GigaScience

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 (Juglans regia L) reference genome
Article Type:	Data Note
Funding Information:	
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.
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 (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 significicnt 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 disjuction 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).

Ll122-125. 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.

Ll126-127. How has the ols assemblly (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 sholuld 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-assmbled 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.

Ll217-247. 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 imbreeding, but could also reflect selective sweeps. Imbreeding does not happpen 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 straregy, 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 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	Yes

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 Marrano[1], amarrano@ucdavis.edu; ***corresponding author***

Monica Britton[2], mtbritton@ucdavis.edu

Paulo A. Zaini[1], pazaini@ucdavis.edu

Aleksey V. Zimin[3,4], alekseyz@jhu.edu

Rachael E. Workman[3], rachael.e.workman@gmail.com

Daniela Puiu[4], dpuiu@jhu.edu

Luca Bianco[5], luca.bianco@fmach.edu

Erica Adele Di Pierro[5], erica.dipierro@fmach.it

Brian J. Allen[1], brallen@ucdavis.edu

Sandeep Chakraborty[1], sanchak@gmail.com

Michela Troggio[5], michela.troggio@fmach.it

Charles A. Leslie[1], caleslie@ucdavis.edu

Winston Timp[3,4], wtimp@jhu.edu

Abhaya Dandekar[1], amdandekar@ucdavis.edu

Steven L. Salzberg[3,4,6], salzberg@jhu.edu

David B. Neale[1], 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.

Introduction

Persian walnut (*Juglans regia* L.) is among the top three most-consumed nuts in the world, and over the last ten years, its global production increased by 37% (International Nut and Dried Fruit Council, 2019). 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 [1]. 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 [2].

The origin and evolution of the Persian walnut are the results of a complex interplay between hybridization, human migration, and biogeographical forces [3]. A recent phylogenomic analysis revealed that Persian walnut (and its landrace *J. sigillata*) arose from an ancient hybridization occurred between American black walnuts and Asian butternuts after a climate-driven range expansion in Eurasia during the Pliocene [4]. Evidence suggests that the mountains of Central Asia were the cradle of domestication of Persian walnut [5], from where 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 (FAOSTAT statistics, 2017). 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 [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 illinoinensis*; 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 miss-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 ($\mathrm{F_{ST}} \geq 0.36$; 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 development, and response to stresses (**Additional file 1: Table S18**). Ten candidate selective sweeps (D_{Asia} = -0.54) were also observed in the Eastern group (**Additional file 1: Figure S14**), which included 57 predicted genes, related to terpenoid biosynthesis, post-embryonic development, and signal transduction (**Additional file 1: Table S19**).

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 [12–14]. We looked to see if 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 [12] are located in two genomic regions on chromosome 3 and 4 with significantly high values of $\mathrm{F_{ST}}$ (**Additional file 1: Table S20**). In both of these regions, Western genotypes presented lower genetic diversity and lower values of Tajima's 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 [14], 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; **Additional file 1: Table S20**). 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 (**Additional file 1: Figure S15**), suggesting differences in the timing of phenological events between these two groups as adaptation to the different climate conditions present in their countries of origin [52].

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* 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 x *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 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 [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 mis-joins, 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 F_1 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 [35]. 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; [60] 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 [61] 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 [62]. 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 [61]. Assembly quality statistics were estimated using QUAST v5.0.2 [63], filtering for contigs with a minimum length of 1Kb. 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 [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^{TM}, 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^{TM} mirVana^{TM} Total RNA Isolation Kit (Applied Biosystems^{TM}, 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^{TM} 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^{TM} 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 (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 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 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^{TM} *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^{TM} [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.

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, Zhukovskiĭ 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^{TM} *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%0Ahttp://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 (*Juglans regia* L.). Tree Genet Genomes. 2018;15:13.

12. Arab MM, Marrano A, Abdollahi-Arpanahi R, Leslie CA, Askari H, Neale DB, 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 700K SNP array. Sci Rep [Internet]. Springer US; 2019;9:6376. Available from: http://www.nature.com/articles/s41598-019-42940-1

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

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

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

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

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

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

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

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

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

22. Zhu T, Wang L, You FM, Rodriguez JC, Deal KR, Chen L, et al. Sequencing a *Juglans regia* × *J . microcarpa* hybrid yields high-quality genome assemblies of parental species. Hortic Res [Internet]. Springer US; 2019;1–16. Available from: http://dx.doi.org/10.1038/s41438-019-0139-

1

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

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

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

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

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

28. Yasodha R, Vasudeva R, Balakrishnan S, Sakthi AR, Abel N, Binai N, 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.

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

Commun. 2018;9:4844.

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

31. Zimin A V., Luo M, Marçais G, Salzberg SL, Yorke JA, Puiu D, 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 [Internet]. 2017;27:787–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28130360%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5411773

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

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

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

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

36. Springer NM, Ying K, Fu Y, Ji T, Yeh C, Jia 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.

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

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

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

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

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

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

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

bioRxiv [Internet]. 2018;307868. Available from: 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

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

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

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

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

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

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

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

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

52. 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

http://www.repeatmasker.org

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

86. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-smith C, Durbin R. BCFtools/RoH : a hidden Markov model approach for detecting autozygosity from next-generation sequencing 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

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

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

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
Number of genes	37,554	32,496	41,188
Average gene length (bp)	5,319	4,358	4,641
Single-exon transcripts	6,613	6,247	6,749

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

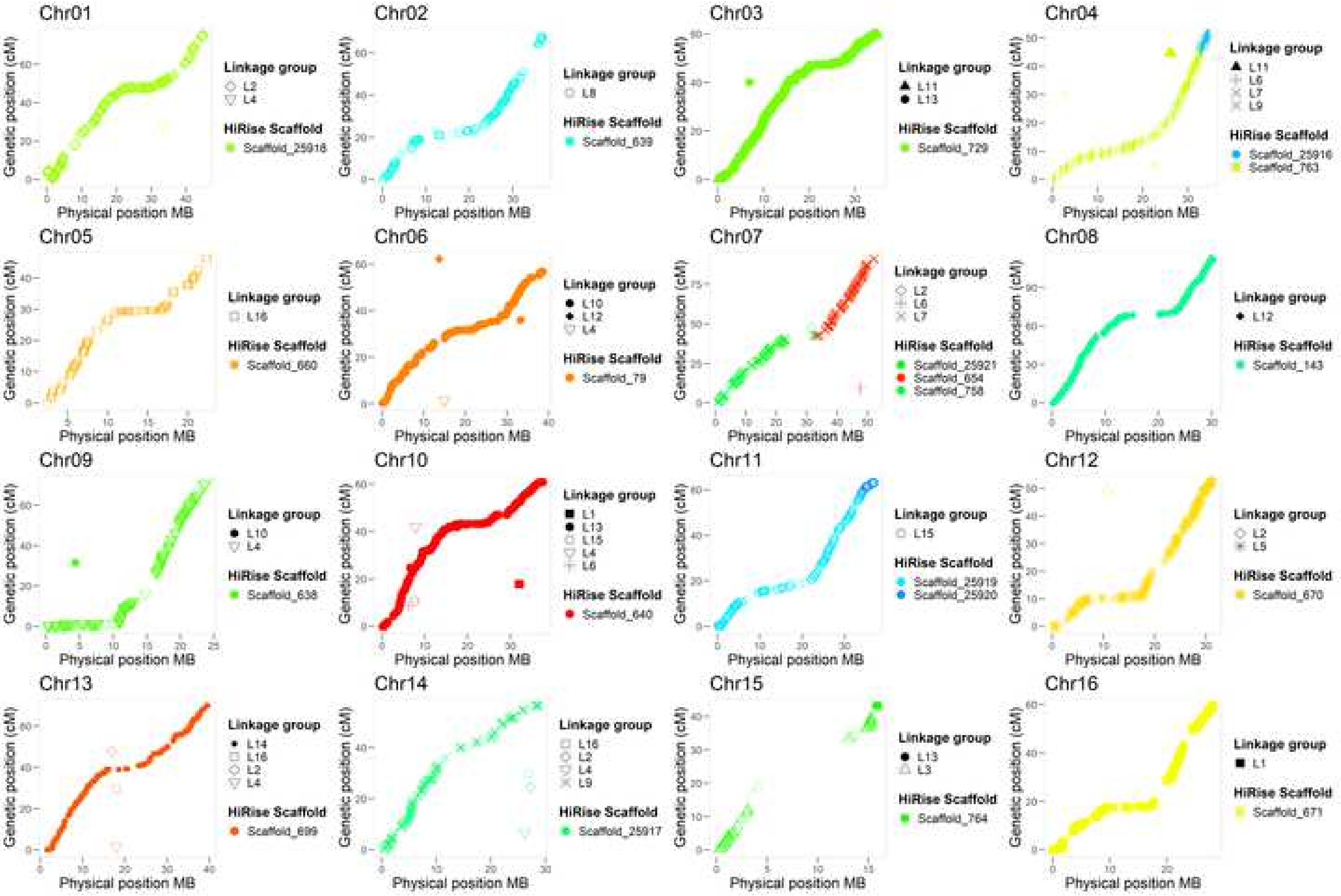
Figure 3. Clustering of the samples used in the proteomic analysis. (**A**) Hierarchical clustering based on Euclidian distances of normalized abundances of detected proteins. Samples are represented in columns and proteins in rows. (**B**) Principal component analysis of the 12 samples analyzed, clustering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

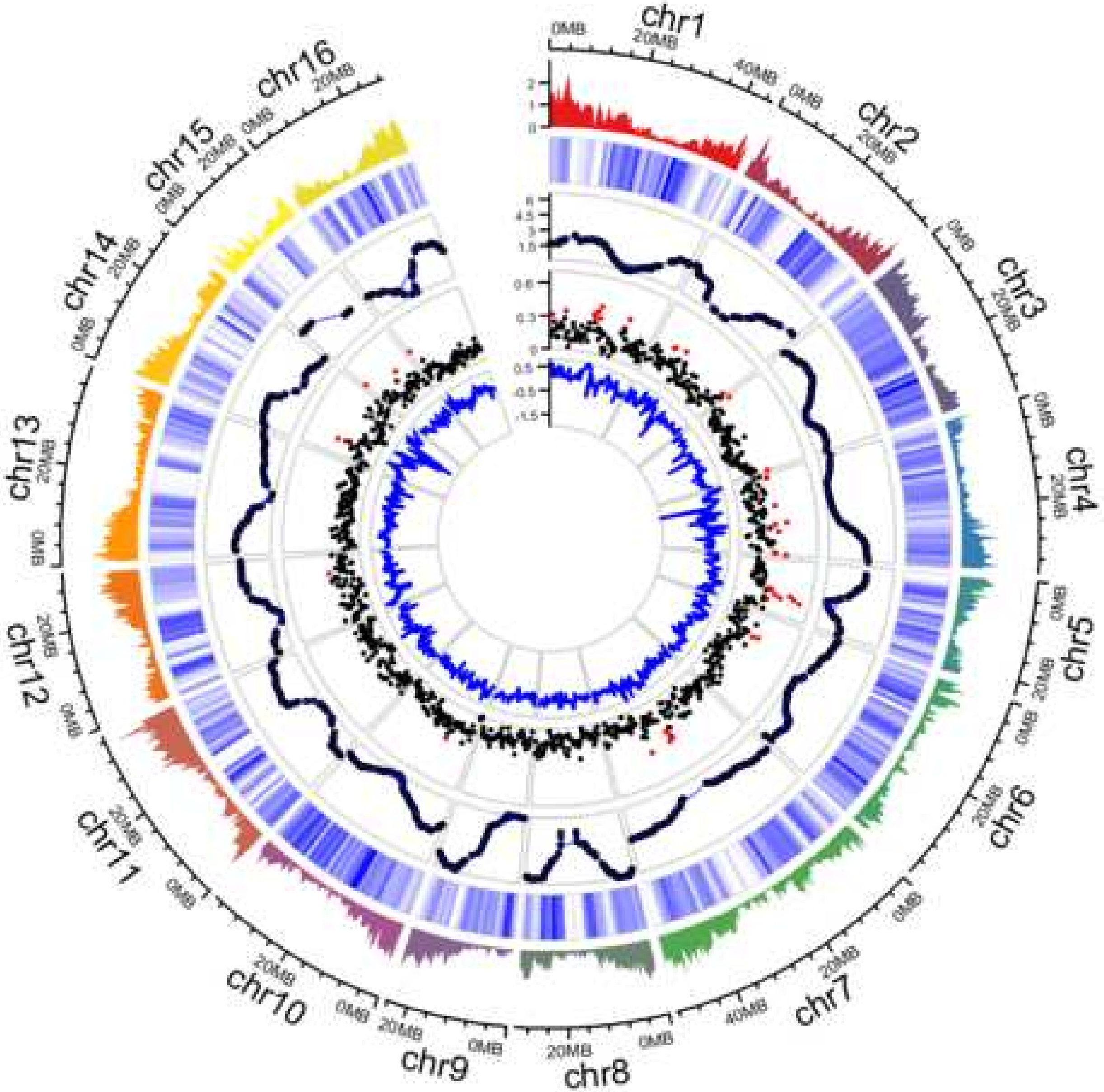
Figure 4. Modeled structure of the putative new allergen encoded by *Jr12_05180*. The compact structure is stabilized by four disulfide bonds, common in other allergenic proteins. The model in

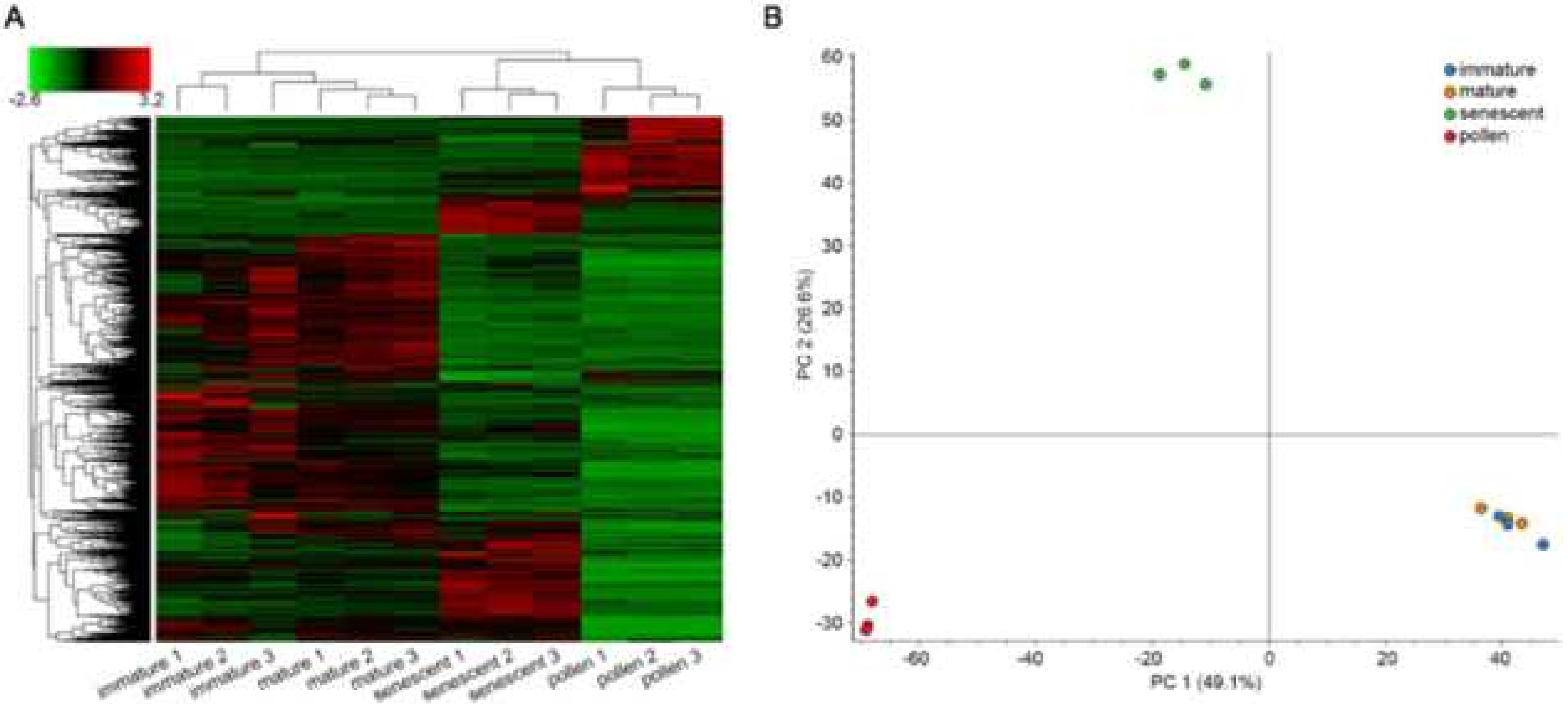
blue is superimposed with a homologous allergen from lentil (PDBid:2MAL) represented in red. Structure rendered with Pymol 2.3.

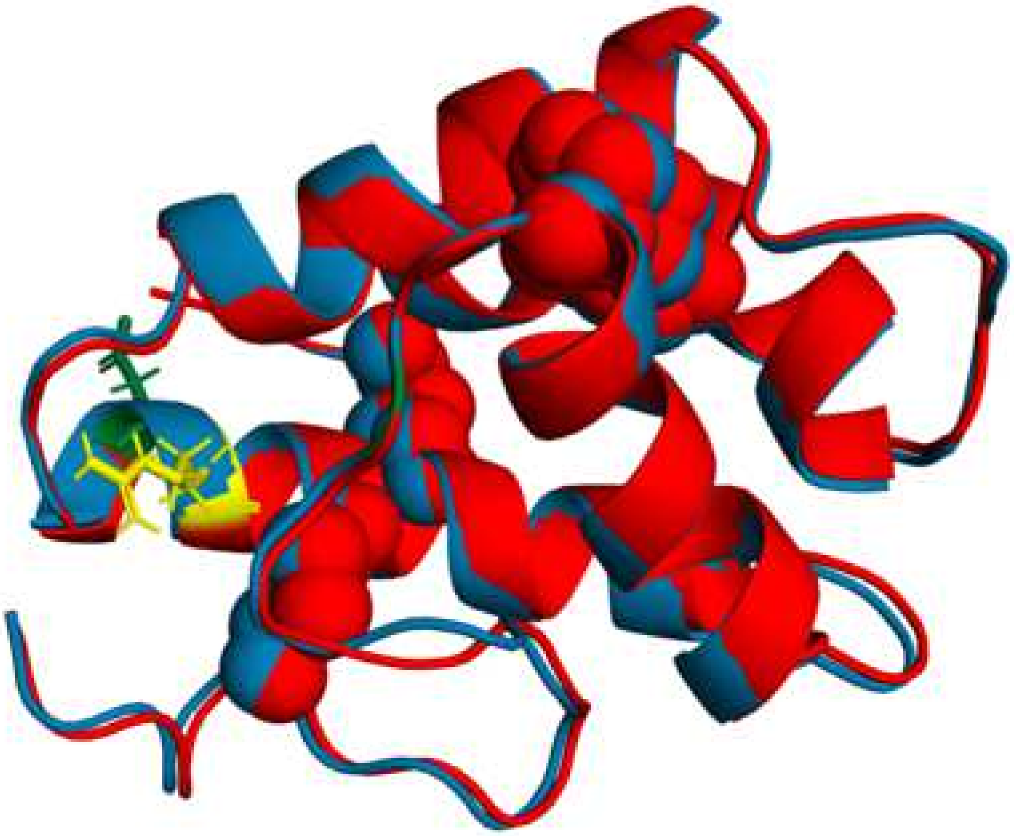
Figure 5. Graphical visualization of haplotype-blocks (HB) inheritance on Chr15 along with the Chandler pedigree. (**A**) The inner-circle highlights in grey two regions of heterozygosity (5 HB the first and 7 HB the second), and in light green two regions of homozygosity (3 HB the first and 4 HB the 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 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.

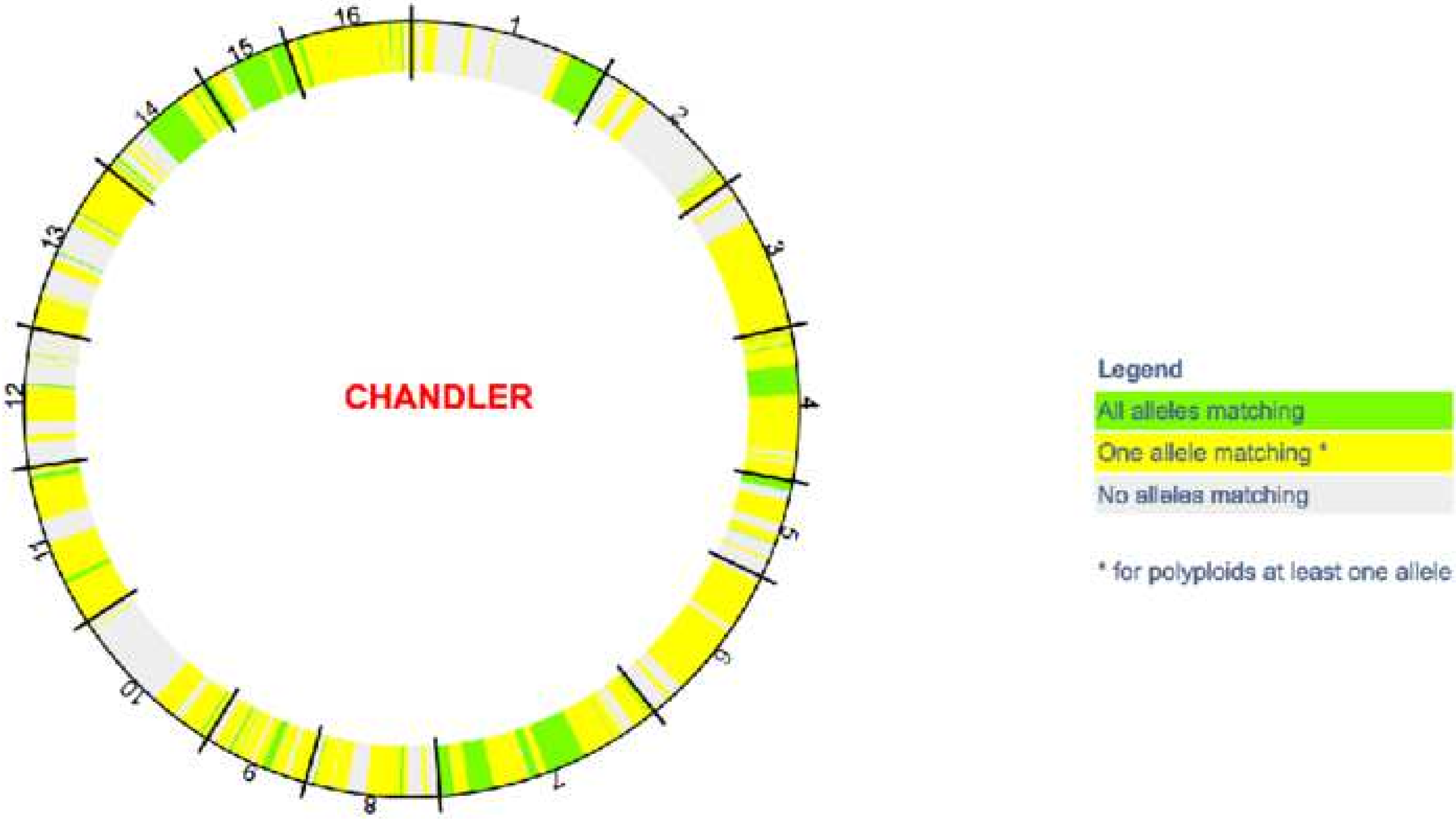
Figure 6. Graphical visualization of allele identity between Chandler and its ancestor Payne for all 16 chromosomes of Chandler.

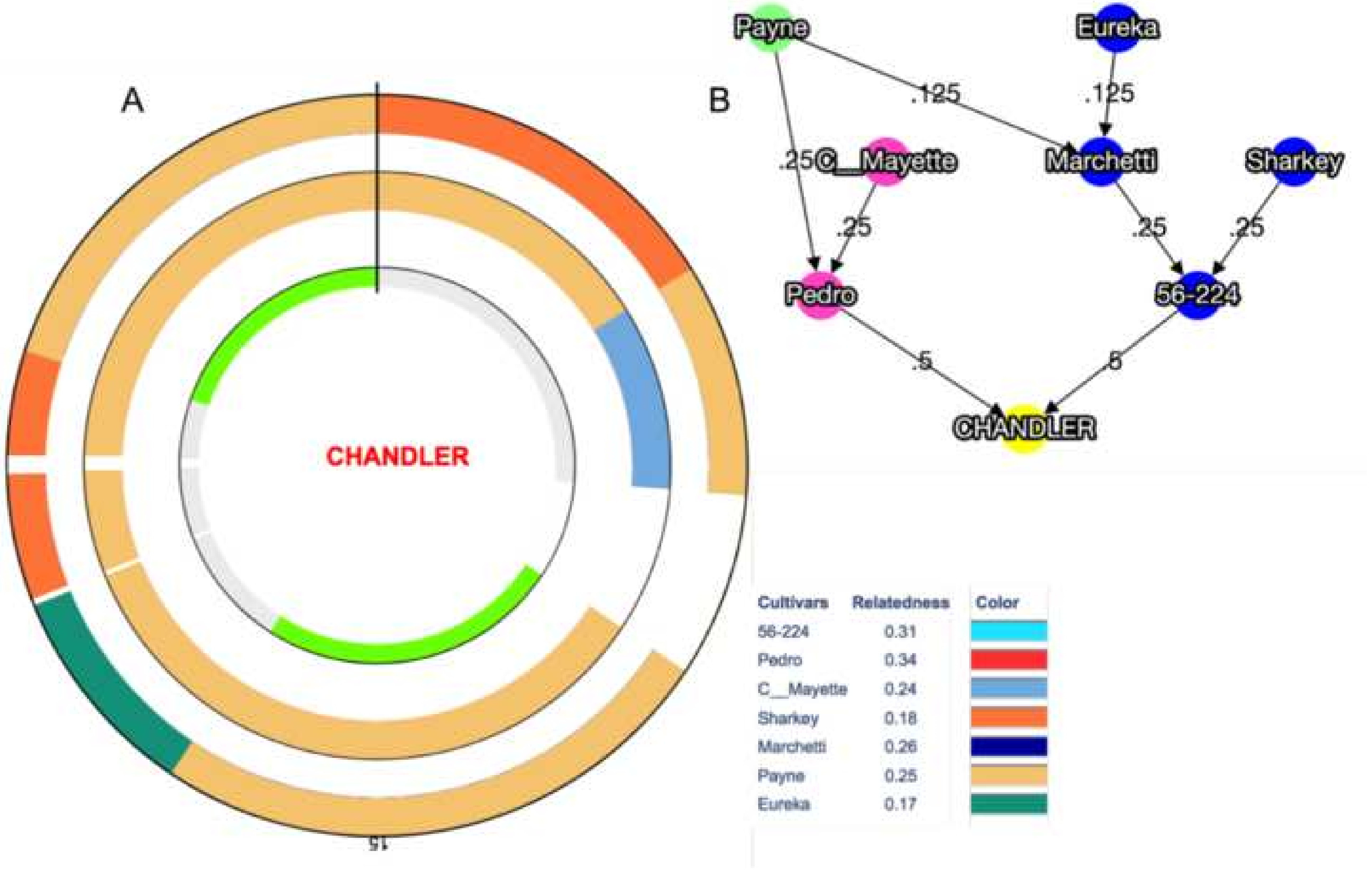

Chr01
Genetic position (cM)
Physical position MB
Linkage group
L2
L4
HiRise Scaffold
Scaffold_25916
Chr02
Linkage group
L8
HiRise Scaffold
Scaffold_639
Chr03
Linkage group
L11
L13
HiRise Scaffold
Scaffold_729
Chr04
Linkage group
L11
L6
L7
L9
HiRise Scaffold
Scaffold_25916
Scaffold_763
Chr05
Linkage group
L16
HiRise Scaffold
Scaffold_660
Chr06
Linkage group
L10
L12
L4
HiRise Scaffold
Scaffold_79
Chr07
Linkage group
L2
L6
L7
HiRise Scaffold
Scaffold_25921
Scaffold_654
Scaffold_758
Chr08
Linkage group
L12
HiRise Scaffold
Scaffold_143
Chr09
Linkage group
L10
L4
HiRise Scaffold
Scaffold_638
Chr10
Linkage group
L1
L13
L15
L4
L6
HiRise Scaffold
Scaffold_640
Chr11
Linkage group
L15
HiRise Scaffold
Scaffold_25919
Scaffold_25920
Chr12
Linkage group
L2
L5
HiRise Scaffold
Scaffold_670
Chr13
Linkage group
L14
L16
L2
L4
HiRise Scaffold
Scaffold_699
Chr14
Linkage group
L16
L2
L4
L9
HiRise Scaffold
Scaffold_25917
Chr15
Linkage group
L13
L3
HiRise Scaffold
Scaffold_764
Chr16
Linkage group
L1
HiRise Scaffold
Scaffold_671


chr1
0MB
20MB
40MB
chr2
0MB
20MB
chr3
0MB
20MB
chr4
0MB
20MB
chr5
0MB
20MB
chr6
0MB
20MB
chr7
0MB
20MB
40MB
chr8
0MB
20MB
chr9
0MB
20MB
chr10
0MB
20MB
chr11
0MB
20MB
chr12
0MB
20MB
chr13
0MB
20MB
chr14
0MB
20MB
chr15
0MB
20MB
chr16
0MB
20MB
2
1
0
6
4.5
3
1.5
0.6
0.3
0
0.5
-0.5
-1.5


A
-2.6
3.2
immature 1
immature 2
immature 3
mature 1
mature 2
mature 3
senescent 1
senescent 2
senescent 3
pollen 1
pollen 2
pollen 3
B
● immature
● mature
● senescent
● pollen
PC 2 (28.6%)
60
50
40
30
20
10
0
-10
-20
-30
-60
-40
-20
0
20
40
PC 1 (49.1%)




CHANDLER
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
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

Supplementary Material

Click here to access/download

Supplementary Material

Additional_File_2.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

Ph.D.
Dept. Plant Sciences, University of California, Davis