome-scale assembly of the walnut reference genome (Chandler v2.0) obtained by combining Oxford Nanopore long-read sequencing with chromosome conformation capture (Hi-C) technology. Relative to the previous reference genome, the new assembly features an 84.4-fold increase in N50 size, with the 16 chromosomal pseudomolecules assembled and representing 95% of its total length. Using full-length transcripts from single-molecule real-time sequencing, we predicted 37,554 gene models, with a mean gene length higher than the previous gene annotations. Most of the new protein-coding genes (90%) presents both start and stop codons, which represents a significant improvement compared to Chandler v1.0 (only 48%). We then tested the potential impact of the new chromosome-level genome on different areas of walnut research. By studying the proteome changes occurring during male flower development, we observed that the virtual proteome obtained from Chandler v2.0 presents fewer artifacts than the previous reference genome, enabling the identification of a new potential pollen allergen in walnut. Also, the new chromosome-scale genome facilitates in-depth studies of intraspecies genetic diversity by revealing previously undetected autozygous regions in Chandler, likely resulting from inbreeding, and 195 genomic regions highly differentiated between Western and Eastern walnut cultivars. Conclusion: Overall, Chandler v2.0 will serve as a valuable resource to understand and explore walnut biology better.
Corresponding Author:	Annarita Marrano, Ph. D. Davis, CA UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Annarita Marrano, Ph. D.
First Author Secondary Information:	
Order of Authors:	Annarita Marrano, Ph. D. Monica Britton Paulo Adriano Zaini Aleksey V. Zimin

	Rachael E. Workman
	Daniela Puiu
	Luca Bianco
	Erica Adele Di Pierro
	Brian J. Allen
	Sandeep Chakraborty
	Michela Troggio
	Charles A. Leslie
	Winston Timp
	Abhaya Dandekar
	Steven L. Salzberg
	David B. Neale
Order of Authors Secondary Information:	
Response to Reviewers:	Reviewer #1: I read the manuscript by Marrano et al. entitled "High-quality chromosome-scale assembly of the walnut (<i>Juglans regia</i> L) reference genome" with interest. The authors describe how they generated a chromosome-scale assembly of <i>J. regia</i> based on ONT, Illumina and Hi-C data. In addition, genetic maps were used to validate and anchor the scaffolds to <i>J. regia</i> linkage groups. The authors performed a wide range of analysis, including gene families of interest and genomic diversity. The article is well written and the methods are sufficiently described. My main concern is the quality of the assembly, although at the chromosome level, the contiguity at the contig level is ten times lower (nearly 1.1Mb) compared to the previously published <i>Juglans</i> genome assemblies (including <i>J. regia</i>). The assembly presented in our manuscript is very high quality, with chromosome-sized scaffolds validated by genetic maps and an N50 contig size of over 1Mb. The assembly was produced on a fairly low budget with older generation Oxford Nanopore reads from the MinION device, which yielded 21.9 Gbp with an average read size of 3.1 kb. (Today we would get reads of 10-20 Kb.) This sequencing data is older and much lower cost than those used by Zhu et al. (2019), who had 57.2 Gbp of PacBio data with an average read size of over 12kb. (see lines 136-140). Even so, a contig N50 size of 1.1 Mb is dramatically larger than our own originally published assembly contigs. I understand that authors choose to use the methods they have developed, however, long-reads assemblies are usually made with dedicated assemblers, which can result in higher assembly quality. Long-read-only data sets are indeed assembled with dedicated methods such as Canu, but when we have both long and short-read data sets, MaSurCA performs better. We did try using the Flye assembler, which is currently one of the best, fastest, and most accurate assemblers for long-read data, and it yielded an assembly with only 530kb N50 contig size, about half of our result. In particular, I was a little surprised by the fact that the v2 assembly contains fewer repetitive elements than the first version of the assembly (L175-176). Generally long-reads assemblies improved the repetitive content of genome assemblies. We re-estimated the repeat content of Chandler v2.0 by running RepeatMasker with the repeat library generated for v1. We found that 58.4% of Chandler v2 is repetitive (we have corrected the main text accordingly), which is higher than what found with the first assembly of the reference genome. Also, Chandler v2.0 is de-duplicated, meaning that almost all of the duplicated regions due to heterozygosity have been removed. These duplicates, which are variants of the same scaffold, were likely mistaken for independent scaffolds in v1.0, overestimating the genome size and, therefore, gene

and repetitive contents. The de-duplication caused a reduction of the genome size from 641 Mb (v1.0) to 573.9 Mb (v2.0), which is now closer to the genome size estimate obtained with Genomescope (488.2 Mb).

The comparison of the Chandler v2 assembly with that provided by Zhu et al. is an important point for the reader, as it will determine which genome will be used for further analysis. As an example, the long-range input data are different (Hi-C vs Optical maps) and maybe specific regions are not of the same quality in both assemblies.

We compared Chandler v2.0 with JSerr_1.0 from Zhu et al (2019). We observed that more than 95% of the two assemblies aligned with sequence identity higher than 98% and exhibited high collinearity, as shown in Figure S4 (see also lines 174-179). This confirms the high quality of both assemblies, even if obtained with different technologies. The availability of both assemblies will facilitate new genomic studies (e.g., pangenome) to understand the morphological and physiological differences among these two walnut varieties better.

Chandler is the most popular cultivar of walnut worldwide. For this reason, it was chosen for sequencing and assembling the first walnut reference genome (Martinez-Garcia et al. 2016), and is used as the standard in many genetic and biological studies (e.g., development of genotyping tools). After four years, we are here proposing a significant improvement of the reference genome. However, a detailed study of the differences between the two assemblies (e.g., structural rearrangements) goes beyond the scope of the present manuscript.

Minor Points:

- * assembly and gene prediction metrics are scattered throughout the manuscript and give a descriptive tone. I think the authors can move these metrics in tables 1 and 2. In addition, contig metrics are not provided in Table 1.

We modified the gene prediction metrics, which changed due to revisions made to conform with NCBI submission requirements. We moved some of the statistics of the gene annotation in Table 2. We also added contig metrics in Table 1.

- * L38: "the full sequence of all 16 chromosomes" : how is this statement validated ?

We changed the sentence to "with the 16 chromosomal pseudomolecules assembled and representing 95% of its total length" (see line 38).

- * L41 and L235: Asserting that the genes are complete based solely on the presence of a start and stop codon is not enough. Please delete the term "full-length". The number of complete BUSCO genes could perhaps be a way to evaluate the proportion of full-length genes.

We removed 'full-length' and added the proportion of BUSCO genes assembled completely (see lines 269-270)

- * L87 and L90: problem with the closing parenthesis.

Done.

- * L97: "...walnut reference genome with unprecedented contiguity...." Please delete this sentence.

Done.

- * L117: a longest read of 992.2Kb is not informative if it does not align.

We are reporting general statistics on the Nanopore sequencing. Such long reads have then been used for scaffolding Illumina reads with MaSurCa.

- * L156-158: The authors should used a kmer approach (Genomescope) to estimate the genome size of both genotypes.

We ran Genomescope using Chandler Illumina pair-end reads generated by Martinez-

Garcia et al., (2016). Results indicate a genome size of 448 Mb, and a level of heterozygosity equal to 0.634%. Our de-duplicated assembly has a total length of 567.2 Mb (> 1 kb), which is a great improvement compared to Chandler v1.0. However, we were unable to run Genomescope for the Serr genome since we could not find any Illumina data on NCBI. We only found 58 PacBio SRA Experiments under the BioProject PRJNA413991 indicated by Zhu et al. (2019). Also, the Illumina reads of Zhu et al. (2019) are of the hybrid 'J.microcarpa x J.regia cv Serr'.

* L239: The proportion of gene models with multiple transcript isoforms is small relative to other plants which may not represent the proportion of genes with alternative splicing. I think the low depth of PACBIO sequencing is the main reason. Please rephrase the sentence to make it clearer.

With the newly revised gene annotation, we observed more gene models with multiple transcript isoforms. We changed the text accordingly, as suggested by RW#1 (see lines 246-248).

* L269-373 : This section is not clear for non-specialist readers.

We modified the paragraph to make it clearer for non-specialist readers (see lines 275-328). We also added Figure S9 showing the different stages of catkin development considered in our analysis.

* L283 : "four developed" ?

We generated proteomes from four tissues (immature catkin, intermediate catkin, mature catkin, and pure pollen). We changed 'developed' to 'analyzed' for clarity (see line 300).

* L343 : Please describe syntelogs.

Done (line 363).

* Figure5A: There may be a problem of alignment between the inner circle and the middle circle (blue region).

We thank the reviewer for highlighting this inconsistency in the figure. We checked our data and modified the image accordingly. The apparent misalignment between the inner circle and the middle circle is actually due to a lack of information – a gap of 4.07 cM – between two haploblocks which follow one another in that region. The inaccuracy was generated by the program used to draw the image, which automatically assigned the inherited segment to the most likely ancestor based on the maximum parsimony algorithm. However, we agree that on some occasions, this can lead to inaccurate results. Therefore, we decided to explicitly represent these regions of missing information using a white color. We also amended the legend of Figure5 to explain what the white sections describe.

* Too many paragraphs end with a sentence such as "support the crucial role of Chandler v2 chromosome-scale assembly".

We removed the sentence and ended the paragraphs differently.

* L463: Please describe how the gaps have been filled.

The assembler MaSurCa used to obtain Chandler hybrid has an internal gap-filling procedure, as described in Zimin et al., 2013 (see line 476-477). In addition, Dovetail uses Illumina reads to close the gaps in the HiRise assembly, as described in lines 500-502.

Reviewer #2: Overall, this genome is a significant advance over the previous one, but there are some points that are discussed in too much breadth, while others are too short for a detailed evaluation. Some of the claims regarding why the new genome assembly is superior over the older one(s) seem rather constructed. The parts that really profit from ONT sequencing - the near-repetitive gene families and the repeat

content have not been expored in detail.

We performed a gene family analysis on Chandler's gene predictions (v1.0, v2.0, and RefSeq). Consistently with the gene prediction results, we observed that v2.0 has, in general, more members per gene family than v1.0, and fewer members than NCBI RefSeq. These results can be due to both the increment of contiguity and the lower gene redundancy of Chandler v2.0 (see lines 254-265).

I realise that this is just a data note, but some clues could help the reader appreciate the current manuscript. What is also missing is a comparison to other published reference genomes in the Fagales s.l..

We compared assembly metrics and BUSCO percentages to other Fagales genomes (see lines 141-142; 192-193; Table S9). However, if the reviewer was referring to comparative genomics studies, then it goes beyond the scope of the present manuscript, which is a Data Note on the improvement of the walnut reference genome. Also, we already provided applications of the new Chandler assembly for proteomic and population genetics studies.

L65. How is this hybridisation possible, given the current disjunction of the populations of the species? Please give a sentence or two as explanation.

Zhang et al. (2019), cited in the text, explained clearly how the contact between American black walnuts and Asian butternuts occurred during the Pliocene. According to their reconstruction, supported by fossil evidence, butternut and black walnuts spread into Eurasia from the late Oligocene to the Miocene-Pliocene. Then, the cooling climate of the Upper Pliocene may have led to range shifts of the butternut and black walnut lineages in Eurasia, permitting the contact required for the hybridization that gave rise to Persian walnut. We add a short sentence describing Zhang et al (2019) results in the text (see line 66).

L87. Actinida is not a tree.

We changed trees to crops (see line 89).

LI122-125. The process of obtaining the megareads is insufficiently described. Please exapand the text and mention also the paramters used.

We added a more detailed description of how MaSurCa builds the mega-reads in the text (lines 466-478).

In addition, please provide statistics for the ONT reads and the illumina reads.

We added the statistics on the ONT reads in Table S1. The Illumina data were generated by Martinez-Garcia et al. (2016), where Illumina sequencing details and statistics are reported.

PLease also mention the library preparation technique used in both cases.

We added details on the ONT library preparation in the method section (Lines 460-464). Regarding the Illumina libraries, their descriptions are reported in Martinez-Garcia et al. (2016).

Please also metnion the known biases associated with ONT sequencing and how strong these were in your raw data.

ONT sequencing tends to have more errors in long homopolymer regions (e.g., runs of all A's). We saw no evidence that these were a particular problem in this genome, although all genomes have such regions. Our use of Illumina reads to compute the final consensus sequence for virtually all positions in the assembly should minimize this problem.

LI126-127. How has the ols assembly (v.1.0) integrated with the new one?

After running the de-duplication module implemented in MaSuRCA, we aligned the v1 scaffolds to v2, identified unaligned regions, and added them to v2. We added this sentence to the text (line 481-483).

L155. Has it been checked, if the unanchored small scaffold are derived from contamination with bacteria/fungi?

All assemblies have contamination, not only from the original samples but from the sequencing lab itself. We ran a thorough contamination screen, aligning every contig and scaffold against an exhaustive set of bacteria, artificial vectors, and other plants and animals, and we removed all contaminants found in this manner. In addition, NCBI runs its own contaminant screen on every submission to GenBank, and that was run on our submission as well.

L170. The identity seems rather low. The possible reasons for this should be given.

We thank the reviewer for noticing this error. We estimated the average sequence identity from the nucmer coord file without filtering. In this way, we considered all alignments, including those between similar but not syntenic regions of the two assemblies. We re-estimated the percentage of sequence identity using only the 1-to-1 best alignments (command dnadiff implemented in MUMmer), and we obtain a value of 99.60. We changed the text accordingly (see lines 181-186).

L172. What was the proportion of unaligned reads? How many reads mapped discordantly?

Over a total of 432,183,992, 2,046,961 reads (0.5%) did not align and 31,169,557 did not pair properly.

L188. This statement cannot be upheld the way it is. Usually the gene space is already well-assembled using only illumina reads (apart from the repetitive genes). The authors should compare the BUSCO scores of several Chandler assembly versions with that of other Fagales genomes, such as oak, beech, and chestnut.

We added Table S9 with BUSCO statistics for the Chandler genome assemblies, J. regia cv Serr, and other Fagales genomes.

L190. There are mapping-based ways to address this. These should be mentioned / applied.

We preferred to remove the sentence instead of following the reviewer's suggestion since revising/applying available methods for improving transcriptome assemblies using short reads is not among the scopes of our work. We aimed to prove how a much contiguous reference genome can improve transcriptome assemblies and gene predictions in walnut.

LI217-247. This seems overly discussed, considering the rather minor differences observed.

We moved some of the statistics in Table 2 and made this paragraph less redundant and descriptive.

L363. This is not necessarily evidence of inbreeding, but could also reflect selective sweeps. Inbreeding does not happen on the sub-genomic level but only on the genomic level.

Sub-genomic inbreeding occurs when an individual inherits the same copy of an allele at one locus from a common ancestor (identical-by-descent; IBD). Chandler parents share Payne as a common ancestor; therefore, there are high probabilities of IBD alleles at a sub-genomic level in Chandler. However, although we found no evidence of a strong selective sweep, it is also possible this pattern was due to direct or indirect selection (see lines 381-384). Future selective sweep studies in larger and more diverse walnut collections could provide more evidence on the high-level of homozygosity in some regions of the Chandler genome.

	L426. Was any surface sterilisation done? Otherwise a lot of contaminant sequences would be expected. In any case, a contamination check should be reported. We could not do surface sterilization since the Nanopore library was built starting from frozen tissue collected at UC Davis and sent to the John Hopkins University. Surface sterilization would have led to tissue degradation and plant production of stress compounds that further impede DNA extraction. Also, we ran a thorough contamination screen of our assembly, to remove contaminations associated with microbes present on the leaf surface and within the tissue (i.e., endophytes). L428. 'g' should be in italics. Done L431. Concentrations/amounts missing. Added. L440. 'was' -> 'were'. Done L456. The assembly strategy, programs and parameters use are not mentioned in sufficient detail (actually hardly any of this is mentioned in the manuscript). We added more details on the assembly strategy in the text (lines 466-478). L531. Do not abbreviate at the beginning of the line. Edit.
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes
Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	
Resources	Yes
A description of all resources used, including antibodies, cell lines, animals and software tools, with enough	

information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

High-quality chromosome-scale assembly of the walnut (*Juglans regia* L.) reference genome

Annarita Marrano¹, amarrano@ucdavis.edu; *corresponding author*

Monica Britton², mtbritton@ucdavis.edu

Paulo A. Zaini¹, pazaini@ucdavis.edu

Aleksey V. Zimin^{3,4}, alekseyz@jhu.edu

Rachael E. Workman³, rachael.e.workman@gmail.com

Daniela Puiu⁴, dpuiu@jhu.edu

Luca Bianco⁵, luca.bianco@fmach.edu

Erica Adele Di Pierro⁵, erica.dipierro@fmach.it

Brian J. Allen¹, brallen@ucdavis.edu

Sandeep Chakraborty¹, sanchak@gmail.com

Michela Troggio⁵, michela.troggio@fmach.it

Charles A. Leslie¹, caleslie@ucdavis.edu

Winston Timp^{3,4}, wtimp@jhu.edu

Abhaya Dandekar¹, amdandekar@ucdavis.edu

Steven L. Salzberg^{3,4,6}, salzberg@jhu.edu

David B. Neale¹, dbneale@ucdavis.edu

¹ Department of Plant Sciences, University of California, Davis, CA 95616, USA

² Bioinformatics Core Facility, Genome Center, University of California Davis, CA 95616, USA

³ Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205, USA

⁴ Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21205, USA

⁵ Research and Innovation Center, Fondazione Edmund Mach, San Michele all'Adige (TN) 38010, Italy

⁶ Departments of Computer Science and Biostatistics, Johns Hopkins University, Baltimore, MD 21218

Abstract

Background: The release of the first reference genome of walnut (*Juglans regia* L.) enabled many achievements in the characterization of walnut genetic and functional variation. However, it is highly fragmented, preventing the integration of genetic, transcriptomic, and proteomic information to fully elucidate walnut biological processes. **Findings:** Here, we report the new chromosome-scale assembly of the walnut reference genome (Chandler v2.0) obtained by combining Oxford Nanopore long-read sequencing with chromosome conformation capture (Hi-C) technology. Relative to the previous reference genome, the new assembly features an 84.4-fold increase in N50 size, with the 16 chromosomal pseudomolecules assembled and representing 95% of its total length. Using full-length transcripts from single-molecule real-time sequencing, we predicted 37,554 gene models, with a mean gene length higher than the previous gene annotations. Most of the new protein-coding genes (90%) presents both start and stop codons, which represents a significant improvement compared to Chandler v1.0 (only 48%). We then tested the potential impact of the new chromosome-level genome on different areas of walnut research. By studying the proteome changes occurring during male flower development, we observed that the virtual proteome obtained from Chandler v2.0 presents fewer artifacts than the previous reference genome, enabling the identification of a new potential pollen allergen in walnut. Also, the new chromosome-scale genome facilitates in-depth studies of intraspecies genetic diversity by revealing previously undetected autozygous regions in Chandler, likely resulting from inbreeding, and 195 genomic regions highly differentiated between Western and Eastern walnut cultivars. **Conclusion:** Overall, Chandler v2.0 will serve as a valuable resource to understand and explore walnut biology better.

Keywords: Nanopore, Hi-C, IsoSeq, gene prediction, genetic diversity, proteome, allergens.

53

54 **Introduction**

55 Persian walnut (*Juglans regia* L.) is among the top three most-consumed nuts in the world, and
56 over the last ten years, its global production increased by 37% (International Nut and Dried Fruit
57 Council, 2019). Its richness in alpha-linolenic acid (ALA), proteins, minerals, and vitamins, along
58 with documented benefits for human health, explains this increased interest in walnut consumption
59 [1]. As suggested by its generic name *Juglans* from the Latin appellation ‘*Jovis glans*’, which
60 loosely means ‘nut of gods’, the culinary and medical value of Persian walnut was already widely
61 prized by ancient civilizations [2].

62 The origin and evolution of the Persian walnut are the results of a complex interplay between
63 hybridization, human migration, and biogeographical forces [3]. A recent phylogenomic analysis
64 revealed that Persian walnut (and its landrace *J. sigillata*) arose from an ancient hybridization
65 occurred between American black walnuts and Asian butternuts after a climate-driven range
66 expansion in Eurasia during the Pliocene [4]. Evidence suggests that the mountains of Central Asia
67 were the cradle of domestication of Persian walnut [5], from where it spread to the rest of Asia,
68 the Balkans, Europe, and, finally, the Americas.

69 Today, walnut is cultivated worldwide in an area of 1,587,566 ha, mostly in China and the USA
70 (FAOSTAT statistics, 2017). Considerable phenotypic and genetic variability can be observed in
71 this wide distribution area, especially in the Eastern countries, where walnuts can still be found in
72 wild fruit forests. Many studies on genetic diversity in walnut have outlined a genetic
73 differentiation between Eastern and Western genotypes [6,7]. Moreover, walnuts from Eastern

Europe, Central Asia, and China exhibit higher genetic diversity and a higher number of rare alleles than the genotypes from Western countries [8].

The release of the first reference genome, Chandler v1.0 [9], enabled the study of walnut genetics at a genome-wide scale. For the first time, it was possible to explore the gene space of Persian walnut with the prediction of 32,498 gene models, providing the basis to untangle complex phenotypic pathways, such as those responsible for the synthesis of phenolic compounds. The availability of a reference genome marked the beginning of a genomics phase in Persian walnut, allowing whole-genome resequencing [4,10], the development of high-density genotyping tools [7,11], and the genetic dissection of important agronomical traits in walnut [12–15]. However, the Chandler v1.0 assembly is highly fragmented, compromising the accuracy of gene prediction and the fulfillment of advanced genomics studies necessary to resolve many, still unanswered questions in walnut research.

The recent introduction of long-read sequencing technologies and long-range scaffolding methods has enabled chromosome-scale assembly for multiple plant species, including highly heterozygous crops such as almond (*Prunus dulcis*; [16]) and kiwifruit (*Actinidia eriantha*; [17]). The availability of genomes with fully assembled chromosomes provides foundations for understanding plant domestication and evolution [16,18,19], the mechanisms governing important traits (e.g., flower color and scent; [20]), as well as the impact of epigenetic modifications on phenotypic variability [21]. Recently, Zhu et al., (2019) assembled the parental genomes of a hybrid *J. microcarpa* × *J. regia* (cv. Serr) at the chromosome-scale using long-read PacBio sequencing and optical mapping. They relied on the haplotype divergence between the two *Juglans* species and demonstrated an ongoing asymmetric fractionation of the two subgenomes present in *Juglans* genomes.

Here we report a new chromosome-level assembly of the walnut reference genome, Chandler v2.0, which we obtained by combining Oxford Nanopore long-read sequencing [23] with chromosome conformation capture (Hi-C) technology [24]. Thanks to the increased contiguity of Chandler v2.0, we were able to substantially improve gene prediction accuracy, with new, longer gene models identified and many fewer artifacts compared to Chandler v1.0. Also, the availability of full, chromosomal sequences reveals new genetic diversity of Chandler, previously inaccessible through standard genotyping tools, and significant genetic differentiation between Western and Eastern walnuts at 195 genomic regions, including also loci involved in nut shape and harvest date. In the present research, we demonstrate the fundamental role of a chromosome-scale reference genome to integrate transcriptomics, population genetics, and proteomics, which in turn enable a better understanding of walnut biology.

Genome long-read sequencing and assembly

To increase the contiguity of the Chandler genome, we first generated deep sequence coverage using Oxford Nanopore Technology (ONT), a cost-effective long-read sequencing approach that determines DNA bases by measuring the changes in electrical conductivity generated while DNA fragments pass a tiny biological pore [25]. Since the release of the first plant genome assembly generated using ONT sequencing [26], this technology has been applied to sequence and obtain chromosome-scale genomes of many other plant species [27–29]. In Persian walnut, ONT sequencing yielded 7,096,311 reads that provided 21.9 Gbp of sequence, or ~35X genome coverage (assuming a genome size of 620 Mb). Read lengths averaged 3.1 kb, and the N50 read length was 6.7 kb, with the longest read being 992.2 kb (**Additional file 1: Table S1**).

One of the major limitations of long-read sequencing technologies is their high error rate, which can range between 5% and 15% for Nanopore sequencing [30]. To overcome this limitation, we

adopted the hybrid assembly technique incorporated into the MaSuRCA assembler, which combines long, high-error reads with shorter but much more accurate Illumina sequencing reads to generate a robust, highly contiguous genome assembly [31]. First, using the Illumina reads, we created 3.7 million 'super-reads' with a total length of 2.9 Gb. We then combined the super-reads with the ONT reads to generate 3.2 million mega-reads with a mean length of 4.7 kb, representing 24X genome coverage (**Additional file 1: Table S2**). Finally, we assembled the mega-reads to obtain the 'hybrid' Illumina-ONT assembly, which comprised 1,498 scaffolds, 258 contigs, and 25,007 old scaffolds from Chandler v1.0 (**Additional file 1: Table S3**).

Even though the total number of scaffolds (> 1 Kb) was reduced by 80% compared to Chandler v1.0 (**Table 1**), the new hybrid assembly was still fragmented. To improve the assembly further and build chromosome-scale scaffolds, we applied Hi-C sequencing, which is based on proximity ligation of DNA fragments in their natural conformation within the nucleus [24]. The HiRise scaffolding pipeline processed 356 million paired-end 100-bp Illumina reads to generate the HiRise assembly (**Table 1**). The top 17 scaffolds from this assembly spanned more than 90% of the total assembly length, with a scaffold length ranging from 19.6 to 45.2 Mb (**Additional file 1: Figures S1-S2**). As shown in **Table 1**, the Chandler genome contiguity increased dramatically compared to the previous assemblies. As compared to the recently published genome assembly of the walnut cultivar Serr [22], Chandler v2.0 was less contiguous at the contig level, with a N50 size of 1.1 Mb against the 15.1 Mb of JrSerr_v1.0. The higher coverage PacBio sequencing data (57.2 Gbp) used to assemble JrSerr_v1.0 may explain this discrepancy in contiguity between the two assemblies. Besides, our assembly presented similar value of contiguity to the recently published genomes of pecan (*Carya illinoensis*; 1.1 Mb [32]), Chinese chestnut (*Castanea mollissima*; 944.4 kb [33]) and oak (*Quercus robur*; 1.35 Mb [34]).

Validation of the HiRise assembly

To assess the quality of the HiRise assembly, we used two independent sources of data. First, we used the single nucleotide polymorphism (SNP) markers mapped on the high-density genetic map of Chandler recently described by [14]. Out of the 8,080 SNPs mapped into 16 linkage groups (LGs), 6,894 had probes aligning uniquely on the HiRise assembly with 98% of identity for more than 95% of their length. A total of 35 scaffolds of the HiRise assembly could be anchored to a chromosomal linkage group by at least one SNP (**Figure 1**). In particular, 13 LGs were spanned by a single HiRise scaffold, while two to three scaffolds each aligned the remaining three LGs.

Second, we anchored the HiRise assembly to the Chandler genetic map used by [35] to construct a walnut physical map. In total, 972 of the mapped markers (1,525 SNPs) aligned uniquely on the same 35 HiRise scaffolds anchored to the linkage map mentioned above. Overall, we observed almost perfect collinearity between the HiRise assembly and both Chandler genetic maps (**Figure 1**, **Additional file 1: Figure S3**). Therefore, we oriented, ordered, and named the HiRise scaffolds consistent with the linkage map of [35], generating the final 16 chromosomal pseudomolecules of *J. regia* Chandler.

These 16 contiguous chromosomal scaffolds account for 95% of the final walnut reference genome v2.0, with an N50 scaffold size of 37 Mb. We identified telomere sequences at both ends for nine of the chromosome scaffolds, on one end of the other seven chromosomes and one end of seven unanchored scaffolds. Also, all 16 chromosomes had centromeric repeats in the middle, alongside regions with low recombination rates (**Figure 2**).

As compared to the previous Chandler genome assemblies (**Table 1**), Chandler v2.0 had a smaller genome size (573.9 Mb), much closer to the Genomescope estimate of 488.2 Mb. This reduction

in genome size represents a great improvement of Chandler v2.0 and can be related to the removal of haplotype variants, likely interpreted and annotated as different scaffolds in the previous genome versions. Compared to the Serr walnut genome (JrSerr_v1.0; 534.7 Mb) [22], Chandler v2.0 had a larger genome size, likely due to structural variation (e.g., copy number and presence/absence variants), whose central role in explaining intraspecific genomic and phenotypic diversity has been reported in different plant species [36,37]. In addition, the higher number of unanchored scaffolds (2,631; 20.9 Mb) in Chandler v2.0 compared to JrSerr_v1.0 can represent autozygous genomic regions of Chandler, devoid of segregating markers and, therefore, difficult to anchor to linkage genetic maps [35], as also suggested by the higher fixation index (F) observed in Chandler (0.03) than Serr (-0.29) in previous genetic surveys [7]. The two walnut assemblies, though, aligned with high sequence identity (over 98% for more than 95% of their total length) and showed high collinearity (**Additional file 1: Figure S4**). Future comparative genomics studies will provide further insights on the functional and structural differences between the two genome assemblies, and their explanatory involvement in the morphological and physiological variation of these two walnut cultivars.

To assess the sequence accuracy of Chandler v2.0, we first compared the scaffold sequences of Chandler v2.0 with the previous version of the walnut reference genome. About 578 Mb of sequence were mutual best alignments, namely best hits of each location between Chandler v2.0 and v1.0 and vice versa, with a sequence identity of 99.6%. We also observed that 135 Mb of Chandler v1.0 (18.9%) aligned to the same locations in Chandler v2.0, suggesting the presence of redundant haplotypes in the previous version of the walnut reference genome that have been removed in our assembly. We then mapped the Illumina whole-genome shotgun data [9] against the new chromosome-scale genome. The alignment resulted in 64,950,691,681 bps mapped, of

which 407,450,406 were single-base mismatches, consistent with an Illumina sequence accuracy rate of 99.5%.

Repeat annotation

More than half (58.4%) of the new Chandler v2.0 is repetitive. This estimate is higher than the previous version of the walnut reference genome (51.19%) and comparable with other *Fagales* genomes [34,38]. As in most plant genomes, interspersed repeats were the most abundant type of repeats, with retrotransposons at 36.45% and DNA transposons at 15.86%. *Gypsies* (10.5%) and *Copias* (7.69%) were the most represented classes of long-terminal retrotransposons (LTR), and, though widely dispersed throughout the genome, they were distributed differently along the 16 chromosomes (Additional file 1: Figure S5): the *Gypsies* LTRs were more abundant alongside the centromeres, where, instead, the density of the *Copia* LTRs decreased, as previously observed in walnut [22]. The long-interspersed nuclear elements (L1/LINE), which possess a poly(A) tail and two open reading frames (ORFs) for autonomous retrotransposition, was the largest class of non-LTRs at 7.14% of the genome. Simple repeats (1.91%) were also found.

PacBio IsoSeq sequencing and gene annotation

A fragmented reference genome can severely hamper the accuracy of gene prediction, because many genes will be broken across multiple small contigs (false negatives), and because heterozygous gene variants may be annotated separately (false positives).

To improve the gene prediction accuracy of Chandler v2.0, we used the “Isoform Sequencing” (Iso-Seq) method, developed by Pacific Biosciences (PacBio), which can generate full-length transcripts up to 10 kb, allowing for accurate determination of exon-intron structure by the alignment of the transcripts to the assembly [39]. The high error rate of PacBio sequencing can be

greatly reduced using circular consensus sequence (CCS), in which a transcript is circularized and then sequenced repeatedly to self-correct the errors. We applied PacBio IsoSeq to sequence full-length transcripts from nine tissues, chosen to cover most of the transcript diversity in walnut (**Additional file 1: Table S4**). Across the four SMRT cells, we obtained 26,328,087 subreads with a mean length of 1,188 bp (**Additional file 1: Table S5**) and CCSs ranging from 13K to 142K per library (**Additional file 1: Table S6**). Out of the 745,730 full-length non-chimeric (FLnc) transcripts, 68,225 were classified as high quality, FL (HQ FL) consensus transcript sequences, with an average length of 1,357 bp (**Additional file 1: Table S6**). Catkin 1-inch elongated (CAT1), shoot, and root yielded the lowest number of HQ FL transcripts, while pollen and leaf had the lowest number of HQ consensus clusters obtained per CCS after polishing (**Additional file 1: Table S6**). These results can be explained by lower cDNA quality or fewer inserts of full-length transcripts from these tissues during the cDNA pooling and library preparation. Nevertheless, more than 99% of the HQ FL transcripts aligned onto the new chromosomal-level walnut reference genome (**Additional file 1: Table S7**).

By combining the HQ FL transcripts with available *Juglans* transcriptome sequences, we identified 37,554 gene models, which are more than those annotated in Chandler v1.0 but fewer than the predicted genes in the NCBI RefSeq *J. regia* annotation generated with the first version of the reference genome (**Table 2**). Thus, the new chromosome-scale genome, along with the availability of full-length transcripts, allowed us to identify genes missed in Chandler v1.0 due to genome fragmentation, as well as to remove false-positive predictions likely caused by heterologous variants of the same locus mistakenly interpreted and annotated as independent scaffolds in Chandler v1.0. Also, the mean gene length in Chandler v2.0 was higher than the previous gene annotations (**Table 2**), a consequence of the increased contiguity of the new chromosome-scale

reference genome. The average gene density of Chandler v2.0 was 19.75 genes per 100 kb, with higher gene content in the proximity of telomeric regions (Figure 2), consistent with other plant genomes [19,40]. The majority of the predicted gene models of Chandler v2.0 was supported by expression data, and showed high similarity with a protein-coding transcript of other plant species (Additional file 1: Table S8). Also, 30,318 models were annotated with 8,243 different Gene Ontology (GO) terms (Additional file 1: Figures S6-S8).

Out of the 40,884 transcripts identified, 84% were multi-exonic, with 5.9 exon each, on average, and longer introns than the previous gene annotations of Chandler (Table 2). The majority of intron/exon junctions were GT/AG-motif (98.2%), even though alternative splicing with non-canonical motifs was also observed (GC/AG – 0.8%; AT/AC – 0.11%). Almost 90% (36,422) of the coding sequences presented both canonical start and stop codons, while 4,462 had either a start or a stop codon. This result represents a great improvement compared to Chandler v1.0, where only 48% of the predicted gene models presented both start and stop codons [9].

Also, we observed that 2,801 gene models had from two to four transcript isoforms each, with a mean length of 9,389 bp. This proportion of gene models with multiple transcript isoforms is smaller compared to other plant species [41,42], likely due to the low depth of coverage of our PacBio sequencing. Out of the 6,437 isoforms identified, 1,448 were covered by FL HQ transcripts in at least one tissue, while 5,689 were expressed in at least one of the 20 tissues [9], which most likely covered higher gene diversity compared to the nine tissues used for PacBio IsoSeq. On average, the Illumina isoforms (9,188bp) were longer than the PacBio isoforms (6,790 bp). By running the EnTAP functional annotation pipeline with the entire NCBI RefSeq plant database [43], we observed that almost all isoforms (98%; 6,287) were annotated with a plant protein.

We also investigated possible gene family expansion and contraction among the three Chandler's gene annotations. Overall, we identified fewer gene families in Chandler v2.0 (5,163 Panther family represented) than v1.0 (5,330) and NCBI RefSeq *J.regia* annotation (5,374). However, when counting the number of members per family, we observed a gene family expansion, in general, in Chandler v2.0 compared to v1.0: 39,357 proteins were assigned to a Panther gene family in Chandler v2.0, with an average of 7.6 members per family, against the 30,639 proteins annotated with a Panther domain in v1.0 (6 members per family on average). On the contrary, we noticed an overall gene family contraction in v2.0 compared to NCBI RefSeq, where 10.4 gene members were assigned to a Panther domain on average. Both the increment of contiguity and the reduction in haplotype redundancy can explain the observed patterns of gene family expansion and contraction among the three Chandler's gene annotations, even if the different methods of gene prediction used in the three studies could also account for these differences.

Most of the 1,440 core genes in the embryophyte dataset from Benchmarking Universal Single-Copy Orthologs (BUSCO) were assembled completely (82.5% single-copy; 12.6% duplicated), similarly to other *Fagales* genome assemblies [32,33,38,44] (Additional file 1: Table S9). Also, 88% of both rosids and green sets of core gene families (coreGFs) were identified in the gene annotation, confirming the high-quality and completeness of the gene space of Chandler v2.0.

Improved assessment of proteomes with the complete genome sequence

After confirming the importance of a chromosome-scale reference genome for the improvement of gene prediction accuracy, we analyzed the impact of a contiguous genome sequence using proteomic analysis. Proteomes are commonly investigated by isolating the total protein complement of a sample and fragmenting those proteins into smaller peptides that are resolved by mass and charge by mass spectrometry. After detection, the peptides' amino acid sequences are

determined by matching their mass and charge to candidate sequences obtained from a reference proteome inferred from the reference genome (virtual proteome). A fragmented assembly of the reference genome can lead to an inaccurate prediction of a species' proteome and, then, a mis-identification of the proteins expressed in specific tissues at particular stages [45].

We isolated proteins of reproductive tissues harvested from mature Chandler walnut trees, focusing on different development stages of the male flower (catkin; **Additional file 1: Figure S9**) and mature pollen grains. We analyzed the proteomic data generated from these samples using the virtual proteomes predicted from the gene annotation of the new chromosome-scale genome and Chandler v1.0 (NCBI RefSeq). Considering all tissues analyzed, we identified fewer unique peptides (43,083) with the new chromosome-scale walnut genome than with Chandler v1.0 (44,679). In addition, 6,966 unique proteins were detected with Chandler v2.0 against the 8,802 found using version 1 as a search database (**Additional file 2: Table S10; Additional file 3**). Most likely, the NCBI proteomic database based on the fragmented Chandler v1.0 included artifacts resulting from an overestimation of the protein-coding genes.

In the example presented below, we focused on the allergenic proteins produced during catkin and pollen development. Approximately 2% of walnut consumers have a high risk of developing allergies to nuts or pollen [46]. Initially, we clustered the samples according to their protein constituents and levels. This revealed a higher similarity between immature and mature catkins and a more distinct pattern of detected proteins between senescent catkins and pure pollen (**Figure 3**).

We then searched the four analyzed proteomes for allergenic proteins listed in the WHO/IUIS Allergen Database (www.allergen.org; **Additional file 2: Table S11**), as well as for additional proteins not yet registered in the allergen database but predicted in Chandler v2.0 as potential

allergens given their predicted structural similarity to known allergens (**Additional file 3**). Four of the eight recognized allergenic proteins were detected in at least one of the catkin developmental stages, with Jug_r_5 (XP_018825777 | *Jr12_10750*) and Jug_r_7 (XP_018808763 | *Jr07_28960*) present in all sample types, including pollen (**Additional file 2: Table S11**). Genes adjacent to known allergen-coding sequences, likely indicating gene duplications, encode three of the new potential allergens (**Additional file 2: Table S11**). Moreover, we discovered that the gene locus *Jr12_05180* encodes a non-specific lipid transfer protein (nsLTP; Jug_r_9 | XP_018813928), a potential allergen highly expressed during catkin maturation and in pollen (**Additional file 2: Table S11-12**). In particular, Jug_r_9 was the most abundant protein in mature and senescent catkins, and the second most abundant in pure pollen. Another interesting allergen similar to Jug_r_9 (same eight cysteine configuration) is XP_018814382 | *Jr03_26970*; it decreases as the catkin matures, and is absent in pollen (**Additional file 2: Table S11-12**). Similarly, polyphenol oxidase (PPO, XP_018858848 | *Jr03_06780*) is high in the immature catkin and almost absent in the pollen.

The integration of this proteomic data with previously published transcriptomic data obtained from 20 walnut tissues [9] shows high reproducibility between the methods. In both datasets, allergens Jug_r_1, 4, and 6 were not detected in catkins, while the new putative allergen Jug_r_9 was highly expressed in catkins (**Additional file 2: Table S12-13**). Also, *Jr12_05180* transcripts were not detected in any of the 20 tissues but catkin, thus confirming the strong specificity of Jug_r_9 for catkin and pollen tissue (**Additional file 2: Table S13**). Modeling the structure of this putative allergen reveals four predicted disulfide bonds, potentially conferring heat and protease-resistance, and further suggesting allergenic properties (**Figure 4**). Future studies will clarify the functional role of this protein and its allergenic nature.

The detection of new potential walnut allergens confirms the positive impact of Chandler v2.0 on proteomic studies in walnut, by providing a clearer and more precise organization of the CDSs within a genomic region than the previous fragmented genome assembly v1.0.

Chandler genomic diversity

By anchoring the HiRise assembly to the Chandler genetic map [14], we observed highly homozygous regions in Chandler, especially on Chr15, where the genetic gap spanned 14.5 cM, corresponding to a physical distance of 9.1 Mb. A large gap on Chr15 (9.23 cM – 1.5 Mb) was also observed by [35], which suggested inbreeding as a possible cause for the lack of segregating loci in this region in Chandler, whose parents shared Payne as an ancestor. To confirm the autozygosity of Chandler on Chr15, we used the Illumina whole-genome shotgun data of Chandler and the identified polymorphisms to study its genetic diversity across the new chromosome-scale genome. We identified 2,205,835 single heterozygous polymorphisms on the 16 chromosomal pseudomolecules, with an SNP density of 4.0 SNPs per kb (**Figure 2; Additional file 1: Table S14**). Fifty-six 1-Mb-regions exhibited less than 377.5 SNPs (10th percentile of the genome-wide SNP number distribution), and chromosomes 15, 1, 7, and 13 were the top four chromosomes in the number of low heterozygous regions (**Additional file 1: Table S15**). In particular, Chr15 presented nine 1-Mb windows with a significantly low number of polymorphisms, five of which span 4 Mb at the end of the chromosome. In these nine low heterozygous regions, we found 1,536 SNPs in total (**Figure 2**), of which only 25 were tiled on the Axiom *J. regia* 700K SNPs array. The absence of these polymorphisms segregating in Chandler in the SNP array could be related to either a failed identification during the SNP calling due to the highly fragmented reference genome v1.0 or with the SNP exclusion during the filtering process applied to build the genotyping array [7]. The low number of Chandler heterozygous SNPs in the array affected the end of Chr15 the

most, causing a reduction in the genetic length of the corresponding linkage group (**Figure 1**), as well as leaving unexplored 4 Mb of Chandler genetic variability, which is now accessible thanks to the new chromosome-scale reference genome. The failure to anchor seven of the scaffolds with telomeric sequences can be explained by the missed detection of terminally located highly homozygous regions during genetic map constructions, due to the absence of crossing-over events with heterozygous flanking markers.

Due to the evidence of whole-genome duplication in *Juglans* genomes [35], we searched for conserved regions of synteny between Chr15 and its homologous regions in the genome, to study their level of divergence and identify other evolutionary forces as possible causes of the localized reduction of heterozygosity on Chr15. Of the 5,739 pairs of paralogous genes (8,701 genes; **Additional file 1: Figure S10**) identified in Chandler v2.0, 448 included genes on Chr15, and 389 of these have their respective paralogues on Chr6 (**Additional file 1: Figure S11**), in line with what was already reported by [35]. The Chr06-Chr15 pairs of paralogous genes showed average values of divergence indexes ($K_S = 0.38$; $K_A = 0.13$) similar to the ones observed genome-wide for other syntelogs ($K_S = 0.4$; $K_A = 0.09$), **which are paralogous genes derived from the same ancestral genomic region**. Similar values of divergence were also observed for the 178 Chr06-Chr15 syntelogs (171 genes) falling within the nine low heterozygous regions on Chr15 ($K_S = 0.4$, $K_A = 0.1$), **excluding different evolutionary rates for these regions**. Other than paralogous genes, we found 393 singleton genes in the low heterozygous regions on Chr15 of Chandler. These genes are involved in different biological processes, many of which related to signal transduction, protein phosphorylation, and response to environmental stimuli (**Additional file 1: Table S16**).

We further investigated the contribution of inbreeding to the high level of autozygosity on Chr15 by visualizing the inheritance of haplotype-blocks (HB; genomic regions with little recombination)

across the Chandler pedigree (**Figure 5B**, **Additional file 1: Figure S12**). We observed that Payne accounts for the entire Chandler genetic makeup (19 HBs for the total length of Chr15) inherited from Pedro (mother), where only one HB (2,08 Mb) shared the same allele of Conway-Mayette (maternal-grandfather; **Figure 5A**). Regarding the paternal genetic makeup of Chandler, 13 out of 19 HBs (9,05 Mb) on Chr15 inherited Payne alleles, providing further evidence of high inbreeding on this chromosome (**Figure 5A**). This is even more evident in assessing the number of alleles matching between Payne and Chandler across the genome: Chr15 (14 HBs for a total of 13,95 Mb; **Figure 6**) shares full allele identity with Payne for almost its entire length. Such allele matching between Chandler and its ancestor Payne also occurs on Chr1 (9 HBs for a total of 8,44 Mb), Chr4 (6 HBs - 7,68 Mb), Chr7 (21 HBs - 21,62Mb) and Chr14 (7 HBs – 12,29 Mb). **These results suggest high level of inbreeding in many genomic regions of Chandler (Additional file 1: Figure S12), even though direct and indirect selection might have caused the observed presence of extended homozygous regions in Chandler's genome.**

Genomic comparison between Eastern and Western walnuts

Even though numerous surveys regarding genetic diversity within walnut germplasm collections have been reported so far [47,48], comparative analyses at the population level and genome scans for signatures of selection are still missing in Persian walnut. The availability of a chromosome-scale reference genome enables exploration of the patterns of intraspecific variation at the genomic level, providing new insight on the extraordinary phenotypic diversity present within *J. regia*.

We used the resequencing data generated for 23 founders of the Walnut Improvement Program of the University of California, Davis (UCD-WIP; **Additional file 1: Table S17**) [10] to study the genome-wide genetic differentiation among walnut genotypes of different geographical provenance. We identified 14,988,422 SNPs, and over 97% of them were distributed on the 16

chromosomal pseudomolecules, with 9.4 polymorphisms per kb. A hierarchical clustering analysis (Additional file 1: Figure S13) divided the 23 founders into two major groups, including genotypes from western countries (USA, France, and Bulgaria) and Asia (China, Japan, Afghanistan), respectively, as previously reported [7,49]. High phenotypic diversity for many traits of interest in walnut, such as phenology, nut quality, and yield, has been observed within and between germplasm collections from Western and Eastern countries [50]. Walnut trees from Asia are noted for their lateral fruitfulness and precocity, rarely observed in the USA and western Europe, so that they have been used as a source of these phenotypes in different walnut breeding programs [51].

At a genomic level, we found a moderate differentiation ($F_{ST} = 0.15$) between Western and Eastern genotypes, except for 195 genomic windows (100 kb) that showed substantially high population differences ($F_{ST} \geq 0.36$; top 5% in the whole genome). In particular, chromosomes 7, 5, 1, 4, and 2 presented about 70% of the divergent sites (Figure 2; Additional file 1: Figure S14). As suggested by the mean reduction of diversity coefficient (ROD) value (0.41), in most of the genomic regions highly differentiated, the UCD-WIP founders from the USA and Europe showed lower nucleotide diversity ($\pi = 2.5 \times 10^{-4}$) than the Asian genotypes ($\pi = 5.0 \times 10^{-4}$), consistent with [8] (Figure 2; Additional file 1: Figure S14). The proximity of our eastern genotypes to the supposed walnut center of domestication in Central Asia can explain the high level of diversity observed in this subgroup.

More than 60% (122) of the highly differentiated windows showed a negative value of Tajima's D in the EU/USA subgroup ($D_{Occ} = -1.12$), thus, suggesting that selection has been likely acting on these genomic regions in the Western genotypes (Additional file 1: Figure S14). Here we found 743 genes, with GO biological categories mostly related to signal transduction, embryo

415 development, and response to stresses (**Additional file 1: Table S18**). Ten candidate selective
416 sweeps ($D_{Asia} = -0.54$) were also observed in the Eastern group (**Additional file 1: Figure S14**),
417 which included 57 predicted genes, related to terpenoid biosynthesis, post-embryonic
418 development, and signal transduction (**Additional file 1: Table S19**).

419 Recently, many marker-trait associations have been reported for different traits of interest in
420 walnut, such as leafing date, nut-related phenotypes, and water use efficiency [12–14]. We looked
421 to see if any of these trait-associated SNPs fell within regions highly differentiated between
422 Western and Eastern genotypes. Three loci associated with shape index, nut roundness, and nut
423 shape [12] are located in two genomic regions on chromosome 3 and 4 with significantly high
424 values of F_{ST} (**Additional file 1: Table S20**). In both of these regions, Western genotypes
425 presented lower genetic diversity and lower values of Tajima's D than the Eastern walnuts. These
426 findings may suggest that, while a selective pressure for nut shape may have occurred in the
427 EU/USA subgroups, higher phenotypic variability can be expected for these traits in the Eastern
428 countries. We also found that the locus AX-170770379, strongly associated with harvesting date
429 [14], falls within a genomic region on Chr1 with an F_{ST} value equal to 0.39 and lower genetic
430 diversity in the western genotypes ($ROD = 0.63$; **Additional file 1: Table S20**). Looking at the
431 phenotypic effect of this SNP on the harvest date of the 23 founders, we observed that most of the
432 western genotypes are later harvesting than the eastern (**Additional file 1: Figure S15**), suggesting
433 differences in the timing of phenological events between these two groups as adaptation to the
434 different climate conditions present in their countries of origin [52].

435 Future resequencing projects involving larger walnut collections and covering a wider area of the
436 global walnut distribution are necessary to confirm and interpret the observed genomic

differentiation between Western and Eastern walnuts, likely helping to understand the role of this genomic divergence in the evolutionary history of Persian walnut.

Methods

Oxford Nanopore sequencing and assembly

High molecular weight (HMW) DNA for Nanopore sequencing (Oxford Nanopore Technologies Inc., UK) was isolated through a nuclei extraction and lysis protocol. First, mature leaf tissue from the same tree used for the original *J. regia* Chandler genome [9] was homogenized with mortar and pestle in liquid nitrogen until well ground, then added to the Nuclei Isolation Buffer [53], and stirred at 4°C for 10 minutes. The cellular homogenate was filtered through 5 layers of Miracloth (Millipore-Sigma) into a 50 mL Falcon tube, then centrifuged at 4°C for 20 minutes at 3000 $\times g$. This speed of centrifugation was selected based on the estimated walnut genome size of 1 Gb [54]. Extracted nuclei were then lysed for 30 minutes at 65°C in the SDS-based lysis buffer described by [55]. Afterwards, 0.3 volumes of 5M potassium acetate were added to the lysate to precipitate residual polysaccharides and proteins. The sample was incubated for 5 minutes at 4°C and then centrifuged at 4°C for 10 minutes at 2400 $\times g$. After removing the supernatant, genomic DNA (gDNA) was ethanol precipitated, and then eluted in 10 mM Tris-Cl. Further purification of the gDNA was then performed using a Zymo Genomic DNA Clean and Concentrate column.

One μg of the isolated gDNA was prepared for sequencing using the Ligation sequencing kit (LSK108, Oxford Nanopore) following manufacturer's protocol with an optimized end repair (100 μl sample, 14 μl enzyme, 6 μl enzyme, incubated at 20°C for 20 minutes then 65°C for 20 minutes). In detail, the gDNA was end polished using the NEBNext® Ultra™ II DNA Library Prep Kit, and then cleaned up with 1X Ampure XP beads (Beckman Coulter). Afterwards, the gDNA was ligated

to Oxford Nanopore specific adapters, followed by an additional cleanup with 0.4X Ampure XP beads. Finally, the libraries were sequenced for 48 hours on six flowcells of the Oxford Nanopore Mk1B MinION platform with the R9.4 chemistry. Raw fast5 data were base-called using Albacore version 1.25.

The ONT data and Illumina reads from [9] were combined to obtain the Chandler hybrid assembly using MaSuRCA v3.2.3 [56]. In detail, MaSurCa first transformed the Illumina pair-end reads in *super-reads* using the super-reads algorithm, which uses k-mers from Illumina reads to extend each Illumina read uniquely in both directions. Then, each ONT read was used as a template to which super-reads can be attached, and the approximate alignments of all super-reads to each ONT read were computed. The best path of the exactly overlapping aligned super-reads on a ONT read was then defined, generating a *mega-read*. The mega-reads typically have a very low error rate (less than 1%) since they are constructed from the super-reads, and most of them span the full length of the long reads. Finally, a customized version of the CABOG assembler [57] was used to assemble the mega-reads along with the Illumina mate pairs, which provide the linking information for the scaffolding. Gaps were closed using the gap-filling procedure implemented in MaSurCa and described by [56]. The de-duplication module implemented in MaSurCa was then applied to remove duplicative sequences (scaffold variants due to heterozygosity).

De-duplicated scaffolds were aligned onto the previously finished *J. regia* chloroplast genome [9] using "minimap2 -x asm5", as well as to a database of 223 finished plant mitochondria (downloaded from NCBI RefSeq) using blastn with default parameters. Finally, Chandler v1.0 was aligned to the de-duplicated hybrid assembly, and the unaligned regions were added to the Chandler hybrid assembly.

Hi-C sequencing

A Hi-C library was prepared by Dovetail Genomics LLC (Santa Cruz, CA, USA) as described previously [58]. Briefly, for each library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Biotin that was not internal to ligated fragments was removed from the purified DNA. Purified DNA was then sheared to ~350 bp mean fragment size. Sequencing libraries were generated using NEBNext® Ultra™ enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were then sequenced on the Illumina HiSeq4000 platform. The hybrid ONT assembly, Illumina shotgun reads [9], and Dovetail Hi-C library reads were used as input data for the scaffolding software HiRise, which uses proximity ligation data to scaffold genome assemblies [59]. Shotgun and Dovetail Hi-C library sequences were aligned to the hybrid ONT assembly using a modified SNAP read mapper. The separations of Dovetail Hi-C read pairs mapped within the ONT scaffolds were analyzed by HiRise to produce a likelihood model for the genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After scaffolding, Illumina shotgun sequences were used to close gaps between contigs, resulting in an improved HiRise assembly.

Validation and anchoring of the HiRise assembly to Chandler genetic maps

The HiRise assembly was first anchored to the Chandler genetic map obtained by [14] from a 312 offspring F₁ population 'Chandler x Idaho' genotyped with the latest Axiom *J. regia* 700K SNP array. SNP probes (71-mers including the SNP site) from the Axiom *J. regia* 700K SNP array were aligned onto the HiRise assembly filtering out alignments with probe/reference identity lower

505 than 98%, covering less than 95% of the probe length or aligning multiple times on the genome.
506 Retained markers with a unique segregation profile were then used to anchor the HiRise scaffolds.
507 The same procedure was also followed to anchor the HiRise assembly to the Chandler genetic map
508 used to construct a walnut bacterial artificial chromosome (BAC) clone-based physical map by
509 [35]. The final ordering of scaffolds was performed by taking into consideration the marker genetic
510 map position, and, in the final sequence, consecutive scaffolds were separated by sequences of
511 100,000 Ns.

512 The tandem repeat finder program (trf v4.09; [60]) was run using the recommended parameters
513 (max mismatch delta PM PI minscore maxperiod, 2 7 7 80 10 50 500 resp.) to identify repeat
514 elements up to 500 bp long. A histogram of repeat unit lengths was generated, and peaks at 7, 29,
515 33, 44, 154, and 308 bp were identified. From this data, a consensus sequence corresponding to
516 each peak was selected. All of these repeat sequences were aligned onto the HiRise assembly using
517 ‘nucmer’ from the MUMmer4 package [61] with a minimum match length of 7 to capture the
518 telomeric repeat. Based on the positions of these alignments along the chromosomes and contigs,
519 we identified the 7-mer as the telomeric repeat and the 154-mer and 308-mer as centromeric
520 repeats.

521 Recombination rate was estimated within sliding windows of 10 Mb with a step of 1 Mb along the
522 chromosome sequence by using the high-density genetic map of Chandler [14] and the
523 R/MareyMap package v 1.3.4 [62]. To evaluate Chandler v2.0 error rate, the two assemblies,
524 Chandler v1.0 and 2.0, were aligned to each other using the ‘nucmer’ program [61]. Assembly
525 quality statistics were estimated using QUAST v5.0.2 [63], filtering for contigs with a minimum
526 length of 1Kb. The haploid size of the walnut genome was estimated by first generating the 24-
527 mer distribution of Illumina paired-end reads (54-fold coverage of the haploid genome) with

Jellyfish v2.2.6 [64], and then uploading it to Genomescope [65]. Comparisons of Chandler v2.0 versus JrSerr_v1.0 and vice versa were performed using ‘nucmer’ [61], and then the function ‘dnadiff’ implemented in MUMmer4 was used to obtain detailed information on the differences between two assemblies.

RNA preparation

Five walnut tissues (leaf, catkin 1-inch elongated; catkin 3-inches elongated, pistillate flower, and pollen) were collected from ‘Chandler’ trees at the UCD walnut orchards. Four additional samples (somatic embryo, callus, shoot, and roots) were taken from tissue culture material of ‘Chandler’. Several grams of each tissue were ground in liquid nitrogen and with insoluble polyvinylpyrrolidone (PVPP; 1% w/w). RNA was isolated using the PureLink™ Plant RNA Reagent (Invitrogen™, Carlsbad, CA) following the manufacturer’s instructions, but with an additional end wash in 1 mL of 75% ethanol. For root tissue only, RNA isolation was performed using the MagMAX™ mirVana™ Total RNA Isolation Kit (Applied Biosystems™, Foster City, CA) as per protocol, except for the lysis step. A different lysis buffer was created adding 100 mg of sodium metabisulfite to 10 mL of guanidine buffer (guanidine thiocyanate 4M, sodium acetate 0.2M, EDTA 25 mM, PVP-40 2.5%, pH 5.0) and 1 mL of nuclease-free water. Then, 100 mg of ground root tissue were lysed in 1 mL of the new lysis buffer using a Tissue Lyser at max frequency for 2 min. The lysate was centrifuged at 4° C for 5 min at max speed. The supernatant (500 µL) was transferred to a new tube for the following steps of RNA isolation as per protocol. RNA samples were then purified, and DNase treated using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). The RNA quality was confirmed by running an aliquot of each sample on an Experion™ Automated Electrophoresis System (Bio-Rad, Hercules, CA).

PacBio IsoSeq sequencing

Full-length cDNA Iso-Seq template libraries for PacBio IsoSeq analysis were constructed and sequenced at the DNA Technologies & Expression Analysis Core Facility of the UC Davis Genome Center. FL double-stranded cDNA was generated from total RNA (2 µg per tissue) using the Lexogen TeloTM prime Full-length cDNA Kit (Lexogen, Inc., Greenland, NH, USA). Tissue-specific cDNAs were first barcoded by PCR (16-19 cycles) using IDT barcoded primers (Integrated DNA Technologies, Inc., Coralville, Iowa), and then bead-size selected with AMPure PB beads (two different size fractions of 1X and 0.4X). The nine cDNAs were pooled in equimolar ratios and used to prepare a SMRTbellTM library using the PacBio Template Prep Kit (PacBio, Menlo Park, CA). The SMRTbellTM library was then sequenced across four Sequel v2 SMRT cells with polymerase 2.1 and chemistry 2.1 (P2.1C2.1).

PacBio raw reads were processed using the Isoseq3 v.3.0 workflow following PacBio recommendations (<https://github.com/PacificBiosciences/IsoSeq3>). Circular consensus sequences (CCSs) were generated using the program ‘ccs’. The CCSs were demultiplexed and cleaned of cDNA primers using the program ‘lima’. Afterward, CCS clustering and polishing was performed using the program ‘isoseq3’, to generate HQ FL sequences for each of the nine tissues. Full-length non-chimeric (FLnc) and HQ clusters were aligned onto the new ‘Chandler’ assembly v2.0 with minimap2 v.2.12-r827, including the parameter ‘-ax splice’ [66].

Repeat annotation

A genome-specific repeat database was created using the ‘basic’ mode implemented in RepeatModeler v.1.0.11 [67]. RepeatMasker v.4.0.7 was then run to mask repeats in the walnut reference genome v.2.0 and generate a GFF file [68].

Gene prediction and functional annotation

Juglans regia RefSeq transcripts and additional *J. regia* transcripts and protein sequences downloaded from NCBI, along with the HQ FL IsoSeq transcripts, were used as input to the PASA pipeline v.2.3.3 [69], to assemble a genome-based transcript annotation. PASA utilizes the aligners BLAT v.35 [70] and GMAP v.2018-07-04 [71], along with TransDecoder v.5.5.0 [72], which predicts open reading frames (ORFs) as genome-based GFF coordinates. The final PASA/TransDecoder GFF3 file was post-processed to name the genes and transcripts by chromosome location consistently. The chloroplast and mitochondrial genomes were annotated using the “CHLOROBOX GeSeq Annotation of Organellar Genomes” tool at <https://chlorobox.mpimp-golm.mpg.de/geseq.html> with default parameters [73]. NCBI accessions NC_028617.1 (*J. regia* chloroplast), KT971339.1 (*Medicago truncatula* mitochondrion), NC_029641.1 (*M. truncatula* mitochondrion) and NC_012119.1 (*Vitis vinifera* mitochondrion) were also input as custom references. The output gff3 files were then post-processed to consistently rename genes.

Functional roles were assigned to predicted peptides using Trinotate v.3.1.1 [74]. In particular, similarity searches were performed against several public databases (i.e., Uniprot/Swiss-Prot, NCBI NR, *Vitis vinifera*.IGGP_12x, *J. regia* RefSeq) using BLAST v.2.8.1, HMMER v.3.1b2, SignalP v.4.1c, and TMHMM v.2.0c. Gene family analysis was performed by running Interproscan v. 5.30-69.0 [75,76] with default parameters on each protein fasta file (v1.0 [9], NCBI RefSeq [GCF_001411555.1_wgs.5d] and v2.0). The PANTHER family ID with the lowest expect value (below expect value threshold of $1.0E^{-11}$) was assigned to each protein.

The completeness and quality of both genome assembly and gene annotation of Chandler v.2.0 were estimated with the BUSCO method v.3 (1,440 core genes in the embryophyte dataset) [77],

and the sets of coreGFs of green plants (2,928 coreGFs) and rosids (6,092 coreGFs) from PLAZA v.2.5 [78]. Also, RNA-Seq data previously generated for 20 tissues (see [9]) were aligned to the reference genome (v1.0 and v2.0) with HISAT2 [79]. The alignments of the 20 RNA-seq data and the FL transcripts along with the new genome annotation v2.0 were then used as input to StringTie v.2.0 [80] to estimate expression levels in both fragments per kilobase per million reads (FPKM) and transcripts per million (TPM) for each transcript in the v2 annotation. The percent identity and coverage of each *J. regia* transcript compared to proteins in the NCBI plant RefSeq database was also determined by running the EnTAP pipeline v.0.9.0 [43].

Label-free shotgun proteomics

Plant tissues of immature, intermediate, mature catkins (Additional file 1, Figure S9) and pure pollen from three individual trees of Chandler at the UCD walnut orchards were collected and frozen immediately in dry ice. Tissues were then further frozen in liquid nitrogen in the laboratory and ground with mortar and pestle. Five hundred milligrams of each sample were used for total protein extraction, following the procedure for recalcitrant plant tissues of [81], with a modification in the final buffer used to resuspend the protein pellet, consisting of 8M urea in 50mM triethylammonium bicarbonate (TEAB). One hundred micrograms of total protein from each sample were then used for proteomics.

Initially, 5 mM dithiothreitol (DTT) was added and incubated at 37°C for 30 min and 1,000 rpm shaking. Next, 15 mM iodoacetamide (IAA) was added, followed by incubation at room temperature for 30 min. The IAA was then neutralized with 30 mM DTT in incubation for 10 min. Lys-C/trypsin then was added (1:25 enzyme: total protein) followed by 4 h incubation at 37°C. After, TEAB (550 µl of 50 mM) was added to dilute the urea and activate trypsin digestion overnight. The digested peptides were desalted with Aspire RP30 Desalting Tips (Thermo

Scientific), vacuum dried, and suspended in 45 μ l of 50 mM TEAB. Peptides were quantified by Pierce quantitative fluorometric assay (Thermo Scientific) and 1 μ g analyzed on a QExactive mass spectrometer (Thermo Scientific) coupled with an Easy-LC source (Thermo Scientific) and a nanospray ionization source. The peptides were loaded onto a Trap (100 microns, C18 100 Å 5U) and desalted online before separation using a reversed-phase (75 microns, C18 200 Å 3U) column. The duration of the peptide separation gradient was 60 min using 0.1% formic acid and 100% acetonitrile (ACN) for solvents A and B, respectively. The data were acquired using a data-dependent MS/MS method, which had a full scan range of 300-1,600 Da and a resolution of 70,000. The resolution of the MS/MS method was 17,500 and the insulation width 2 m/z with a normalized collision energy of 27. The nanospray source was operated using a spray voltage of 2.2 KV and a transfer capillary temperature heated to 250°C. Samples were analyzed at the UC Davis Proteomics Core.

The raw data were analyzed using X! Tandem and viewed using the Scaffold Software v.4. (Proteome Software, Inc.). Samples were searched against UniProt databases appended with the cRAP database, which recognizes common laboratory contaminants. Reverse decoy databases were also applied to the database before the X! Tandem searches. The protein-coding sequences (CDS) annotated in Chandler v1.0 (NCBI accession PRJNA350852) and v2.0 were used as a reference for identification of proteins from the mass spectrometry data. The proteins identified were filtered in the Scaffold software based on the following criteria: 1.0% FDR (false discovery rate) at protein level (following the prophet algorithm: <http://proteinprophet.sourceforge.net/>), the minimum number of 2 peptides and 0.1% FDR at the peptide level. Structure of the walnut allergen (Jug r 9) was modelled using SWISS-MODEL [82] based on the structure of a homologous

allergen from lentil (PDBid:2MAL). Structures were superimposed using MUSTANG (2MAL:in red, walnut in blue) [83].

Chandler genomic diversity

Illumina whole-genome shotgun data of Chandler were aligned on the Chandler v2.0 with BWA [84] with standard parameters. SNP calling was performed using SAMtools v1.9 [85] and BCFtools v.2.1 [86]. SNP density for windows of 1 Mb was estimated using the command ‘SNPdensity’ implemented in VCFtools v0.1.16 [87]. Self-collinearity analysis to detect duplicated regions in Chandler v2.0 was performed with MCScanX [88], using a simplified GFF file of the new gene annotation and a self-BLASTP as input. To improve the power of collinearity detection, tandem duplications were excluded after running the function ‘detect_collinear_tandem_arrays’ implemented in MCScanX. Synonymous (K_S) and nonsynonymous (K_A) changes for syntenic protein-coding gene pairs were measured using the Perl script “add_ka_and_ks_to_collinearity.pl” implemented in MCScanX.

To explore the inbreeding level across the 16 chromosomal pseudomolecules of Chandler, haplotypes were built for 55 individuals of the UCD-WIP, including 25 founders and several commercially relevant walnut cultivars (e.g., Chandler, Howard, Tulare, Vina, Franquette) along with their parents and progenitors. All individuals were genotyped using the latest Axiom™ *J. regia* 700K SNP array as described in [7]. To define SNP HBs, 26,544 unique and robust SNPs were selected and ordered according to the Chandler genome v2.0 physical map. Subsequently, for each SNP markers and individual, phasing and identification of closely linked groups of SNPs, without recombination in most of the pedigree, was performed using the software FlexQTL™ [89] and PediHaplotyper [90] following the approach described in [91] and [90]. In particular, HBs were defined by recombination sites detected in ancestral generation of Chandler.

Genomic comparison between Eastern and Western walnuts

The resequencing data of 23 founders of the UCD-WIP (**Additional file 26**)[10] were mapped onto the Chandler v2.0 with BWA, and SNPs were called following the same procedure described above for Chandler. SNPs with no missing data and minor allele frequency (MAF) higher than 10% were retained for the following genetic analyses (7,269,224 SNPs out of the 14,988,422 identified). Hierarchical cluster analysis on a dissimilarity matrix of the 23 UCD-WIP founders was performed using R/SNPRelate v.1.18.0 [92]. Fixation index (F_{ST}) was measured between genotypes from EU/USA and Asia with VCFtools v0.1.16, setting windows of 100kb and 500kb. Genomic windows with the top 5% of F_{ST} values were selected as candidate regions for further analysis. The empirical cutoff with a low false discovery rate (5%) was verified by performing whole-genome permutation test (1000) with a custom Python script. Nucleotide diversity (π) and Tajima's D [93] were also computed along the whole genome in 100-kb and 500-kb windows using VCFtools. Reduction of diversity coefficient (ROD) was estimated as $1 - (\pi_{Occ} / \pi_{Asia})$. The new walnut gene annotation v.2.0 was used to identify predicted genes in the candidate regions under selection. The distribution of the identified genes into different biological processes was evaluated using the weight01 method provided by the R/topGO [94]. The Kolmogorov–Smirnov-like test was performed to assess the significance of over-representation of GO categories compared with all genes in the walnut gene prediction. Plots were obtained using the R/circlize v.0.4.6 and R/ggplot2 v.3.5.3 packages.

Availability of supporting data

All raw and processed sequencing data generated in this study have been submitted to the NCBI BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject/>) under accession number PRJNA291087. All SNP data have been submitted to Hardwood Genomics (<https://hardwoodgenomics.org/Genome-assembly/2539069>).

Additional files

Additional file 1. Tables S1-S9; Table S14-S20; Figures S1-S15.

Additional file 2. Tables S10-S13

Additional file 3. Mass-spectrometry proteome data of catkins and pollen tissues. Three samples of each tissue type (immature catkin, mature catkin, senescent catkin, and pure pollen) were analyzed using v1.0 and v2.0 reference walnut genome assemblies. Total intensity of matching peptides, number of spectra and percentage of protein covered by the identified peptides are reported.

Competing interests

The authors declare no conflict of interest.

Funding:

This project has been funded by the Californian Walnut Board.

Author's contribution

DBN and AM conceived and coordinated the research. REW and WT performed the HMW DNA extraction and Nanopore sequencing. AVZ, DP and SLS assembled the hybrid Illumina-ONT assembly. LB, MT, DP and SLS validated and anchored the HiRise assembly to the genetic maps. AM and BJA collected and extracted all RNA samples. MB analyzed the PacBio IsoSeq results and performed the repeat and gene annotation. AD conceived the design of the proteomic analyses; PAZ and SC generated and analyzed the proteomic data. LB called the SNPs in Chandler and the 23 UCD WIP founders, while AM carried out the analyses on walnut genomic diversity. EAD, LB and MT built and analyzed the SNP haplotypes. CAL provided all the plant material. AM wrote the manuscript, which has been revised by all coauthors.

Acknowledgments

We are grateful to Sriema Walawage for assistance with RNA extraction, and Brett Phinney for preparing the raw proteome data.

References

1. Martínez ML, Labuckas DO, Lamarque AL, Maestri DM. Walnut (*Juglans regia* L.): Genetic resources, chemistry, by-products. J Sci Food Agric. 2010;90:1959–67.
2. McGranahan G, Leslie C. Walnut. In: Badenes ML, Byrne DH, editors. Fruit Breed. Springer Science+Business Media, LLC; 2012. p. 827–46.
3. Pollegioni P, Woeste K, Chiocchini F, Del Lungo S, Ciolfi M, Olimpieri I, et al. Rethinking the history of common walnut (*Juglans regia* L.) in Europe: Its origins and human interactions.

727 PLoS One. 2017;12:1–24.

728 4. Zhang B, Xu L, Li N, Yan P, Jiang X, Woeste KE, et al. Phylogenomics Reveals an Ancient
729 Hybrid Origin of the Persian Walnut. *Mol Biol Evol.* 2019;1–11.

730 5. Zeven A, Zhukovskii PM. Dictionary of cultivated plants and their centres of diversity,
731 excluding ornamentals, forest trees, and lower plants [Internet]. Cent. Agric. Publ. Doc.
732 Wageningen. 1975. Available from: <https://core.ac.uk/download/pdf/29387092.pdf>

733 6. Ebrahimi A, Zarei A, Lawson S, Woeste KE, Smulders MJM. Genetic diversity and genetic
734 structure of Persian walnut (*Juglans regia*) accessions from 14 European, African, and Asian
735 countries using SSR markers. *Tree Genet Genomes* [Internet]. Tree Genetics & Genomes;
736 2016;12:114. Available from: <http://link.springer.com/10.1007/s11295-016-1075-y>

737 7. Marrano A, Martínez-García PJ, Bianco L, Sideli GM, Di Pierro EA, Leslie CA, et al. A new
738 genomic tool for walnut (*Juglans regia* L.): development and validation of the high-density
739 Axiom™ *J. regia* 700K SNP genotyping array. *Plant Biotechnol J* [Internet]. 2018;1–10.
740 Available from: <http://doi.wiley.com/10.1111/pbi.13034>

741 8. Bernard A, Barreneche T, Lheureux F, Dirlewanger E. Analysis of genetic diversity and
742 structure in a worldwide walnut (*Juglans regia* L.) germplasm using SSR markers. *PLoS One.*
743 2018;13:1–19.

744 9. Martínez-García PJ, Crepeau MW, Puiu D, Gonzalez-Ibeas D, Whalen J, Stevens KA, et al.
745 The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the
746 biosynthesis of non-structural polyphenols. *Plant J.* 2016;87:507–32.

747 10. Stevens KA, Woeste K, Chakraborty S, Crepeau MW, Leslie CA, Martínez-García PJ, et al.

748 Genomic Variation Among and Within Six *Juglans* Species. G3 Genes|Genomes|Genetics
749 [Internet]. 2018;8:1–37. Available from:
750 <http://www.ncbi.nlm.nih.gov/pubmed/29792315>
751 <http://g3journal.org/lookup/doi/10.1534/g3.118.200030>

752 11. Kefayati S, Ikhsan AS, Sutyemez M, Paizila A, Topcu H, Bukubu SB, et al. First simple
753 sequence repeat-based genetic linkage map reveals a major QTL for leafing time in walnut
754 (*Juglans regia* L.). Tree Genet Genomes. 2018;15:13.

755 12. Arab MM, Marrano A, Abdollahi-Arpanahi R, Leslie CA, Askari H, Neale DB, et al.
756 Genome-wide patterns of population structure and association mapping of nut-related traits in
757 Persian walnut populations from Iran using the Axiom J. regia 700K SNP array. Sci Rep
758 [Internet]. Springer US; 2019;9:6376. Available from: [http://www.nature.com/articles/s41598-](http://www.nature.com/articles/s41598-019-42940-1)
759 [019-42940-1](http://www.nature.com/articles/s41598-019-42940-1)

760 13. Famula RA, Richards JH, Famula TR, Neale DB. Association Genetics of Carbon Isotope
761 Discrimination in the Founding Individuals of a Breeding Population of *Juglans regia* L. Tree
762 Genet Genomes [Internet]. Tree Genetics & Genomes; 2019;15:6. Available from:
763 <https://doi.org/10.1007/s11295-018-1307-4>

764 14. Marrano A, Sideli GM, Leslie CA, Cheng H, Neale DB. Deciphering of the genetic control
765 of phenology, yield and pellicle color in Persian walnut (*Juglans regia* L.). Front Plant Sci.
766 2019;10:1–14.

767 15. Bernard A, Marrano A, Donkpegan A, Brown PJ, Leslie CA, Neale DB, et al. Association
768 and linkage mapping to unravel genetic architecture of phenological traits and lateral bearing in
769 Persian walnut (*Juglans regia* L.). BMC Genomics. BMC Genomics; 2020;21:1–25.

- 770 16. Sánchez-Pérez R, Pavan S, Mazzeo R, Moldovan C, Aiese Cigliano R, Del Cueto J, et al.
771 Mutation of a bHLH transcription factor allowed almond domestication. *Science* (80-)
772 [Internet]. 2019;364:1095–8. Available from:
773 <http://www.sciencemag.org/lookup/doi/10.1126/science.aav8197>
- 774 17. Tang W, Sun X, Yue J, Tang X, Jiao C, Yang Y, et al. Chromosome-scale genome assembly
775 of kiwifruit *Actinidia eriantha* with single-molecule sequencing and chromatin interaction
776 mapping. *Gigascience*. Oxford University Press; 2019;8:1–10.
- 777 18. Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, et al. The genome of
778 *Chenopodium quinoa*. *Nature*. 2017;542:307–12.
- 779 19. Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, et al. Durum
780 wheat genome highlights past domestication signatures and future improvement targets. *Nat*
781 *Genet*. 2019;51:885–95.
- 782 20. Raymond O, Gouzy J, Just J, Badouin H, Verdenaud M, Lemainque A, et al. The *Rosa*
783 genome provides new insights into the domestication of modern roses. *Nat Genet*. 2018;50:772–
784 7.
- 785 21. Daccord N, Celton JM, Linsmith G, Becker C, Choisne N, Schijlen E, et al. High-quality de
786 novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat*
787 *Genet* [Internet]. Nature Publishing Group; 2017;49:1099–106. Available from:
788 <http://dx.doi.org/10.1038/ng.3886>
- 789 22. Zhu T, Wang L, You FM, Rodriguez JC, Deal KR, Chen L, et al. Sequencing a *Juglans regia*
790 $\times$ *J. microcarpa* hybrid yields high-quality genome assemblies of parental species. *Hortic Res*
791 [Internet]. Springer US; 2019;1–16. Available from: <http://dx.doi.org/10.1038/s41438-019-0139->

792 1

793 23. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly.
794 Genomics, Proteomics Bioinforma [Internet]. Beijing Institute of Genomics, Chinese Academy
795 of Sciences and Genetics Society of China; 2016;14:265–79. Available from:
796 <http://dx.doi.org/10.1016/j.gpb.2016.05.004>

797 24. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: A comprehensive
798 technique to capture the conformation of genomes. Methods [Internet]. Elsevier Inc.;
799 2012;58:268–76. Available from: <http://dx.doi.org/10.1016/j.ymeth.2012.05.001>

800 25. Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. J Ex.
801 2017;68:5419–29.

802 26. Schmidt MH-W, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, et al. De Novo
803 Assembly of a New *Solanum pennellii* Accession Using Nanopore Sequencing . Plant Cell.
804 2017;29:2336–48.

805 27. Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, et al. Chromosome-scale
806 assemblies of plant genomes using nanopore long reads and optical maps. Nat Plants [Internet].
807 Springer US; 2018;4:879–87. Available from: <http://dx.doi.org/10.1038/s41477-018-0289-4>

808 28. Yasodha R, Vasudeva R, Balakrishnan S, Sakthi AR, Abel N, Binai N, et al. Draft genome of
809 a high value tropical timber tree, Teak (*Tectona grandis* L. f): insights into SSR diversity,
810 phylogeny and conservation. DNA Res. 2018;25:409–19.

811 29. Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, et al. A chromosome-scale
812 assembly of the sorghum genome using nanopore sequencing and optical mapping. Nat

813 Commun. 2018;9:4844.

814 30. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: Computational
815 approaches for improving nanopore sequencing read accuracy. *Genome Biol. Genome Biology*;
816 2018;19:1–11.

817 31. Zimin A V., Luo M, Marçais G, Salzberg SL, Yorke JA, Puiu D, et al. Hybrid assembly of
818 the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the
819 MaSuRCA mega-reads algorithm. *Genome Res* [Internet]. 2017;27:787–92. Available from:
820 <http://www.ncbi.nlm.nih.gov/pubmed/28130360>
821 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5411773>

822 32. Huang Y, Xiao L, Zhang Z, Zhang R, Wang Z, Huang C, et al. The genomes of pecan and
823 Chinese hickory provide insights into *Carya* evolution and nut nutrition. *Gigascience*. 2019;8:1–
824 17.

825 33. Xing Y, Liu Y, Zhang Q, Nie X, Sun Y, Zhang Z, et al. Hybrid de novo genome assembly of
826 Chinese chestnut (*Castanea mollissima*). *Gigascience*. 2019;8:1–7.

827 34. Plomion C, Aury JM, Amselem J, Leroy T, Murat F, Duplessis S, et al. Oak genome reveals
828 facets of long lifespan. *Nat Plants*. Springer US; 2018;4:440–52.

829 35. Luo M-C, You FM, Li P, Wang J-R, Zhu T, Dandekar AM, et al. Synteny analysis in Rosids
830 with a walnut physical map reveals slow genome evolution in long-lived woody perennials.
831 *BMC Genomics* [Internet]. *BMC Genomics*; 2015;16:1–17. Available from:
832 <http://www.biomedcentral.com/1471-2164/16/707>

833 36. Springer NM, Ying K, Fu Y, Ji T, Yeh C, Jia Y, et al. Maize Inbreds Exhibit High Levels of

834 Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content.
835 PLoS Genet. 2009;5.

836 37. Marroni F, Pinosio S, Morgante M. Structural variation and genome complexity : is
837 dispensable really dispensable ? Curr Opin Plant Biol [Internet]. Elsevier Ltd; 2014;18:31–6.
838 Available from: <http://dx.doi.org/10.1016/j.pbi.2014.01.003>

839 38. Mishra B, Gupta DK, Pfenninger M, Hickler T, Langer E, Nam B, et al. A reference genome
840 of the European beech (*Fagus sylvatica* L.). Gigascience. 2018;7.

841 39. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics, Proteomics
842 Bioinforma [Internet]. Beijing Institute of Genomics, Chinese Academy of Sciences and
843 Genetics Society of China; 2015;13:278–89. Available from:
844 <http://dx.doi.org/10.1016/j.gpb.2015.08.002>

845 40. Linsmith G, Rombauts S, Montanari S, Deng CH, Celton JM, Guérif P, et al. Pseudo-
846 chromosome-length genome assembly of a double haploid “bartlett” pear (*Pyrus communis* L.).
847 Gigascience. 2019;8:1–17.

848 41. Vanechoutte D, Estrada AR, Lin YC, Loraine AE, Vandepoele K. Genome-wide
849 characterization of differential transcript usage in *Arabidopsis thaliana*. Plant J. 2017;92:1218–
850 31.

851 42. Clark S, Yu F, Gu L, Min XJ. Expanding alternative splicing identification by integrating
852 multiple sources of transcription data in tomato. Front Plant Sci. 2019;10:1–12.

853 43. Hart AJ, Ginzburg S, Xu M (Sam), Fisher CR, Rahmatpour N, Mitton JB, et al. EnTAP:
854 bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes.

855 bioRxiv [Internet]. 2018;307868. Available from:
 856 <https://www.biorxiv.org/content/biorxiv/early/2018/04/28/307868.full.pdf%0Ahttps://www.biorxiv.org/content/early/2018/04/24/307868%0Ahttps://www.biorxiv.org/content/early/2018/04/24/307868>
 857
 858 307868

859 44. Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I. A chromosome-scale genome
 860 assembly of European Hazel (*Corylus avellana* L.) reveals targets for crop improvement.
 861 bioRxiv. 2019;44.

862 45. Jamet E, Santoni V. Editorial for Special Issue: 2017 Plant Proteomics. proteomes.
 863 2018;6:28.

864 46. Costa J, Carrapatoso I, Oliveira MBPP, Mafra I. Walnut allergens: Molecular
 865 characterization, detection and clinical relevance. Clin Exp Allergy. 2014;44:319–41.

866 47. Aradhya M, Woeste K, Velasco D. Genetic diversity, structure and differentiation in
 867 cultivated walnut (*Juglans regia* L.). Acta Hortic. 2010;861:127–32.

868 48. Ruiz-Garcia L, Lopez-Ortega G, Fuentes Denia a., Frutos Tomas D. Identification of a
 869 walnut (*Juglans regia* L.) germplasm collection and evaluation of their genetic variability by
 870 microsatellite markers. Spanish J Agric Res. 2011;9:179–92.

871 49. Dangl GS, Woeste K, Aradhya MK, Koehmstedt A, Simon C, Potter D, et al.
 872 Characterization of 14 Microsatellite Markers for Genetic Analysis and Cultivar Identification of
 873 Walnut. J Am Soc Hortic Sci [Internet]. 2005;130:348–54. Available from:
 874 [http://journal.ashspublications.org/content/130/3/348%5Cnhttp://journal.ashspublications.org/co](http://journal.ashspublications.org/content/130/3/348%5Cnhttp://journal.ashspublications.org/content/130/3/348.full.pdf)
 875 [ntent/130/3/348.full.pdf](http://journal.ashspublications.org/content/130/3/348.full.pdf)

876 50. McGranahan GH, Leslie CA. Walnuts. In: Moore JN, Ballington JRJ, editors. Genet Resour
877 Temp Fruit Nut Crop. International Society for Horticultural Science; 1991. p. 907–18.

878 51. Bernard A, Lheureux F, Dirlewanger E. Walnut: past and future of genetic improvement.
879 Tree Genet Genomes. Tree Genetics & Genomes; 2018;14:1–28.

880 52. Gauthier MM, Jacobs DF. Walnut (*Juglans* spp.) ecophysiology in response to environmental
881 stresses and potential acclimation to climate change. Ann For Sci. 2011;68:1277–90.

882 53. Workman R, Fedak R, Kilburn D, Hao S, Liu K, Timp W. High Molecular Weight DNA
883 Extraction from Recalcitrant Plant Species for Third Generation Sequencing. Protoc Exch.
884 2018;1–12.

885 54. Zhang M, Zhang Y, Scheuring CF, Wu CC, Dong JJ, Zhang H Bin. Preparation of megabase-
886 sized DNA from a variety of organisms using the nuclei method for advanced genomics
887 research. Nat Protoc [Internet]. Nature Publishing Group; 2012;7:467–78. Available from:
888 <http://dx.doi.org/10.1038/nprot.2011.455>

889 55. Mayjonade B, Gouzy J, Donnadiou C, Pouilly N, Marande W, Callot C, et al. Extraction of
890 high-molecular-weight genomic DNA for long-read sequencing of single molecules.
891 Biotechniques. 2017;62:xv.

892 56. Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome
893 assembler. Bioinformatics. 2013;29:2669–77.

894 57. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, et al. Aggressive
895 assembly of pyrosequencing reads with mates. Bioinformatics. 2008;

896 58. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al.

897 Comprehensive mapping of long-range interactions reveals folding principles of the human
898 genome. *Science* (80-). 2009;

899 59. Putnam NH, O’Connell, Brendan L. Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, et
900 al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage.
901 *Genome Res.* 2016;26:342–50.

902 60. Benson G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids*
903 *Res.* 1999;27:573–80.

904 61. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast
905 and versatile genome alignment system. *PLoS Comput Biol.* 2018;14:1–14.

906 62. Rezvoy C, Charif D, Guéguen L, Marais GAB. MareyMap: An R-based tool with graphical
907 interface for estimating recombination rates. *Bioinformatics.* 2007;23:2188–9.

908 63. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST : quality assessment tool for genome
909 assemblies. *Bioinformatics.* 2013;29:1072–5.

910 64. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of
911 occurrences of k-mers. *Bioinformatics.* 2011;

912 65. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al.
913 *GenomeScope* : fast reference-free genome profiling from short reads. *Bioinformatics.*
914 2017;33:2202–4.

915 66. Li H. Minimap2 : pairwise alignment for nucleotide sequences. *Bioinformatics.*
916 2018;34:3094–100.

917 67. Smit A, Hubley R. RepeatModeler Open-1.0. [Internet]. 2008. Available from:

918 <http://www.repeatmasker.org>

919 68. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. [Internet]. 2013. Available from:
 920 <http://www.repeatmasker.org>

921 69. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, et al. Improving
 922 the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic*
 923 *Acids Res.* 2003;31:5654–66.

924 70. Kent WJ. BLAT — The BLAST -Like Alignment Tool. *Genome Res.* 2002;12:656–64.

925 71. Wu TD, Watanabe CK. GMAP : a genomic mapping and alignment program for mRNA and
 926 EST sequences. *Bioinformatics.* 2005;21:1859–75.

927 72. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo
 928 transcript sequence reconstruction from RNA-Seq: reference generation and analysis with
 929 Trinity. *Nat Protoc.* 2013;8:1–43.

930 73. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq -
 931 Versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 2017;

932 74. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length
 933 transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011;

934 75. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan:
 935 Protein domains identifier. *Nucleic Acids Res.* 2005;

936 76. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: Genome-
 937 scale protein function classification. *Bioinformatics.* 2014;

938 77. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO:

939 Assessing genome assembly and annotation completeness with single-copy orthologs.
 940 Bioinformatics. 2015;31:3210–2.

941 78. Veeckman E, Ruttink T, Vandepoele K. Are We There Yet? Reliably Estimating the
 942 Completeness of Plant Genome Sequences. *Plant Cell*. 2016;28:1759–68.

943 79. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory
 944 requirements. *Nat Methods*. 2015;

945 80. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of
 946 RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. 2016;

947 81. Valerie M, Catherine D, Michel Z, Hervé T, Faurobert M, Pelpoir E, et al. Phenol Extraction
 948 of Proteins for Proteomic Studies of Recalcitrant Plant Tissues. *Plant Proteomics*. 2006.

949 82. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: A web-based
 950 environment for protein structure homology modelling. *Bioinformatics*. 2006;

951 83. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. MUSTANG: A multiple structural
 952 alignment algorithm. *Proteins Struct Funct Genet*. 2006;

953 84. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
 954 *Bioinformatics*. 2009;25:1754–60.

955 85. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
 956 Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.

957 86. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-smith C, Durbin R. BCFtools/RoH : a
 958 hidden Markov model approach for detecting autozygosity from next-generation sequencing
 959 data. *Bioinformatics*. 2016;32:1749–51.

87. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156–8.
88. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 2012;40:1–14.
89. Bink MCAM, Jansen J, Madduri M, Voorrips RE, Durel CE, Kouassi AB, et al. Bayesian QTL analyses using pedigreed families of an outcrossing species, with application to fruit firmness in apple. *Theor Appl Genet*. 2014;127:1073–90.
90. Voorrips RE, Bink MCAM, Kruisselbrink JW, Koehorst-van Putten HJJ, van de Weg WE. *PediHaplotyper: software for consistent assignment of marker haplotypes in pedigrees*. Mol Breed. Springer Netherlands; 2016;36.
91. Vanderzande S, Howard NP, Cai L, Da Silva Linge C, Antanaviciute L, Bink MCAM, et al. High-quality, genome-wide SNP genotypic data for pedigreed germplasm of the diploid outbreeding species apple, peach, and sweet cherry through a common workflow. *PLoS One*. 2019;
92. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*. 2012;28:3326–8.
93. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989;123:585–95.
94. Alexa A. Gene set enrichment analysis with topGO. 2015;47–53.

Tables

981 **Table 1. Comparison among the four assemblies of Chandler.** Scaffolds shorter than 1,000 bp are not
982 included in these totals.

983

Statistics	Chandler v1.0	Chandler v1.5	Chandler hybrid	Chandler HiRise	Chandler v2.0
<i>Number of scaffolds</i>	27,032	4,401	3,497	2,656	2,643
<i>N50 length (scaffolds) (bp)</i>	304,423	637,984	1,640,935	32,655,472	37,114,715
<i>L50 (scaffolds)</i>	344	272	89	8	7
<i>Total length of assembled scaffolds (bp)</i>	667,299,356	650,478,320	567,378,842	567,480,142	567,796,851
<i>Number of contigs</i>	53,156	7,411	3,592	3,700	3,684
<i>N50 length (contigs) (bp)</i>	42,417	317,751	1,512,354	1,083,883	1,083,883
<i>L50 (contigs)</i>	3,630	482	97	144	144
<i>Total size of assembled contigs (bp)</i>	641,521,787	617,088,256	567,276,004	567,276,244	567,192,099

984

985

986 **Table 2.** Statistics on the gene annotation of Chandler v2.0 compared to the previous gene annotations of
987 the Chandler genome.

Statistics	Chandler v2.0	Chandler v1.0	Chandler RefSeq v1.0
<i>Number of genes</i>	37,554	32,496	41,188
<i>Average gene length (bp)</i>	5,319	4,358	4,641
<i>Single-exon transcripts</i>	6,613	6,247	6,749

<i>Average CDS length (bp)</i>	1,335	1,222	1,336
<i>Number of exons</i>	242,208	172,273	230,261
<i>Average exon length (bp)</i>	257.8	229.5	314
<i>Number of Introns</i>	201,290	139,775	181,419
<i>Average intron length</i>	853.9	730	835
<i>Introns per gene</i>	5.9	5.3	4.4

Figure legends

Figure 1. Collinearity between the high-density ‘Chandler’ genetic map of [14] and the 16 chromosomal pseudomolecules of Chandler v2.0.

Figure 2. Summary of gene distribution and genetic diversity across the 16 chromosomes of Chandler v2.0. Tracks from outside to inside: **(i)** gene density of Chandler v2.0 in 1-Mb windows; **(ii)** Chandler heterozygosity in 1-Mb windows (white = low heterozygosity; blue = high heterozygosity); **(iii)** Recombination rate for sliding windows of 10 Mb (average = 2.63 cM/Mb); **(iv)** F_{ST} in 500-kb windows. Windows in the 95 percentiles of the F_{ST} distribution are highlighted in red; **(v)** ROD values for 500-kb windows.

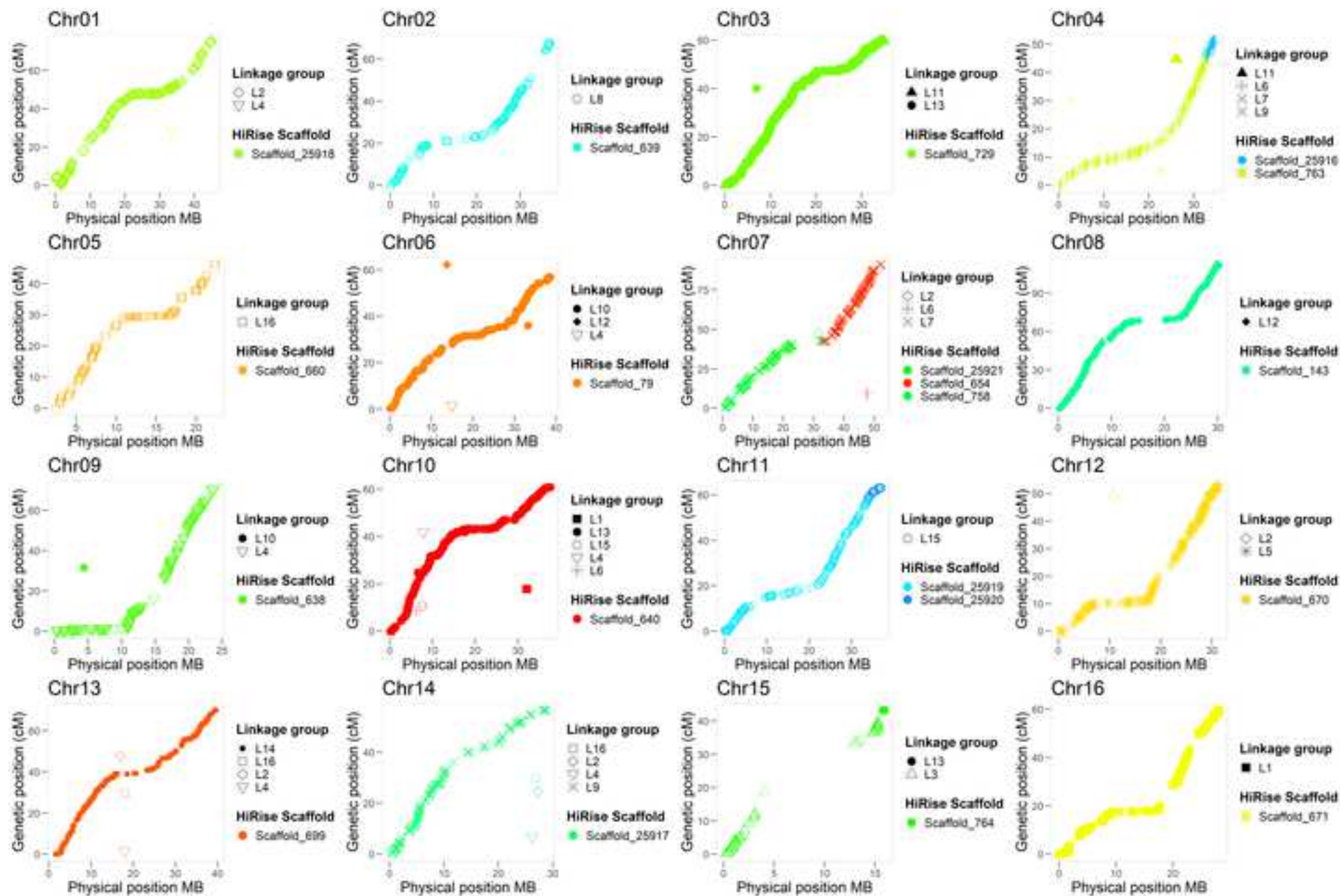
Figure 3. Clustering of the samples used in the proteomic analysis. **(A)** Hierarchical clustering based on Euclidian distances of normalized abundances of detected proteins. Samples are represented in columns and proteins in rows. **(B)** Principal component analysis of the 12 samples analyzed, clustering according to tissue type.

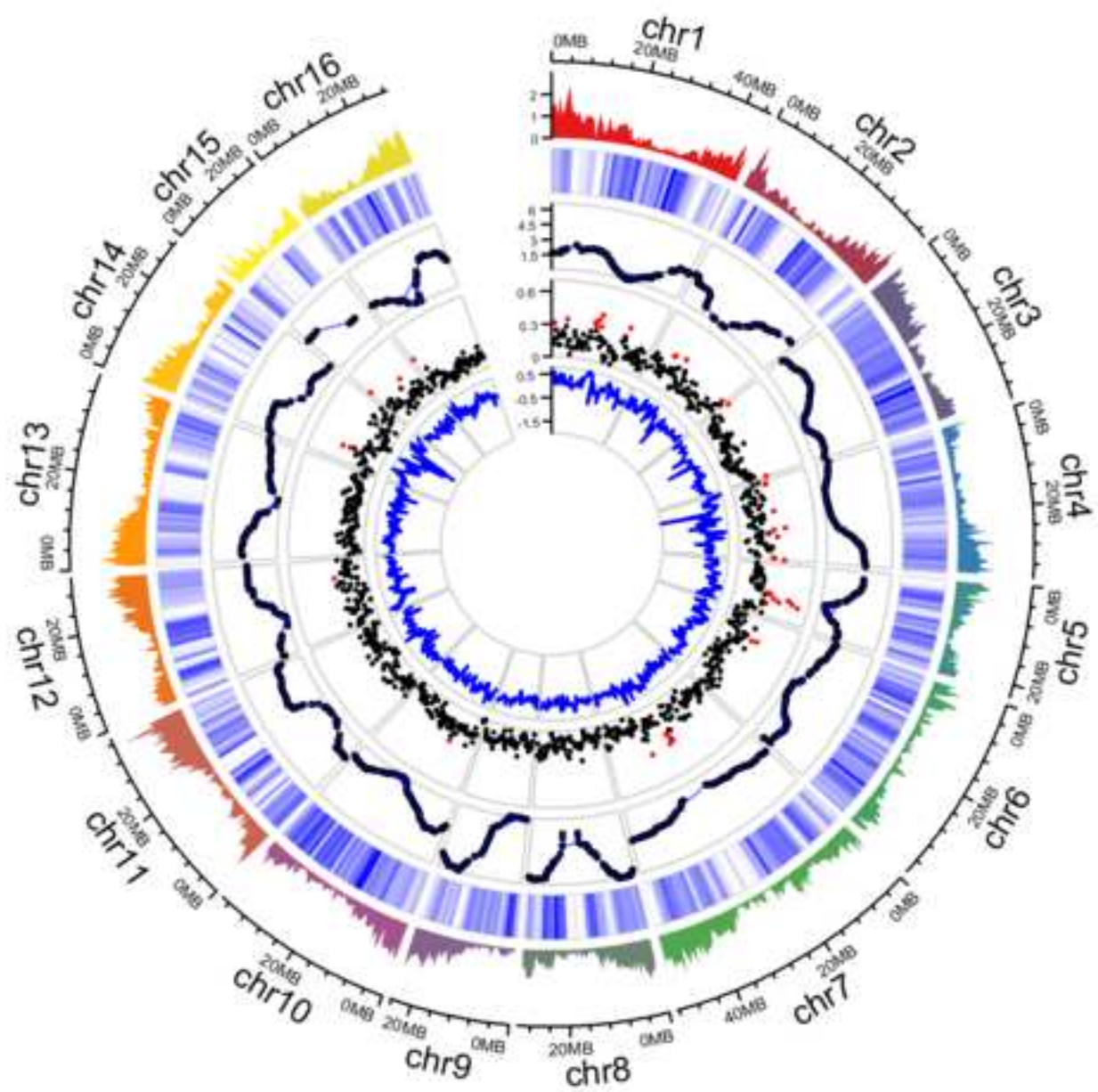
Figure 4. Modeled structure of the putative new allergen encoded by *Jr12_05180*. The compact structure is stabilized by four disulfide bonds, common in other allergenic proteins. The model in

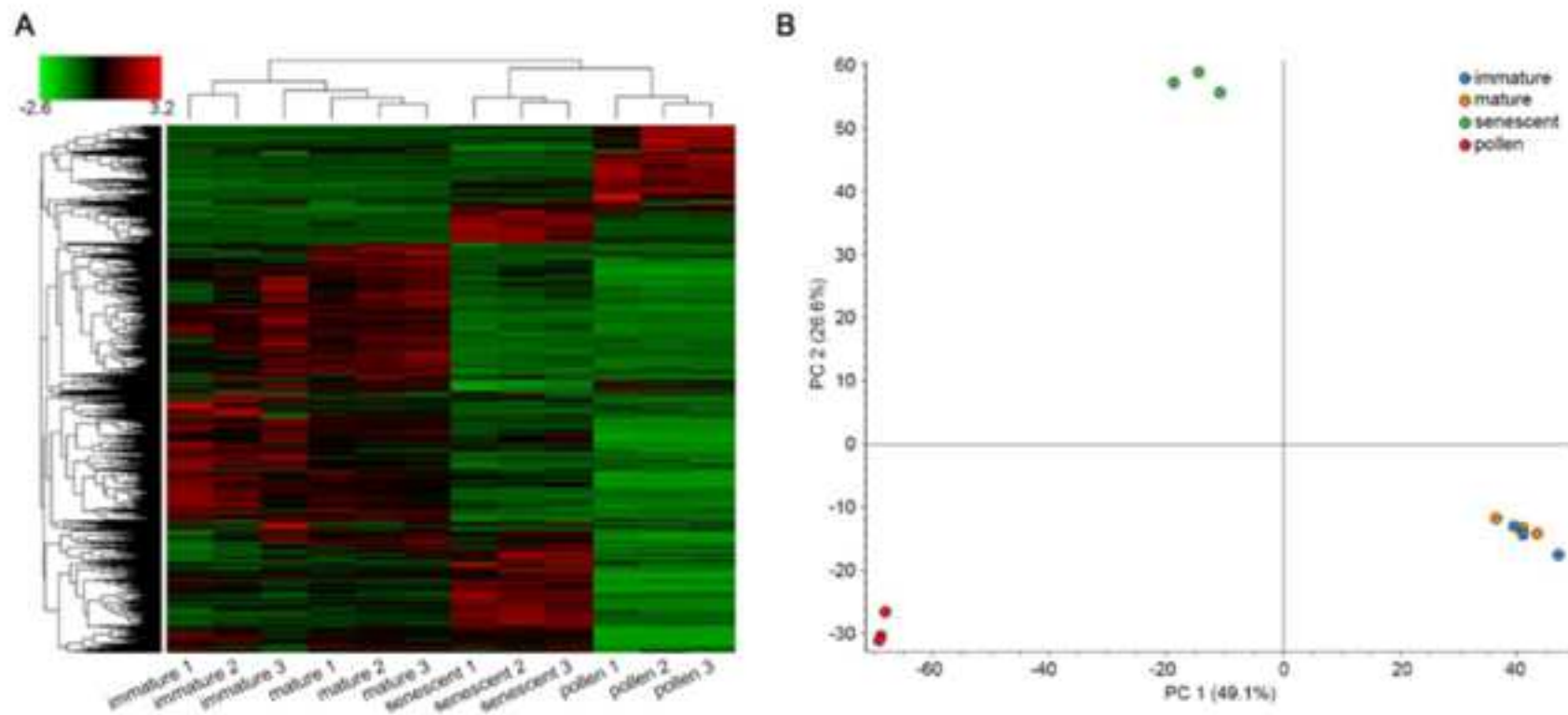
1004 blue is superimposed with a homologous allergen from lentil (PDBid:2MAL) represented in red.
1005 Structure rendered with Pymol 2.3.

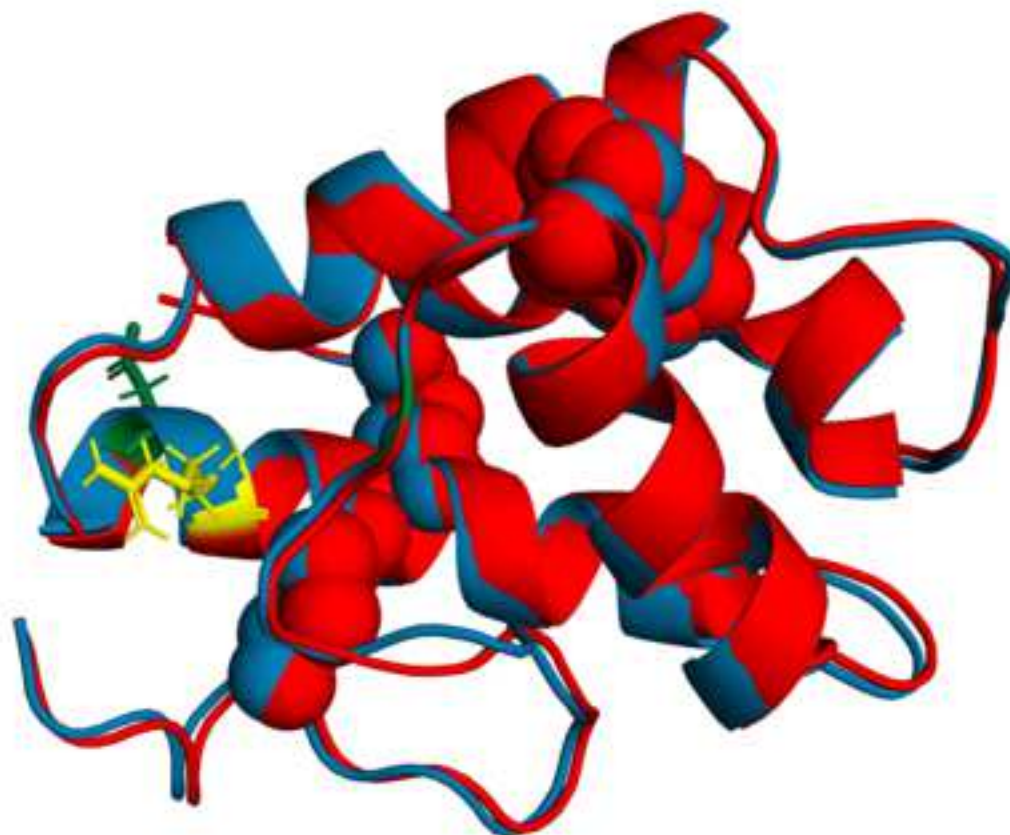
1006 **Figure 5.** Graphical visualization of haplotype-blocks (HB) inheritance on Chr15 along with the
1007 Chandler pedigree. **(A)** The inner-circle highlights in grey two regions of heterozygosity (5 HB
1008 the first and 7 HB the second), and in light green two regions of homozygosity (3 HB the first and
1009 4 HB the second). The circle in the middle shows maternally inherited HBs, while the HBs
1010 inherited through the paternal line are visualized in the outer circle. Payne's haplotypes are clearly
1011 present in both parental lines. White spaces represent segments of missing haplotype information.
1012 **(B)** Chandler pedigree, where Pedro is the maternal line and 56-224, the paternal line.

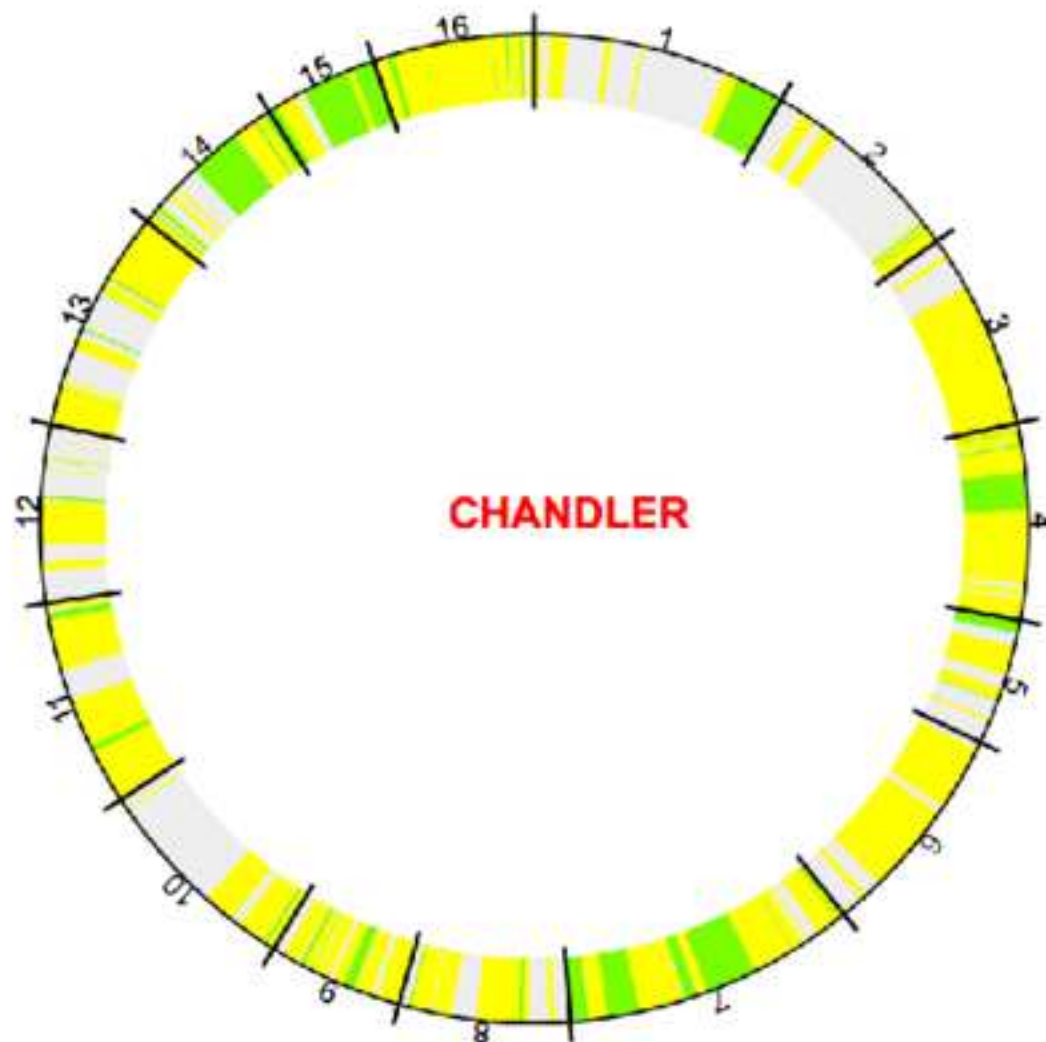
1013 **Figure 6.** Graphical visualization of allele identity between Chandler and its ancestor Payne for
1014 all 16 chromosomes of Chandler.









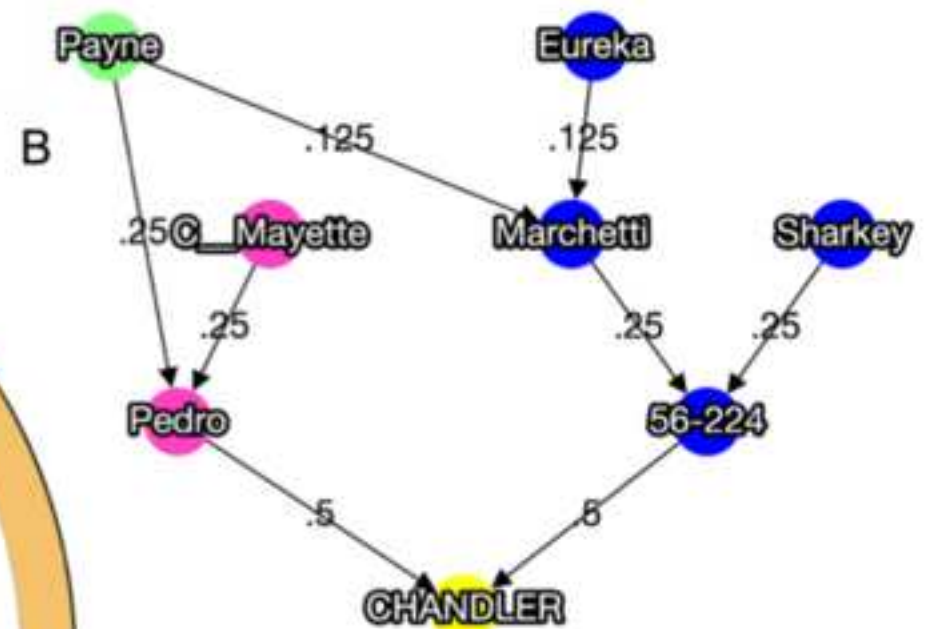
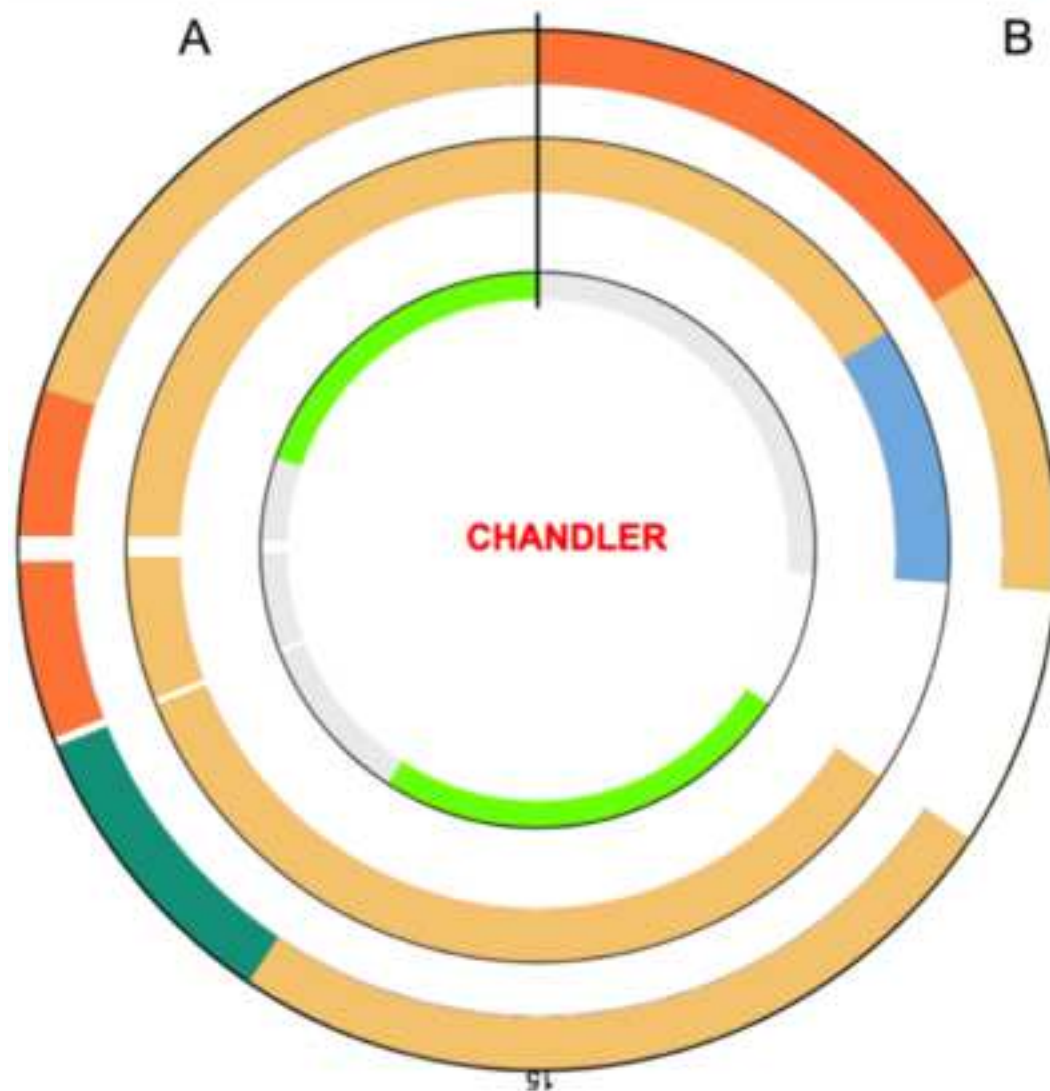
**Legend**

All alleles matching

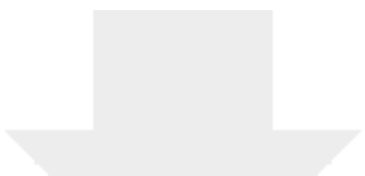
One allele matching *

No alleles matching

* for polyploids at least one allele



Cultivars	Relatedness	Color
56-224	0.31	
Pedro	0.34	
C_Mayette	0.24	
Sharkey	0.18	
Marchetti	0.26	
Payne	0.25	
Eureka	0.17	



Click here to access/download
Supplementary Material
Additional_File_1.docx





Click here to access/download
Supplementary Material
Additional_File_2.xlsx



Click here to access/download
Supplementary Material
Additional_File_3.xlsx

UNIVERSITY OF CALIFORNIA, DAVIS

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO



SANTA BARBARA • SANTA CRUZ

DEPARTMENT OF PLANT SCIENCES
MAIL STOP 6
UNIVERSITY OF CALIFORNIA
ONE SHIELDS AVE
DAVIS, CALIFORNIA 95616-8780
TELEPHONE: 530-752-1703
FAX: 530-752-1819

COLLEGE OF AGRICULTURAL AND
ENVIRONMENTAL SCIENCES
AGRICULTURAL EXPERIMENT STATION
COOPERATIVE EXTENSION

March 13th, 2020

Dear Editor,

With this letter, we are enclosing a revised version of the manuscript entitled “High-quality chromosome-scale assembly of the walnut (*Juglans regia* L.) reference genome” (submission id GIGA-D-19-00363) for publication as a Data Note in *GigaScience*. We are also attaching a detailed response to the referees, a revised version of repeat and gene predictions, as well as three additional files, which comprise all supplementary tables and figures. All authors have approved the revised manuscript for submission. We declare no conflict of interest, and that all previously published work has been thoroughly acknowledged in our manuscript.

We thank you and the reviewers for the valuable comments provided, which we have thoroughly implemented in our manuscripts. In particular, the revised manuscript includes the following major changes:

- More details on the assembly strategy used to obtain the new chromosome-scale reference genome of walnut (in the main text and as supplementary tables);
- Improved annotation of the repeat content of Chandler v2.0;
- A revised version of the gene prediction, along with the analysis of gene family expansion and contraction between Chandler v1.0 and v2.0;
- A comparison with the genome assembly of the *J. regia* cv Serr;
- A comparison with other *Fagales* genomes in terms of genome contiguity, repeat content, and BUSCO completeness;
- Edited versions of Tables 1, 2, and Figure 5.

We believe that our manuscript has significantly improved, and it will be of interest to the audience of *Gigascience*.

Sincerely,

Annarita Marrano

A handwritten signature in blue ink, appearing to read 'Annarita Marrano', written in a cursive style.

Ph.D.

Dept. Plant Sciences, University of California, Davis

GigaScience

High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome

--Manuscript Draft--

Manuscript Number:	GIGA-D-19-00363R2
Full Title:	High-quality chromosome-scale assembly of the walnut (<i>Juglans regia</i> L) reference genome
Article Type:	Data Note
Funding Information:	
Abstract:	Background: The release of the first reference genome of walnut (<i>Juglans regia</i> L.) enabled many achievements in the characterization of walnut genetic and functional variation. However, it is highly fragmented, preventing the integration of genetic, transcriptomic, and proteomic information to fully elucidate walnut biological processes. Findings: Here, we report the new chromosome-scale assembly of the walnut reference genome (Chandler v2.0) obtained by combining Oxford Nanopore long-read sequencing with chromosome conformation capture (Hi-C) technology. Relative to the previous reference genome, the new assembly features an 84.4-fold increase in N50 size, with the 16 chromosomal pseudomolecules assembled and representing 95% of its total length. Using full-length transcripts from single-molecule real-time sequencing, we predicted 37,554 gene models, with a mean gene length higher than the previous gene annotations. Most of the new protein-coding genes (90%) presents both start and stop codons, which represents a significant improvement compared to Chandler v1.0 (only 48%). We then tested the potential impact of the new chromosome-level genome on different areas of walnut research. By studying the proteome changes occurring during male flower development, we observed that the virtual proteome obtained from Chandler v2.0 presents fewer artifacts than the previous reference genome, enabling the identification of a new potential pollen allergen in walnut. Also, the new chromosome-scale genome facilitates in-depth studies of intraspecies genetic diversity by revealing previously undetected autozygous regions in Chandler, likely resulting from inbreeding, and 195 genomic regions highly differentiated between Western and Eastern walnut cultivars. Conclusion: Overall, Chandler v2.0 will serve as a valuable resource to understand and explore walnut biology better.
Corresponding Author:	Annarita Marrano, Ph. D. Davis, CA UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Annarita Marrano, Ph. D.
First Author Secondary Information:	
Order of Authors:	Annarita Marrano, Ph. D.
	Monica Britton
	Paulo Adriano Zaini
	Aleksey V. Zimin
	Rachael E. Workman
	Daniela Puiu
	Luca Bianco
	Erica Adele Di Pierro
	Brian J. Allen
	Sandeep Chakraborty

	Michela Troggio
	Charles A. Leslie
	Winston Timp
	Abhaya Dandekar
	Steven L. Salzberg
	David B. Neale
Order of Authors Secondary Information:	
Response to Reviewers:	Reviewer #1: I read the revised manuscript by Marrano et al. entitled "High-quality chromosome-scale assembly of the walnut (<i>Juglans regia</i> L) reference genome", my comments have been addressed and the authors modified the manuscript accordingly. I still have a few minor points that need to be checked before publication. * I maintain that the comparison of the Chandler v2 assembly with that provided by Zhu et al. is an important point for the reader and I would expect to have a comparison of the two assemblies in Table 1. Done * Please check how MaSurCA is written (MaSuRCA or MaSurCa) We named MaSuRCA consistently across the manuscript. * Table S3 : ON instead of ONT Changed. Reviewer #2: I found the manuscript to be a pleasant read, with the major methodological unclarities resolved. There are two minor suggestions from my side. 1. To promote Table S9 to a full table (Table 1 and 2 could probably be condensed into one table) with the inclusion of the <i>Quercus lobata</i> and <i>Quercus robur</i> genomes, as I feel that for readers it is always good to have some upfront benchmarking so they can put the assembly into the context of other genome assemblies. We moved Table S9 as full table in the main text (now Table 3), and we also included the BUSCO metrics for <i>Quercus lobata</i> and <i>Q. robur</i>. We moved Figure 4 (now Figure S10) to the supplementary in place of Table S9. 2. To promote Figure S12 to a full figure (if possible). We moved Figure S12 as a full figure (now Figure 5) in place of Figure 6 (now Figure S13).
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No

Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

High-quality chromosome-scale assembly of the walnut (*Juglans regia* L.) reference genome

Annarita Marrano¹, amarrano@ucdavis.edu; *corresponding author*

Monica Britton², mtbritton@ucdavis.edu

Paulo A. Zaini¹, pazaini@ucdavis.edu

Aleksey V. Zimin^{3,4}, alekseyz@jhu.edu

Rachael E. Workman³, rachael.e.workman@gmail.com

Daniela Puiu⁴, dpuiu@jhu.edu

Luca Bianco⁵, luca.bianco@fmach.edu

Erica Adele Di Pierro⁵, erica.dipierro@fmach.it

Brian J. Allen¹, brallen@ucdavis.edu

Sandeep Chakraborty¹, sanchak@gmail.com

Michela Troggio⁵, michela.troggio@fmach.it

Charles A. Leslie¹, caleslie@ucdavis.edu

Winston Timp^{3,4}, wtimp@jhu.edu

Abhaya Dandekar¹, amdandekar@ucdavis.edu

Steven L. Salzberg^{3,4,6}, salzberg@jhu.edu

David B. Neale¹, dbneale@ucdavis.edu

¹ Department of Plant Sciences, University of California, Davis, CA 95616, USA

² Bioinformatics Core Facility, Genome Center, University of California Davis, CA 95616, USA

³ Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205, USA

⁴ Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21205, USA

⁵ Research and Innovation Center, Fondazione Edmund Mach, San Michele all'Adige (TN) 38010, Italy

⁶ Departments of Computer Science and Biostatistics, Johns Hopkins University, Baltimore, MD 21218

Abstract

Background: The release of the first reference genome of walnut (*Juglans regia* L.) enabled many achievements in the characterization of walnut genetic and functional variation. However, it is highly fragmented, preventing the integration of genetic, transcriptomic, and proteomic information to fully elucidate walnut biological processes. **Findings:** Here, we report the new chromosome-scale assembly of the walnut reference genome (Chandler v2.0) obtained by combining Oxford Nanopore long-read sequencing with chromosome conformation capture (Hi-C) technology. Relative to the previous reference genome, the new assembly features an 84.4-fold increase in N50 size, with the 16 chromosomal pseudomolecules assembled and representing 95% of its total length. Using full-length transcripts from single-molecule real-time sequencing, we predicted 37,554 gene models, with a mean gene length higher than the previous gene annotations. Most of the new protein-coding genes (90%) presents both start and stop codons, which represents a significant improvement compared to Chandler v1.0 (only 48%). We then tested the potential impact of the new chromosome-level genome on different areas of walnut research. By studying the proteome changes occurring during male flower development, we observed that the virtual proteome obtained from Chandler v2.0 presents fewer artifacts than the previous reference genome, enabling the identification of a new potential pollen allergen in walnut. Also, the new chromosome-scale genome facilitates in-depth studies of intraspecies genetic diversity by revealing previously undetected autozygous regions in Chandler, likely resulting from inbreeding, and 195 genomic regions highly differentiated between Western and Eastern walnut cultivars. **Conclusion:** Overall, Chandler v2.0 will serve as a valuable resource to understand and explore walnut biology better.

Keywords: Nanopore, Hi-C, IsoSeq, gene prediction, genetic diversity, proteome, allergens.

53

54 **Introduction**

55 Persian walnut (*Juglans regia* L.) is among the top three most-consumed nuts in the world, and
56 over the last ten years, its global production increased by 37% (International Nut and Dried Fruit
57 Council, 2019). Its richness in alpha-linolenic acid (ALA), proteins, minerals, and vitamins, along
58 with documented benefits for human health, explains this increased interest in walnut consumption
59 [1]. As suggested by its generic name *Juglans* from the Latin appellation ‘*Jovis glans*’, which
60 loosely means ‘nut of gods’, the culinary and medical value of Persian walnut was already widely
61 prized by ancient civilizations [2].

62 The origin and evolution of the Persian walnut are the results of a complex interplay between
63 hybridization, human migration, and biogeographical forces [3]. A recent phylogenomic analysis
64 revealed that Persian walnut (and its landrace *J. sigillata*) arose from an ancient hybridization
65 occurred between American black walnuts and Asian butternuts after a climate-driven range
66 expansion in Eurasia during the Pliocene [4]. Evidence suggests that the mountains of Central Asia
67 were the cradle of domestication of Persian walnut [5], from where it spread to the rest of Asia,
68 the Balkans, Europe, and, finally, the Americas.

69 Today, walnut is cultivated worldwide in an area of 1,587,566 ha, mostly in China and the USA
70 (FAOSTAT statistics, 2017). Considerable phenotypic and genetic variability can be observed in
71 this wide distribution area, especially in the Eastern countries, where walnuts can still be found in
72 wild fruit forests. Many studies on genetic diversity in walnut have outlined a genetic
73 differentiation between Eastern and Western genotypes [6,7]. Moreover, walnuts from Eastern

Europe, Central Asia, and China exhibit higher genetic diversity and a higher number of rare alleles than the genotypes from Western countries [8].

The release of the first reference genome, Chandler v1.0 [9], enabled the study of walnut genetics at a genome-wide scale. For the first time, it was possible to explore the gene space of Persian walnut with the prediction of 32,498 gene models, providing the basis to untangle complex phenotypic pathways, such as those responsible for the synthesis of phenolic compounds. The availability of a reference genome marked the beginning of a genomics phase in Persian walnut, allowing whole-genome resequencing [4,10], the development of high-density genotyping tools [7,11], and the genetic dissection of important agronomical traits in walnut [12–15]. However, the Chandler v1.0 assembly is highly fragmented, compromising the accuracy of gene prediction and the fulfillment of advanced genomics studies necessary to resolve many, still unanswered questions in walnut research.

The recent introduction of long-read sequencing technologies and long-range scaffolding methods has enabled chromosome-scale assembly for multiple plant species, including highly heterozygous crops such as almond (*Prunus dulcis*; [16]) and kiwifruit (*Actinidia eriantha*; [17]). The availability of genomes with fully assembled chromosomes provides foundations for understanding plant domestication and evolution [16,18,19], the mechanisms governing important traits (e.g., flower color and scent; [20]), as well as the impact of epigenetic modifications on phenotypic variability [21]. Recently, Zhu et al., (2019) assembled the parental genomes of a hybrid *J. microcarpa* × *J. regia* (cv. Serr) at the chromosome-scale using long-read PacBio sequencing and optical mapping. They relied on the haplotype divergence between the two *Juglans* species and demonstrated an ongoing asymmetric fractionation of the two subgenomes present in *Juglans* genomes.

Here we report a new chromosome-level assembly of the walnut reference genome, Chandler v2.0, which we obtained by combining Oxford Nanopore long-read sequencing [23] with chromosome conformation capture (Hi-C) technology [24]. Thanks to the increased contiguity of Chandler v2.0, we were able to substantially improve gene prediction accuracy, with new, longer gene models identified and many fewer artifacts compared to Chandler v1.0. Also, the availability of full, chromosomal sequences reveals new genetic diversity of Chandler, previously inaccessible through standard genotyping tools, and significant genetic differentiation between Western and Eastern walnuts at 195 genomic regions, including also loci involved in nut shape and harvest date. In the present research, we demonstrate the fundamental role of a chromosome-scale reference genome to integrate transcriptomics, population genetics, and proteomics, which in turn enable a better understanding of walnut biology.

Genome long-read sequencing and assembly

To increase the contiguity of the Chandler genome, we first generated deep sequence coverage using Oxford Nanopore Technology (ONT), a cost-effective long-read sequencing approach that determines DNA bases by measuring the changes in electrical conductivity generated while DNA fragments pass a tiny biological pore [25]. Since the release of the first plant genome assembly generated using ONT sequencing [26], this technology has been applied to sequence and obtain chromosome-scale genomes of many other plant species [27–29]. In Persian walnut, ONT sequencing yielded 7,096,311 reads that provided 21.9 Gbp of sequence, or ~35X genome coverage (assuming a genome size of 620 Mb). Read lengths averaged 3.1 kb, and the N50 read length was 6.7 kb, with the longest read being 992.2 kb (**Additional file 1: Table S1**).

One of the major limitations of long-read sequencing technologies is their high error rate, which can range between 5% and 15% for Nanopore sequencing [30]. To overcome this limitation, we

adopted the hybrid assembly technique incorporated into the **MaSuRCA** assembler, which combines long, high-error reads with shorter but much more accurate Illumina sequencing reads to generate a robust, highly contiguous genome assembly [31]. First, using the Illumina reads, we created 3.7 million 'super-reads' with a total length of 2.9 Gb. We then combined the super-reads with the ONT reads to generate 3.2 million mega-reads with a mean length of 4.7 kb, representing 24X genome coverage (**Additional file 1: Table S2**). Finally, we assembled the mega-reads to obtain the 'hybrid' Illumina-ONT assembly, which comprised 1,498 scaffolds, 258 contigs, and 25,007 old scaffolds from Chandler v1.0 (**Additional file 1: Table S3**).

Even though the total number of scaffolds (> 1 Kb) was reduced by 80% compared to Chandler v1.0 (**Table 1**), the new hybrid assembly was still fragmented. To improve the assembly further and build chromosome-scale scaffolds, we applied Hi-C sequencing, which is based on proximity ligation of DNA fragments in their natural conformation within the nucleus [24]. The HiRise scaffolding pipeline processed 356 million paired-end 100-bp Illumina reads to generate the HiRise assembly (**Table 1**). The top 17 scaffolds from this assembly spanned more than 90% of the total assembly length, with a scaffold length ranging from 19.6 to 45.2 Mb (**Additional file 1: Figures S1-S2**). As shown in **Table 1**, the Chandler genome contiguity increased dramatically compared to the previous assemblies. As compared to the recently published genome assembly of the walnut cultivar Serr [22], Chandler v2.0 was less contiguous at the contig level, with a N50 size of 1.1 Mb against the 15.1 Mb of JrSerr_v1.0. The higher coverage PacBio sequencing data (57.2 Gbp) used to assemble JrSerr_v1.0 may explain this discrepancy in contiguity between the two assemblies. Besides, our assembly presented similar value of contiguity to the recently published genomes of pecan (*Carya illinoensis*; 1.1 Mb [32]), Chinese chestnut (*Castanea mollissima*; 944.4 kb [33]) and oak (*Quercus robur*; 1.35 Mb [34]).

Validation of the HiRise assembly

To assess the quality of the HiRise assembly, we used two independent sources of data. First, we used the single nucleotide polymorphism (SNP) markers mapped on the high-density genetic map of Chandler recently described by [14]. Out of the 8,080 SNPs mapped into 16 linkage groups (LGs), 6,894 had probes aligning uniquely on the HiRise assembly with 98% of identity for more than 95% of their length. A total of 35 scaffolds of the HiRise assembly could be anchored to a chromosomal linkage group by at least one SNP (**Figure 1**). In particular, 13 LGs were spanned by a single HiRise scaffold, while two to three scaffolds each aligned the remaining three LGs.

Second, we anchored the HiRise assembly to the Chandler genetic map used by [35] to construct a walnut physical map. In total, 972 of the mapped markers (1,525 SNPs) aligned uniquely on the same 35 HiRise scaffolds anchored to the linkage map mentioned above. Overall, we observed almost perfect collinearity between the HiRise assembly and both Chandler genetic maps (**Figure 1, Additional file 1: Figure S3**). Therefore, we oriented, ordered, and named the HiRise scaffolds consistent with the linkage map of [35], generating the final 16 chromosomal pseudomolecules of *J. regia* Chandler.

These 16 contiguous chromosomal scaffolds account for 95% of the final walnut reference genome v2.0, with an N50 scaffold size of 37 Mb. We identified telomere sequences at both ends for nine of the chromosome scaffolds, on one end of the other seven chromosomes and one end of seven unanchored scaffolds. Also, all 16 chromosomes had centromeric repeats in the middle, alongside regions with low recombination rates (**Figure 2**).

As compared to the previous Chandler genome assemblies (**Table 1**), Chandler v2.0 had a smaller genome size (573.9 Mb), much closer to the Genomescope estimate of 488.2 Mb. This reduction

in genome size represents a great improvement of Chandler v2.0 and can be related to the removal of haplotype variants, likely interpreted and annotated as different scaffolds in the previous genome versions. Compared to the Serr walnut genome (JrSerr_v1.0; 534.7 Mb) [22], Chandler v2.0 had a larger genome size, likely due to structural variation (e.g., copy number and presence/absence variants), whose central role in explaining intraspecific genomic and phenotypic diversity has been reported in different plant species [36,37]. In addition, the higher number of unanchored scaffolds (2,631; 20.9 Mb) in Chandler v2.0 compared to JrSerr_v1.0 can represent autozygous genomic regions of Chandler, devoid of segregating markers and, therefore, difficult to anchor to linkage genetic maps [35], as also suggested by the higher fixation index (F) observed in Chandler (0.03) than Serr (-0.29) in previous genetic surveys [7]. The two walnut assemblies, though, aligned with high sequence identity (over 98% for more than 95% of their total length) and showed high collinearity (**Additional file 1: Figure S4**). Future comparative genomics studies will provide further insights on the functional and structural differences between the two genome assemblies, and their explanatory involvement in the morphological and physiological variation of these two walnut cultivars.

To assess the sequence accuracy of Chandler v2.0, we first compared the scaffold sequences of Chandler v2.0 with the previous version of the walnut reference genome. About 578 Mb of sequence were mutual best alignments, namely best hits of each location between Chandler v2.0 and v1.0 and vice versa, with a sequence identity of 99.6%. We also observed that 135 Mb of Chandler v1.0 (18.9%) aligned to the same locations in Chandler v2.0, suggesting the presence of redundant haplotypes in the previous version of the walnut reference genome that have been removed in our assembly. We then mapped the Illumina whole-genome shotgun data [9] against the new chromosome-scale genome. The alignment resulted in 64,950,691,681 bps mapped, of

which 407,450,406 were single-base mismatches, consistent with an Illumina sequence accuracy rate of 99.5%.

Repeat annotation

More than half (58.4%) of the new Chandler v2.0 is repetitive. This estimate is higher than the previous version of the walnut reference genome (51.19%) and comparable with other *Fagales* genomes [34,38]. As in most plant genomes, interspersed repeats were the most abundant type of repeats, with retrotransposons at 36.45% and DNA transposons at 15.86%. *Gypsies* (10.5%) and *Copias* (7.69%) were the most represented classes of long-terminal retrotransposons (LTR), and, though widely dispersed throughout the genome, they were distributed differently along the 16 chromosomes (**Additional file 1: Figure S5**): the *Gypsies* LTRs were more abundant alongside the centromeres, where, instead, the density of the *Copia* LTRs decreased, as previously observed in walnut [22]. The long-interspersed nuclear elements (L1/LINE), which possess a poly(A) tail and two open reading frames (ORFs) for autonomous retrotransposition, was the largest class of non-LTRs at 7.14% of the genome. Simple repeats (1.91%) were also found.

PacBio IsoSeq sequencing and gene annotation

A fragmented reference genome can severely hamper the accuracy of gene prediction, because many genes will be broken across multiple small contigs (false negatives), and because heterozygous gene variants may be annotated separately (false positives).

To improve the gene prediction accuracy of Chandler v2.0, we used the “Isoform Sequencing” (Iso-Seq) method, developed by Pacific Biosciences (PacBio), which can generate full-length transcripts up to 10 kb, allowing for accurate determination of exon-intron structure by the alignment of the transcripts to the assembly [39]. The high error rate of PacBio sequencing can be

greatly reduced using circular consensus sequence (CCS), in which a transcript is circularized and then sequenced repeatedly to self-correct the errors. We applied PacBio IsoSeq to sequence full-length transcripts from nine tissues, chosen to cover most of the transcript diversity in walnut (**Additional file 1: Table S4**). Across the four SMRT cells, we obtained 26,328,087 subreads with a mean length of 1,188 bp (**Additional file 1: Table S5**) and CCSs ranging from 13K to 142K per library (**Additional file 1: Table S6**). Out of the 745,730 full-length non-chimeric (FLnc) transcripts, 68,225 were classified as high quality, FL (HQ FL) consensus transcript sequences, with an average length of 1,357 bp (**Additional file 1: Table S6**). Catkin 1-inch elongated (CAT1), shoot, and root yielded the lowest number of HQ FL transcripts, while pollen and leaf had the lowest number of HQ consensus clusters obtained per CCS after polishing (**Additional file 1: Table S6**). These results can be explained by lower cDNA quality or fewer inserts of full-length transcripts from these tissues during the cDNA pooling and library preparation. Nevertheless, more than 99% of the HQ FL transcripts aligned onto the new chromosomal-level walnut reference genome (**Additional file 1: Table S7**).

By combining the HQ FL transcripts with available *Juglans* transcriptome sequences, we identified 37,554 gene models, which are more than those annotated in Chandler v1.0 but fewer than the predicted genes in the NCBI RefSeq *J. regia* annotation generated with the first version of the reference genome (**Table 2**). Thus, the new chromosome-scale genome, along with the availability of full-length transcripts, allowed us to identify genes missed in Chandler v1.0 due to genome fragmentation, as well as to remove false-positive predictions likely caused by heterologous variants of the same locus mistakenly interpreted and annotated as independent scaffolds in Chandler v1.0. Also, the mean gene length in Chandler v2.0 was higher than the previous gene annotations (**Table 2**), a consequence of the increased contiguity of the new chromosome-scale

reference genome. The average gene density of Chandler v2.0 was 19.75 genes per 100 kb, with higher gene content in the proximity of telomeric regions (**Figure 2**), consistent with other plant genomes [19,40]. The majority of the predicted gene models of Chandler v2.0 was supported by expression data, and showed high similarity with a protein-coding transcript of other plant species (**Additional file 1: Table S8**). Also, 30,318 models were annotated with 8,243 different Gene Ontology (GO) terms (**Additional file 1: Figures S6-S8**).

Out of the 40,884 transcripts identified, 84% were multi-exonic, with 5.9 exon each, on average, and longer introns than the previous gene annotations of Chandler (**Table 2**). The majority of intron/exon junctions were GT/AG-motif (98.2%), even though alternative splicing with non-canonical motifs was also observed (GC/AG – 0.8%; AT/AC – 0.11%). Almost 90% (36,422) of the coding sequences presented both canonical start and stop codons, while 4,462 had either a start or a stop codon. This result represents a great improvement compared to Chandler v1.0, where only 48% of the predicted gene models presented both start and stop codons [9].

Also, we observed that 2,801 gene models had from two to four transcript isoforms each, with a mean length of 9,389 bp. This proportion of gene models with multiple transcript isoforms is smaller compared to other plant species [41,42], likely due to the low depth of coverage of our PacBio sequencing. Out of the 6,437 isoforms identified, 1,448 were covered by FL HQ transcripts in at least one tissue, while 5,689 were expressed in at least one of the 20 tissues [9], which most likely covered higher gene diversity compared to the nine tissues used for PacBio IsoSeq. On average, the Illumina isoforms (9,188bp) were longer than the PacBio isoforms (6,790 bp). By running the EnTAP functional annotation pipeline with the entire NCBI RefSeq plant database [43], we observed that almost all isoforms (98%; 6,287) were annotated with a plant protein.

We also investigated possible gene family expansion and contraction among the three Chandler's gene annotations. Overall, we identified fewer gene families in Chandler v2.0 (5,163 Panther family represented) than v1.0 (5,330) and NCBI RefSeq *J.regia* annotation (5,374). However, when counting the number of members per family, we observed a gene family expansion, in general, in Chandler v2.0 compared to v1.0: 39,357 proteins were assigned to a Panther gene family in Chandler v2.0, with an average of 7.6 members per family, against the 30,639 proteins annotated with a Panther domain in v1.0 (6 members per family on average). On the contrary, we noticed an overall gene family contraction in v2.0 compared to NCBI RefSeq, where 10.4 gene members were assigned to a Panther domain on average. Both the increment of contiguity and the reduction in haplotype redundancy can explain the observed patterns of gene family expansion and contraction among the three Chandler's gene annotations, even if the different methods of gene prediction used in the three studies could also account for these differences.

Most of the 1,440 core genes in the embryophyte dataset from Benchmarking Universal Single-Copy Orthologs (BUSCO) were assembled completely (82.5% single-copy; 12.6% duplicated), similarly to other *Fagales* genome assemblies (**Table 3**) [32,33,38,44,45]. Also, 88% of both rosids and green sets of core gene families (coreGFs) were identified in the gene annotation, confirming the high-quality and completeness of the gene space of Chandler v2.0.

Improved assessment of proteomes with the complete genome sequence

After confirming the importance of a chromosome-scale reference genome for the improvement of gene prediction accuracy, we analyzed the impact of a contiguous genome sequence using proteomic analysis. Proteomes are commonly investigated by isolating the total protein complement of a sample and fragmenting those proteins into smaller peptides that are resolved by mass and charge by mass spectrometry. After detection, the peptides' amino acid sequences are

determined by matching their mass and charge to candidate sequences obtained from a reference proteome inferred from the reference genome (virtual proteome). A fragmented assembly of the reference genome can lead to an inaccurate prediction of a species' proteome and, then, a miss-identification of the proteins expressed in specific tissues at particular stages [46].

We isolated proteins of reproductive tissues harvested from mature Chandler walnut trees, focusing on different development stages of the male flower (catkin; **Additional file 1: Figure S9**) and mature pollen grains. We analyzed the proteomic data generated from these samples using the virtual proteomes predicted from the gene annotation of the new chromosome-scale genome and Chandler v1.0 (NCBI RefSeq). Considering all tissues analyzed, we identified fewer unique peptides (43,083) with the new chromosome-scale walnut genome than with Chandler v1.0 (44,679). In addition, 6,966 unique proteins were detected with Chandler v2.0 against the 8,802 found using version 1 as a search database (**Additional file 2: Table S9; Additional file 3**). Most likely, the NCBI proteomic database based on the fragmented Chandler v1.0 included artifacts resulting from an overestimation of the protein-coding genes.

In the example presented below, we focused on the allergenic proteins produced during catkin and pollen development. Approximately 2% of walnut consumers have a high risk of developing allergies to nuts or pollen [47]. Initially, we clustered the samples according to their protein constituents and levels. This revealed a higher similarity between immature and mature catkins and a more distinct pattern of detected proteins between senescent catkins and pure pollen (**Figure 3**).

We then searched the four analyzed proteomes for allergenic proteins listed in the WHO/IUIS Allergen Database (www.allergen.org; **Additional file 2: Table S10**), as well as for additional proteins not yet registered in the allergen database but predicted in Chandler v2.0 as potential

allergens given their predicted structural similarity to known allergens (**Additional file 3**). Four of the eight recognized allergenic proteins were detected in at least one of the catkin developmental stages, with Jug_r_5 (XP_018825777 | *Jr12_10750*) and Jug_r_7 (XP_018808763 | *Jr07_28960*) present in all sample types, including pollen (**Additional file 2: Table S10**). Genes adjacent to known allergen-coding sequences, likely indicating gene duplications, encode three of the new potential allergens (**Additional file 2: Table S10**). Moreover, we discovered that the gene locus *Jr12_05180* encodes a non-specific lipid transfer protein (nsLTP; Jug_r_9 | XP_018813928), a potential allergen highly expressed during catkin maturation and in pollen (**Additional file 2: Table S10-11**). In particular, Jug_r_9 was the most abundant protein in mature and senescent catkins, and the second most abundant in pure pollen. Another interesting allergen similar to Jug_r_9 (same eight cysteine configuration) is XP_018814382 | *Jr03_26970*; it decreases as the catkin matures, and is absent in pollen (**Additional file 2: Table S10-11**). Similarly, polyphenol oxidase (PPO, XP_018858848 | *Jr03_06780*) is high in the immature catkin and almost absent in the pollen.

The integration of this proteomic data with previously published transcriptomic data obtained from 20 walnut tissues [9] shows high reproducibility between the methods. In both datasets, allergens Jug_r_1, 4, and 6 were not detected in catkins, while the new putative allergen Jug_r_9 was highly expressed in catkins (**Additional file 2: Table S11-12**). Also, *Jr12_05180* transcripts were not detected in any of the 20 tissues but catkin, thus confirming the strong specificity of Jug_r_9 for catkin and pollen tissue (**Additional file 2: Table S12**). Modeling the structure of this putative allergen reveals four predicted disulfide bonds, potentially conferring heat and protease-resistance, and further suggesting allergenic properties (**Figure S10**). Future studies will clarify the functional role of this protein and its allergenic nature.

The detection of new potential walnut allergens confirms the positive impact of Chandler v2.0 on proteomic studies in walnut, by providing a clearer and more precise organization of the CDSs within a genomic region than the previous fragmented genome assembly v1.0.

Chandler genomic diversity

By anchoring the HiRise assembly to the Chandler genetic map [14], we observed highly homozygous regions in Chandler, especially on Chr15, where the genetic gap spanned 14.5 cM, corresponding to a physical distance of 9.1 Mb. A large gap on Chr15 (9.23 cM – 1.5 Mb) was also observed by [35], which suggested inbreeding as a possible cause for the lack of segregating loci in this region in Chandler, whose parents shared Payne as an ancestor. To confirm the autozygosity of Chandler on Chr15, we used the Illumina whole-genome shotgun data of Chandler and the identified polymorphisms to study its genetic diversity across the new chromosome-scale genome. We identified 2,205,835 single heterozygous polymorphisms on the 16 chromosomal pseudomolecules, with an SNP density of 4.0 SNPs per kb (**Figure 2; Additional file 1: Table S13**). Fifty-six 1-Mb-regions exhibited less than 377.5 SNPs (10th percentile of the genome-wide SNP number distribution), and chromosomes 15, 1, 7, and 13 were the top four chromosomes in the number of low heterozygous regions (**Additional file 1: Table S14**). In particular, Chr15 presented nine 1-Mb windows with a significantly low number of polymorphisms, five of which span 4 Mb at the end of the chromosome. In these nine low heterozygous regions, we found 1,536 SNPs in total (**Figure 2**), of which only 25 were tiled on the Axiom *J. regia* 700K SNPs array. The absence of these polymorphisms segregating in Chandler in the SNP array could be related to either a failed identification during the SNP calling due to the highly fragmented reference genome v1.0 or with the SNP exclusion during the filtering process applied to build the genotyping array [7]. The low number of Chandler heterozygous SNPs in the array affected the end of Chr15 the

most, causing a reduction in the genetic length of the corresponding linkage group (**Figure 1**), as well as leaving unexplored 4 Mb of Chandler genetic variability, which is now accessible thanks to the new chromosome-scale reference genome. The failure to anchor seven of the scaffolds with telomeric sequences can be explained by the missed detection of terminally located highly homozygous regions during genetic map constructions, due to the absence of crossing-over events with heterozygous flanking markers.

Due to the evidence of whole-genome duplication in *Juglans* genomes [35], we searched for conserved regions of synteny between Chr15 and its homologous regions in the genome, to study their level of divergence and identify other evolutionary forces as possible causes of the localized reduction of heterozygosity on Chr15. Of the 5,739 pairs of paralogous genes (8,701 genes; **Additional file 1: Figure S11**) identified in Chandler v2.0, 448 included genes on Chr15, and 389 of these have their respective paralogues on Chr6 (**Additional file 1: Figure S12**), in line with what was already reported by [35]. The Chr06-Chr15 pairs of paralogous genes showed average values of divergence indexes ($K_S = 0.38$; $K_A = 0.13$) similar to the ones observed genome-wide for other syntelogs ($K_S = 0.4$; $K_A = 0.09$), which are paralogous genes derived from the same ancestral genomic region. Similar values of divergence were also observed for the 178 Chr06-Chr15 syntelogs (171 genes) falling within the nine low heterozygous regions on Chr15 ($K_S = 0.4$, $K_A = 0.1$), excluding different evolutionary rates for these regions. Other than paralogous genes, we found 393 singleton genes in the low heterozygous regions on Chr15 of Chandler. These genes are involved in different biological processes, many of which related to signal transduction, protein phosphorylation, and response to environmental stimuli (**Additional file 1: Table S15**).

We further investigated the contribution of inbreeding to the high level of autozygosity on Chr15 by visualizing the inheritance of haplotype-blocks (HB; genomic regions with little recombination)

across the Chandler pedigree (**Figure 4B**). We observed that Payne accounts for the entire Chandler genetic makeup (19 HBs for the total length of Chr15) inherited from Pedro (mother), where only one HB (2,08 Mb) shared the same allele of Conway-Mayette (maternal-grandfather; **Figure 4A**). Regarding the paternal genetic makeup of Chandler, 13 out of 19 HBs (9,05 Mb) on Chr15 inherited Payne alleles, providing further evidence of high inbreeding on this chromosome (**Figure 4A**). This is even more evident in assessing the number of alleles matching between Payne and Chandler across the genome: Chr15 (14 HBs for a total of 13,95 Mb; **Figure S13**) shares full allele identity with Payne for almost its entire length. Such allele matching between Chandler and its ancestor Payne also occurs on Chr1 (9 HBs for a total of 8,44 Mb), Chr4 (6 HBs - 7,68 Mb), Chr7 (21 HBs - 21,62Mb) and Chr14 (7 HBs – 12,29 Mb; **Figure 5**). These results suggest high level of inbreeding in many genomic regions of Chandler (**Figure 5**), even though direct and indirect selection might have caused the observed presence of extended homozygous regions in Chandler's genome.

Genomic comparison between Eastern and Western walnuts

Even though numerous surveys regarding genetic diversity within walnut germplasm collections have been reported so far [48,49], comparative analyses at the population level and genome scans for signatures of selection are still missing in Persian walnut. The availability of a chromosome-scale reference genome enables exploration of the patterns of intraspecific variation at the genomic level, providing new insight on the extraordinary phenotypic diversity present within *J. regia*.

We used the resequencing data generated for 23 founders of the Walnut Improvement Program of the University of California, Davis (UCD-WIP; **Additional file 1: Table S16**) [10] to study the genome-wide genetic differentiation among walnut genotypes of different geographical provenance. We identified 14,988,422 SNPs, and over 97% of them were distributed on the 16

chromosomal pseudomolecules, with 9.4 polymorphisms per kb. A hierarchical clustering analysis (Additional file 1: Figure S14) divided the 23 founders into two major groups, including genotypes from western countries (USA, France, and Bulgaria) and Asia (China, Japan, Afghanistan), respectively, as previously reported [7,50]. High phenotypic diversity for many traits of interest in walnut, such as phenology, nut quality, and yield, has been observed within and between germplasm collections from Western and Eastern countries [51]. Walnut trees from Asia are noted for their lateral fruitfulness and precocity, rarely observed in the USA and western Europe, so that they have been used as a source of these phenotypes in different walnut breeding programs [52].

At a genomic level, we found a moderate differentiation ($F_{ST} = 0.15$) between Western and Eastern genotypes, except for 195 genomic windows (100 kb) that showed substantially high population differences ($F_{ST} \geq 0.36$; top 5% in the whole genome). In particular, chromosomes 7, 5, 1, 4, and 2 presented about 70% of the divergent sites (Figure 2; Additional file 1: Figure S15). As suggested by the mean reduction of diversity coefficient (ROD) value (0.41), in most of the genomic regions highly differentiated, the UCD-WIP founders from the USA and Europe showed lower nucleotide diversity ($\pi = 2.5 \times 10^{-4}$) than the Asian genotypes ($\pi = 5.0 \times 10^{-4}$), consistent with [8] (Figure 2; Additional file 1: Figure S15). The proximity of our eastern genotypes to the supposed walnut center of domestication in Central Asia can explain the high level of diversity observed in this subgroup.

More than 60% (122) of the highly differentiated windows showed a negative value of Tajima's D in the EU/USA subgroup ($D_{Occ} = -1.12$), thus, suggesting that selection has been likely acting on these genomic regions in the Western genotypes (Additional file 1: Figure S15). Here we found 743 genes, with GO biological categories mostly related to signal transduction, embryo

415 development, and response to stresses (**Additional file 1: Table S17**). Ten candidate selective
416 sweeps ($D_{Asia} = -0.54$) were also observed in the Eastern group (**Additional file 1: Figure S15**),
417 which included 57 predicted genes, related to terpenoid biosynthesis, post-embryonic
418 development, and signal transduction (**Additional file 1: Table S18**).

419 Recently, many marker-trait associations have been reported for different traits of interest in
420 walnut, such as leafing date, nut-related phenotypes, and water use efficiency [12–14]. We looked
421 to see if any of these trait-associated SNPs fell within regions highly differentiated between
422 Western and Eastern genotypes. Three loci associated with shape index, nut roundness, and nut
423 shape [12] are located in two genomic regions on chromosome 3 and 4 with significantly high
424 values of F_{ST} (**Additional file 1: Table S19**). In both of these regions, Western genotypes
425 presented lower genetic diversity and lower values of Tajima's D than the Eastern walnuts. These
426 findings may suggest that, while a selective pressure for nut shape may have occurred in the
427 EU/USA subgroups, higher phenotypic variability can be expected for these traits in the Eastern
428 countries. We also found that the locus AX-170770379, strongly associated with harvesting date
429 [14], falls within a genomic region on Chr1 with an F_{ST} value equal to 0.39 and lower genetic
430 diversity in the western genotypes ($ROD = 0.63$; **Additional file 1: Table S19**). Looking at the
431 phenotypic effect of this SNP on the harvest date of the 23 founders, we observed that most of the
432 western genotypes are later harvesting than the eastern (**Additional file 1: Figure S16**), suggesting
433 differences in the timing of phenological events between these two groups as adaptation to the
434 different climate conditions present in their countries of origin [53].

435 Future resequencing projects involving larger walnut collections and covering a wider area of the
436 global walnut distribution are necessary to confirm and interpret the observed genomic

differentiation between Western and Eastern walnuts, likely helping to understand the role of this genomic divergence in the evolutionary history of Persian walnut.

Methods

Oxford Nanopore sequencing and assembly

High molecular weight (HMW) DNA for Nanopore sequencing (Oxford Nanopore Technologies Inc., UK) was isolated through a nuclei extraction and lysis protocol. First, mature leaf tissue from the same tree used for the original *J. regia* Chandler genome [9] was homogenized with mortar and pestle in liquid nitrogen until well ground, then added to the Nuclei Isolation Buffer [54], and stirred at 4°C for 10 minutes. The cellular homogenate was filtered through 5 layers of Miracloth (Millipore-Sigma) into a 50 mL Falcon tube, then centrifuged at 4°C for 20 minutes at 3000 x *g*. This speed of centrifugation was selected based on the estimated walnut genome size of 1 Gb [55]. Extracted nuclei were then lysed for 30 minutes at 65°C in the SDS-based lysis buffer described by [56]. Afterwards, 0.3 volumes of 5M potassium acetate were added to the lysate to precipitate residual polysaccharides and proteins. The sample was incubated for 5 minutes at 4°C and then centrifuged at 4°C for 10 minutes at 2400 x *g*. After removing the supernatant, genomic DNA (gDNA) was ethanol precipitated, and then eluted in 10 mM Tris-Cl. Further purification of the gDNA was then performed using a Zymo Genomic DNA Clean and Concentrate column.

One µg of the isolated gDNA was prepared for sequencing using the Ligation sequencing kit (LSK108, Oxford Nanopore) following manufacturer's protocol with an optimized end repair (100 µl sample, 14 µl enzyme, 6 µl enzyme, incubated at 20°C for 20 minutes then 65°C for 20 minutes). In detail, the gDNA was end polished using the NEBNext® Ultra™ II DNA Library Prep Kit, and then cleaned up with 1X Ampure XP beads (Beckman Coulter). Afterwards, the gDNA was ligated

to Oxford Nanopore specific adapters, followed by an additional cleanup with 0.4X Ampure XP beads. Finally, the libraries were sequenced for 48 hours on six flowcells of the Oxford Nanopore Mk1B MinION platform with the R9.4 chemistry. Raw fast5 data were base-called using Albacore version 1.25.

The ONT data and Illumina reads from [9] were combined to obtain the Chandler hybrid assembly using **MaSuRCA** v3.2.3 [57]. In detail, **MaSuRCA** first transformed the Illumina pair-end reads in *super-reads* using the super-reads algorithm, which uses k-mers from Illumina reads to extend each Illumina read uniquely in both directions. Then, each ONT read was used as a template to which super-reads can be attached, and the approximate alignments of all super-reads to each ONT read were computed. The best path of the exactly overlapping aligned super-reads on a ONT read was then defined, generating a *mega-read*. The mega-reads typically have a very low error rate (less than 1%) since they are constructed from the super-reads, and most of them span the full length of the long reads. Finally, a customized version of the CABOG assembler [58] was used to assemble the mega-reads along with the Illumina mate pairs, which provide the linking information for the scaffolding. Gaps were closed using the gap-filling procedure implemented in **MaSuRCA** and described by [57]. The de-duplication module implemented in **MaSuRCA** was then applied to remove duplicative sequences (scaffold variants due to heterozygosity).

De-duplicated scaffolds were aligned onto the previously finished *J. regia* chloroplast genome [9] using "minimap2 -x asm5", as well as to a database of 223 finished plant mitochondria (downloaded from NCBI RefSeq) using blastn with default parameters. Finally, Chandler v1.0 was aligned to the de-duplicated hybrid assembly, and the unaligned regions were added to the Chandler hybrid assembly.

Hi-C sequencing

A Hi-C library was prepared by Dovetail Genomics LLC (Santa Cruz, CA, USA) as described previously [59]. Briefly, for each library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Biotin that was not internal to ligated fragments was removed from the purified DNA. Purified DNA was then sheared to ~350 bp mean fragment size. Sequencing libraries were generated using NEBNext® Ultra™ enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were then sequenced on the Illumina HiSeq4000 platform. The hybrid ONT assembly, Illumina shotgun reads [9], and Dovetail Hi-C library reads were used as input data for the scaffolding software HiRise, which uses proximity ligation data to scaffold genome assemblies [60]. Shotgun and Dovetail Hi-C library sequences were aligned to the hybrid ONT assembly using a modified SNAP read mapper. The separations of Dovetail Hi-C read pairs mapped within the ONT scaffolds were analyzed by HiRise to produce a likelihood model for the genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After scaffolding, Illumina shotgun sequences were used to close gaps between contigs, resulting in an improved HiRise assembly.

Validation and anchoring of the HiRise assembly to Chandler genetic maps

The HiRise assembly was first anchored to the Chandler genetic map obtained by [14] from a 312 offspring F₁ population 'Chandler x Idaho' genotyped with the latest Axiom *J. regia* 700K SNP array. SNP probes (71-mers including the SNP site) from the Axiom *J. regia* 700K SNP array were aligned onto the HiRise assembly filtering out alignments with probe/reference identity lower

505 than 98%, covering less than 95% of the probe length or aligning multiple times on the genome.
506 Retained markers with a unique segregation profile were then used to anchor the HiRise scaffolds.
507 The same procedure was also followed to anchor the HiRise assembly to the Chandler genetic map
508 used to construct a walnut bacterial artificial chromosome (BAC) clone-based physical map by
509 [35]. The final ordering of scaffolds was performed by taking into consideration the marker genetic
510 map position, and, in the final sequence, consecutive scaffolds were separated by sequences of
511 100,000 Ns.

512 The tandem repeat finder program (trf v4.09; [61]) was run using the recommended parameters
513 (max mismatch delta PM PI minscore maxperiod, 2 7 7 80 10 50 500 resp.) to identify repeat
514 elements up to 500 bp long. A histogram of repeat unit lengths was generated, and peaks at 7, 29,
515 33, 44, 154, and 308 bp were identified. From this data, a consensus sequence corresponding to
516 each peak was selected. All of these repeat sequences were aligned onto the HiRise assembly using
517 ‘nucmer’ from the MUMmer4 package [62] with a minimum match length of 7 to capture the
518 telomeric repeat. Based on the positions of these alignments along the chromosomes and contigs,
519 we identified the 7-mer as the telomeric repeat and the 154-mer and 308-mer as centromeric
520 repeats.

521 Recombination rate was estimated within sliding windows of 10 Mb with a step of 1 Mb along the
522 chromosome sequence by using the high-density genetic map of Chandler [14] and the
523 R/MareyMap package v 1.3.4 [63]. To evaluate Chandler v2.0 error rate, the two assemblies,
524 Chandler v1.0 and 2.0, were aligned to each other using the ‘nucmer’ program [62]. Assembly
525 quality statistics were estimated using QUAST v5.0.2 [64], filtering for contigs with a minimum
526 length of 1Kb. The haploid size of the walnut genome was estimated by first generating the 24-
527 mer distribution of Illumina paired-end reads (54-fold coverage of the haploid genome) with

Jellyfish v2.2.6 [65], and then uploading it to Genomescope [66]. Comparisons of Chandler v2.0 versus JrSerr_v1.0 and vice versa were performed using ‘nucmer’ [62], and then the function ‘dnadiff’ implemented in MUMmer4 was used to obtain detailed information on the differences between two assemblies.

RNA preparation

Five walnut tissues (leaf, catkin 1-inch elongated; catkin 3-inches elongated, pistillate flower, and pollen) were collected from ‘Chandler’ trees at the UCD walnut orchards. Four additional samples (somatic embryo, callus, shoot, and roots) were taken from tissue culture material of ‘Chandler’. Several grams of each tissue were ground in liquid nitrogen and with insoluble polyvinylpyrrolidone (PVPP; 1% w/w). RNA was isolated using the PureLink™ Plant RNA Reagent (Invitrogen™, Carlsbad, CA) following the manufacturer’s instructions, but with an additional end wash in 1 mL of 75% ethanol. For root tissue only, RNA isolation was performed using the MagMAX™ mirVana™ Total RNA Isolation Kit (Applied Biosystems™, Foster City, CA) as per protocol, except for the lysis step. A different lysis buffer was created adding 100 mg of sodium metabisulfite to 10 mL of guanidine buffer (guanidine thiocyanate 4M, sodium acetate 0.2M, EDTA 25 mM, PVP-40 2.5%, pH 5.0) and 1 mL of nuclease-free water. Then, 100 mg of ground root tissue were lysed in 1 mL of the new lysis buffer using a Tissue Lyser at max frequency for 2 min. The lysate was centrifuged at 4° C for 5 min at max speed. The supernatant (500 µL) was transferred to a new tube for the following steps of RNA isolation as per protocol. RNA samples were then purified, and DNase treated using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). The RNA quality was confirmed by running an aliquot of each sample on an Experion™ Automated Electrophoresis System (Bio-Rad, Hercules, CA).

PacBio IsoSeq sequencing

Full-length cDNA Iso-Seq template libraries for PacBio IsoSeq analysis were constructed and sequenced at the DNA Technologies & Expression Analysis Core Facility of the UC Davis Genome Center. FL double-stranded cDNA was generated from total RNA (2 µg per tissue) using the Lexogen TeloTM prime Full-length cDNA Kit (Lexogen, Inc., Greenland, NH, USA). Tissue-specific cDNAs were first barcoded by PCR (16-19 cycles) using IDT barcoded primers (Integrated DNA Technologies, Inc., Coralville, Iowa), and then bead-size selected with AMPure PB beads (two different size fractions of 1X and 0.4X). The nine cDNAs were pooled in equimolar ratios and used to prepare a SMRTbellTM library using the PacBio Template Prep Kit (PacBio, Menlo Park, CA). The SMRTbellTM library was then sequenced across four Sequel v2 SMRT cells with polymerase 2.1 and chemistry 2.1 (P2.1C2.1).

PacBio raw reads were processed using the Isoseq3 v.3.0 workflow following PacBio recommendations (<https://github.com/PacificBiosciences/IsoSeq3>). Circular consensus sequences (CCSs) were generated using the program ‘ccs’. The CCSs were demultiplexed and cleaned of cDNA primers using the program ‘lima’. Afterward, CCS clustering and polishing was performed using the program ‘isoseq3’, to generate HQ FL sequences for each of the nine tissues. Full-length non-chimeric (FLnc) and HQ clusters were aligned onto the new ‘Chandler’ assembly v2.0 with minimap2 v.2.12-r827, including the parameter ‘-ax splice’ [67].

Repeat annotation

A genome-specific repeat database was created using the ‘basic’ mode implemented in RepeatModeler v.1.0.11 [68]. RepeatMasker v.4.0.7 was then run to mask repeats in the walnut reference genome v.2.0 and generate a GFF file [69].

Gene prediction and functional annotation

Juglans regia RefSeq transcripts and additional *J. regia* transcripts and protein sequences downloaded from NCBI, along with the HQ FL IsoSeq transcripts, were used as input to the PASA pipeline v.2.3.3 [70], to assemble a genome-based transcript annotation. PASA utilizes the aligners BLAT v.35 [71] and GMAP v.2018-07-04 [72], along with TransDecoder v.5.5.0 [73], which predicts open reading frames (ORFs) as genome-based GFF coordinates. The final PASA/TransDecoder GFF3 file was post-processed to name the genes and transcripts by chromosome location consistently. The chloroplast and mitochondrial genomes were annotated using the “CHLOROBOX GeSeq Annotation of Organellar Genomes” tool at <https://chlorobox.mpimp-golm.mpg.de/geseq.html> with default parameters [74]. NCBI accessions NC_028617.1 (*J. regia* chloroplast), KT971339.1 (*Medicago truncatula* mitochondrion), NC_029641.1 (*M. truncatula* mitochondrion) and NC_012119.1 (*Vitis vinifera* mitochondrion) were also input as custom references. The output gff3 files were then post-processed to consistently rename genes.

Functional roles were assigned to predicted peptides using Trinotate v.3.1.1 [75]. In particular, similarity searches were performed against several public databases (i.e., Uniprot/Swiss-Prot, NCBI NR, *Vitis vinifera*.IGGP_12x, *J. regia* RefSeq) using BLAST v.2.8.1, HMMER v.3.1b2, SignalP v.4.1c, and TMHMM v.2.0c. Gene family analysis was performed by running Interproscan v. 5.30-69.0 [76,77] with default parameters on each protein fasta file (v1.0 [9], NCBI RefSeq [GCF_001411555.1_wgs.5d] and v2.0). The PANTHER family ID with the lowest expect value (below expect value threshold of $1.0E^{-11}$) was assigned to each protein.

The completeness and quality of both genome assembly and gene annotation of Chandler v.2.0 were estimated with the BUSCO method v.3 (1,440 core genes in the embryophyte dataset) [78],

and the sets of coreGFs of green plants (2,928 coreGFs) and rosids (6,092 coreGFs) from PLAZA v.2.5 [79]. Also, RNA-Seq data previously generated for 20 tissues (see [9]) were aligned to the reference genome (v1.0 and v2.0) with HISAT2 [80]. The alignments of the 20 RNA-seq data and the FL transcripts along with the new genome annotation v2.0 were then used as input to StringTie v.2.0 [81] to estimate expression levels in both fragments per kilobase per million reads (FPKM) and transcripts per million (TPM) for each transcript in the v2 annotation. The percent identity and coverage of each *J. regia* transcript compared to proteins in the NCBI plant RefSeq database was also determined by running the EnTAP pipeline v.0.9.0 [43].

Label-free shotgun proteomics

Plant tissues of immature, intermediate, mature catkins (**Additional file 1, Figure S9**) and pure pollen from three individual trees of Chandler at the UCD walnut orchards were collected and frozen immediately in dry ice. Tissues were then further frozen in liquid nitrogen in the laboratory and ground with mortar and pestle. Five hundred milligrams of each sample were used for total protein extraction, following the procedure for recalcitrant plant tissues of [82], with a modification in the final buffer used to resuspend the protein pellet, consisting of 8M urea in 50mM triethylammonium bicarbonate (TEAB). One hundred micrograms of total protein from each sample were then used for proteomics.

Initially, 5 mM dithiothreitol (DTT) was added and incubated at 37°C for 30 min and 1,000 rpm shaking. Next, 15 mM iodoacetamide (IAA) was added, followed by incubation at room temperature for 30 min. The IAA was then neutralized with 30 mM DTT in incubation for 10 min. Lys-C/trypsin then was added (1:25 enzyme: total protein) followed by 4 h incubation at 37°C. After, TEAB (550 µl of 50 mM) was added to dilute the urea and activate trypsin digestion overnight. The digested peptides were desalted with Aspire RP30 Desalting Tips (Thermo

Scientific), vacuum dried, and suspended in 45 μ l of 50 mM TEAB. Peptides were quantified by Pierce quantitative fluorometric assay (Thermo Scientific) and 1 μ g analyzed on a QExactive mass spectrometer (Thermo Scientific) coupled with an Easy-LC source (Thermo Scientific) and a nanospray ionization source. The peptides were loaded onto a Trap (100 microns, C18 100 Å 5U) and desalted online before separation using a reversed-phase (75 microns, C18 200 Å 3U) column. The duration of the peptide separation gradient was 60 min using 0.1% formic acid and 100% acetonitrile (ACN) for solvents A and B, respectively. The data were acquired using a data-dependent MS/MS method, which had a full scan range of 300-1,600 Da and a resolution of 70,000. The resolution of the MS/MS method was 17,500 and the insulation width 2 m/z with a normalized collision energy of 27. The nanospray source was operated using a spray voltage of 2.2 KV and a transfer capillary temperature heated to 250°C. Samples were analyzed at the UC Davis Proteomics Core.

The raw data were analyzed using X! Tandem and viewed using the Scaffold Software v.4. (Proteome Software, Inc.). Samples were searched against UniProt databases appended with the cRAP database, which recognizes common laboratory contaminants. Reverse decoy databases were also applied to the database before the X! Tandem searches. The protein-coding sequences (CDS) annotated in Chandler v1.0 (NCBI accession PRJNA350852) and v2.0 were used as a reference for identification of proteins from the mass spectrometry data. The proteins identified were filtered in the Scaffold software based on the following criteria: 1.0% FDR (false discovery rate) at protein level (following the prophet algorithm: <http://proteinprophet.sourceforge.net/>), the minimum number of 2 peptides and 0.1% FDR at the peptide level. Structure of the walnut allergen (Jug r 9) was modelled using SWISS-MODEL [83] based on the structure of a homologous

allergen from lentil (PDBid:2MAL). Structures were superimposed using MUSTANG (2MAL:in red, walnut in blue) [84].

Chandler genomic diversity

Illumina whole-genome shotgun data of Chandler were aligned on the Chandler v2.0 with BWA [85] with standard parameters. SNP calling was performed using SAMtools v1.9 [86] and BCFtools v.2.1 [87]. SNP density for windows of 1 Mb was estimated using the command ‘SNPdensity’ implemented in VCFtools v0.1.16 [88]. Self-collinearity analysis to detect duplicated regions in Chandler v2.0 was performed with MCScanX [89], using a simplified GFF file of the new gene annotation and a self-BLASTP as input. To improve the power of collinearity detection, tandem duplications were excluded after running the function ‘detect_collinear_tandem_arrays’ implemented in MCScanX. Synonymous (K_S) and nonsynonymous (K_A) changes for syntenic protein-coding gene pairs were measured using the Perl script “add_ka_and_ks_to_collinearity.pl” implemented in MCScanX.

To explore the inbreeding level across the 16 chromosomal pseudomolecules of Chandler, haplotypes were built for 55 individuals of the UCD-WIP, including 25 founders and several commercially relevant walnut cultivars (e.g., Chandler, Howard, Tulare, Vina, Franquette) along with their parents and progenitors. All individuals were genotyped using the latest Axiom™ *J. regia* 700K SNP array as described in [7]. To define SNP HBs, 26,544 unique and robust SNPs were selected and ordered according to the Chandler genome v2.0 physical map. Subsequently, for each SNP markers and individual, phasing and identification of closely linked groups of SNPs, without recombination in most of the pedigree, was performed using the software FlexQTL™ [90] and PediHaplotyper [91] following the approach described in [92] and [91]. In particular, HBs were defined by recombination sites detected in ancestral generation of Chandler.

Genomic comparison between Eastern and Western walnuts

The resequencing data of 23 founders of the UCD-WIP (**Additional file1, Table S16**)[10] were mapped onto the Chandler v2.0 with BWA, and SNPs were called following the same procedure described above for Chandler. SNPs with no missing data and minor allele frequency (MAF) higher than 10% were retained for the following genetic analyses (7,269,224 SNPs out of the 14,988,422 identified). Hierarchical cluster analysis on a dissimilarity matrix of the 23 UCD-WIP founders was performed using R/SNPRelate v.1.18.0 [93]. Fixation index (F_{ST}) was measured between genotypes from EU/USA and Asia with VCFtools v0.1.16, setting windows of 100kb and 500kb. Genomic windows with the top 5% of F_{ST} values were selected as candidate regions for further analysis. The empirical cutoff with a low false discovery rate (5%) was verified by performing whole-genome permutation test (1000) with a custom Python script. Nucleotide diversity (π) and Tajima's D [94] were also computed along the whole genome in 100-kb and 500-kb windows using VCFtools. Reduction of diversity coefficient (ROD) was estimated as $1 - (\pi_{Occ} / \pi_{Asia})$. The new walnut gene annotation v.2.0 was used to identify predicted genes in the candidate regions under selection. The distribution of the identified genes into different biological processes was evaluated using the weight01 method provided by the R/topGO [95]. The Kolmogorov–Smirnov-like test was performed to assess the significance of over-representation of GO categories compared with all genes in the walnut gene prediction. Plots were obtained using the R/circlize v.0.4.6 and R/ggplot2 v.3.5.3 packages.

Availability of supporting data

All raw and processed sequencing data generated in this study have been submitted to the NCBI BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject/>) under accession number PRJNA291087. All SNP data have been submitted to Hardwood Genomics (<https://hardwoodgenomics.org/Genome-assembly/2539069>). Data further supporting this work are openly available in the GigaScience repository, GigaDB [96].

Additional files

Additional file 1. Tables S1-S8; Table S13-S19; Figures S1-S16.

Additional file 2. Tables S9-S12

Additional file 3. Mass-spectrometry proteome data of catkins and pollen tissues. Three samples of each tissue type (immature catkin, mature catkin, senescent catkin, and pure pollen) were analyzed using v1.0 and v2.0 reference walnut genome assemblies. Total intensity of matching peptides, number of spectra and percentage of protein covered by the identified peptides are reported.

Competing interests

The authors declare no conflict of interest.

Funding:

This project has been funded by the Californian Walnut Board.

Author's contribution

DBN and AM conceived and coordinated the research. REW and WT performed the HMW DNA extraction and Nanopore sequencing. AVZ, DP and SLS assembled the hybrid Illumina-ONT assembly. LB, MT, DP and SLS validated and anchored the HiRise assembly to the genetic maps. AM and BJA collected and extracted all RNA samples. MB analyzed the PacBio IsoSeq results and performed the repeat and gene annotation. AD conceived the design of the proteomic analyses; PAZ and SC generated and analyzed the proteomic data. LB called the SNPs in Chandler and the 23 UCD WIP founders, while AM carried out the analyses on walnut genomic diversity. EAD, LB and MT built and analyzed the SNP haplotypes. CAL provided all the plant material. AM wrote the manuscript, which has been revised by all coauthors.

Acknowledgments

We are grateful to Sriema Walawage for assistance with RNA extraction, and Brett Phinney for preparing the raw proteome data.

References

1. Martínez ML, Labuckas DO, Lamarque AL, Maestri DM. Walnut (*Juglans regia* L.): Genetic resources, chemistry, by-products. *J Sci Food Agric*. 2010;90:1959–67.
2. McGranahan G, Leslie C. Walnut. In: Badenes ML, Byrne DH, editors. *Fruit Breed*. Springer Science+Business Media, LLC; 2012. p. 827–46.
3. Pollegioni P, Woeste K, Chiocchini F, Del Lungo S, Ciolfi M, Olimpieri I, et al. Rethinking

727 the history of common walnut (*Juglans regia* L.) in Europe: Its origins and human interactions.
 728 PLoS One. 2017;12:1–24.

729 4. Zhang B, Xu L, Li N, Yan P, Jiang X, Woeste KE, et al. Phylogenomics Reveals an Ancient
 730 Hybrid Origin of the Persian Walnut. *Mol Biol Evol.* 2019;1–11.

731 5. Zeven A, Zhukovskii PM. Dictionary of cultivated plants and their centres of diversity,
 732 excluding ornamentals, forest trees, and lower plants [Internet]. Cent. Agric. Publ. Doc.
 733 Wageningen. 1975. Available from: <https://core.ac.uk/download/pdf/29387092.pdf>

734 6. Ebrahimi A, Zarei A, Lawson S, Woeste KE, Smulders MJM. Genetic diversity and genetic
 735 structure of Persian walnut (*Juglans regia*) accessions from 14 European, African, and Asian
 736 countries using SSR markers. *Tree Genet Genomes* [Internet]. *Tree Genetics & Genomes*;
 737 2016;12:114. Available from: <http://link.springer.com/10.1007/s11295-016-1075-y>

738 7. Marrano A, Martínez-García PJ, Bianco L, Sideli GM, Di Pierro EA, Leslie CA, et al. A new
 739 genomic tool for walnut (*Juglans regia* L.): development and validation of the high-density
 740 Axiom™ *J. regia* 700K SNP genotyping array. *Plant Biotechnol J* [Internet]. 2018;1–10.
 741 Available from: <http://doi.wiley.com/10.1111/pbi.13034>

742 8. Bernard A, Barreneche T, Lheureux F, Dirlewanger E. Analysis of genetic diversity and
 743 structure in a worldwide walnut (*Juglans regia* L.) germplasm using SSR markers. *PLoS One.*
 744 2018;13:1–19.

745 9. Martínez-García PJ, Crepeau MW, Puiu D, Gonzalez-Ibeas D, Whalen J, Stevens KA, et al.
 746 The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the
 747 biosynthesis of non-structural polyphenols. *Plant J.* 2016;87:507–32.

- 748 10. Stevens KA, Woeste K, Chakraborty S, Crepeau MW, Leslie CA, Martínez-García PJ, et al.
749 Genomic Variation Among and Within Six *Juglans* Species. *G3 Genes|Genomes|Genetics*
750 [Internet]. 2018;8:1–37. Available from:
751 <http://www.ncbi.nlm.nih.gov/pubmed/29792315>
752 118.200030
- 753 11. Kefayati S, Ikhsan AS, Sutyemez M, Paizila A, Topcu H, Bukubu SB, et al. First simple
754 sequence repeat-based genetic linkage map reveals a major QTL for leafing time in walnut
755 (*Juglans regia* L.). *Tree Genet Genomes*. 2018;15:13.
- 756 12. Arab MM, Marrano A, Abdollahi-Arpanahi R, Leslie CA, Askari H, Neale DB, et al.
757 Genome-wide patterns of population structure and association mapping of nut-related traits in
758 Persian walnut populations from Iran using the Axiom *J. regia* 700K SNP array. *Sci Rep*
759 [Internet]. Springer US; 2019;9:6376. Available from: [http://www.nature.com/articles/s41598-](http://www.nature.com/articles/s41598-019-42940-1)
760 019-42940-1
- 761 13. Famula RA, Richards JH, Famula TR, Neale DB. Association Genetics of Carbon Isotope
762 Discrimination in the Founding Individuals of a Breeding Population of *Juglans regia* L. *Tree*
763 *Genet Genomes* [Internet]. *Tree Genetics & Genomes*; 2019;15:6. Available from:
764 <https://doi.org/10.1007/s11295-018-1307-4>
- 765 14. Marrano A, Sideli GM, Leslie CA, Cheng H, Neale DB. Deciphering of the genetic control
766 of phenology, yield and pellicle color in Persian walnut (*Juglans regia* L.). *Front Plant Sci*.
767 2019;10:1–14.
- 768 15. Bernard A, Marrano A, Donkpegan A, Brown PJ, Leslie CA, Neale DB, et al. Association
769 and linkage mapping to unravel genetic architecture of phenological traits and lateral bearing in

770 Persian walnut (*Juglans regia* L .). BMC Genomics. BMC Genomics; 2020;21:1–25.

771 16. Sánchez-Pérez R, Pavan S, Mazzeo R, Moldovan C, Aiese Cigliano R, Del Cueto J, et al.
772 Mutation of a bHLH transcription factor allowed almond domestication. Science (80-)
773 [Internet]. 2019;364:1095–8. Available from:
774 <http://www.sciencemag.org/lookup/doi/10.1126/science.aav8197>

775 17. Tang W, Sun X, Yue J, Tang X, Jiao C, Yang Y, et al. Chromosome-scale genome assembly
776 of kiwifruit *Actinidia eriantha* with single-molecule sequencing and chromatin interaction
777 mapping. Gigascience. Oxford University Press; 2019;8:1–10.

778 18. Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, et al. The genome of
779 *Chenopodium quinoa*. Nature. 2017;542:307–12.

780 19. Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, et al. Durum
781 wheat genome highlights past domestication signatures and future improvement targets. Nat
782 Genet. 2019;51:885–95.

783 20. Raymond O, Gouzy J, Just J, Badouin H, Verdenaud M, Lemainque A, et al. The Rosa
784 genome provides new insights into the domestication of modern roses. Nat Genet. 2018;50:772–
785 7.

786 21. Daccord N, Celton JM, Linsmith G, Becker C, Choisne N, Schijlen E, et al. High-quality de
787 novo assembly of the apple genome and methylome dynamics of early fruit development. Nat
788 Genet [Internet]. Nature Publishing Group; 2017;49:1099–106. Available from:
789 <http://dx.doi.org/10.1038/ng.3886>

790 22. Zhu T, Wang L, You FM, Rodriguez JC, Deal KR, Chen L, et al. Sequencing a *Juglans regia*

791 × J . microcarpa hybrid yields high-quality genome assemblies of parental species. Hortic Res
 792 [Internet]. Springer US; 2019;1–16. Available from: [http://dx.doi.org/10.1038/s41438-019-0139-](http://dx.doi.org/10.1038/s41438-019-0139-1)
 793 1

794 23. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly.
 795 Genomics, Proteomics Bioinforma [Internet]. Beijing Institute of Genomics, Chinese Academy
 796 of Sciences and Genetics Society of China; 2016;14:265–79. Available from:
 797 <http://dx.doi.org/10.1016/j.gpb.2016.05.004>

798 24. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: A comprehensive
 799 technique to capture the conformation of genomes. Methods [Internet]. Elsevier Inc.;
 800 2012;58:268–76. Available from: <http://dx.doi.org/10.1016/j.ymeth.2012.05.001>

801 25. Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. J Ex.
 802 2017;68:5419–29.

803 26. Schmidt MH-W, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, et al. De Novo
 804 Assembly of a New *Solanum pennellii* Accession Using Nanopore Sequencing . Plant Cell.
 805 2017;29:2336–48.

806 27. Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, et al. Chromosome-scale
 807 assemblies of plant genomes using nanopore long reads and optical maps. Nat Plants [Internet].
 808 Springer US; 2018;4:879–87. Available from: <http://dx.doi.org/10.1038/s41477-018-0289-4>

809 28. Yasodha R, Vasudeva R, Balakrishnan S, Sakthi AR, Abel N, Binai N, et al. Draft genome of
 810 a high value tropical timber tree, Teak (*Tectona grandis* L. f): insights into SSR diversity,
 811 phylogeny and conservation. DNA Res. 2018;25:409–19.

- 812 29. Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, et al. A chromosome-scale
813 assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat*
814 *Commun.* 2018;9:4844.
- 815 30. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: Computational
816 approaches for improving nanopore sequencing read accuracy. *Genome Biol. Genome Biology*;
817 2018;19:1–11.
- 818 31. Zimin A V., Luo M, Marçais G, Salzberg SL, Yorke JA, Puiu D, et al. Hybrid assembly of
819 the large and highly repetitive genome of *Aegilops tauschii* , a progenitor of bread wheat, with
820 the MaSuRCA mega-reads algorithm. *Genome Res* [Internet]. 2017;27:787–92. Available from:
821 <http://www.ncbi.nlm.nih.gov/pubmed/28130360>
822 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5411773>
- 823 32. Huang Y, Xiao L, Zhang Z, Zhang R, Wang Z, Huang C, et al. The genomes of pecan and
824 Chinese hickory provide insights into *Carya* evolution and nut nutrition. *Gigascience.* 2019;8:1–
825 17.
- 826 33. Xing Y, Liu Y, Zhang Q, Nie X, Sun Y, Zhang Z, et al. Hybrid de novo genome assembly of
827 Chinese chestnut (*Castanea mollissima*). *Gigascience.* 2019;8:1–7.
- 828 34. Plomion C, Aury JM, Amselem J, Leroy T, Murat F, Duplessis S, et al. Oak genome reveals
829 facets of long lifespan. *Nat Plants.* Springer US; 2018;4:440–52.
- 830 35. Luo M-C, You FM, Li P, Wang J-R, Zhu T, Dandekar AM, et al. Synteny analysis in Rosids
831 with a walnut physical map reveals slow genome evolution in long-lived woody perennials.
832 *BMC Genomics* [Internet]. *BMC Genomics*; 2015;16:1–17. Available from:
833 <http://www.biomedcentral.com/1471-2164/16/707>

- 834 36. Springer NM, Ying K, Fu Y, Ji T, Yeh C, Jia Y, et al. Maize Inbreds Exhibit High Levels of
835 Copy Number Variation (CNV) and Presence / Absence Variation (PAV) in Genome Content.
836 PLoS Genet. 2009;5.
- 837 37. Marroni F, Pinosio S, Morgante M. Structural variation and genome complexity : is
838 dispensable really dispensable ? Curr Opin Plant Biol [Internet]. Elsevier Ltd; 2014;18:31–6.
839 Available from: <http://dx.doi.org/10.1016/j.pbi.2014.01.003>
- 840 38. Mishra B, Gupta DK, Pfenninger M, Hickler T, Langer E, Nam B, et al. A reference genome
841 of the European beech (*Fagus sylvatica* L.). Gigascience. 2018;7.
- 842 39. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics, Proteomics
843 Bioinforma [Internet]. Beijing Institute of Genomics, Chinese Academy of Sciences and
844 Genetics Society of China; 2015;13:278–89. Available from:
845 <http://dx.doi.org/10.1016/j.gpb.2015.08.002>
- 846 40. Linsmith G, Rombauts S, Montanari S, Deng CH, Celton JM, Guérif P, et al. Pseudo-
847 chromosome-length genome assembly of a double haploid “bartlett” pear (*Pyrus communis* L.).
848 Gigascience. 2019;8:1–17.
- 849 41. Vaneechoutte D, Estrada AR, Lin YC, Loraine AE, Vandepoele K. Genome-wide
850 characterization of differential transcript usage in *Arabidopsis thaliana*. Plant J. 2017;92:1218–
851 31.
- 852 42. Clark S, Yu F, Gu L, Min XJ. Expanding alternative splicing identification by integrating
853 multiple sources of transcription data in tomato. Front Plant Sci. 2019;10:1–12.
- 854 43. Hart AJ, Ginzburg S, Xu M (Sam), Fisher CR, Rahmatpour N, Mitton JB, et al. EnTAP:

855 bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes.
856 bioRxiv [Internet]. 2018;307868. Available from:
857 <https://www.biorxiv.org/content/biorxiv/early/2018/04/28/307868.full.pdf%0Ahttps://www.biorxiv.org/content/early/2018/04/24/307868%0Ahttps://www.biorxiv.org/content/early/2018/04/24/307868>
859 307868

860 44. Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I. A chromosome-scale genome
861 assembly of European Hazel (*Corylus avellana* L .) reveals targets for crop improvement.
862 bioRxiv. 2019;44.

863 45. Sork VL, Fitz-Gibbon ST, Puiu D, Crepeau M, Gugger PF, Sherman R, et al. First draft
864 assembly and annotation of the genome of a California endemic oak *Quercus lobata* Née
865 (Fagaceae). *G3 Genes, Genomes, Genet.* 2016;6:3485–95.

866 46. Jamet E, Santoni V. Editorial for Special Issue: 2017 Plant Proteomics. *proteomes.*
867 2018;6:28.

868 47. Costa J, Carrapatoso I, Oliveira MBPP, Mafra I. Walnut allergens: Molecular
869 characterization, detection and clinical relevance. *Clin Exp Allergy.* 2014;44:319–41.

870 48. Aradhya M, Woeste K, Velasco D. Genetic diversity, structure and differentiation in
871 cultivated walnut (*Juglans regia* L.). *Acta Hortic.* 2010;861:127–32.

872 49. Ruiz-Garcia L, Lopez-Ortega G, Fuentes Denia a., Frutos Tomas D. Identification of a
873 walnut (*Juglans regia* L.) germplasm collection and evaluation of their genetic variability by
874 microsatellite markers. *Spanish J Agric Res.* 2011;9:179–92.

875 50. Dangl GS, Woeste K, Aradhya MK, Koehmstedt A, Simon C, Potter D, et al.

876 Characterization of 14 Microsatellite Markers for Genetic Analysis and Cultivar Identification of
877 Walnut. *J Am Soc Hortic Sci* [Internet]. 2005;130:348–54. Available from:
878 <http://journal.ashspublications.org/content/130/3/348%5Cnhttp://journal.ashspublications.org/co>
879 [tent/130/3/348.full.pdf](http://journal.ashspublications.org/content/130/3/348.full.pdf)

880 51. McGranahan GH, Leslie CA. Walnuts. In: Moore JN, Ballington JRJ, editors. *Genet Resour*
881 *Temp Fruit Nut Crop*. International Society for Horticultural Science; 1991. p. 907–18.

882 52. Bernard A, Lheureux F, Dirlewanger E. Walnut: past and future of genetic improvement.
883 *Tree Genet Genomes*. *Tree Genetics & Genomes*; 2018;14:1–28.

884 53. Gauthier MM, Jacobs DF. Walnut (*Juglans* spp.) ecophysiology in response to environmental
885 stresses and potential acclimation to climate change. *Ann For Sci*. 2011;68:1277–90.

886 54. Workman R, Fedak R, Kilburn D, Hao S, Liu K, Timp W. High Molecular Weight DNA
887 Extraction from Recalcitrant Plant Species for Third Generation Sequencing. *Protoc Exch*.
888 2018;1–12.

889 55. Zhang M, Zhang Y, Scheuring CF, Wu CC, Dong JJ, Zhang H Bin. Preparation of megabase-
890 sized DNA from a variety of organisms using the nuclei method for advanced genomics
891 research. *Nat Protoc* [Internet]. Nature Publishing Group; 2012;7:467–78. Available from:
892 <http://dx.doi.org/10.1038/nprot.2011.455>

893 56. Mayjonade B, Gouzy J, Donnadiou C, Pouilly N, Marande W, Callot C, et al. Extraction of
894 high-molecular-weight genomic DNA for long-read sequencing of single molecules.
895 *Biotechniques*. 2017;62:xv.

896 57. Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome

897 assembler. *Bioinformatics*. 2013;29:2669–77.

898 58. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, et al. Aggressive
899 assembly of pyrosequencing reads with mates. *Bioinformatics*. 2008;

900 59. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al.
901 Comprehensive mapping of long-range interactions reveals folding principles of the human
902 genome. *Science* (80-). 2009;

903 60. Putnam NH, O’Connell, Brendan L. Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, et
904 al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage.
905 *Genome Res*. 2016;26:342–50.

906 61. Benson G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids*
907 *Res*. 1999;27:573–80.

908 62. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast
909 and versatile genome alignment system. *PLoS Comput Biol*. 2018;14:1–14.

910 63. Rezvoy C, Charif D, Guéguen L, Marais GAB. MareyMap: An R-based tool with graphical
911 interface for estimating recombination rates. *Bioinformatics*. 2007;23:2188–9.

912 64. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST : quality assessment tool for genome
913 assemblies. *Bioinformatics*. 2013;29:1072–5.

914 65. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of
915 occurrences of k-mers. *Bioinformatics*. 2011;

916 66. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al.
917 *GenomeScope* : fast reference-free genome profiling from short reads. *Bioinformatics*.

918 2017;33:2202–4.

919 67. Li H. Minimap2 : pairwise alignment for nucleotide sequences. *Bioinformatics*.
920 2018;34:3094–100.

921 68. Smit A, Hubley R. RepeatModeler Open-1.0. [Internet]. 2008. Available from:
922 <http://www.repeatmasker.org>

923 69. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. [Internet]. 2013. Available from:
924 <http://www.repeatmasker.org>

925 70. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, et al. Improving
926 the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic
927 Acids Res*. 2003;31:5654–66.

928 71. Kent WJ. BLAT — The BLAST -Like Alignment Tool. *Genome Res*. 2002;12:656–64.

929 72. Wu TD, Watanabe CK. GMAP : a genomic mapping and alignment program for mRNA and
930 EST sequences. *Bioinformatics*. 2005;21:1859–75.

931 73. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo
932 transcript sequence reconstruction from RNA-Seq: reference generation and analysis with
933 Trinity. *Nat Protoc*. 2013;8:1–43.

934 74. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq -
935 Versatile and accurate annotation of organelle genomes. *Nucleic Acids Res*. 2017;

936 75. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length
937 transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;

938 76. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan:

939 Protein domains identifier. *Nucleic Acids Res.* 2005;

940 77. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: Genome-
941 scale protein function classification. *Bioinformatics.* 2014;

942 78. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO:
943 Assessing genome assembly and annotation completeness with single-copy orthologs.
944 *Bioinformatics.* 2015;31:3210–2.

945 79. Veeckman E, Ruttink T, Vandepoele K. Are We There Yet? Reliably Estimating the
946 Completeness of Plant Genome Sequences. *Plant Cell.* 2016;28:1759–68.

947 80. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory
948 requirements. *Nat Methods.* 2015;

949 81. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of
950 RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc.* 2016;

951 82. Valerie M, Catherine D, Michel Z, Hervé T, Faurobert M, Pelpoir E, et al. Phenol Extraction
952 of Proteins for Proteomic Studies of Recalcitrant Plant Tissues. *Plant Proteomics.* 2006.

953 83. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: A web-based
954 environment for protein structure homology modelling. *Bioinformatics.* 2006;

955 84. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. MUSTANG: A multiple structural
956 alignment algorithm. *Proteins Struct Funct Genet.* 2006;

957 85. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
958 *Bioinformatics.* 2009;25:1754–60.

959 86. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence

960 Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.

961 87. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-smith C, Durbin R. BCFtools/RoH : a
962 hidden Markov model approach for detecting autozygosity from next-generation sequencing
963 data. *Bioinformatics*. 2016;32:1749–51.

964 88. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call
965 format and VCFtools. *Bioinformatics*. 2011;27:2156–8.

966 89. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: A toolkit for detection
967 and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 2012;40:1–14.

968 90. Bink MCAM, Jansen J, Madduri M, Voorrips RE, Durel CE, Kouassi AB, et al. Bayesian
969 QTL analyses using pedigreed families of an outcrossing species, with application to fruit
970 firmness in apple. *Theor Appl Genet*. 2014;127:1073–90.

971 91. Voorrips RE, Bink MCAM, Kruisselbrink JW, Koehorst-van Putten HJJ, van de Weg WE.
972 PediHaplotyper: software for consistent assignment of marker haplotypes in pedigrees. *Mol*
973 *Breed*. Springer Netherlands; 2016;36.

974 92. Vanderzande S, Howard NP, Cai L, Da Silva Linge C, Antanaviciute L, Bink MCAM, et al.
975 High-quality, genome-wide SNP genotypic data for pedigreed germplasm of the diploid
976 outbreeding species apple, peach, and sweet cherry through a common workflow. *PLoS One*.
977 2019;

978 93. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance
979 computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*.
980 2012;28:3326–8.

981 94. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA
 982 polymorphism. Genetics. 1989;123:585–95.

983 95. Alexa A. Gene set enrichment analysis with topGO. 2015;47–53.

984 96. Marrano A; Britton M; Zaini PA; Zimin AV; Workman RE; Puiu D; Bianco L; Di Pierro EA;
 985 Allen BJ; Chakraborty S; Troggio M; Leslie CA; Timp W; Dandekar A; Salzberg SL; Neale DB:
 986 Supporting data for "High-quality chromosome-scale assembly of the walnut (*Juglans regia*
 987 L) reference genome" GigaScience Database.2020. <http://dx.doi.org/10.5524/100735>

988 **Tables**

989 **Table 1. Comparison among the four assemblies of Chandler.** Scaffolds shorter than 1,000 bp are not
 990 included in these totals.

Statistics	Chandler v1.0	Chandler v1.5	Chandler hybrid	Chandler HiRise	Chandler v2.0	JrSerr_v1.0
<i>Number of scaffolds</i>	27,032	4,401	3,497	2,656	2,643	73
<i>N50 length (scaffolds) (bp)</i>	304,423	637,984	1,640,935	32,655,472	37,114,715	35,197,335
<i>L50 (scaffolds)</i>	344	272	89	8	7	7
<i>Total length of assembled scaffolds (bp)</i>	667,299,356	650,478,320	567,378,842	567,480,142	567,796,851	534,671,929
<i>Number of contigs</i>	53,156	7,411	3,592	3,700	3,684	127
<i>N50 length (contigs) (bp)</i>	42,417	317,751	1,512,354	1,083,883	1,083,883	15,066,219
<i>L50 (contigs)</i>	3,630	482	97	144	144	13
<i>Total size of assembled contigs (bp)</i>	641,521,787	617,088,256	567,276,004	567,276,244	567,192,099	530,618,363

Table 2. Statistics on the gene annotation of Chandler v2.0 compared to the previous gene annotations of the Chandler genome.

Statistics	Chandler v2.0	Chandler v1.0	Chandler RefSeq v1.0
<i>Number of genes</i>	37,554	32,496	41,188
<i>Average gene length (bp)</i>	5,319	4,358	4,641
<i>Single-exon transcripts</i>	6,613	6,247	6,749
<i>Average CDS length (bp)</i>	1,335	1,222	1,336
<i>Number of exons</i>	242,208	172,273	230,261
<i>Average exon length (bp)</i>	257.8	229.5	314
<i>Number of Introns</i>	201,290	139,775	181,419
<i>Average intron length</i>	853.9	730	835
<i>Introns per gene</i>	5.9	5.3	4.4

Table 3. Statistics of the completeness of Chandler v2.0 assessed with BUSCO and compared to other Fagales genomes.

Genome	BUSCO complete (%)	BUSCO duplicated (%)	BUSCO fragmented (%)	BUSCO missing (%)	Reference
<i>Juglans regia</i> cv. 'Chandler' v2.0	95.1	12.6	1.3	3.6	This genome
<i>Juglans regia</i> cv. 'Chandler' v1.0	94.8	13.8	1.2	4.0	[9]
<i>Juglans regia</i> cv. 'Serr' v1.0	94.5	11.1	1.5	4.0	[22]

<i>Fagus sylvatica</i> v1.2	94	19	1.7	3.6	[34]
<i>Castanea mollissima</i>	96.7	7.7	1.4	1.9	[33]
<i>Carya illinoensis</i> v1	94	23	1.4	3.6	[32]
<i>Corylus avellane</i> cv. 'Tombul'	96	6	1	3	[44]
<i>Quercus lobata</i> v1.0	90	52	4	6	[45]
<i>Quercus robur</i>	93	49	3	4	[34]

Figure legends

Figure 1. Collinearity between the high-density 'Chandler' genetic map of [14] and the 16 chromosomal pseudomolecules of Chandler v2.0.

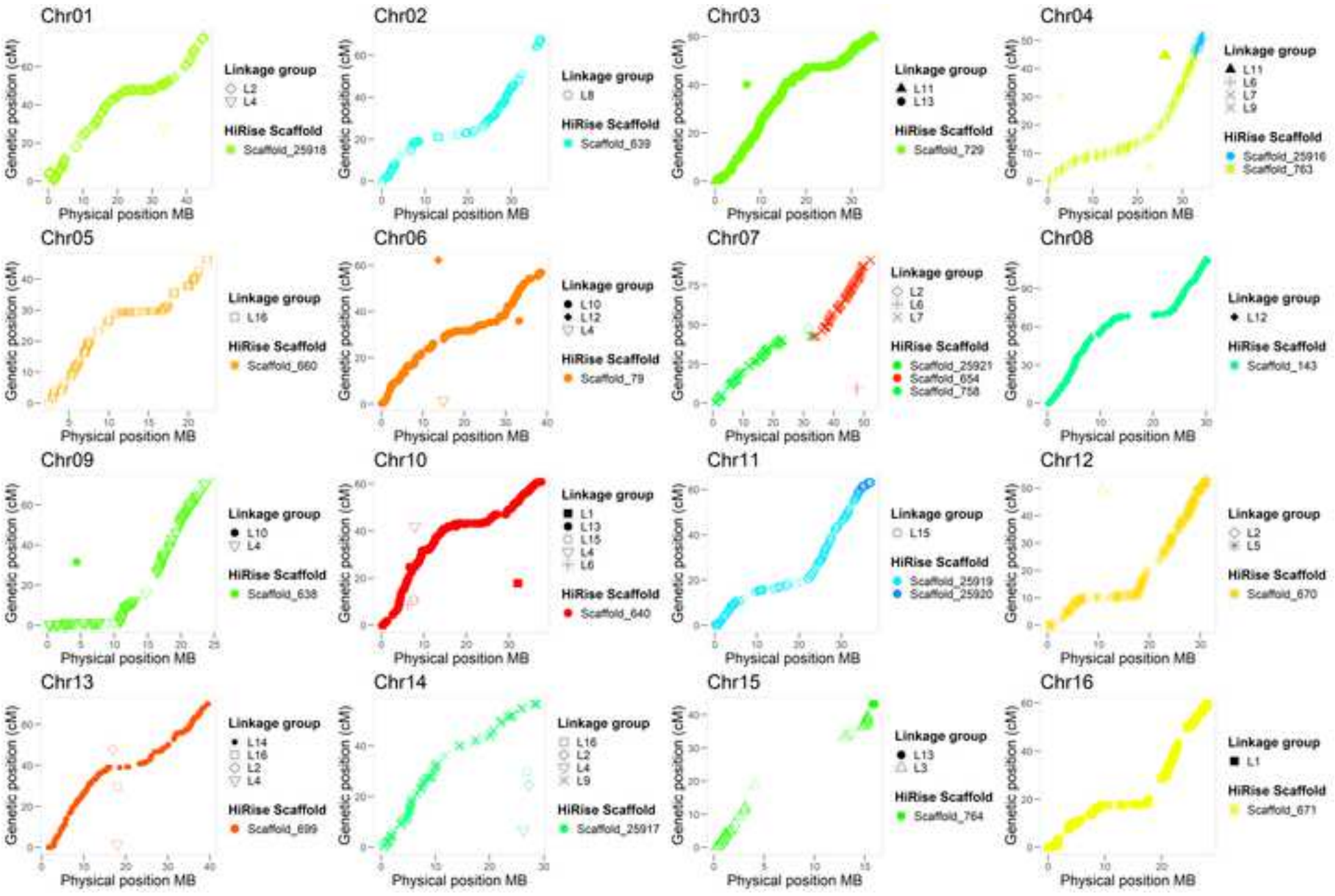
Figure 2. Summary of gene distribution and genetic diversity across the 16 chromosomes of Chandler v2.0. Tracks from outside to inside: **(i)** gene density of Chandler v2.0 in 1-Mb windows; **(ii)** Chandler heterozygosity in 1-Mb windows (white = low heterozygosity; blue = high heterozygosity); **(iii)** Recombination rate for sliding windows of 10 Mb (average = 2.63 cM/Mb); **(iv)** F_{ST} in 500-kb windows. Windows in the 95 percentiles of the F_{ST} distribution are highlighted in red; **(v)** ROD values for 500-kb windows.

Figure 3. Clustering of the samples used in the proteomic analysis. **(A)** Hierarchical clustering based on Euclidian distances of normalized abundances of detected proteins. Samples are represented in columns and proteins in rows. **(B)** Principal component analysis of the 12 samples analyzed, clustering according to tissue type.

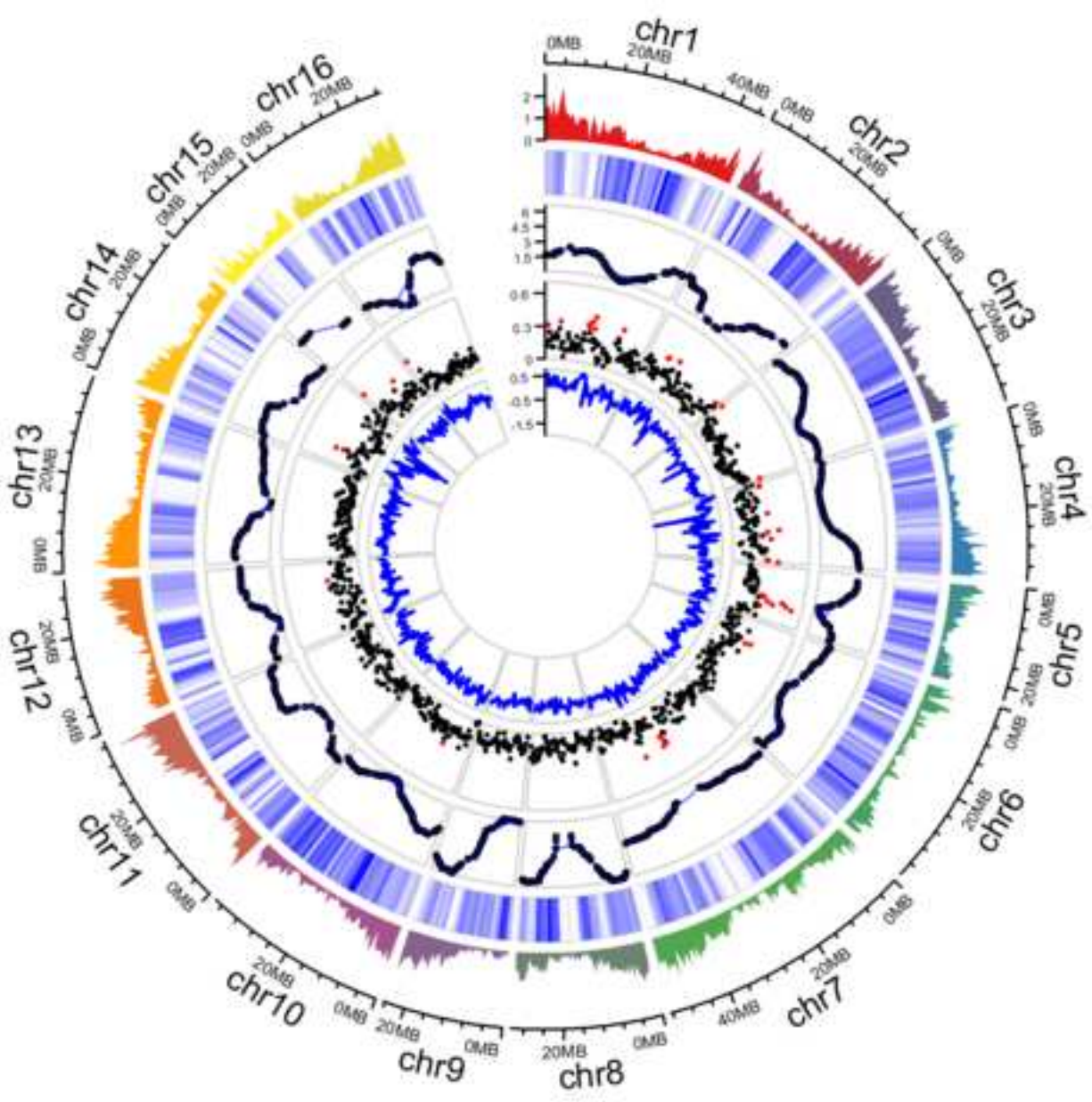
Figure 4. Graphical visualization of haplotype-blocks (HB) inheritance on Chr15 along with the Chandler pedigree. **(A)** The inner-circle highlights in grey two regions of heterozygosity (5 HB the first and 7 HB the second), and in light green two regions of homozygosity (3 HB the first and 4 HB the second). The circle in the middle shows maternally inherited HBs, while the HBs inherited through the paternal line are visualized in the outer circle. Payne's haplotypes are clearly present in both parental lines. White spaces represent segments of missing haplotype information. **(B)** Chandler pedigree, where Pedro is the maternal line and 56-224, the paternal line.

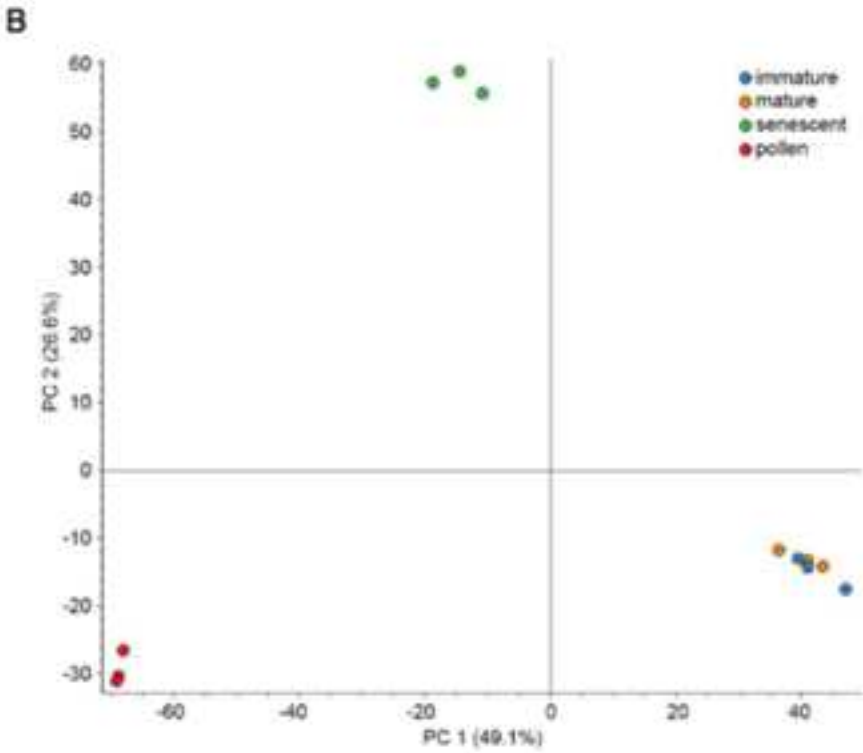
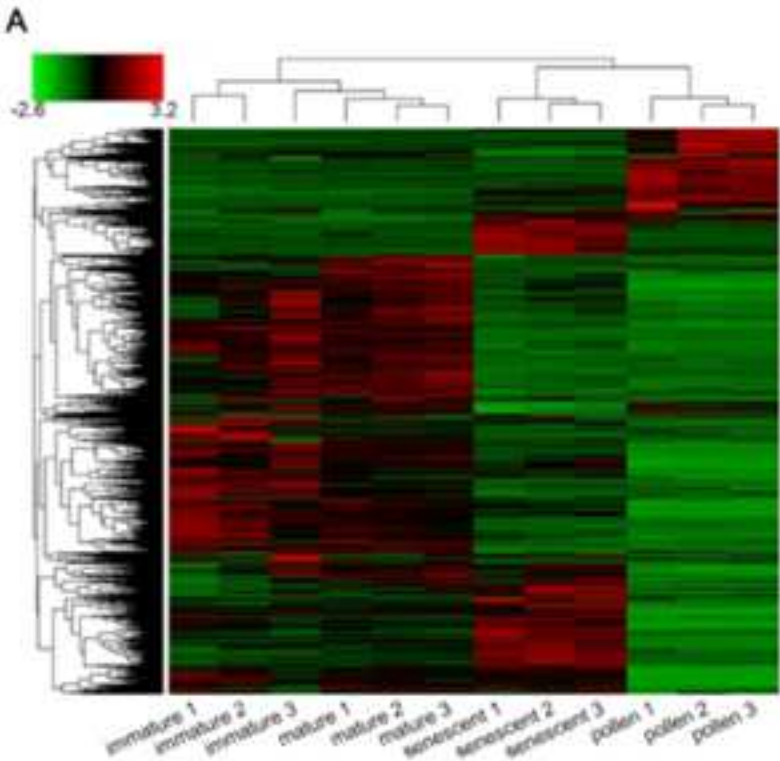
Figure 5. Graphical visualization of the haplotype-blocks (HB) inheritance across Chandler pedigree in the 16 chromosomes. The inner circle highlights in grey the regions of heterozygosity and in light green the regions of homozygosity for each chromosome. The circle in the middle shows the maternally inherited HBs, while the HBs inherited from the paternal line are visualized in outer circle. In both parental line circles, missing data are highlighted in grey. Payne haplotypes are inherited along both parental lines in all chromosomes, but Chr5, Chr9, Chr10, Chr14 and Chr16. Chandler pedigree is represented on the side, where Pedro is the maternal line and 56-224 the paternal line.

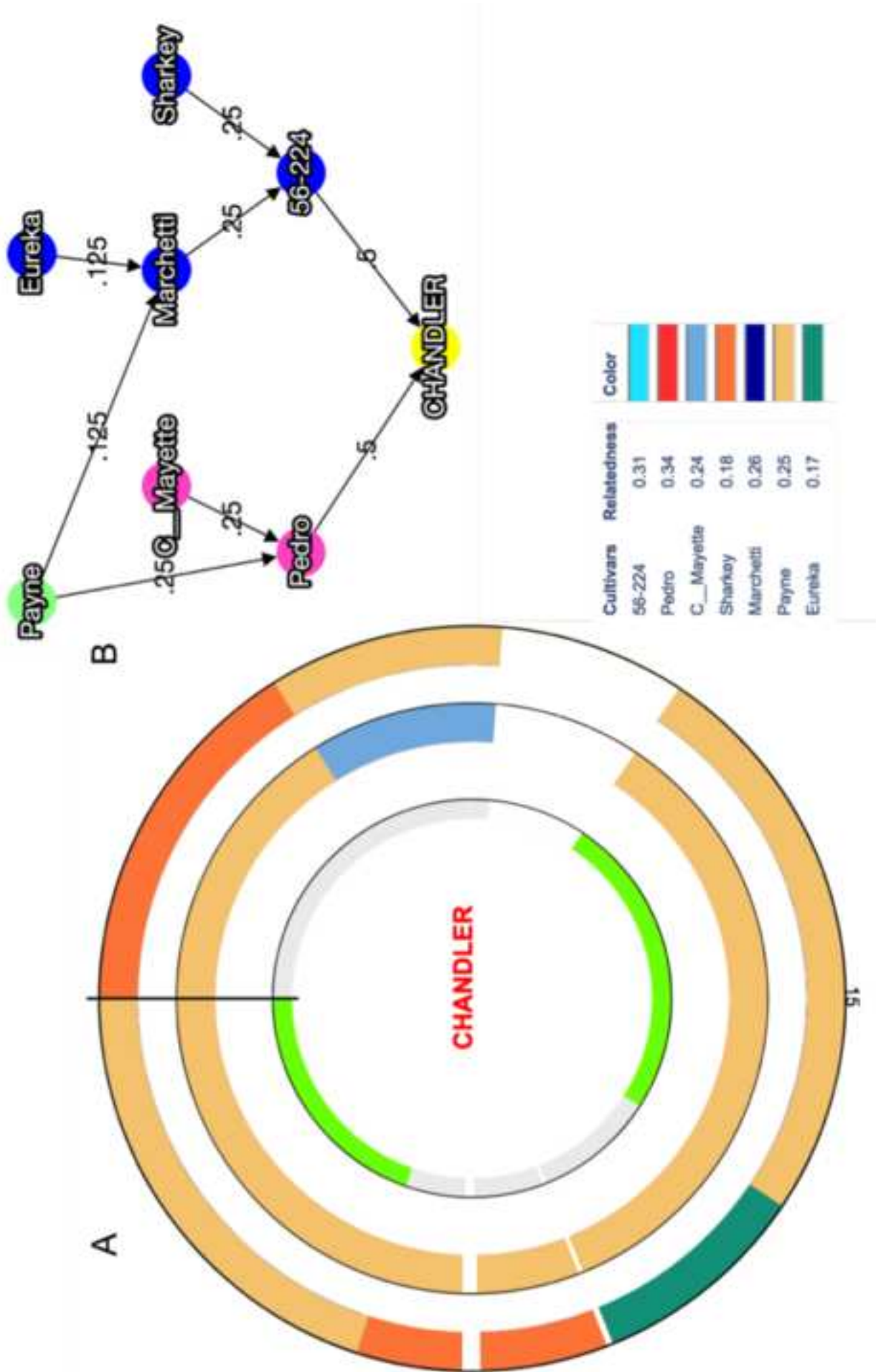
Figure

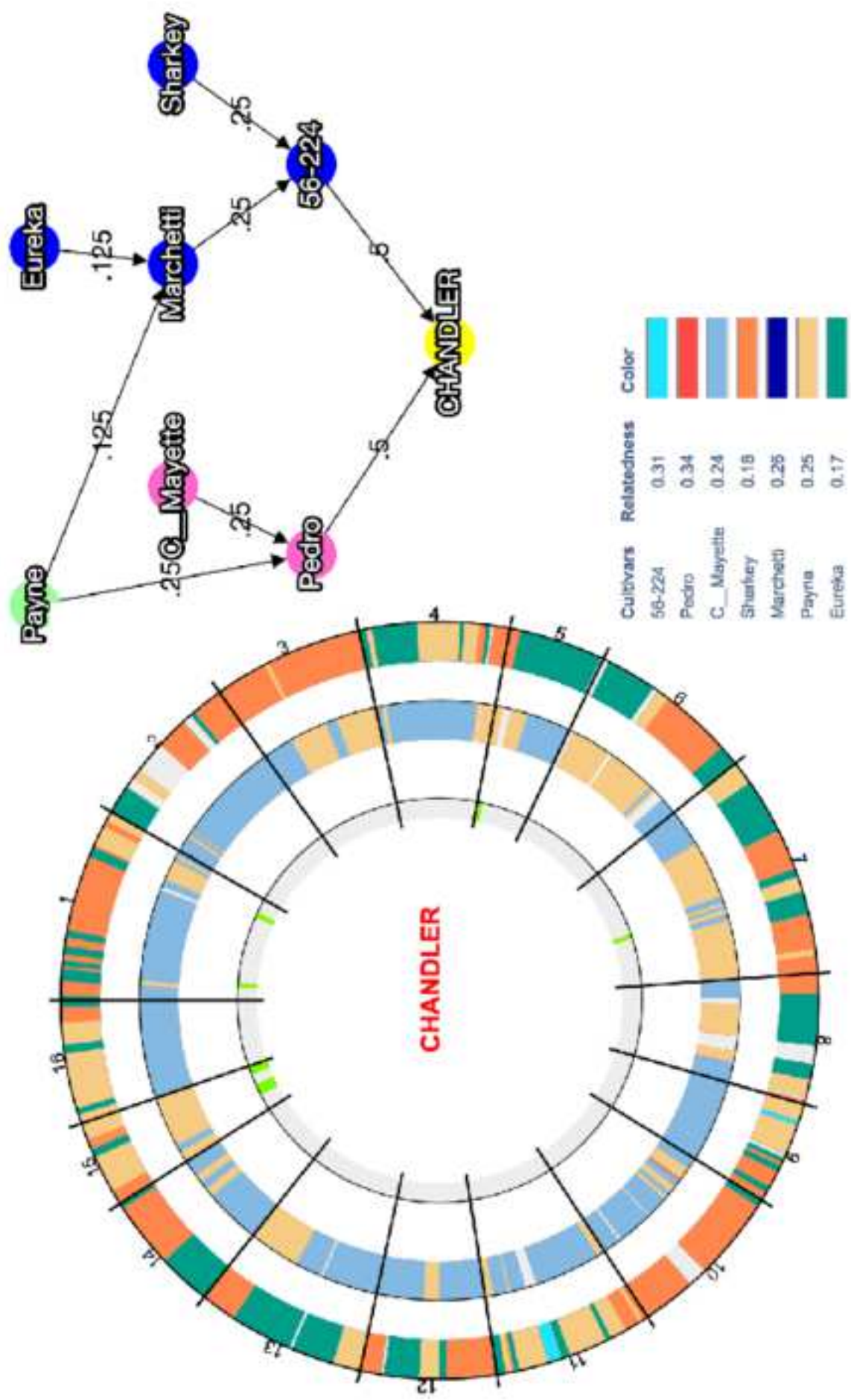


Figure



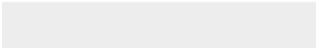
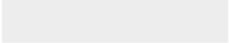


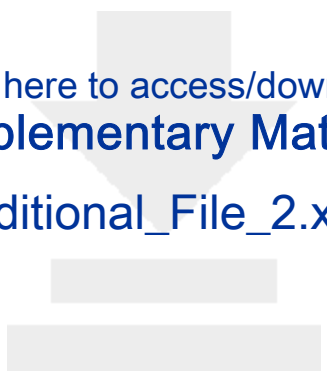






Click here to access/download
Supplementary Material
Additional_File_1_Revised.docx





Click here to access/download
Supplementary Material
Additional_File_2.xlsx



Click here to access/download
Supplementary Material
Additional_File_3.xlsx

UNIVERSITY OF CALIFORNIA, DAVIS

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO



SANTA BARBARA • SANTA CRUZ

DEPARTMENT OF PLANT SCIENCES
MAIL STOP 6
UNIVERSITY OF CALIFORNIA
ONE SHIELDS AVE
DAVIS, CALIFORNIA 95616-8780
TELEPHONE: 530-752-1703
FAX: 530-752-1819

COLLEGE OF AGRICULTURAL AND
ENVIRONMENTAL SCIENCES
AGRICULTURAL EXPERIMENT STATION
COOPERATIVE EXTENSION

March 13th, 2020

Dear Editor,

With this letter, we are enclosing a revised version of the manuscript entitled “High-quality chromosome-scale assembly of the walnut (*Juglans regia* L.) reference genome” (submission id GIGA-D-19-00363R1) for publication as a Data Note in *GigaScience*. We are also attaching a response to the referees and a revised version of the Additional files 1 and 2. All authors have approved the revised manuscript for submission. We declare no conflict of interest, and that all previously published work has been thoroughly acknowledged in our manuscript.

We have implemented the latest minor revisions suggested by the referees in our manuscripts. Among these changes, we made Table S9 a full table and Figure S12 a full figure, while we moved Figure 4 and Figure 6 to the Additional file 1. We changed the table and figure numbering accordingly.

We believe that our manuscript has significantly improved, and it will be of interest to the audience of *Gigascience*.

Sincerely,

Annarita Marrano

A handwritten signature in blue ink, appearing to read 'Annarita Marrano'.

Ph.D.

Dept. Plant Sciences, University of California, Davis

Author's Response To Reviewer Comments

Close

Reviewer #1: I read the manuscript by Marrano et al. entitled "High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome" with interest. The authors describe how they generated a chromosome-scale assembly of *J. regia* based on ONT, Illumina and Hi-C data. In addition, genetic maps were used to validate and anchor the scaffolds to *J. regia* linkage groups. The authors performed a wide range of analysis, including gene families of interest and genomic diversity. The article is well written and the methods are sufficiently described.

My main concern is the quality of the assembly, although at the chromosome level, the contiguity at the contig level is ten times lower (nearly 1.1Mb) compared to the previously published *Juglans* genome assemblies (including *J. regia*).

The assembly presented in our manuscript is very high quality, with chromosome-sized scaffolds validated by genetic maps and an N50 contig size of over 1Mb.

The assembly was produced on a fairly low budget with older generation Oxford Nanopore reads from the MinION device, which yielded 21.9 Gbp with an average read size of 3.1 kb. (Today we would get reads of 10-20 Kb.) This sequencing data is older and much lower cost than those used by Zhu et al. (2019), who had 57.2 Gbp of PacBio data with an average read size of over 12kb. (see lines 136-140). Even so, a contig N50 size of 1.1 Mb is dramatically larger than our own originally published assembly contigs.

I understand that authors choose to use the methods they have developed, however, long-reads assemblies are usually made with dedicated assemblers, which can result in higher assembly quality.

Long-read-only data sets are indeed assembled with dedicated methods such as Canu, but when we have both long and short-read data sets, MaSurCA performs better. We did try using the Flye assembler, which is currently one of the best, fastest, and most accurate assemblers for long-read data, and it yielded an assembly with only 530kb N50 contig size, about half of our result.

In particular, I was a little surprised by the fact that the v2 assembly contains fewer repetitive elements than the first version of the assembly (L175-176). Generally long-reads assemblies improved the repetitive content of genome assemblies.

We re-estimated the repeat content of Chandler v2.0 by running RepeatMasker with the repeat library generated for v1. We found that 58.4% of Chandler v2 is repetitive (we have corrected the main text accordingly), which is higher than what found with the first assembly of the reference genome. Also, Chandler v2.0 is de-duplicated, meaning that almost all of the duplicated regions due to heterozygosity have been removed. These duplicates, which are variants of the same scaffold, were likely mistaken for independent scaffolds in v1.0, overestimating the genome size and, therefore, gene and repetitive contents. The de-duplication caused a reduction of the genome size from 641 Mb (v1.0) to 573.9 Mb (v2.0), which is now closer to the genome size estimate obtained with Genomescope (488.2 Mb).

The comparison of the Chandler v2 assembly with that provided by Zhu et al. is an important point for the reader, as it will determine which genome will be used for further analysis. As an example, the long-range input data are different (Hi-C vs Optical maps) and maybe specific regions are not of the same quality in both assemblies.

We compared Chandler v2.0 with JSerr_1.0 from Zhu et al (2019). We observed that more than 95% of the two assemblies aligned with sequence identity higher than 98% and exhibited high collinearity, as shown in Figure S4 (see also lines 174-179). This confirms the high quality of both assemblies, even if obtained with different technologies. The availability of both assemblies will facilitate new genomic studies (e.g., pangenome) to understand the morphological and physiological differences among these two walnut varieties better.

Chandler is the most popular cultivar of walnut worldwide. For this reason, it was chosen for sequencing

and assembling the first walnut reference genome (Martinez-Garcia et al. 2016), and is used as the standard in many genetic and biological studies (e.g., development of genotyping tools). After four years, we are here proposing a significant improvement of the reference genome. However, a detailed study of the differences between the two assemblies (e.g., structural rearrangements) goes beyond the scope of the present manuscript.

Minor Points:

* assembly and gene prediction metrics are scattered throughout the manuscript and give a descriptive tone. I think the authors can move these metrics in tables 1 and 2. In addition, contig metrics are not provided in Table 1.

We modified the gene prediction metrics, which changed due to revisions made to conform with NCBI submission requirements. We moved some of the statistics of the gene annotation in Table 2. We also added contig metrics in Table 1.

* L38: "the full sequence of all 16 chromosomes" : how is this statement validated ?

We changed the sentence to "with the 16 chromosomal pseudomolecules assembled and representing 95% of its total length" (see line 38).

* L41 and L235: Asserting that the genes are complete based solely on the presence of a start and stop codon is not enough. Please delete the term "full-length". The number of complete BUSCO genes could perhaps be a way to evaluate the proportion of full-length genes.

We removed 'full-length' and added the proportion of BUSCO genes assembled completely (see lines 269-270)

* L87 and L90: problem with the closing parenthesis.

Done.

* L97: "...walnut reference genome with unprecedented contiguity...." Please delete this sentence.

Done.

* L117: a longest read of 992.2Kb is not informative if it does not align.

We are reporting general statistics on the Nanopore sequencing. Such long reads have then been used for scaffolding Illumina reads with MaSurCa.

* L156-158: The authors should used a kmer approach (Genomescope) to estimate the genome size of both genotypes.

We ran Genomescope using Chandler Illumina pair-end reads generated by Martinez-Garcia et al., (2016). Results indicate a genome size of 448 Mb, and a level of heterozygosity equal to 0.634%. Our de-duplicated assembly has a total length of 567.2 Mb (> 1 kb), which is a great improvement compared to Chandler v1.0. However, we were unable to run Genomescope for the Serr genome since we could not find any Illumina data on NCBI. We only found 58 PacBio SRA Experiments under the BioProject PRJNA413991 indicated by Zhu et al. (2019). Also, the Illumina reads of Zhu et al. (2019) are of the hybrid 'J.microcarpa x J.regia cv Serr'.

* L239: The proportion of gene models with multiple transcript isoforms is small relative to other plants which may not represent the proportion of genes with alternative splicing. I think the low depth of PACBIO sequencing is the main reason. Please rephrase the sentence to make it clearer.

With the newly revised gene annotation, we observed more gene models with multiple transcript isoforms. We changed the text accordingly, as suggested by RW#1 (see lines 246-248).

* L269-373 : This section is not clear for non-specialist readers.

We modified the paragraph to make it clearer for non-specialist readers (see lines 275-328). We also added Figure S9 showing the different stages of catkin development considered in our analysis.

* L283 : "four developed" ?

We generated proteomes from four tissues (immature catkin, intermediate catkin, mature catkin, and pure pollen). We changed 'developed' to 'analyzed' for clarity (see line 300).

* L343 : Please describe syntelogs.

Done (line 363).

* Figure5A: There may be a problem of alignment between the inner circle and the middle circle (blue region).

We thank the reviewer for highlighting this inconsistency in the figure. We checked our data and modified the image accordingly. The apparent misalignment between the inner circle and the middle circle is actually due to a lack of information – a gap of 4.07 cM – between two haploblocks which follow one another in that region. The inaccuracy was generated by the program used to draw the image, which automatically assigned the inherited segment to the most likely ancestor based on the maximum parsimony algorithm. However, we agree that on some occasions, this can lead to inaccurate results. Therefore, we decided to explicitly represent these regions of missing information using a white color. We also amended the legend of Figure5 to explain what the white sections describe.

* Too many paragraphs end with a sentence such as "support the crucial role of Chandler v2 chromosome-scale assembly".

We removed the sentence and ended the paragraphs differently.

* L463: Please describe how the gaps have been filled.

The assembler MaSurCa used to obtain Chandler hybrid has an internal gap-filling procedure, as described in Zimin et al., 2013 (see line 476-477). In addition, Dovetail uses Illumina reads to close the gaps in the HiRise assembly, as described in lines 500-502.

Reviewer #2: Overall, this genome is a significant advance over the previous one, but there are some points that are discussed in too much breadth, while others are too short for a detailed evaluation. Some of the claims regarding why the new genome assembly is superior over the older one(s) seem rather constructed. The parts that really profit from ONT sequencing - the near-repetitive gene families and the repeat content have not been explored in detail.

We performed a gene family analysis on Chandler's gene predictions (v1.0, v2.0, and RefSeq). Consistently with the gene prediction results, we observed that v2.0 has, in general, more members per gene family than v1.0, and fewer members than NCBI RefSeq. These results can be due to both the increment of contiguity and the lower gene redundancy of Chandler v2.0 (see lines 254-265).

I realise that this is just a data note, but some clues could help the reader appreciate the current manuscript. What is also missing is a comparison to other published reference genomes in the Fagales s.l..

We compared assembly metrics and BUSCO percentages to other Fagales genomes (see lines 141-142; 192-193; Table S9). However, if the reviewer was referring to comparative genomics studies, then it goes beyond the scope of the present manuscript, which is a Data Note on the improvement of the walnut reference genome. Also, we already provided applications of the new Chandler assembly for proteomic and population genetics studies.

L65. How is this hybridisation possible, given the current disjunction of the populations of the species? Please give a sentence or two as explanation.

Zhang et al. (2019), cited in the text, explained clearly how the contact between American black walnuts and Asian butternuts occurred during the Pliocene. According to their reconstruction, supported by fossil evidence, butternut and black walnuts spread into Eurasia from the late Oligocene to the Miocene-Pliocene. Then, the cooling climate of the Upper Pliocene may have led to range shifts of the butternut and black walnut lineages in Eurasia, permitting the contact required for the hybridization that gave rise

to Persian walnut. We add a short sentence describing Zhang et al (2019) results in the text (see line 66).

L87. Actinida is not a tree.

We changed trees to crops (see line 89).

L122-125. The process of obtaining the megareads is insufficiently described. Please expand the text and mention also the parameters used.

We added a more detailed description of how MaSurCa builds the mega-reads in the text (lines 466-478).

In addition, please provide statistics for the ONT reads and the Illumina reads.

We added the statistics on the ONT reads in Table S1. The Illumina data were generated by Martinez-Garcia et al. (2016), where Illumina sequencing details and statistics are reported.

Please also mention the library preparation technique used in both cases.

We added details on the ONT library preparation in the method section (Lines 460-464). Regarding the Illumina libraries, their descriptions are reported in Martinez-Garcia et al. (2016).

Please also mention the known biases associated with ONT sequencing and how strong these were in your raw data.

ONT sequencing tends to have more errors in long homopolymer regions (e.g., runs of all A's). We saw no evidence that these were a particular problem in this genome, although all genomes have such regions. Our use of Illumina reads to compute the final consensus sequence for virtually all positions in the assembly should minimize this problem.

L126-127. How has the old assembly (v.1.0) integrated with the new one?

After running the de-duplication module implemented in MaSuRCA, we aligned the v1 scaffolds to v2, identified unaligned regions, and added them to v2. We added this sentence to the text (line 481-483).

L155. Has it been checked, if the unanchored small scaffold are derived from contamination with bacteria/fungi?

All assemblies have contamination, not only from the original samples but from the sequencing lab itself. We ran a thorough contamination screen, aligning every contig and scaffold against an exhaustive set of bacteria, artificial vectors, and other plants and animals, and we removed all contaminants found in this manner. In addition, NCBI runs its own contaminant screen on every submission to GenBank, and that was run on our submission as well.

L170. The identity seems rather low. The possible reasons for this should be given.

We thank the reviewer for noticing this error. We estimated the average sequence identity from the nucmer coord file without filtering. In this way, we considered all alignments, including those between similar but not syntenic regions of the two assemblies. We re-estimated the percentage of sequence identity using only the 1-to-1 best alignments (command dnadiff implemented in MUMmer), and we obtain a value of 99.60. We changed the text accordingly (see lines 181-186).

L172. What was the proportion of unaligned reads? How many reads mapped discordantly?

Over a total of 432,183,992, 2,046,961 reads (0.5%) did not align and 31,169,557 did not pair properly.

L188. This statement cannot be upheld the way it is. Usually the gene space is already well-assembled using only Illumina reads (apart from the repetitive genes). The authors should compare the BUSCO scores of several Chandler assembly versions with that of other Fagales genomes, such as oak, beech, and chestnut.

We added Table S9 with BUSCO statistics for the Chandler genome assemblies, *J. regia* cv Serr, and other Fagales genomes.

L190. There are mapping-based ways to address this. These should be mentioned / applied.

We preferred to remove the sentence instead of following the reviewer's suggestion since revising/applying available methods for improving transcriptome assemblies using short reads is not among the scopes of our work. We aimed to prove how a much contiguous reference genome can improve transcriptome assemblies and gene predictions in walnut.

L1217-247. This seems overly discussed, considering the rather minor differences observed.

We moved some of the statistics in Table 2 and made this paragraph less redundant and descriptive.

L363. This is not necessarily evidence of inbreeding, but could also reflect selective sweeps. Inbreeding does not happen on the sub-genomic level but only on the genomic level.

Sub-genomic inbreeding occurs when an individual inherits the same copy of an allele at one locus from a common ancestor (identical-by-descent; IBD). Chandler parents share Payne as a common ancestor; therefore, there are high probabilities of IBD alleles at a sub-genomic level in Chandler. However, although we found no evidence of a strong selective sweep, it is also possible this pattern was due to direct or indirect selection (see lines 381-384). Future selective sweep studies in larger and more diverse walnut collections could provide more evidence on the high-level of homozygosity in some regions of the Chandler genome.

L426. Was any surface sterilisation done? Otherwise a lot of contaminant sequences would be expected. In any case, a contamination check should be reported.

We could not do surface sterilization since the Nanopore library was built starting from frozen tissue collected at UC Davis and sent to the John Hopkins University. Surface sterilization would have led to tissue degradation and plant production of stress compounds that further impede DNA extraction. Also, we ran a thorough contamination screen of our assembly, to remove contaminations associated with microbes present on the leaf surface and within the tissue (i.e., endophytes).

L428. 'g' should be in italics.

Done

L431. Concentrations/amounts missing.

Added.

L440. 'was' -> 'were'.

Done

L456. The assembly strategy, programs and parameters use are not mentioned in sufficient detail (actually hardly any of this is mentioned in the manuscript).

We added more details on the assembly strategy in the text (lines 466-478).

L531. Do not abbreviate at the beginning of the line.

Edit.

Close

Author's Response To Reviewer Comments

Close

Reviewer #1: I read the revised manuscript by Marrano et al. entitled "High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome", my comments have been addressed and the authors modified the manuscript accordingly. I still have a few minor points that need to be checked before publication.

* I maintain that the comparison of the Chandler v2 assembly with that provided by Zhu et al. is an important point for the reader and I would expect to have a comparison of the two assemblies in Table 1.

Done

* Please check how MaSurCA is written (MaSuRCA or MaSurCa)

We named MaSuRCA consistently across the manuscript.

* Table S3 : ON instead of ONT

Changed.

Reviewer #2: I found the manuscript to be a pleasant read, with the major methodological unclarities resolved. There are two minor suggestions from my side.

1. To promote Table S9 to a full table (Table 1 and 2 could probably be condensed into one table) with the inclusion of the *Quercus lobata* and *Quercus robur* genomes, as I feel that for readers it is always good to have some upfront benchmarking so they can put the assembly into the context of other genome assemblies.

We moved Table S9 as full table in the main text (now Table 3), and we also included the BUSCO metrics for *Quercus lobata* and *Q. robur*. We moved Figure 4 (now Figure S10) to the supplementary in place of Table S9.

2. To promote Figure S12 to a full figure (if possible).

We moved Figure S12 as a full figure (now Figure 5) in place of Figure 6 (now Figure S13).

Close

Reviewer Report

Title: High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome

Version: Original Submission **Date:** 11/27/2019

Reviewer name: Jean-Marc Aury

Reviewer Comments to Author:

I read the manuscript by Marrano et al. entitled "High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome" with interest. The authors describe how they generated a chromosome-scale assembly of *J. regia* based on ONT, Illumina and Hi-C data. In addition, genetic maps were used to validate and anchor the scaffolds to *J. regia* linkage groups. The authors performed a wide range of analysis, including gene families of interest and genomic diversity. The article is well written and the methods are sufficiently described.

My main concern is the quality of the assembly, although at the chromosome level, the contiguity at the contig level is ten times lower (nearly 1.1Mb) compared to the previously published *Juglans* genome assemblies (including *J. regia*). I understand that authors choose to use the methods they have developed, however, long-reads assemblies are usually made with dedicated assemblers, which can result in higher assembly quality. In particular, I was a little surprised by the fact that the v2 assembly contains fewer repetitive elements than the first version of the assembly (L175-176). Generally long-reads assemblies improved the repetitive content of genome assemblies.

The comparison of the Chandler v2 assembly with that provided by Zhu et al. is an important point for the reader, as it will determine which genome will be used for further analysis. As an example, the long-range input data are different (Hi-C vs Optical maps) and maybe specific regions are not of the same quality in both assemblies.

Minor Points:

- * assembly and gene prediction metrics are scattered throughout the manuscript and give a descriptive tone. I think the authors can move these metrics in tables 1 and 2. In addition, contig metrics are not provided in Table 1.

- * L38: "the full sequence of all 16 chromosomes" : how is this statement validated ?

- * L41 and L235: Asserting that the genes are complete based solely on the presence of a start and stop codon is not enough. Please delete the term "full-length". The number of complete BUSCO genes could perhaps be a way to evaluate the proportion of full-length genes.

- * L87 and L90: problem with the closing parenthesis.

- * L97: "...walnut reference genome with unprecedented contiguity...." Please delete this sentence.

- * L117: a longest read of 992.2Kb is not informative if it does not align.

- * L156-158: The authors should used a kmer approach (Genomescope) to estimate the genome size of both genotypes.

- * L239: The proportion of gene models with multiple transcript isoforms is small relative to other plants which may not represent the proportion of genes with alternative splicing. I think the low depth of PACBIO sequencing is the main reason. Please rephrase the sentence to make it clearer.

- * L269-373 : This section is not clear for non-specialist readers.
- * L283 : "four developed" ?
- * L343 : Please describe syntelogs.
- * Figure5A: There may be a problem of alignment between the inner circle and the middle circle (blue region).
- * Too many paragraphs end with a sentence such as "support the crucial role of Chandler v2 chromosome-scale assembly".
- * L463: Please describe how the gaps have been filled.

Level of Interest

Please indicate how interesting you found the manuscript: Choose an item.

Quality of Written English

Please indicate the quality of language in the manuscript: Choose an item.

Declaration of Competing Interests

Please complete a declaration of competing interests, considering the following questions:

- Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
- Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
- Do you hold or are you currently applying for any patents relating to the content of the manuscript?
- Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?
- Do you have any other financial competing interests?
- Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare no competing interests, however I received travel and accommodation expenses to speak at ONT conferences.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any

attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (<http://creativecommons.org/licenses/by/4.0/>). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

Choose an item.

To further support our reviewers, we have joined with Publons, where you can gain additional credit to further highlight your hard work (see: <https://publons.com/journal/530/gigascience>). On publication of this paper, your review will be automatically added to Publons, you can then choose whether or not to claim your Publons credit. I understand this statement.

Yes Choose an item.

Reviewer Report

Title: High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome

Version: Revision 1 **Date:** 3/24/2020

Reviewer name: Jean-Marc Aury

Reviewer Comments to Author:

I read the revised manuscript by Marrano et al. entitled "High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome", my comments have been addressed and the authors modified the manuscript accordingly. I still have a few minor points that need to be checked before publication.

* I maintain that the comparison of the Chandler v2 assembly with that provided by Zhu et al. is an important point for the reader and I would expect to have a comparison of the two assemblies in Table 1.

* Please check how MaSurCA is written (MaSuRCA or MaSurCa)

* Table S3 : ON instead of ONT

Level of Interest

Please indicate how interesting you found the manuscript: Choose an item.

Quality of Written English

Please indicate the quality of language in the manuscript: Choose an item.

Declaration of Competing Interests

Please complete a declaration of competing interests, considering the following questions:

- Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
- Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
- Do you hold or are you currently applying for any patents relating to the content of the manuscript?
- Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?
- Do you have any other financial competing interests?
- Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (<http://creativecommons.org/licenses/by/4.0/>). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

Choose an item.

To further support our reviewers, we have joined with Publons, where you can gain additional credit to further highlight your hard work (see: <https://publons.com/journal/530/gigascience>). On publication of this paper, your review will be automatically added to Publons, you can then choose whether or not to claim your Publons credit. I understand this statement.

Yes Choose an item.

Reviewer Report

Title: High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome

Version: Original Submission **Date:** 12/16/2019

Reviewer name: Marco Thines

Reviewer Comments to Author:

Overall, this genome is a significant advance over the previous one, but there are some points that are discussed in too much breadth, while others are too short for a detailed evaluation. Some of the claims regarding why the new genome assembly is superior over the older one(s) seem rather constructed. The parts that really profit from ONT sequencing - the near-repetitive gene families and the repeat content have not been explored in detail. I realise that this is just a data note, but some clues could help the reader appreciate the current manuscript. What is also missing is a comparison to other published reference genomes in the Fagales s.l..

L65. How is this hybridisation possible, given the current disjunction of the populations of the species? Please give a sentence or two as explanation.

L87. Actinida is not a tree.

L122-125. The process of obtaining the megareads is insufficiently described. Please expand the text and mention also the parameters used. In addition, please provide statistics for the ONT reads and the Illumina reads. Please also mention the library preparation technique used in both cases. Please also mention the known biases associated with ONT sequencing and how strong these were in your raw data.

L126-127. How has the old assembly (v.1.0) integrated with the new one?

L155. Has it been checked, if the unanchored small scaffolds are derived from contamination with bacteria/fungi?

L170. The identity seems rather low. The possible reasons for this should be given.

L172. What was the proportion of unaligned reads? How many reads mapped discordantly?

L188. This statement cannot be upheld the way it is. Usually the gene space is already well-assembled using only Illumina reads (apart from the repetitive genes). The authors should compare the BUSCO scores of several Chandler assembly versions with that of other Fagales genomes, such as oak, beech, and chestnut.

L190. There are mapping-based ways to address this. These should be mentioned / applied.

L217-247. This seems overly discussed, considering the rather minor differences observed.

L363. This is not necessarily evidence of inbreeding, but could also reflect selective sweeps. Inbreeding does not happen on the sub-genomic level but only on the genomic level.

L426. Was any surface sterilisation done? Otherwise a lot of contaminant sequences would be expected. In any case, a contamination check should be reported.

L428. 'g' should be in italics.

L431. Concentrations/amounts missing.

L440. 'was' -> 'were'.

L456. The assembly strategy, programs and parameters use are not mentioned in sufficient detail (actually hardly any of this is mentioned in the manuscript).

L531. Do not abbreviate at the beginning of the line.

Level of Interest

Please indicate how interesting you found the manuscript: Choose an item.

Quality of Written English

Please indicate the quality of language in the manuscript: Choose an item.

Declaration of Competing Interests

Please complete a declaration of competing interests, considering the following questions:

- Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
- Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
- Do you hold or are you currently applying for any patents relating to the content of the manuscript?
- Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?
- Do you have any other financial competing interests?
- Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (<http://creativecommons.org/licenses/by/4.0/>). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

Choose an item.

To further support our reviewers, we have joined with Publons, where you can gain additional credit to further highlight your hard work (see: <https://publons.com/journal/530/gigascience>). On publication of this paper, your review will be automatically added to Publons, you can then choose whether or not to claim your Publons credit. I understand this statement.

Yes Choose an item.

Reviewer Report

Title: High-quality chromosome-scale assembly of the walnut (*Juglans regia* L) reference genome

Version: Revision 1 **Date:** 3/23/2020

Reviewer name: Marco Thines

Reviewer Comments to Author:

I found the manuscript to be a pleasant read, with the major methodological unclarities resolved. There are two minor suggestions from my side.

1. To promote Table S9 to a full table (Table 1 and 2 could probably be condensed into one table) with the inclusion of the *Quercus lobata* and *Quercus robur* genomes, as I feel that for readers it is always good to have some upfront benchmarking so they can put the assembly into the context of other genome assemblies.
2. To promote Figure S12 to a full figure (if possible).

Level of Interest

Please indicate how interesting you found the manuscript: Choose an item.

Quality of Written English

Please indicate the quality of language in the manuscript: Choose an item.

Declaration of Competing Interests

Please complete a declaration of competing interests, considering the following questions:

- Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
- Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
- Do you hold or are you currently applying for any patents relating to the content of the manuscript?
- Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?
- Do you have any other financial competing interests?
- Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (<http://creativecommons.org/licenses/by/4.0/>). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

Choose an item.

To further support our reviewers, we have joined with Publons, where you can gain additional credit to further highlight your hard work (see: <https://publons.com/journal/530/gigascience>). On publication of this paper, your review will be automatically added to Publons, you can then choose whether or not to claim your Publons credit. I understand this statement.

Yes Choose an item.

Table S1. Statistics on the six Nanopore flowcells used to sequence Chandler’s genome.

Flowcell	Number of reads	Mean Length (bp)	Max Length (bp)	Total Length (Mbp)	N50 size (bp)	Mean Gc%	Mean read quality
170615	104,477	4,891	726,206	511	16,260	39.99	14.66
170710	625,785	6,332	208,758	3,963	9,813	37.38	14.7
170714	1,639,729	2,163	145,375	3,547.5	2,789	41.41	14.3
170714.n	3,109,204	1,265	184,491	3,933.5	1,496	45.86	12.8
170908	28,997	4,309	992,241	125	9,018	37.37	12.12
170911	1,588,119	6,197	472,064	9,842.2	9,605	37.07	15.48
Total	7,096,311	3,089	992,241	21,922.4	6,763	-	-

Table S2. Statistics on k-unitigs, super-reads and mega-reads obtained with the MaSuRCA assembler on ONT and Illumina reads.

	Number	Average length (bp)	N50 (bp)	Coverage
K-unitigs	11,818,995	170	160	3.25
Super-reads	3,679,106	777	982	4.61
Mega-reads	3,160,807	4,716.12	8,158	24.04

Table S3. Characteristics of the Chandler **ONT** assembly.

	Number	Total Size (bp)	N50 (bp)
ONT Scaffolds	1,498	560,575,767	1,663,083
ONT Contigs	258	1,592,429	15,764
Scaffolds “Chandler” v1.0	25,007	11,648,497	731
Total (nuclear genome)	26,763	573,816,693	1,609,923

Table S4. List of tissues used for PacBio IsoSeq.

Tissue	Developmental Stage	Source	Abbreviation
Callus	-	In vitro	CALS
Catkin (1-inch)	Vegetative	Tree	CAT1
Catkin (3-inches)	Vegetative	Tree	CAT3
Leaf	Vegetative	Tree	LEAF

Pistillate Flower	Vegetative	Tree	PISF
Pollen	Vegetative	Tree	POLN
Root	Vegetative	Pot	ROOT
Shoot	Vegetative	In vitro	SHOOT
Somatic Embryo	Immature	In vitro	SOME

Table S5. Statistics on the PacBio IsoSeq sequencing per flow-cell.

SMRT Cell	First	Second	Third	Fourth	Total
Subreads	12,615,171	4,762,871	3,856,461	5,093,584	26,328,087
Subreads Total Length	15,022,521,907	6,207,229,047	4,290,409,550	5,756,344,948	31,276,505,452
Average Subread Length	1,191	1,303	1,113	1,130	1,188
Used for CCS	12,367,203	4,702,605	3,797,625	5,013,184	25,880,617
CCS generated	488,157	174,534	134,471	174,331	971,493
CCS Total Length	698,640,953	280,847,348	169,362,257	227,109,058	1,375,959,616
Average CCS Length	1,431	1,609	1,259	1,303	1,416

Table S6. Statistics on CCSs, FLnc and HQ FL transcripts obtained per tissue with PacBio IsoSeq.

Tissue ¹	N. of CCSs ²	Total Length (CCS)	Average Length (CCS)	N. of FLnc	Total Length (FLnc)	Average Length (FLnc)	HQ Clusters	Total Length (HQ FL)	Average Length (HQ FL)
CALS	129,653	163,156,475	1,258	129,535	160,363,808	1,238	11,828	15,284,033	1,292
CAT1	13,631	18,596,784	1,364	13,613	18,287,125	1,343	1,512	1,987,615	1,315
CAT3	125,219	161,569,359	1,290	125,083	158,780,204	1,269	11,739	15,170,585	1,292
LEAF	142,314	195,702,629	1,375	142,168	192,587,202	1,355	11,940	16,713,706	1,400
PISF	108,068	145,732,504	1,349	107,989	143,491,487	1,329	10,501	14,204,376	1,353
POLN	74,310	108,230,041	1,456	74,179	106,535,501	1,436	4,573	6,685,431	1,462
ROOT	17,897	23,874,571	1,334	17,875	23,480,651	1,314	2,252	2,852,647	1,267
SHOOT	14,508	19,327,322	1,332	14,494	19,003,828	1,311	1,400	1,743,863	1,246
SOME	120,935	187,564,762	1,551	120,794	184,845,798	1,530	12,480	19,783,294	1,585
Total	746,535	1,023,754,447	1,368	745,730	1,007,375,604	1,347	68,225	94,425,550	1,357

¹ Abbreviations are described in Table S4

² above threshold

Table S7. Percentage of FLnc and HQ FL transcripts aligned on the new assembly per tissue.

Tissue	FLnc mapped on Chandler v2.0	HQ clusters mapped on Chandler v2.0
CALS	99.88%	100.00%
CAT1	99.76%	100.00%
CAT3	99.77%	99.96%
LEAF	99.90%	99.99%
PISF	99.89%	99.99%
POLN	97.84%	94.18%
ROOT	98.69%	99.88%
SHOOT	99.86%	100.00%
SOME	99.90%	99.99%

Table S8. Validation of gene annotation with expression data and public plant protein databases.

	v2 transcripts (40,884)	v2 gene models (37,554)
supported by RNA-seq data from [9]	37,202	34,203
Supported by PacBio IsoSeq data	19,124	18,880
With a hit to a <i>J. regia</i> RefSeq v1 accessions	40,203	36,683
Annotate with a plant protein by EnTAP using the entire NCBI RefSeq plant database	40,318	36,824

Table S13. Number of SNPs and SNP density in 100-kb windows per chromosome in Chandler v2.0.

Chr	Length	Total Number of SNPs	Mean Number of SNPs/100Kb
Chr01	45,207,397	173,403	382.79
Chr02	37,821,870	145,689	385.42
Chr03	35,064,427	194,559	554.30
Chr04	34,823,025	151,361	433.70
Chr05	22,562,875	92,795	412.42
Chr06	39,020,271	189,076	483.57

Chr07	52,418,484	195,627	373.33
Chr08	30,564,197	128,621	420.33
Chr09	24,263,475	97,202	400.01
Chr10	37,707,155	216,161	571.85
Chr11	37,114,715	140,513	377.72
Chr12	31,492,331	120,082	381.21
Chr13	39,757,759	127,886	321.32
Chr14	28,841,373	77,099	266.78
Chr15	20,407,330	40,447	198.27
Chr16	28,711,772	115,314	400.40
Tot.	545,778,456	2,205,835	6,363

Table S14. Regions (1 Mb) with less than 377.5 SNPs (10th percentile of the SNP number distribution) in Chandler.

Chr	Window start	Window end	SNP count	Variants/kb
Chr01	5000000	6000000	11	0.011
Chr01	6000000	7000000	145	0.145
Chr01	11000000	12000000	99	0.099
Chr01	12000000	13000000	153	0.153
Chr01	37000000	38000000	278	0.278
Chr01	38000000	39000000	75	0.075
Chr01	43000000	44000000	163	0.163
Chr01	45000000	46000000	334	0.334
Chr02	14000000	15000000	184	0.184
Chr02	16000000	17000000	357	0.357
Chr02	18000000	19000000	203	0.203
Chr02	33000000	34000000	30	0.03
Chr02	34000000	35000000	83	0.083
Chr03	35000000	36000000	256	0.256
Chr05	1000000	2000000	123	0.123
Chr06	10000000	11000000	174	0.174
Chr06	39000000	40000000	3	0.003
Chr07	10000000	11000000	227	0.227
Chr07	26000000	27000000	112	0.112
Chr07	27000000	28000000	240	0.24
Chr07	29000000	30000000	181	0.181

Chr07	30000000	31000000	119	0.119
Chr07	51000000	52000000	92	0.092
Chr08	16000000	17000000	116	0.116
Chr08	18000000	19000000	293	0.293
Chr08	19000000	20000000	359	0.359
Chr09	15000000	16000000	110	0.11
Chr09	24000000	25000000	65	0.065
Chr11	8000000	9000000	109	0.109
Chr11	18000000	19000000	173	0.173
Chr11	20000000	21000000	109	0.109
Chr11	21000000	22000000	191	0.191
Chr11	37000000	38000000	326	0.326
Chr12	1000000	2000000	171	0.171
Chr12	20000000	21000000	73	0.073
Chr12	21000000	22000000	79	0.079
Chr13	0	1000000	244	0.244
Chr13	17000000	18000000	352	0.352
Chr13	19000000	20000000	261	0.261
Chr13	20000000	21000000	376	0.376
Chr13	21000000	22000000	189	0.189
Chr13	27000000	28000000	227	0.227
Chr14	12000000	13000000	219	0.219
Chr14	13000000	14000000	213	0.213
Chr14	16000000	17000000	358	0.358
Chr15	4000000	5000000	247	0.247
Chr15	6000000	7000000	91	0.091
Chr15	9000000	10000000	190	0.19
Chr15	11000000	12000000	351	0.351
Chr15	16000000	17000000	199	0.199
Chr15	17000000	18000000	59	0.059
Chr15	18000000	19000000	202	0.202
Chr15	19000000	20000000	116	0.116
Chr15	20000000	21000000	81	0.081
Chr16	6000000	7000000	106	0.106
Chr16	23000000	24000000	184	0.184

Table S15. Top 50 biological process GO terms for the 393 singletons genes in the low heterozygous regions on Chr15 of Chandler.

GO.ID	Term	Total annotated genes	Number of significant genes on Chr15	P-value
-------	------	-----------------------	--------------------------------------	---------

GO:0007165	signal transduction	2,682	49	9.00E-22
GO:0009626	plant-type hypersensitive response	366	14	3.60E-19
GO:0006468	protein phosphorylation	1,154	24	3.10E-10
GO:0045088	regulation of innate immune response	212	2	3.10E-08
GO:0051258	protein polymerization	105	2	4.60E-08
GO:0048544	recognition of pollen	120	10	1.30E-06
GO:0000086	G2/M transition of mitotic cell cycle	44	1	1.80E-06
GO:0009308	amine metabolic process	152	2	2.40E-06
GO:0072593	reactive oxygen species metabolic process	236	4	7.60E-06
GO:0009833	plant-type primary cell wall biogenesis	54	1	9.50E-06
GO:0009793	embryo development ending in seed dormancy	506	4	9.60E-06
GO:0001709	cell fate determination	35	1	1.00E-05
GO:0009267	cellular response to starvation	202	1	1.10E-05
GO:0009636	response to toxic substance	306	4	1.20E-05
GO:0016311	dephosphorylation	194	6	3.20E-05
GO:0016236	macroautophagy	115	2	3.40E-05
GO:0015074	DNA integration	724	14	4.00E-05
GO:0009791	post-embryonic development	2,153	20	4.80E-05
GO:0009409	response to cold	484	6	9.10E-05
GO:0016114	terpenoid biosynthetic process	234	5	0.00014
GO:0046777	protein autophosphorylation	423	6	0.0002
GO:0006790	sulfur compound metabolic process	421	4	0.00025
GO:0000272	polysaccharide catabolic process	294	7	0.00027

Table S16. List of 23 UCD-WIP founders used for the selective sweep analysis [10].

Accession Name	Origin	Abbreviation
UC-85-008	China	XJG6
UC-85-043-1	Bulgaria	SH_SD
UC-86-011	France/Germany	J_PUR
UC-91-013-5	China	ZL5
UC-91-031-8	China	AH85
UC-91-041-12	China	AK67
UC-91-056-9	China	PRLY
Conway-Mayette	USA	CMAY
Eureka	USA	EURK
Hartley	USA	HART
Idaho	USA	IDA
Lara	France	LARA

Manregian (PI18256)	China	MAREG
Marchetti	USA	MARCH
Meylan	France	MEYL
Payne	USA	PAYN
PI159568	Afghanistan	PI15_8
Scharsch Franquette	USA/France	S_FRA
Sharkey	USA	SHARK
Sinensis #5	Japan	SIN5
Soleze	France	SLZ
Waterloo	USA	WAT
UC-64-057	USA	64-057

Table S17. Top 50 biological process GO terms for the 122 windows (100 kb) with negative value of Tajima's D in the Western genotypes.

GO.ID	Term	Total annotated genes	Number of genes in the regions with negative D_{occ}	P-value
GO:0007165	signal transduction	2,682	43	4.30E-21
GO:0009626	plant-type hypersensitive response	366	3	5.30E-19
GO:0006468	protein phosphorylation	1,154	14	4.90E-10
GO:0051258	protein polymerization	105	7	4.10E-08
GO:0045088	regulation of innate immune response	212	2	1.50E-07
GO:0000086	G2/M transition of mitotic cell cycle	44	2	1.80E-06
GO:0009308	amine metabolic process	152	5	3.90E-06
GO:0072593	reactive oxygen species metabolic process	236	4	6.70E-06
GO:0009636	response to toxic substance	306	5	8.50E-06
GO:0009625	response to insect	28	1	1.20E-05
GO:0006081	cellular aldehyde metabolic process	135	1	1.90E-05
GO:0080027	response to herbivore	4	1	2.00E-05
GO:0016311	dephosphorylation	194	2	2.50E-05
GO:0009791	post-embryonic development	2,153	36	3.60E-05
GO:0009793	embryo development ending in seed dormancy	506	8	3.90E-05
GO:0009861	jasmonic acid and ethylene-dependent systemic resistance	45	1	4.60E-05
GO:0046777	protein autophosphorylation	423	4	8.80E-05
GO:0015074	DNA integration	724	12	0.0001
GO:0046839	phospholipid dephosphorylation	31	1	0.0001
GO:0016114	terpenoid biosynthetic process	234	2	0.00011
GO:0071555	cell wall organization	677	7	0.00018

GO:0030244	cellulose biosynthetic process	120	2	0.0002
------------	--------------------------------	-----	---	--------

Table S18. Top 50 biological process GO terms for the 122 windows (100 kb) with negative value of Tajima's D in the Eastern genotypes.

GO.ID	Term	Total annotated genes	Number of genes in the regions with negative D_{Asia}	P-value
GO:0007165	signal transduction	2,682	2	7.00E-22
GO:0009793	embryo development ending in seed dormancy	506	1	9.40E-06
GO:0009636	response to toxic substance	306	1	1.20E-05
GO:0006081	cellular aldehyde metabolic process	135	2	2.40E-05
GO:0009791	post-embryonic development	2,153	3	4.70E-05
GO:0016114	terpenoid biosynthetic process	234	5	0.00014

Table S19. Marker-trait associations identified within genomic regions highly differentiated between Western and Eastern walnuts.

Trait	SNP	Chr	Position (bp)	P-value	Window Start	Window end	N. SNPs	F _{ST}	π_{Asia}	π_{Occ}	ROD	D _{asia}	D _{Occ}
Harvest Date	AX-170770379 ¹	1	6506363	9.2E-13	6500001	6600000	1526	0.394	0.006	0.002	0.629	1.877	0.771
Shape index	AX-171158510 ²	3	280565	1.82E-08	200001	300000	659	0.373	0.002	0.001	0.392	1.318	0.105
Round index	AX-171158510	3	280565	4.38E-07	200001	300000	659	0.373	0.002	0.001	0.392	1.318	0.105
Nut shape	AX-171105437 ²	4	6315247	1.07E-05	6300001	6400000	2189	0.371	0.010	0.002	0.791	2.652	-0.533
Nut shape	AX-171105430 ²	4	6326737	1.07E-05	6300001	6400000	2189	0.371	0.010	0.002	0.791	2.652	-0.533

¹ [14]

² [12]

Figure S1. Contiguity of the Chandler ON assembly and the final HiRise scaffolds. Each curve shows the fraction of the total length of the assembly present in scaffolds of a given length or smaller. The fraction of the assembly is indicated on the Y-axis and the scaffold length in base pairs is given on the X-axis. The two dashed lines mark the N50 and N90 lengths of each assembly. Scaffolds less than 1 kb are excluded.

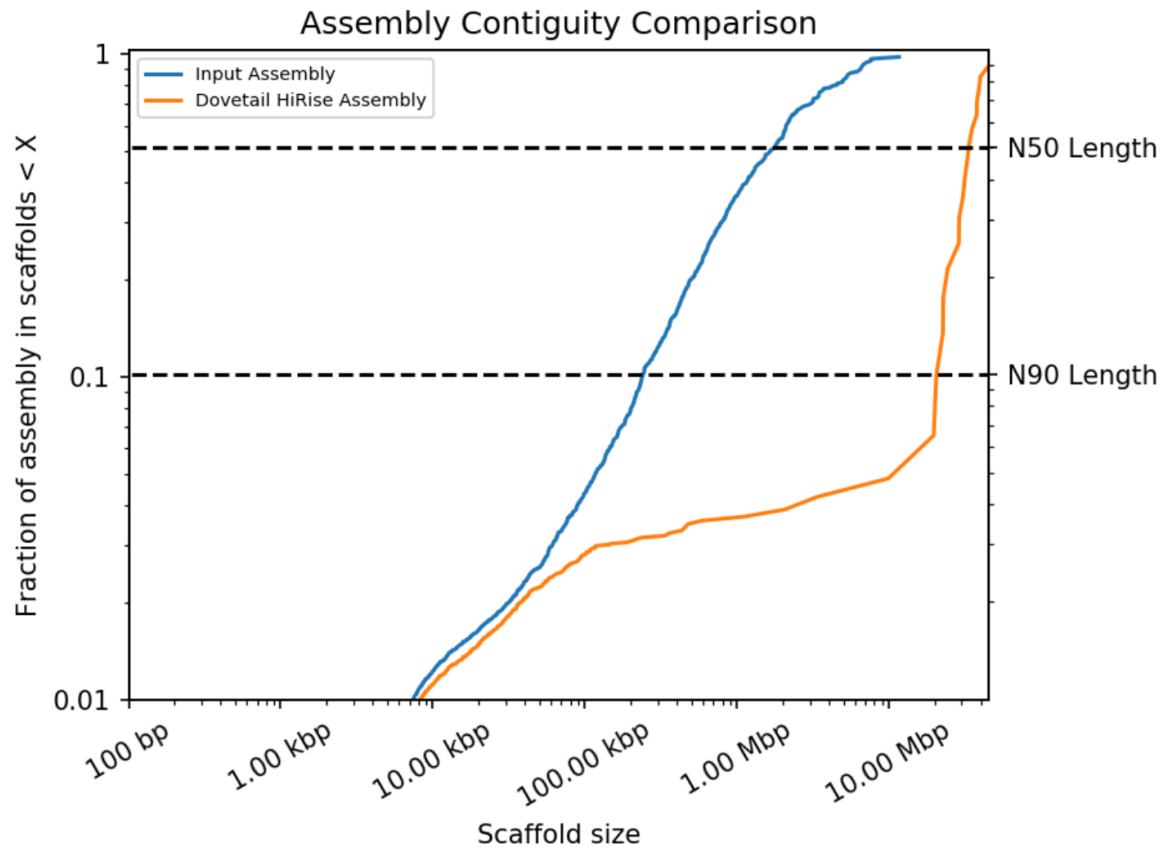


Figure S2. Mapping positions of the first and second read in the read pair respectively, grouped into bins. The color of each square gives the number of read pairs within that bin. White vertical and black horizontal lines show the borders between scaffolds. Scaffolds less than 1 Mb are excluded.

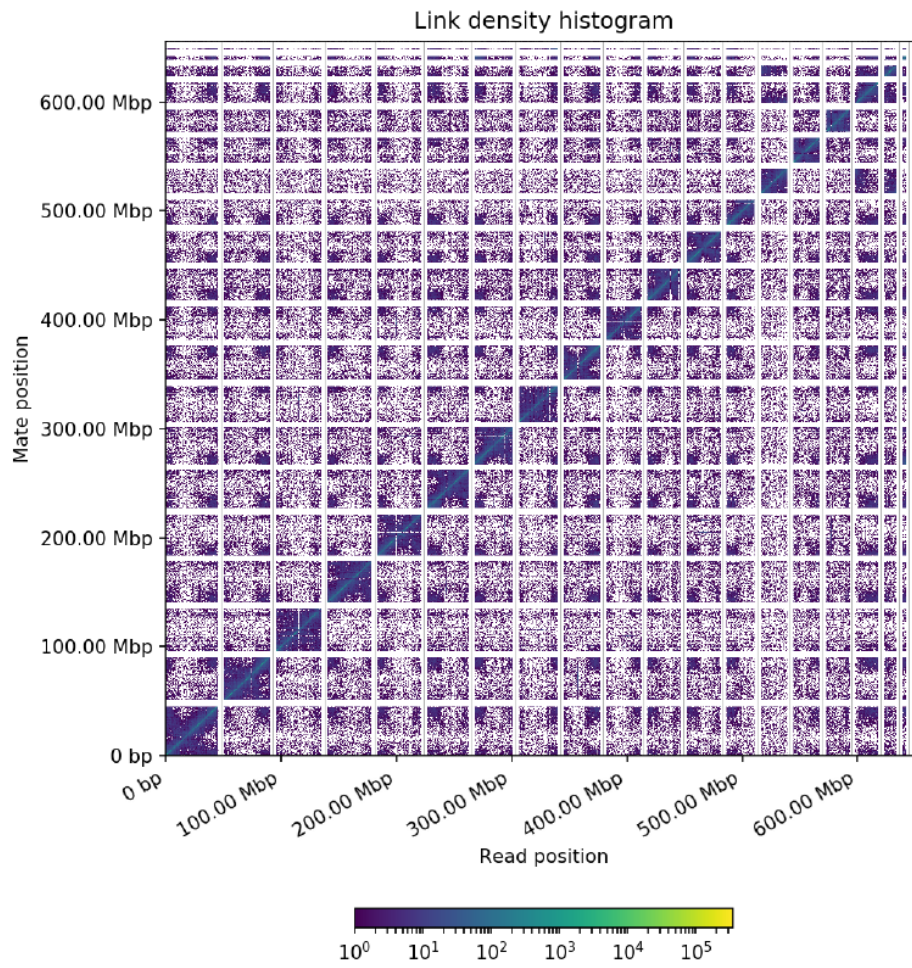


Figure S3. Collinearity between the ‘Chandler’ genetic map of [35] and the 16 chromosomal pseudomolecules of Chandler v2.0.

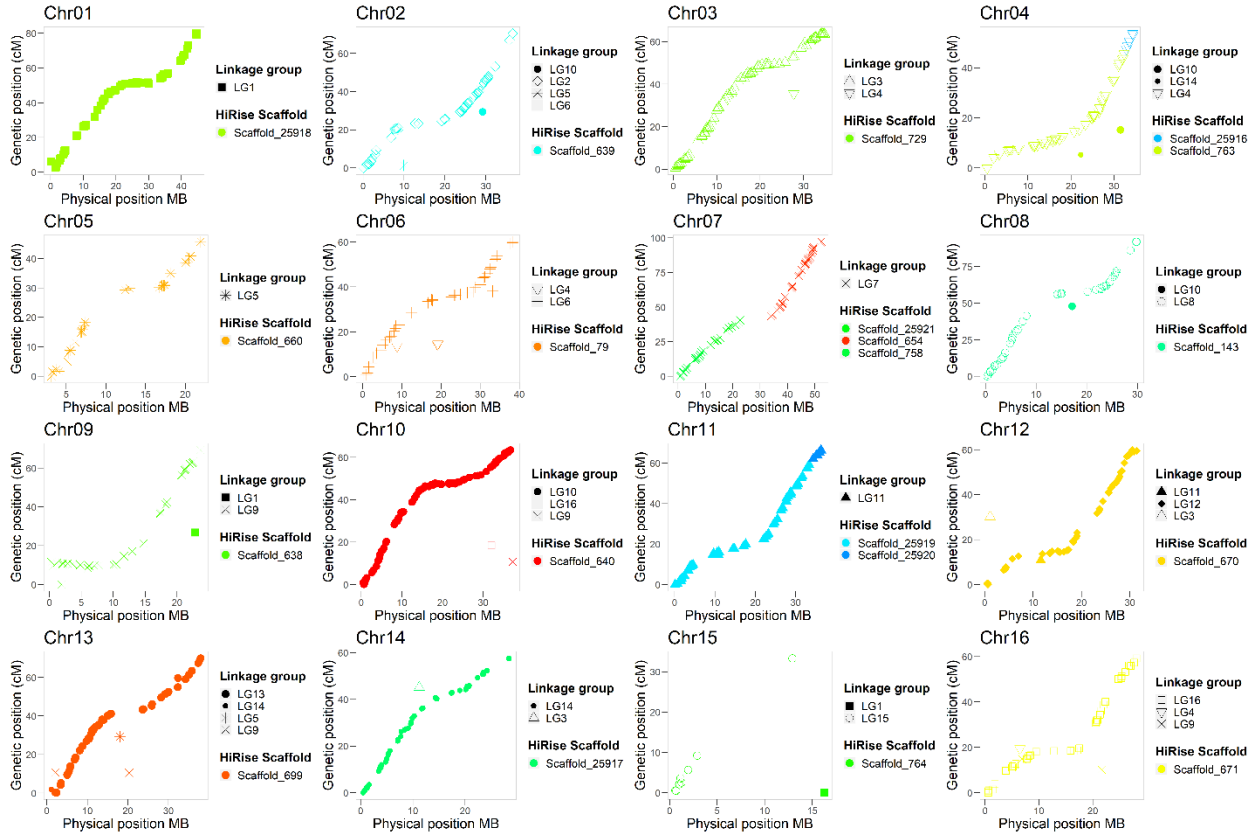


Figure S4. Whole-genome comparison between Chandler v2.0 and JrSerr_v1.0 [22].

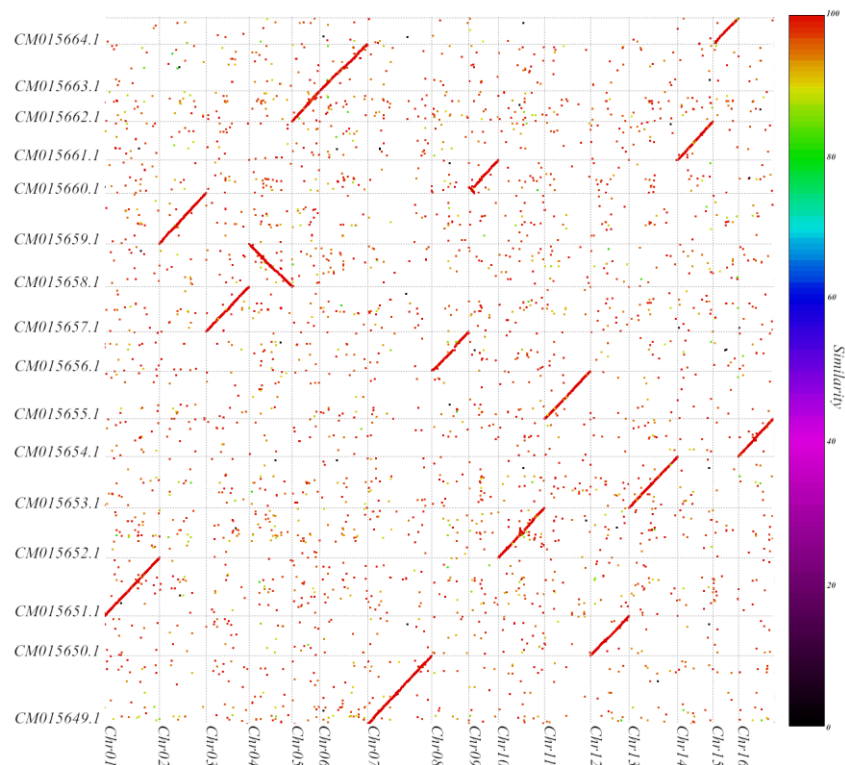


Figure S5. Retrotransposons distribution across the 16 chromosomes of Chandler v2.0.

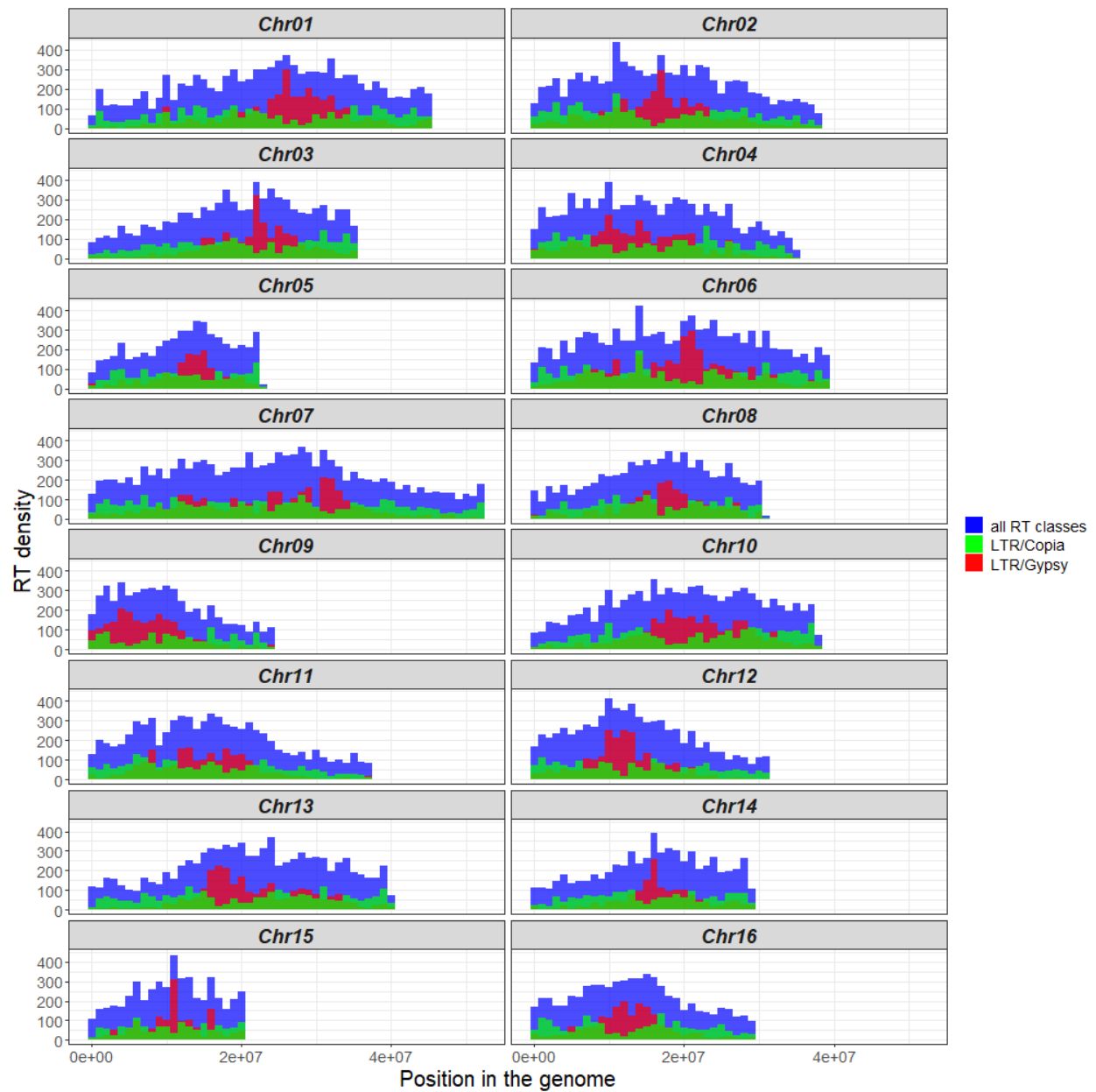


Figure S6. Top 20 biological process GO terms.

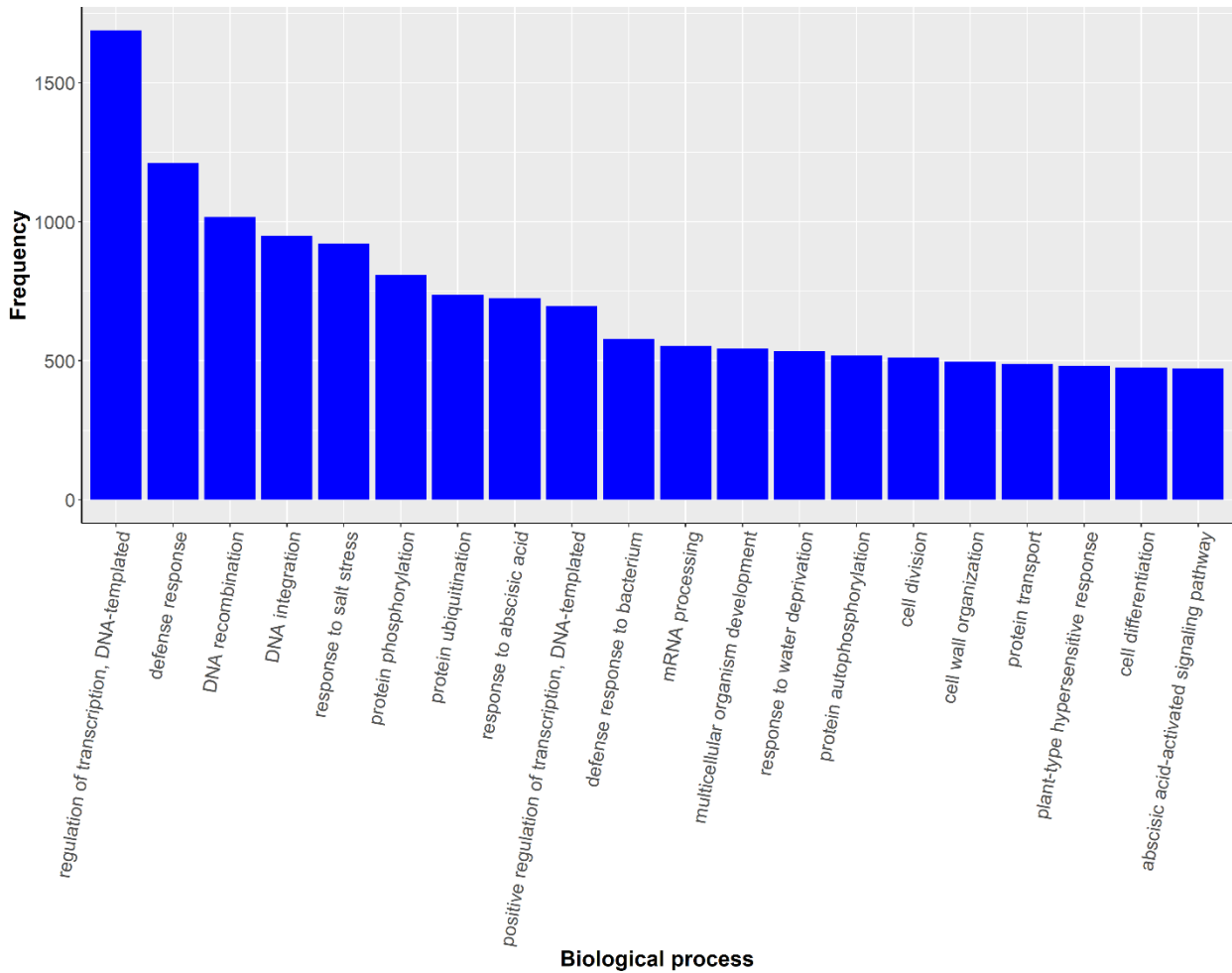


Figure S7. Top 20 molecular function GO terms.

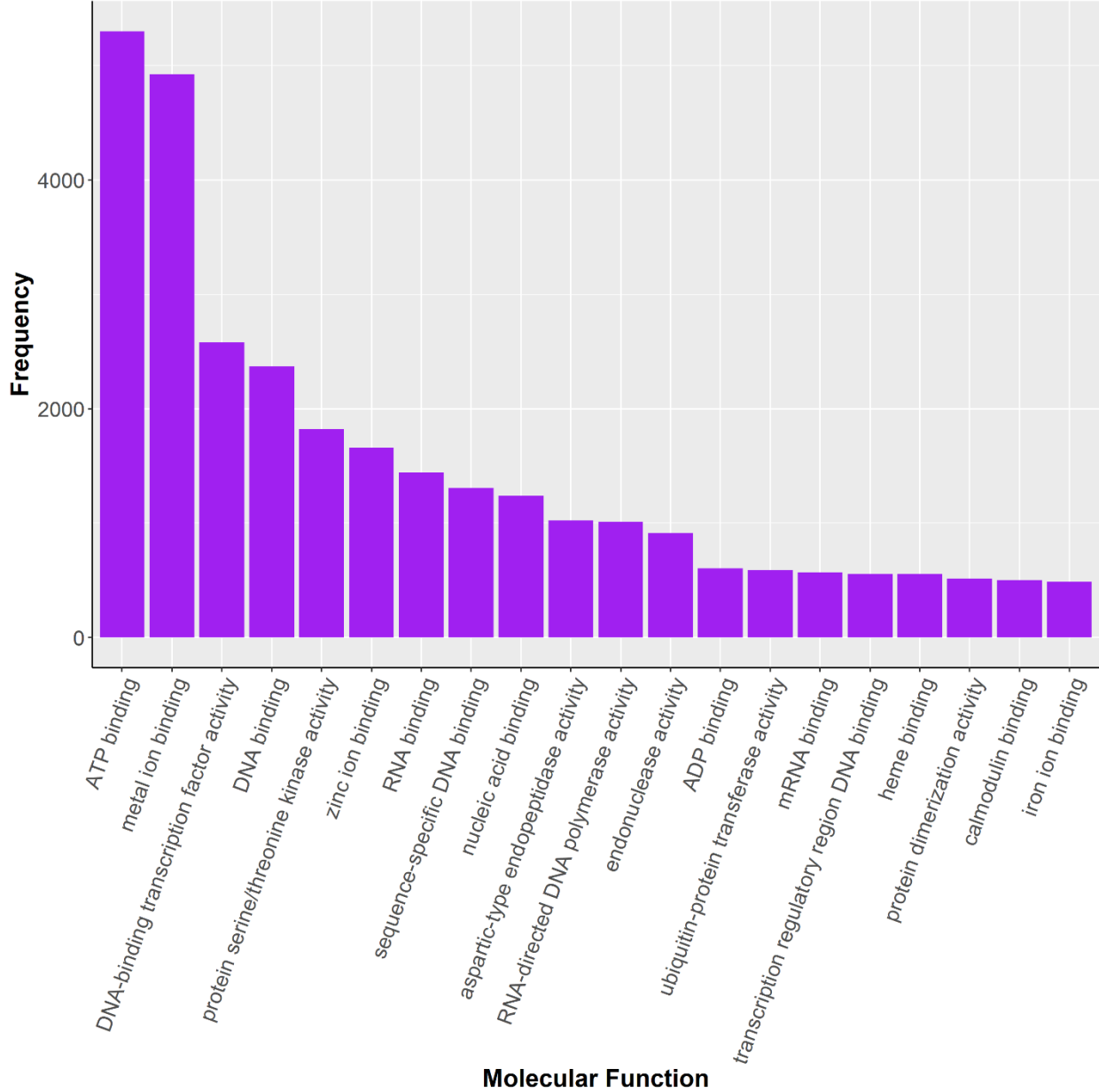


Figure S8. Top 20 cellular component GO terms.

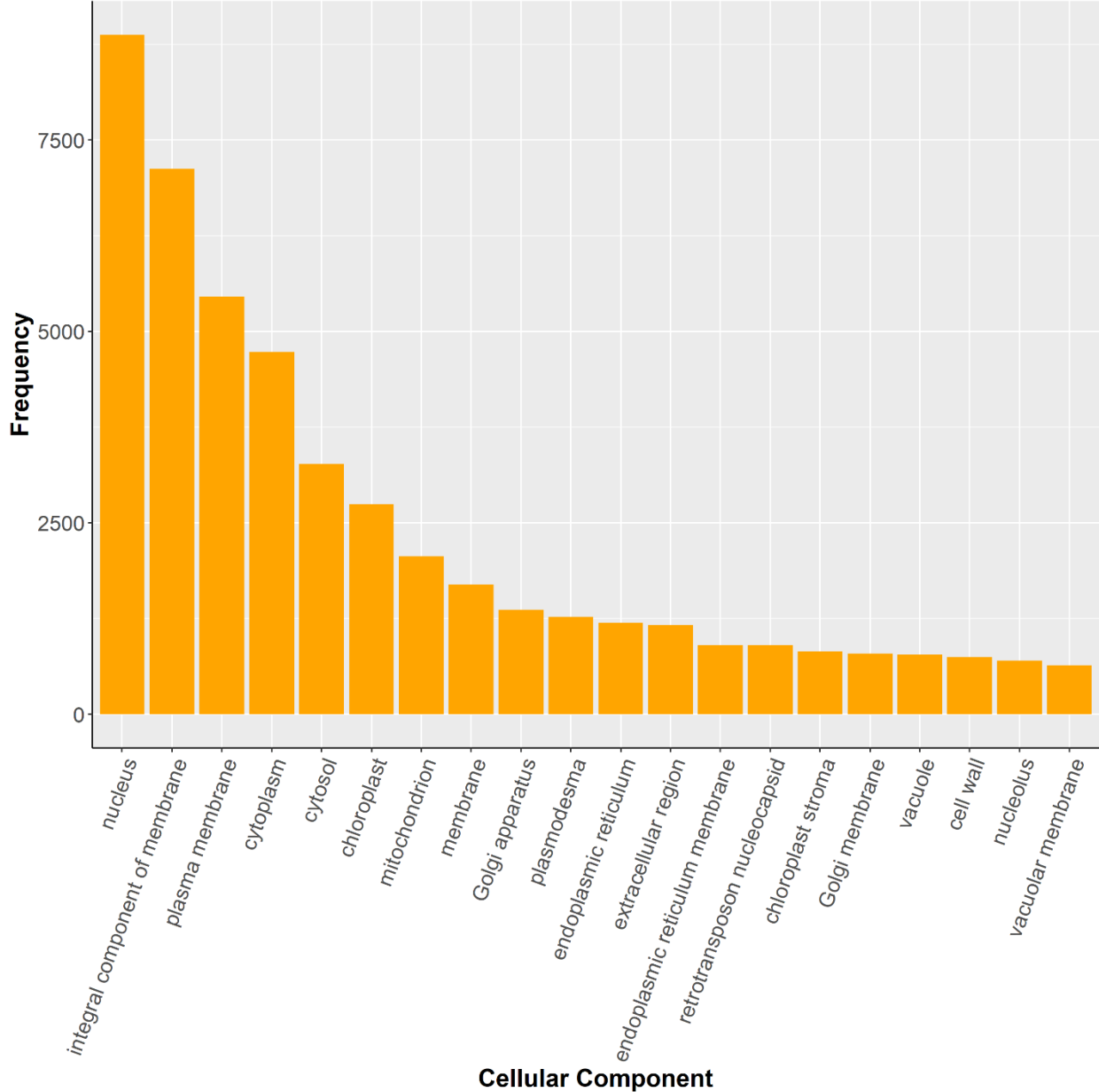


Figure S9. Catkin development stages sampled for the label-shotgun proteomic study (pollen is not shown).

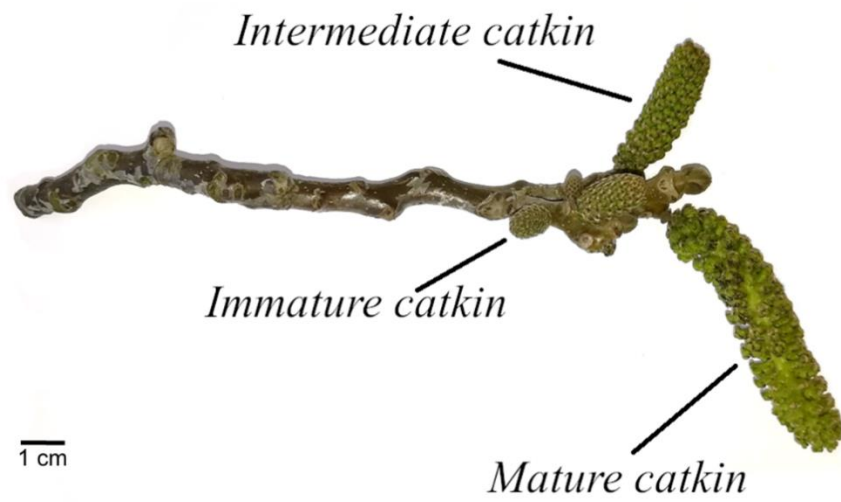


Figure S10. Modeled structure of the putative new allergen encoded by *Jr12_05180*. The compact structure is stabilized by four disulfide bonds, common in other allergenic proteins. The model in blue is superimposed with a homologous allergen from lentil (PDBid:2MAL) represented in red. Structure rendered with Pymol 2.3.

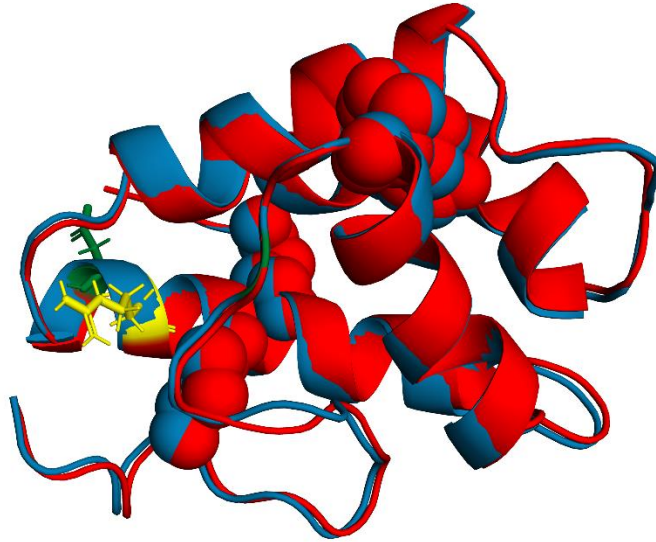


Figure S11. Collinear blocks among the 16 chromosomal pseudomolecules of Chandler v2.0. Different colors of dots represent different pairs of paralogous genes.

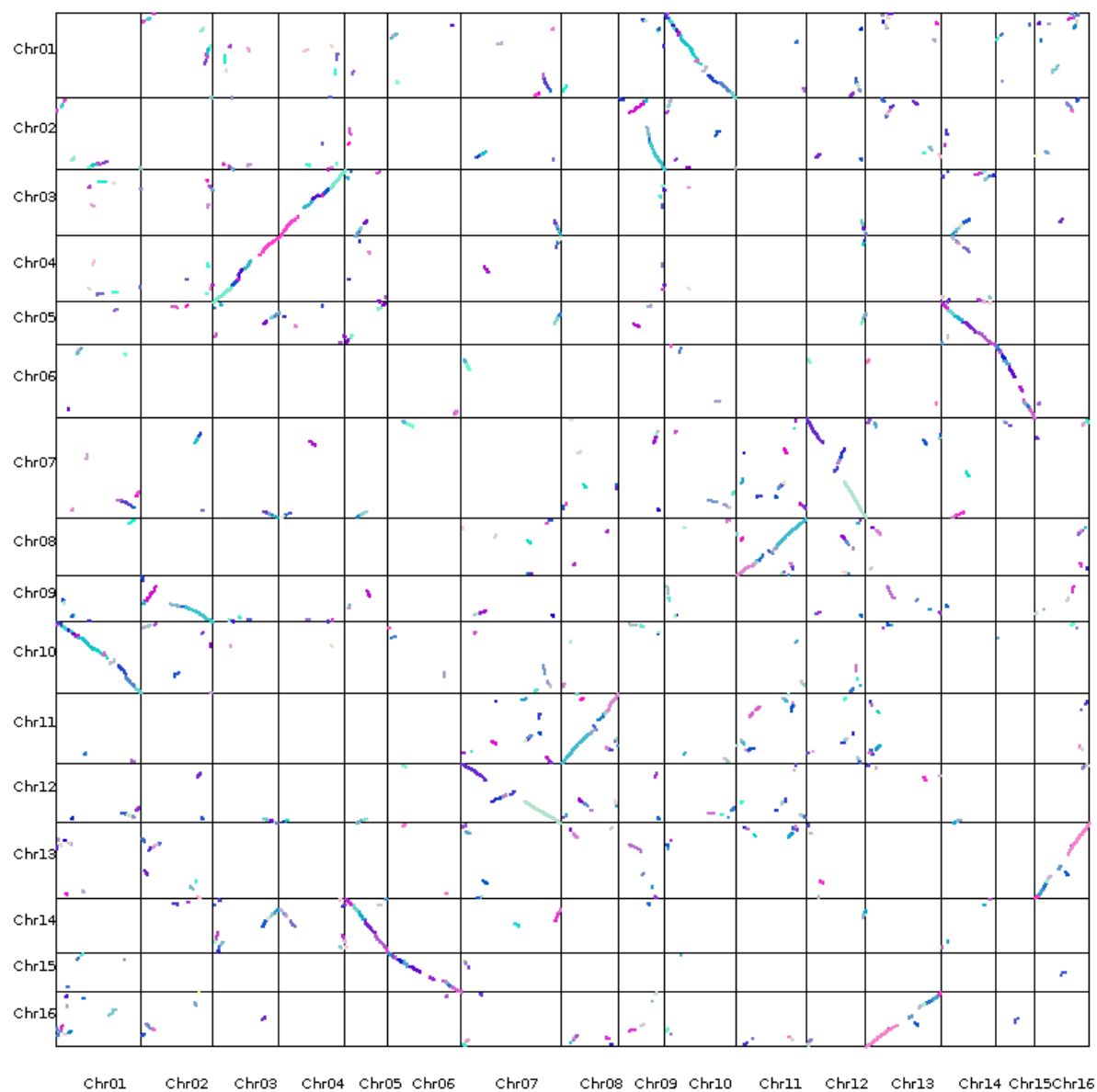


Figure S12. Dual syntenic plot between Chr06 a- Chr15 of Chandler v2.0.

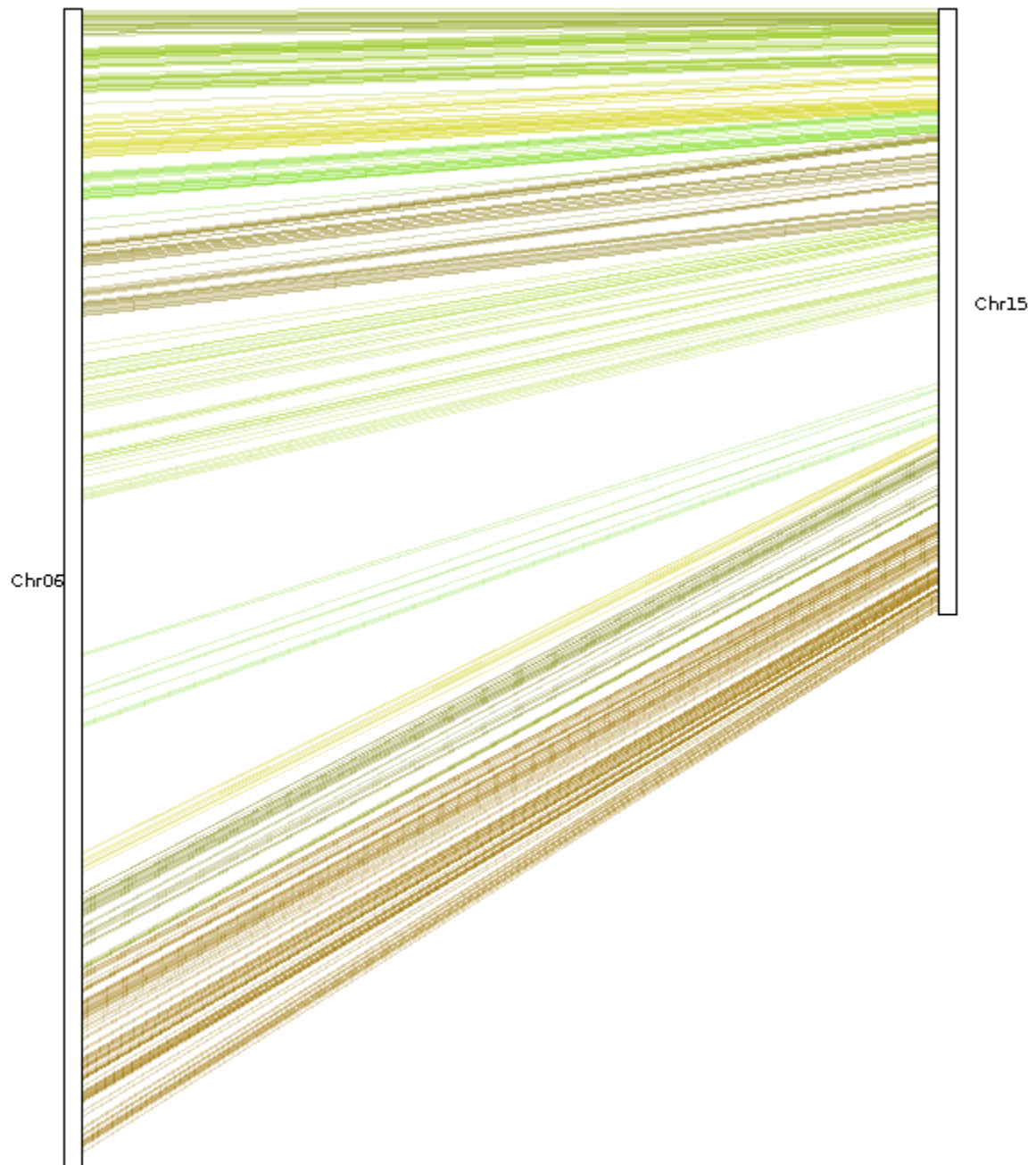


Figure S13. Graphical visualization of allele identity between Chandler and its ancestor Payne for all 16 chromosomes of Chandler.

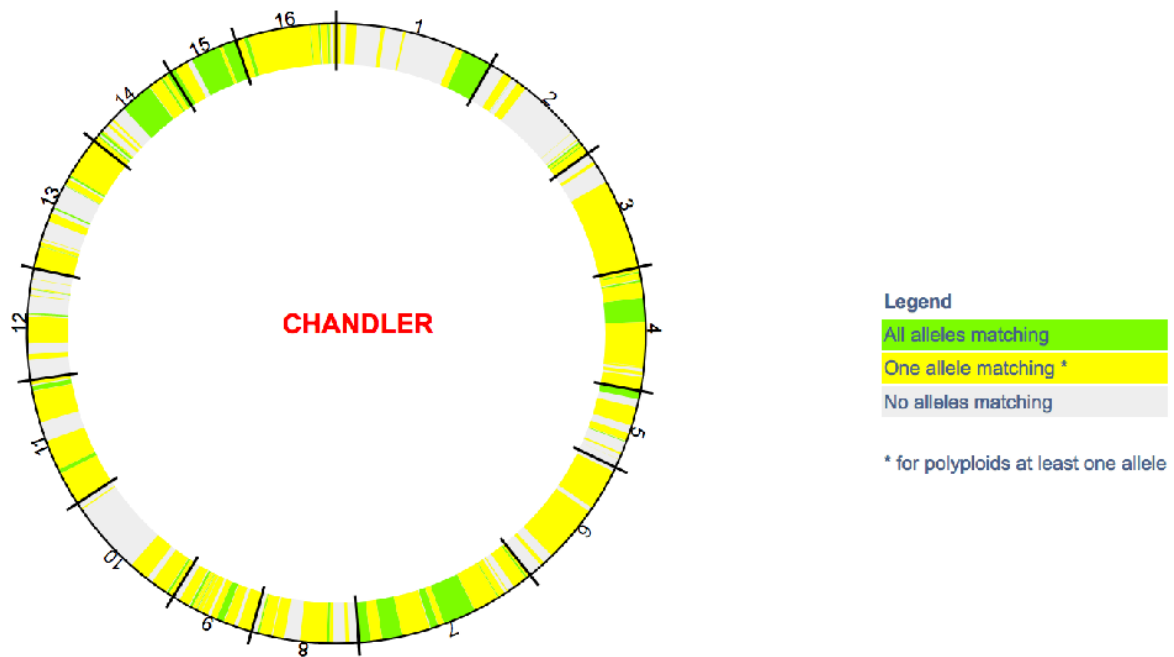


Figure S14. Hierarchical clustering analysis among the 23 re-sequenced founders of the UCD-WIP.

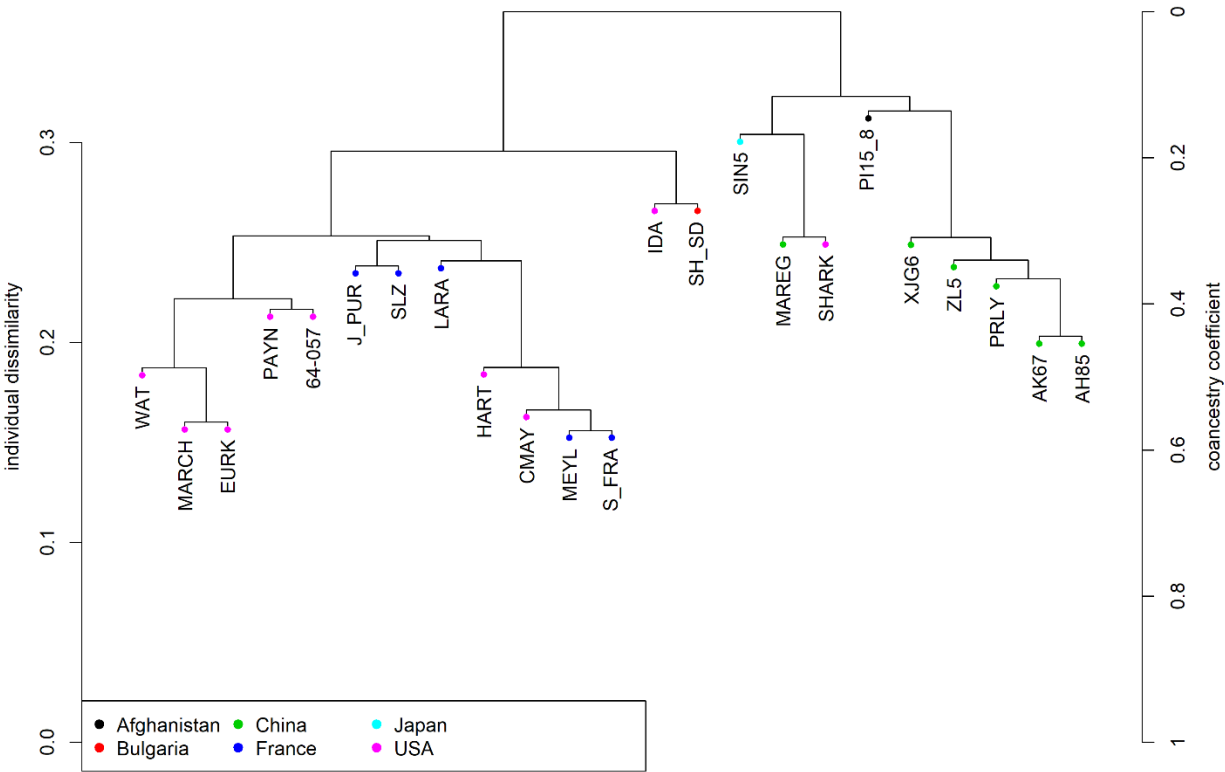


Figure S15. Genome scan for selective sweeps between walnuts from EU/USA and Asia. Tracks from outside to inside: **(i)** F_{ST} in 500-kb windows. Windows in the 95 percentiles of the F_{ST} distribution are highlighted in red; **(ii)** ROD values for 500-kb windows; **(iii)** Tajima's D in 500-kb windows for genotypes from Europe a- USA; **(iv)** Tajima's D in 500-kb windows for genotypes from Asia.

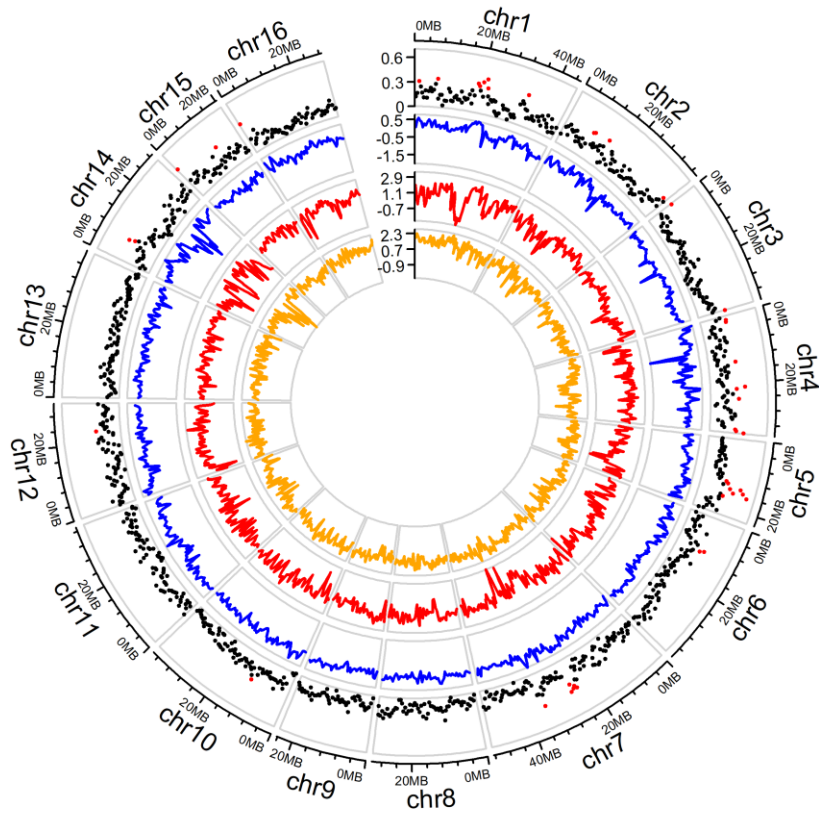


Figure S16. Phenotypic differences of harvesting date observed among the three genotypic classes of the marker AX-170770379 significantly associated to harvest date (Marrano et al., 2019).

