

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

Manuscript Number:	
Full Title:	High-quality chromosome-scale assembly of the walnut (<i>Juglans regia</i> L) reference genome
Article Type:	Data Note
Funding Information:	
Abstract:	<p>Background</p> <p>The release of the first reference genome of walnut (<i>Juglans regia</i> L.) enabled many achievements in the characterization of walnut genetic and functional variation. However, it is highly fragmented, preventing the integration of genetic, transcriptomic, and proteomic information to fully elucidate walnut biological processes. Findings</p> <p>Here, we report the new chromosome-scale assembly of the walnut reference genome (Chandler v2.0) obtained by combining Oxford Nanopore long-read sequencing with chromosome conformation capture (Hi-C) technology. Relative to the previous reference genome, the new assembly features an 84.4-fold increase in N50 size and the full sequence of all 16 chromosomal pseudomolecules. Using full-length transcripts from single-molecule real-time sequencing, we predicted 40,491 gene models, with a mean gene length higher than the previous gene annotations. Most of the new protein-coding genes (90%) are full-length, which represents a significant improvement compared to Chandler v1.0 (only 48%). We then tested the potential impact of the new chromosome-level genome on different areas of walnut research. By studying the proteome changes occurring during catkin development, we observed that the virtual proteome obtained from Chandler v2.0 presents fewer artifacts than the previous reference genome, enabling the identification of a new potential pollen allergen in walnut. Also, the new chromosome-scale genome facilitates in-depth studies of intraspecies genetic diversity by revealing previously undetected autozygous regions in Chandler, likely resulting from inbreeding, and 195 genomic regions highly differentiated between Western and Eastern walnut cultivars. Conclusion</p> <p>Overall, Chandler v2.0 will serve as a valuable resource to understand and explore walnut biology better.</p>
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 Troglio
	Charles A. Leslie
	Winston Timp
	Abhaya Dandekar
	Steven L. Salzberg
	David B. Neale
Order of Authors Secondary Information:	
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
<p>Experimental design and statistics</p> <p>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.</p> <p>Have you included all the information requested in your manuscript?</p>	Yes
<p>Resources</p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	Yes

<p>Availability of data and materials</p> <p>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.</p> <p>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>
---	------------



[Click here to view linked References](#)

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

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

Monica Britton², mtbritton@ucdavis.edu

Paulo A. Zaini¹, pazaini@ucdavis.edu

Aleksey V. Zimin³, alekseyz@jhu.edu

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

Daniela Puiu⁴, dpuiu@jhu.edu

Luca Bianco⁵, luca.bianco@fmach.edu

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

Brian J. Allen¹, brallen@ucdavis.edu

Sandeep Chakraborty¹, sanchak@gmail.com

Michela Troggio⁵, michela.troggio@fmach.edu

Charles A. Leslie¹, caleslie@ucdavis.edu

Winston Timp³, wtimp@jhu.edu

Abhaya Dandekar¹, amdandekar@ucdavis.edu

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

David B. Neale¹, dbneale@ucdavis.edu

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

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

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

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

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

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

30 **Abstract**

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

51

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

53

54 **Introduction**

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

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

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

75 The release of the first reference genome, Chandler v1.0 [9], enabled the study of walnut genetics
76 at a genome-wide scale. For the first time, it was possible to explore the gene space of Persian
77 walnut with the prediction of 32,498 gene models, providing the basis to untangle complex
78 phenotypic pathways, such as those responsible for the synthesis of phenolic compounds. The
79 availability of a reference genome marked the beginning of a genomics phase in Persian walnut,
80 allowing whole-genome resequencing [4,10], the development of high-density genotyping tools
81 [7,11], and the genetic dissection of important agronomical traits in walnut [12–14]. However, the
82 Chandler v1.0 assembly is highly fragmented, compromising the accuracy of gene prediction and
83 the fulfillment of advanced genomics studies necessary to resolve many, still unanswered
84 questions in walnut research.

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

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

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

108 **Genome long-read sequencing and assembly**

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

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

121 combines long, high-error reads with shorter but much more accurate Illumina sequencing reads
122 to generate a robust, highly contiguous genome assembly [30]. First, using the Illumina reads, we
123 created 3.7 million 'super-reads' with a total length of 2.9 Gb. We then combined the super-reads
124 with the ONT reads to generate 3.2 million mega-reads with a mean length of 4.7 kb, representing
125 24X genome coverage (**Additional file 1**). Finally, we assembled the mega-reads to obtain the
126 'hybrid' Illumina-ONT assembly, which comprised 1,498 scaffolds, 258 contigs, and 25,007 old
127 scaffolds from Chandler v1.0 (**Table 1; Additional file 2**).

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

138 **Validation of the HiRise assembly**

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

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

146 Second, we anchored the HiRise assembly to the Chandler genetic map used by [31] to construct
147 a walnut physical map. In total, 972 of the mapped markers (1,525 SNPs) aligned uniquely on the
148 same 35 HiRise scaffolds anchored to the linkage map mentioned above. Overall, we observed
149 almost perfect collinearity between the HiRise assembly and both Chandler genetic maps (**Figure**
150 **1, Additional file 5**). Therefore, we oriented, ordered, and named the HiRise scaffolds consistent
151 with the linkage map of [31], generating the final 16 chromosomal pseudomolecules of *J. regia*
152 Chandler.

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

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

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

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

174 **Repeat annotation**

175 Almost half (49.68%) of the new Chandler v2.0 is repetitive, similarly to the previous version of
176 the walnut reference genome (51.19%). As in most plant genomes, interspersed repeats (45.13%)
177 were the most abundant type of repeats, with retrotransposons at 18.55% and DNA transposons at
178 3.05%. *Gypsies* (6.58%) and *Copias* (4.1%) were the most represented classes of long-terminal
179 retrotransposons (LTR), and, though widely dispersed throughout the genome, they were
180 distributed differently along the 16 chromosomes (**Additional file 6**): the *Gypsies* LTRs were more
181 abundant alongside the centromeres, where, instead, the density of the *Copia* LTRs decreased,
182 consistent with [21]. L1/LINE (long-interspersed nuclear elements), which possess a poly(A) tail
183 and two open reading frames (ORFs) for autonomous retrotransposition, was the largest class of
184 non-LTRs at 6.98% of the genome. Simple repeats (4.29%) and low-complexity regions (0.26%)
185 were also found.

186 **PacBio IsoSeq sequencing and gene annotation**

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

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

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

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

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

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

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

248 Of the 41,103 gene models, 83% were annotated with a plant protein, and 84% had a known Pfam
249 domain. Also, 33,034 models were annotated with 8,244 different Gene Ontology (GO) terms. The
250 three most common biological processes were regulation of transcription (2%), defense response
251 (1.43%), and DNA recombination (1.2%; **Additional file 12**). ATP binding (7.7%), metal ion
252 binding (7.1%) and DNA-binding transcription factor activity (3.7%; **Additional file 13**) were the
253 most abundant molecular functions, while nucleus (13%), integral component of membrane
254 (10.3%) and plasma membrane (8%) were the top three cellular components (**Additional file 14**).

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

260 **Improved assessment of proteomes with the complete genome sequence**

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

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

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

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

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

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

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

310 **Chandler genomic diversity**

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

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

336 Due to the evidence of whole-genome duplication in *Juglans* genomes [31], we searched for
337 conserved regions of synteny between Chr15 and its homologous regions in the genome, to study
338 their level of divergence and identify other evolutionary forces as possible causes of the localized
339 reduction of heterozygosity on Chr15. Of the 5,739 pairs of paralogous genes (8,701 genes;
340 **Additional file 22**) identified in Chandler v2.0, 448 included genes on Chr15, and 389 of these
341 have their respective paralogues on Chr6 (**Additional file 23**), in line with what was already
342 reported by [31]. The Chr06-Chr15 pairs of paralogous genes showed average values of divergence
343 indexes ($K_S = 0.38$; $K_A = 0.13$) similar to the ones observed genome-wide for other syntelogs (K_S
344 $= 0.4$; $K_A = 0.09$). Similar values of divergence were also observed for the 178 Chr06-Chr15
345 syntelogs (171 genes) falling within the nine low heterozygous regions on Chr15 ($K_S = 0.4$, $K_A =$

346 0.1), excluding different evolutionary rates or positive selection for these regions, and leaving
347 inbreeding as the most reasonable explanation. Other than paralogous genes, we found 393
348 singletons genes in the low heterozygous regions on Chr15 of Chandler. These genes are involved
349 in different biological processes, many of which related to signal transduction, protein
350 phosphorylation, and response to environmental stimuli (**Additional file 24**).

351 We further investigated the contribution of inbreeding to the high level of autozygosity on Chr 15
352 by visualizing the inheritance of haplotype-blocks (HB; genomic regions with little recombination)
353 across the Chandler pedigree (**Figure 5B, Additional file 25**). We observed that Payne accounts
354 for the entire Chandler genetic makeup (19 HBs for the total length of Chr15) inherited from Pedro
355 (mother), where only one HB (2,08 Mb) shared the same allele of Conway-Mayette (maternal-
356 grandfather; **Figure 5A**). Regarding the paternal genetic makeup of Chandler, 13 out of 19 HBs
357 (9,05 Mb) on Chr15 inherited Payne alleles, providing further evidence of high inbreeding on this
358 chromosome (**Figure 5A**). This is even more evident in assessing the number of alleles matching
359 between Payne and Chandler across the genome: Chr15 (14 HBs for a total of 13,95 Mb; **Figure**
360 **6**) shares full allele identity with Payne for almost its entire length. Such allele matching between
361 Chandler and its ancestor Payne also occurs on Chr1 (9 HBs for a total of 8,44 Mb), Chr4 (6 HBs
362 - 7,68 Mb), Chr7 (21 HBs - 21,62Mb) and Chr14 (7 HBs – 12,29 Mb). These results confirm a
363 high level of inbreeding in many genomic regions of Chandler (**Additional file 25**) and support
364 the crucial role of Chandler v2.0 for understanding trait genetic inheritance in walnut.

365 **Genomic comparison between Eastern and Western walnuts**

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

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

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

383 At a genomic level, we found a moderate differentiation ($F_{ST} = 0.15$) between Western and Eastern
384 genotypes, except for 195 genomic windows (100 kb) that showed substantially high population
385 differences ($F_{ST} \geq 0.36$; top 5% in the whole genome). In particular, chromosomes 7, 5, 1, 4, and
386 2 presented about 70% of the divergent sites (**Figure 2; Additional file 28**). As suggested by the
387 mean reduction of diversity coefficient (ROD) value (0.41), in most of the genomic regions highly
388 differentiated, the UCD-WIP founders from the USA and Europe showed lower nucleotide
389 diversity ($\pi = 2.5 \times 10^{-4}$) than the Asian genotypes ($\pi = 5.0 \times 10^{-4}$), consistent with [8] (**Figure 2;**
390 **Additional file 28**). The proximity of our eastern genotypes to the supposed walnut center of
391 domestication in Central Asia can explain the high level of diversity observed in this subgroup.

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

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

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

421 **Methods**

422 **Oxford Nanopore sequencing and assembly**

423 High molecular weight (HMW) DNA for Nanopore sequencing (Oxford Nanopore Technologies
424 Inc., UK) was isolated through a nuclei extraction and lysis protocol. First, mature leaf tissue from
425 the same tree used for the original *J. regia* genome [9] was homogenized with mortar and pestle
426 in liquid nitrogen until well ground, then added to the Nuclei Isolation Buffer [46], and stirred at
427 4°C for 10 minutes. The cellular homogenate was filtered through 5 layers of Miracloth (Millipore-
428 Sigma) into a 50 mL Falcon tube, then centrifuged at 4°C for 20 minutes at 3000 x g. This speed
429 of centrifugation was selected based on the estimated walnut genome size of 1 Gb [47]. Extracted
430 nuclei were then lysed for 30 minutes at 65°C in the SDS-based lysis buffer described by [48].
431 Potassium acetate was added to the lysate to precipitate residual polysaccharides and proteins. The
432 sample was incubated for 5 minutes at 4°C and then centrifuged at 4°C for 10 minutes at 2400 x
433 g. After removing the supernatant, genomic DNA (gDNA) was ethanol precipitated, and then
434 eluted in 10 mM Tris-Cl. Further purification of the gDNA was then performed using a Zymo
435 Genomic DNA Clean and Concentrate column.

436 One μg of the isolated gDNA was prepared for sequencing using the Ligation sequencing kit
437 (LSK108, Oxford Nanopore) following manufacturer's protocol with an optimized end repair (100
438 μl sample, 14 μl enzyme, 6 μl enzyme, incubated at 20°C for 20 minutes then 65°C for 20 minutes).
439 Libraries were sequenced for 48 hours on the Oxford Nanopore Mk1B MinION platform with the
440 R9.4 chemistry on eight flowcells. Raw fast5 data was base-called using Albacore version 1.25.
441 The ONT data and Illumina reads from [9] were combined using the assembly algorithm
442 implemented in MaSuRCA v3.2.2 [49]. Super-reads were constructed using a k-mer size of 41 bp.
443 De-duplicated scaffolds were aligned onto the previously finished *J. regia* chloroplast genome [9]
444 using "minimap2 -x asm5", as well as to a database of 223 finished plant mitochondria
445 (downloaded from NCBI RefSeq) using blastn with default parameters.

446 **Hi-C sequencing**

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

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

465 **Validation and anchoring of the HiRise assembly to Chandler genetic maps**

466 The HiRise assembly was first anchored to the Chandler genetic map obtained by [14] from a 312
467 offspring F₁ population ‘Chandler x Idaho’ genotyped with the latest Axiom *J. regia* 700K SNP
468 array. SNP probes (71-mers including the SNP site) from the Axiom *J. regia* 700K SNP array
469 were aligned onto the HiRise assembly filtering out alignments with probe/reference identity lower
470 than 98%, covering less than 95% of the probe length or aligning multiple times on the genome.
471 Retained markers with a unique segregation profile were then used to anchor the HiRise scaffolds.
472 The same procedure was also followed to anchor the HiRise assembly to the Chandler genetic map
473 used to construct a walnut bacterial artificial chromosome (BAC) clone-based physical map by
474 [31]. The final ordering of scaffolds was performed by taking into consideration the marker genetic
475 map position, and, in the final sequence, consecutive scaffolds were separated by sequences of
476 100,000 Ns.

477 The tandem repeat finder program (trf v4.09; [52]) was run using the recommended parameters
478 (max mismatch delta PM PI minscore maxperiod, 2 7 7 80 10 50 500 resp.) to identify repeat
479 elements up to 500 bp long. A histogram of repeat unit lengths was generated, and peaks at 7, 29,
480 33, 44, 154, and 308 bp were identified. From this data, a consensus sequence corresponding to

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

486 Recombination rate was estimated within sliding windows of 10 Mb with a step of 1 Mb along the
487 chromosome sequence by using the high-density genetic map of Chandler [14] and the
488 R/MareyMap package v 1.3.4 [54]. To evaluate Chandler v2.0 error rate, the two assemblies,
489 Chandler v1.0 and 2.0, were aligned to each other using the ‘nucmer’ program (Marçais et al.,
490 2018).

491 **RNA preparation**

492 Five walnut tissues (leaf, catkin 1-inch elongated; catkin 3-inches elongated, pistillate flower, and
493 pollen) were collected from ‘Chandler’ trees at the UCD walnut orchards. Four additional samples
494 (somatic embryo, callus, shoot, and roots) were taken from tissue culture material of ‘Chandler’.
495 Several grams of each tissue were ground in liquid nitrogen and with insoluble
496 polyvinylpyrrolidone (PVPP; 1% w/w). RNA was isolated using the PureLink™ Plant RNA
497 Reagent (Invitrogen™, Carlsbad, CA) following the manufacturer’s instructions, but with an
498 additional end wash in 1 mL of 75% Ethanol. For root tissue only, RNA isolation was performed
499 using the MagMAX™ mirVana™ Total RNA Isolation Kit (Applied Biosystems™, Foster City,
500 CA) as per protocol, except for the lysis step. A different lysis buffer was created adding 100 mg
501 of sodium metabisulfite to 10 mL of guanidine buffer (guanidine thiocyanate 4M, sodium acetate
502 0.2M, EDTA 25 mM, PVP-40 2.5%, pH 5.0) and 1 mL of nuclease-free water. Then, 100 mg of
503 ground root tissue were lysed in 1 mL of the new lysis buffer using a Tissue Lyser at max frequency

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

509 **PacBio IsoSeq sequencing**

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

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

526 **Repeat annotation**

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

530 **Gene prediction and functional annotation**

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

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

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

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

551 **Label-free shotgun proteomics**

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

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

569 nanospray ionization source. The peptides were loaded onto a Trap (100 microns, C18 100 Å 5U)
570 and desalted online before separation using a reversed-phase (75 microns, C18 200 Å 3U) column.
571 The duration of the peptide separation gradient was 60 min using 0.1% formic acid and 100%
572 acetonitrile (ACN) for solvents A and B, respectively. The data were acquired using a data-
573 dependent MS/MS method, which had a full scan range of 300-1,600 Da and a resolution of
574 70,000. The resolution of the MS/MS method was 17,500 and the insulation width 2 m/z with a
575 normalized collision energy of 27. The nanospray source was operated using a spray voltage of
576 2.2 KV and a transfer capillary temperature heated to 250°C. Samples were analyzed at the UC
577 Davis Proteome Core.

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

590 **Chandler genomic diversity**

591 Illumina whole-genome shotgun data of Chandler were aligned on the Chandler v2.0 with BWA
592 [70] with standard parameters. SNP calling was performed using SAMtools v1.9 [71] and
593 BCFtools v.2.1 [72]. SNP density for windows of 1 Mb was estimated using the command
594 ‘SNPdensity’ implemented in VCFtools v0.1.16 [73]. Self-collinearity analysis to detect
595 duplicated regions in Chandler v2.0 was performed with MCScanX [74], using a simplified GFF
596 file of the new gene annotation and a self-BLASTP as input. To improve the power of collinearity
597 detection, tandem duplications were excluded after running the function
598 ‘detect_collinear_tandem_arrays’ implemented in MCScanX. Synonymous (K_S) and
599 nonsynonymous (K_A) changes for syntenic protein-coding gene pairs were measured using the Perl
600 script “add_ka_and_ks_to_collinearity.pl” implemented in MCScanX.

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

611 **Genomic comparison between Eastern and Western walnuts**

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

614 above for Chandler. SNPs with no missing data and minor allele frequency (MAF) higher than
615 10% were retained for the following genetic analyses (7,269,224 SNPs out of the 14,988,422
616 identified). Hierarchical cluster analysis on a dissimilarity matrix of the 23 UCD-WIP founders
617 was performed using R/SNPRelate v.1.18.0 [78]. Fixation index (F_{ST}) was measured between
618 genotypes from EU/USA and Asia with VCFtools v0.1.16, setting windows of 100kb and 500kb.
619 Genomic windows with the top 5% of F_{ST} values were selected as candidate regions for further
620 analysis. The empirical cutoff with a low false discovery rate (5%) was verified by performing
621 whole-genome permutation test (1000) with a custom Python script. Nucleotide diversity (π) and
622 Tajima's D [79] were also computed along the whole genome in 100-kb and 500-kb windows
623 using VCFtools. Reduction of diversity coefficient (ROD) was estimated as $1 - (\pi_{Occ} / \pi_{Asia})$. The
624 new walnut gene annotation v.2.0 was used to identify predicted genes in the candidate regions
625 under selection. The distribution of the identified genes into different biological processes was
626 evaluated using the weight01 method provided by the R/topGO [80]. The Kolmogorov–Smirnov-
627 like test was performed to assess the significance of over-representation of GO categories
628 compared with all genes in the walnut gene prediction. Plots were obtained using the R/circlize
629 v.0.4.6 and R/ggplot2 v.3.5.3 packages.

630

631 **Availability of supporting data**

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

636 **Additional files**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

696 500-kb windows for genotypes from Europe a- USA; (iv) Tajima's D in 500-kb windows for
697 genotypes from Asia.

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

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

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

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

706

707

708 **Competing interests**

709 The authors declare no conflict of interest.

710 **Funding:**

711 This project has been funded by the Californian Walnut Board.

712 **Author's contribution**

713 DBN and AM conceived and coordinated the research. REW and WT performed the HMW DNA
714 extraction and Nanopore sequencing. AVZ, DP and SLS assembled the hybrid Illumina-ONT
715 assembly. LB, MT, DP and SLS validated and anchored the HiRise assembly to the genetic maps.

716 AM and BJA collected and extracted all RNA samples. MB analyzed the PacBio IsoSeq results

717 and performed the repeat and gene annotation. AD conceived the design of the proteomic analyses;
718 PAZ and SC generated and analyzed the proteomic data. LB called the SNPs in Chandler and the
719 23 UCD WIP founders, while AM carried out the analyses on walnut genomic diversity. EAD, LB
720 and MT built and analyzed the SNP haplotypes. CAL provided all the plant material. AM wrote
721 the manuscript, which has been revised by all coauthors.

722 **Acknowledgments**

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

725

726 **References**

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

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

739 6. Ebrahimi A, Zarei A, Lawson S, Woeste KE, Smulders MJM. Genetic diversity and genetic
740 structure of Persian walnut (*Juglans regia*) accessions from 14 European, African, and Asian
741 countries using SSR markers. *Tree Genet Genomes* [Internet]. *Tree Genetics & Genomes*;
742 2016;12:114. Available from: <http://link.springer.com/10.1007/s11295-016-1075-y>

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

747 8. Bernard A, Barreneche T, Lheureux F, Dirlewanger E. Analysis of genetic diversity and
748 structure in a worldwide walnut (*Juglans regia* L.) germplasm using SSR markers. *PLoS One*.
749 2018;13:1–19.

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

753 10. Stevens KA, Woeste K, Chakraborty S, Crepeau MW, Leslie CA, Martínez-García PJ, et al.
754 Genomic Variation Among and Within Six *Juglans* Species. *G3 Genes|Genomes|Genetics*
755 [Internet]. 2018;8:1–37. Available from:
756 <http://www.ncbi.nlm.nih.gov/pubmed/29792315>
[http://g3journal.org/lookup/doi/10.1534/g3.
757 118.200030](http://g3journal.org/lookup/doi/10.1534/g3.118.200030)

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

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

761 12. Arab MM, Marrano A, Abdollahi-Arpanahi R, Leslie CA, Askari H, Neale DB, et al.
762 Genome-wide patterns of population structure and association mapping of nut-related traits in
763 Persian walnut populations from Iran using the Axiom *J. regia* 700K SNP array. *Sci Rep*
764 [Internet]. Springer US; 2019;9:6376. Available from: [http://www.nature.com/articles/s41598-](http://www.nature.com/articles/s41598-019-42940-1)
765 019-42940-1

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

770 14. Marrano A, Sideli GM, Leslie CA, Cheng H, Neale DB. Deciphering of the genetic control
771 of phenology, yield and pellicle color in Persian walnut (*Juglans regia* L.). *Front Plant Sci*.
772 2019;10:1–14.

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

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

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

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

804 2017;68:5419–29.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- 889 52. Benson G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids*
890 *Res.* 1999;27:573–80.
- 891 53. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast
892 and versatile genome alignment system. *PLoS Comput Biol.* 2018;14:1–14.
- 893 54. Rezvoy C, Charif D, Guéguen L, Marais GAB. MareyMap: An R-based tool with graphical
894 interface for estimating recombination rates. *Bioinformatics.* 2007;23:2188–9.
- 895 55. Li H. Minimap2 : pairwise alignment for nucleotide sequences. *Bioinformatics.*
896 2018;34:3094–100.
- 897 56. Smit A, Hubley R. RepeatModeler Open-1.0. [Internet]. 2008. Available from:
898 <http://www.repeatmasker.org>
- 899 57. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. [Internet]. 2013. Available from:
900 <http://www.repeatmasker.org>
- 901 58. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, et al. Improving
902 the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic*
903 *Acids Res.* 2003;31:5654–66.
- 904 59. Kent WJ. BLAT — The BLAST -Like Alignment Tool. *Genome Res.* 2002;12:656–64.
- 905 60. Wu TD, Watanabe CK. GMAP : a genomic mapping and alignment program for mRNA and
906 EST sequences. *Bioinformatics.* 2005;21:1859–75.
- 907 61. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo
908 transcript sequence reconstruction from RNA-Seq: reference generation and analysis with
909 Trinity. *Nat Protoc.* 2013;8:1–43.

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

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

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

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

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

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

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

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

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

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

- 931 72. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-smith C, Durbin R. BCFtools/RoH : a
932 hidden Markov model approach for detecting autozygosity from next-generation sequencing
933 data. *Bioinformatics*. 2016;32:1749–51.
- 934 73. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call
935 format and VCFtools. *Bioinformatics*. 2011;27:2156–8.
- 936 74. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCSScanX: A toolkit for detection
937 and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 2012;40:1–14.
- 938 75. Bink MCAM, Jansen J, Madduri M, Voorrips RE, Durel CE, Kouassi AB, et al. Bayesian
939 QTL analyses using pedigreed families of an outcrossing species, with application to fruit
940 firmness in apple. *Theor Appl Genet*. 2014;127:1073–90.
- 941 76. Voorrips RE, Bink MCAM, Kruisselbrink JW, Koehorst-van Putten HJJ, van de Weg WE.
942 *PediHaplotyper*: software for consistent assignment of marker haplotypes in pedigrees. *Mol*
943 *Breed*. Springer Netherlands; 2016;36.
- 944 77. Vanderzande S, Howard NP, Cai L, Da Silva Linge C, Antanaviciute L, Bink MCAM, et al.
945 High-quality, genome-wide SNP genotypic data for pedigreed germplasm of the diploid
946 outbreeding species apple, peach, and sweet cherry through a common workflow. *PLoS One*.
947 2019;
- 948 78. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance
949 computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*.
950 2012;28:3326–8.
- 951 79. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA

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

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

954 **Tables**

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

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

957

958

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

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

961

962

963 **Figure legends**

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

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

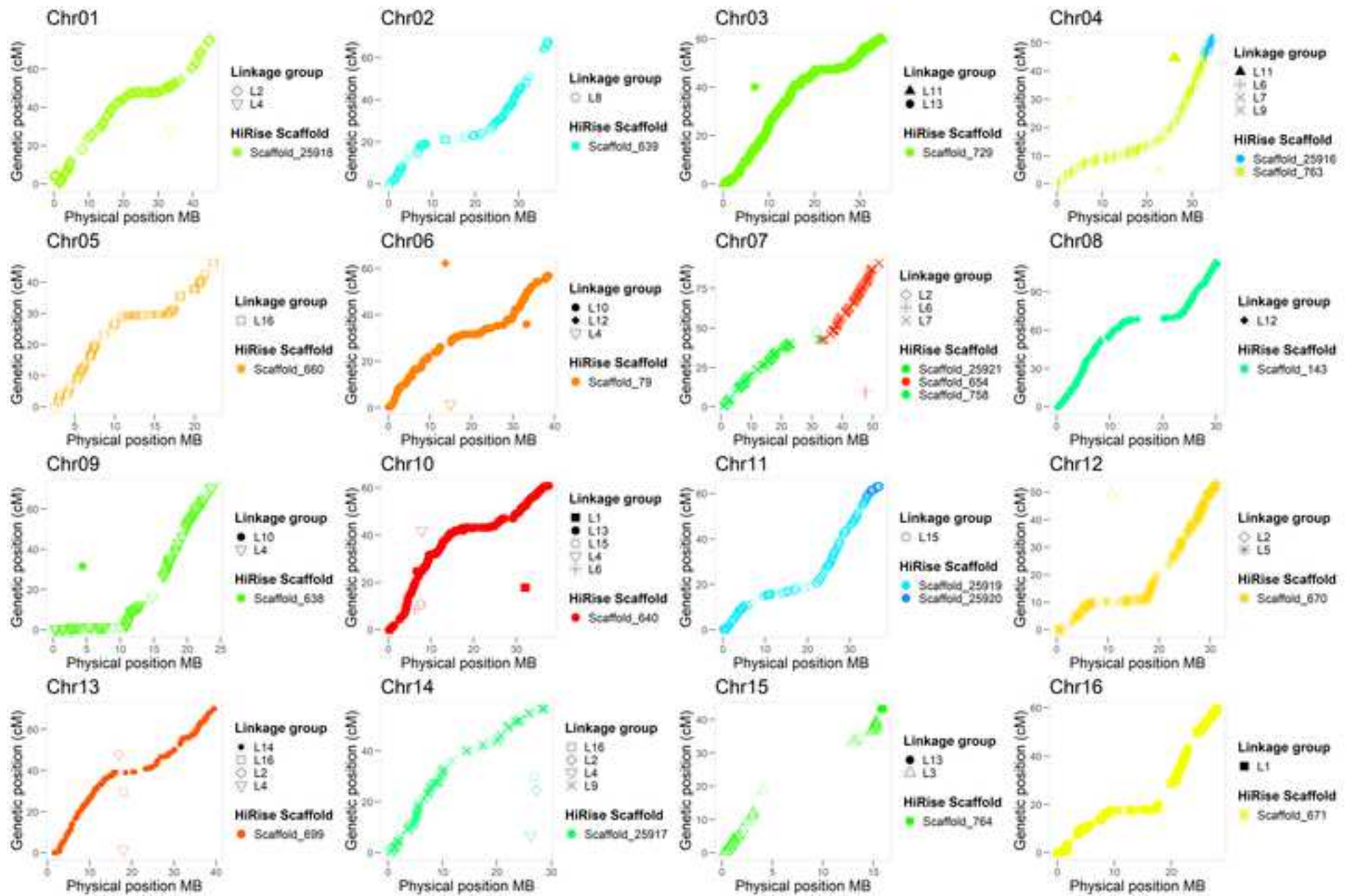
972 **Figure 3.** Clustering of the samples used in the proteomic analysis. **(A)** Hierarchical clustering
973 based on Euclidian distances of normalized abundances of detected proteins. Samples are
974 represented in columns and proteins in rows. **(B)** Principal component analysis of the 12 samples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

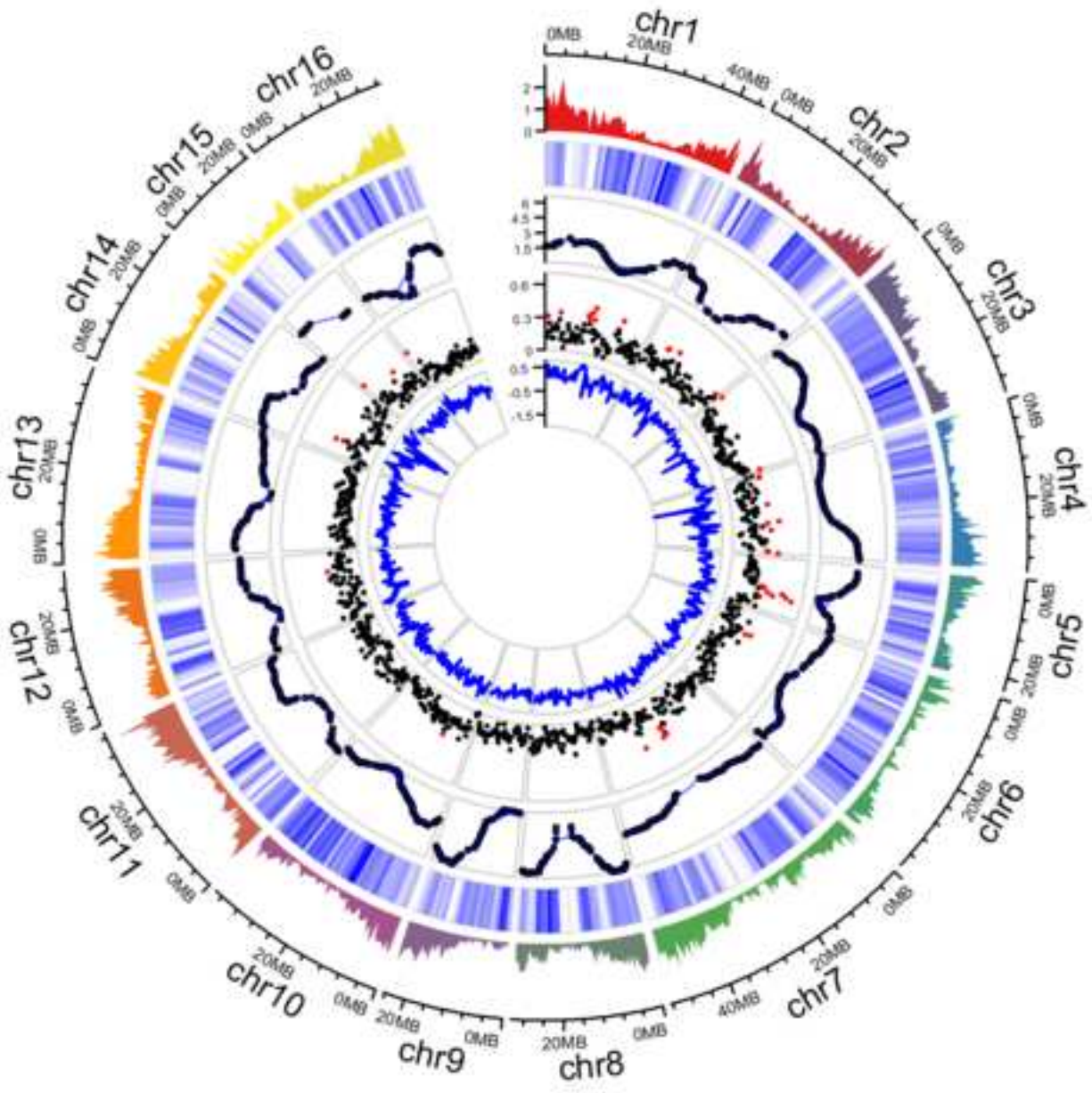
976 **Figure 4.** Modeled structure of the putative new allergen encoded by *Jr12_05180*. The compact
977 structure is stabilized by four disulfide bonds, common in other allergenic proteins. The model in
978 blue is superimposed with a homologous allergen from lentil (PDBid:2MAL) represented in red.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

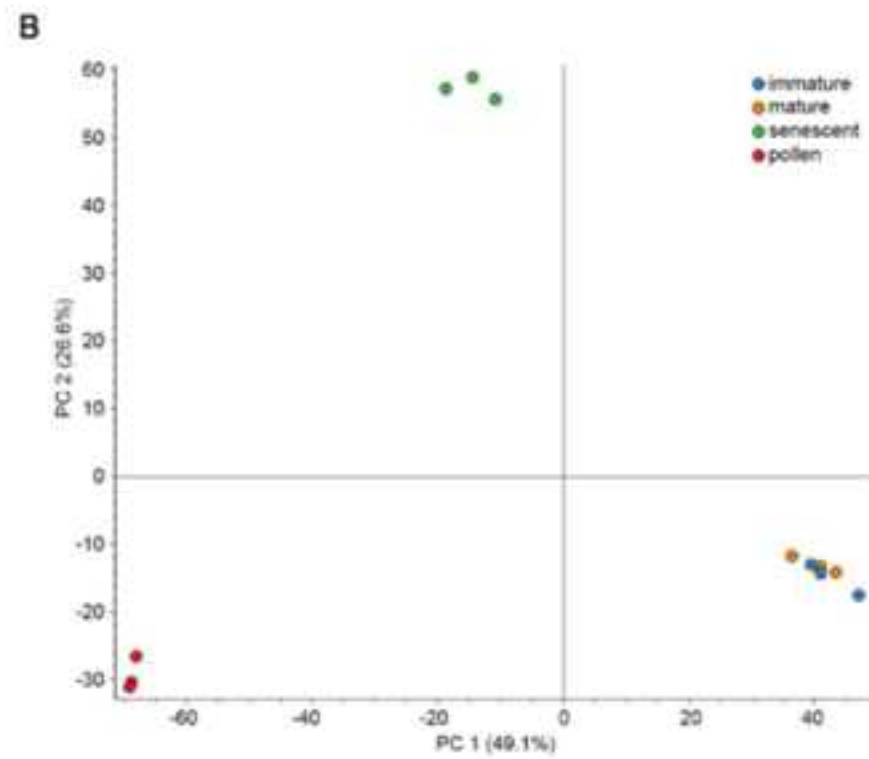
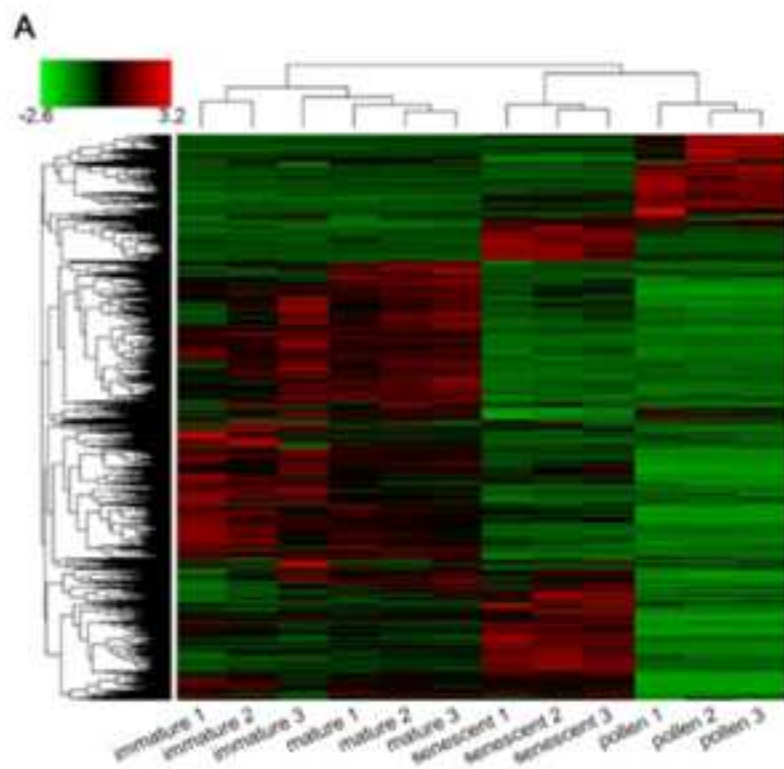
980 **Figure 5.** Graphical visualization of haplotype-blocks (HB) inheritance on Chr15 along with the
981 Chandler pedigree. **(A)** The inner-circle highlights in grey two regions of heterozygosity (5 HB
982 the first and 7 HB the second), and in light green two regions of homozygosity (3 HB the first and
983 4 HB the second). The circle in the middle shows maternally inherited HBs, while the HBs
984 inherited through the paternal line are visualized in the outer circle. Payne’s haplotypes are clearly

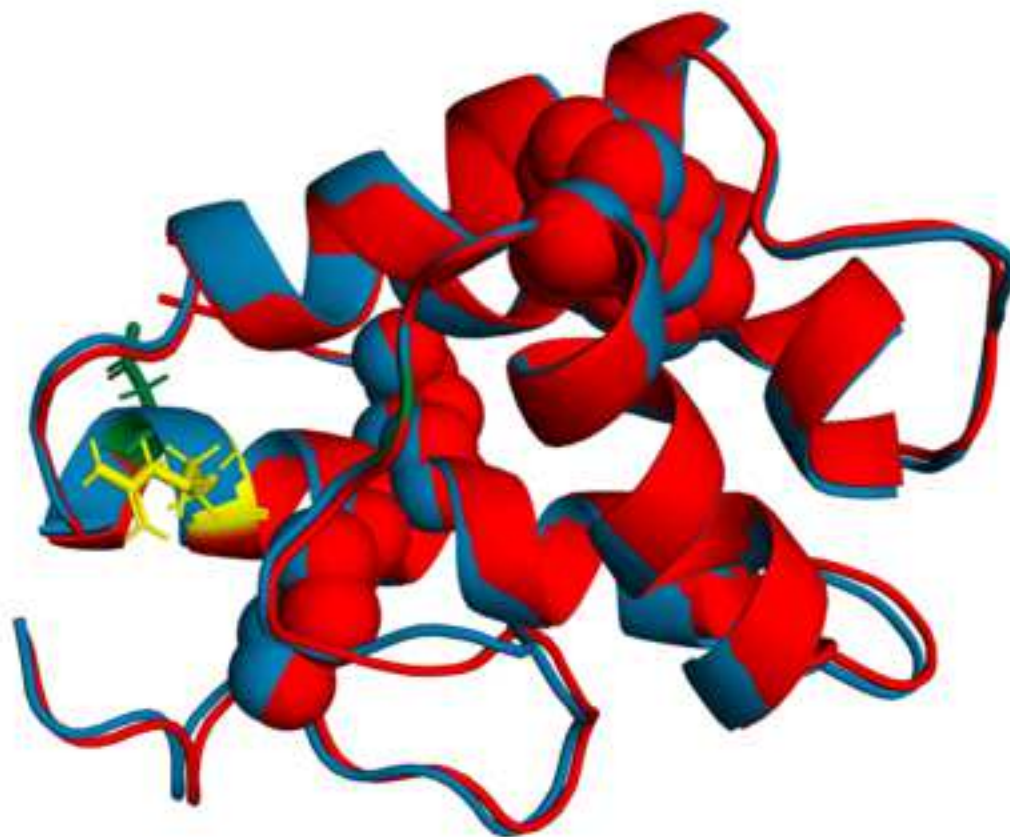
985 present in both parental lines. **(B)** Chandler pedigree, where Pedro is the maternal line and 56-224,
986 the paternal line.

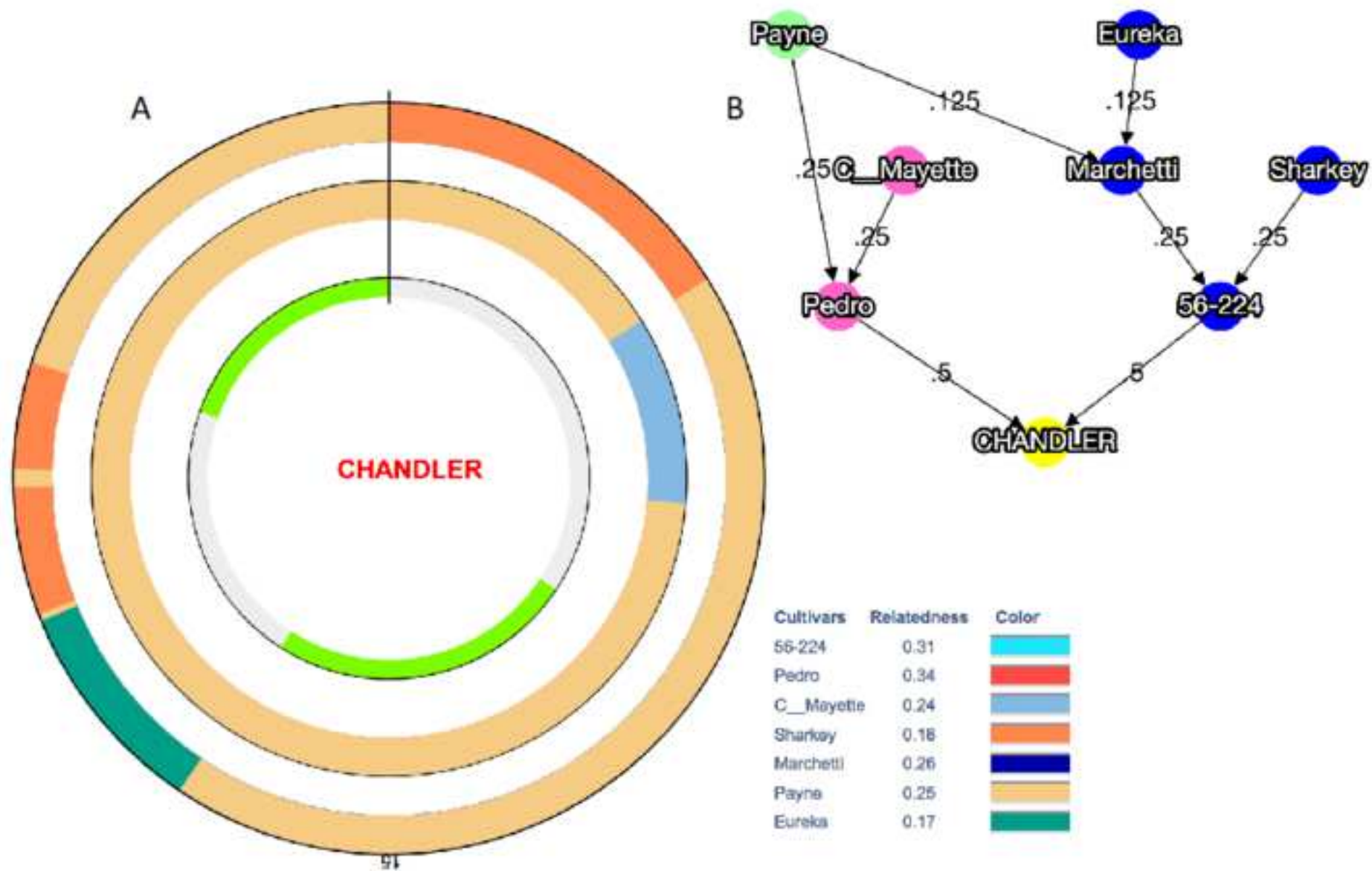
987 **Figure 6.** Graphical visualization of allele identity between Chandler and its ancestor Payne for
988 all 16 chromosomes of Chandler.

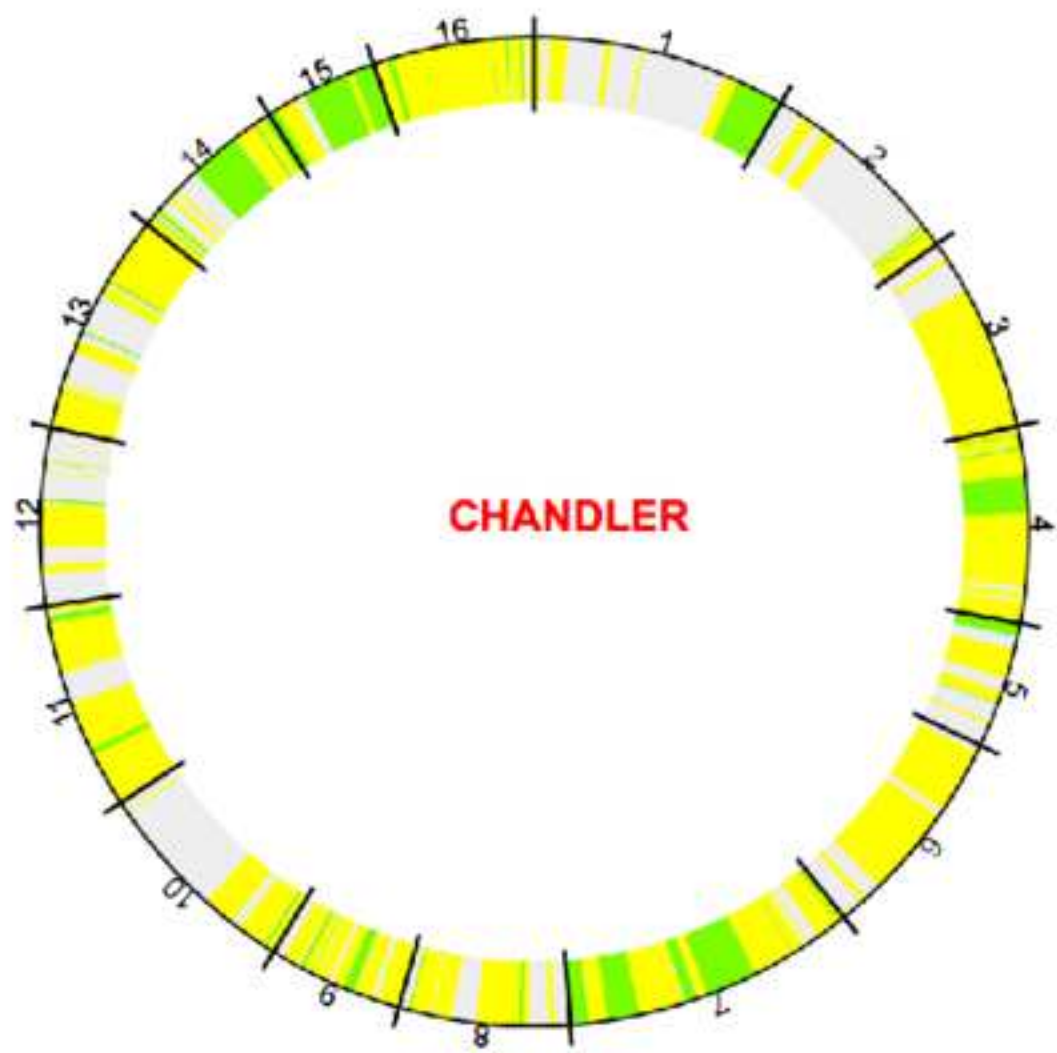




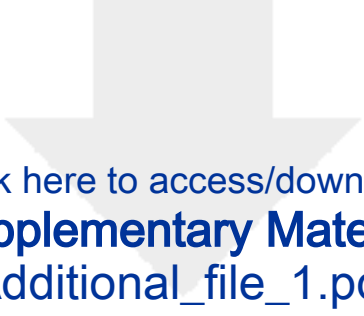







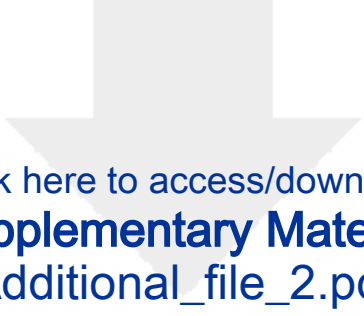


Legend
All alleles matching
One allele matching *
No alleles matching
* for polyploids at least one allele





Click here to access/download
Supplementary Material
Additional_file_1.pdf



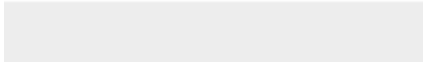




Click here to access/download
Supplementary Material
Additional_file_2.pdf






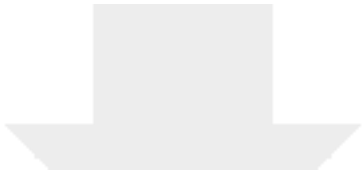
Click here to access/download
Supplementary Material
Additional_file_3.tif



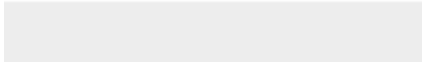



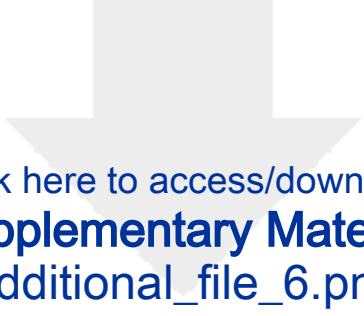
Click here to access/download
Supplementary Material
Additional_file_4.tif



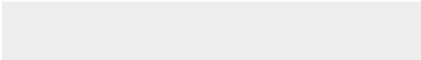



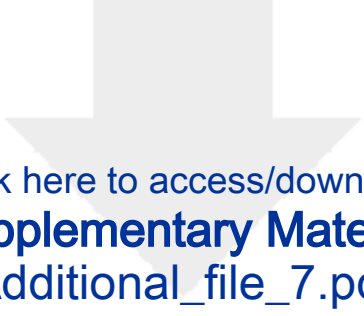
Click here to access/download
Supplementary Material
Additional_file_5.tiff






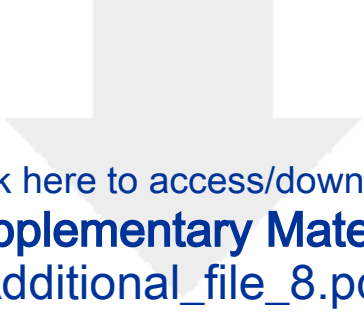
Click here to access/download
Supplementary Material
Additional_file_6.png



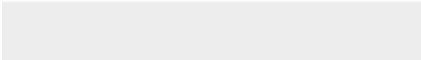



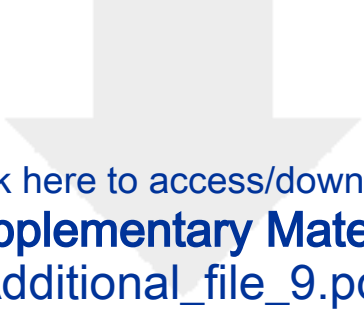
Click here to access/download
Supplementary Material
Additional_file_7.pdf






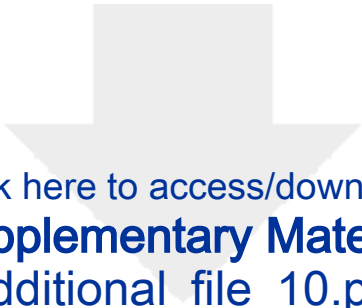
Click here to access/download
Supplementary Material
Additional_file_8.pdf



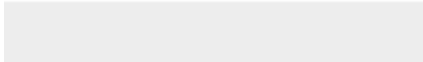



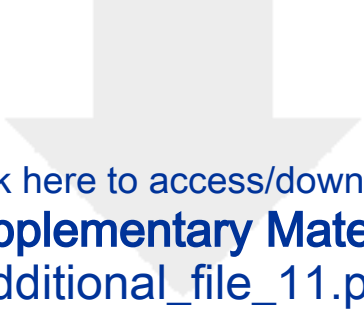
Click here to access/download
Supplementary Material
Additional_file_9.pdf






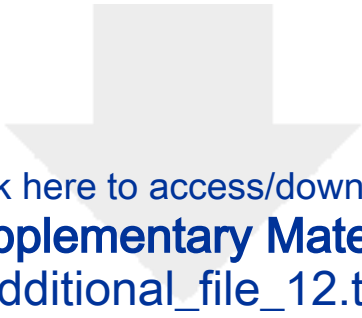
Click here to access/download
Supplementary Material
Additional_file_10.pdf



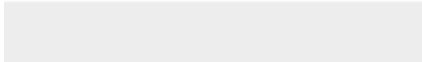



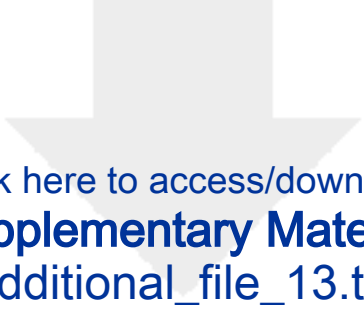
Click here to access/download
Supplementary Material
Additional_file_11.pdf






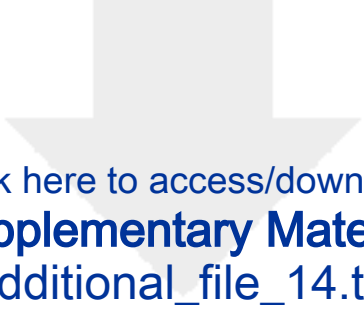
Click here to access/download
Supplementary Material
Additional_file_12.tiff



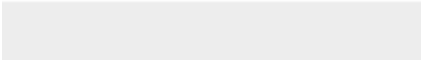



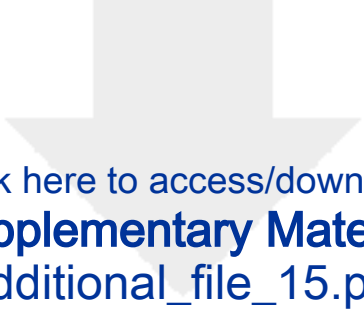
Click here to access/download
Supplementary Material
Additional_file_13.tiff



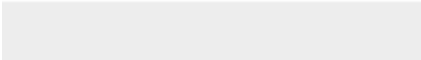



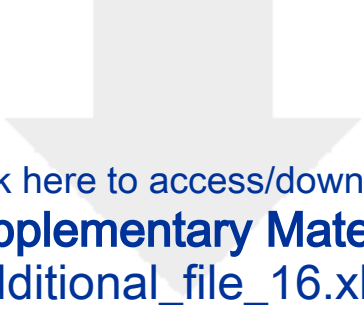
Click here to access/download
Supplementary Material
Additional_file_14.tiff






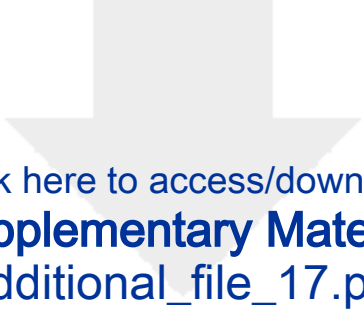
Click here to access/download
Supplementary Material
Additional_file_15.pdf



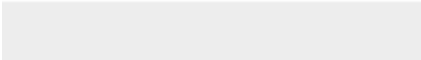



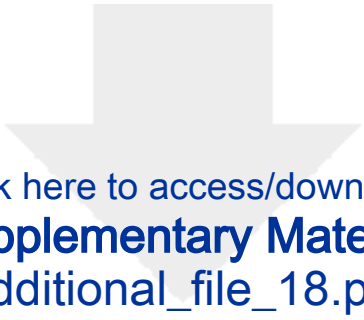
Click here to access/download
Supplementary Material
Additional_file_16.xlsx







Click here to access/download
Supplementary Material
Additional_file_17.pdf






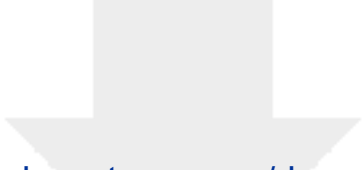
Click here to access/download
Supplementary Material
Additional_file_18.pdf






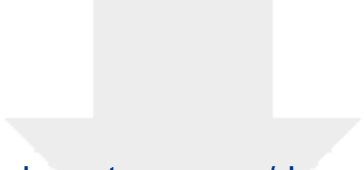
Click here to access/download
Supplementary Material
Additional_file_19.xlsx






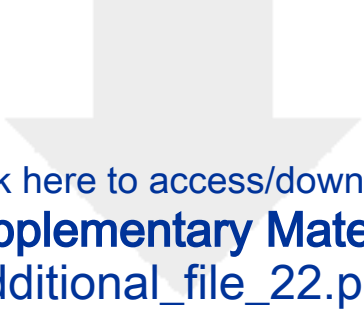
Click here to access/download
Supplementary Material
Additional_file_20.pdf



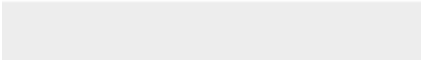



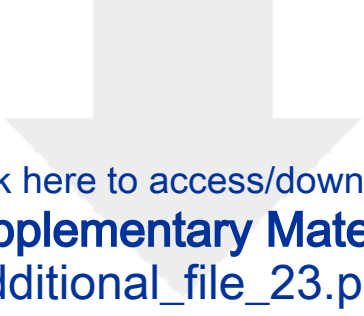
Click here to access/download
Supplementary Material
Additional_file_21.pdf



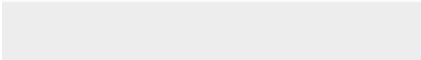



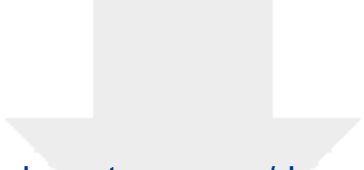
Click here to access/download
Supplementary Material
Additional_file_22.png






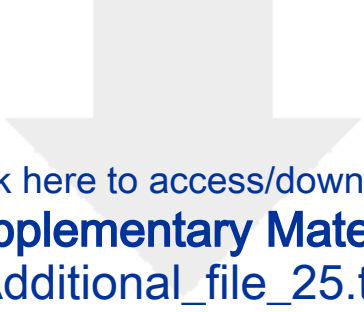
Click here to access/download
Supplementary Material
Additional_file_23.png



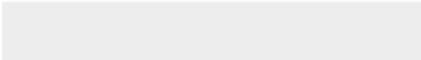



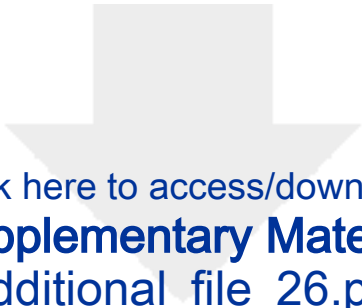
Click here to access/download
Supplementary Material
Additional_file_24.pdf



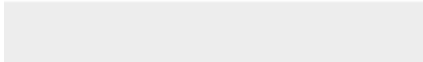



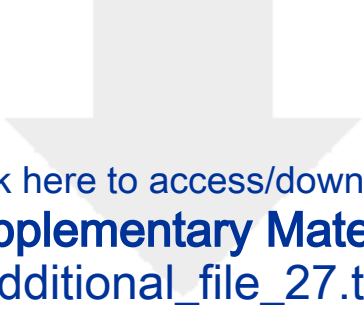
Click here to access/download
Supplementary Material
Additional_file_25.tif



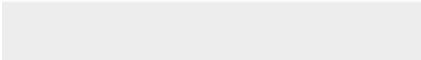




Click here to access/download
Supplementary Material
Additional_file_26.pdf






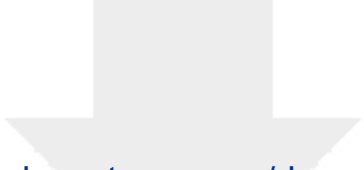
Click here to access/download
Supplementary Material
Additional_file_27.tiff



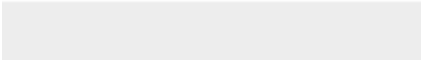



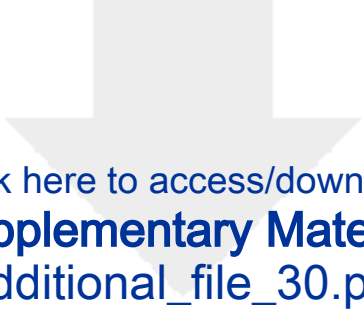
Click here to access/download
Supplementary Material
Additional_file_28.png



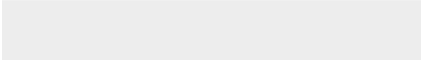



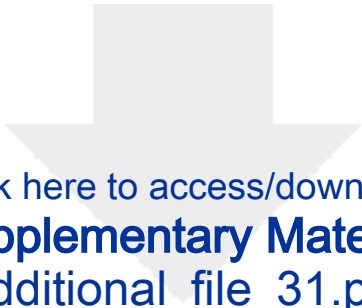
Click here to access/download
Supplementary Material
Additional_file_29.pdf



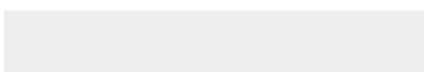
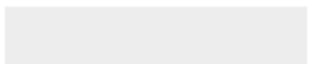


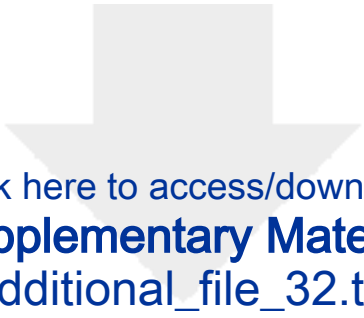
Click here to access/download
Supplementary Material
Additional_file_30.pdf



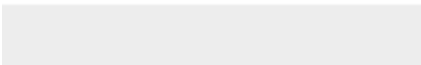



Click here to access/download
Supplementary Material
Additional_file_31.pdf





Click here to access/download
Supplementary Material
Additional_file_32.tiff



UNIVERSITY OF CALIFORNIA, DAVIS

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



SANTA BARBARA • SANTA CRUZ

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

COLLEGE OF AGRICULTURAL AND
ENVIRONMENTAL SCIENCES
AGRICULTURAL EXPERIMENT STATION
COOPERATIVE EXTENSION

October 19th, 2018

Dear Editor,

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

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

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

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

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

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

We kindly request to exclude the following referees:

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

We suggest the following referees:

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

Sincerely,

Annarita Marrano



Ph.D.

Dept. Plant Sciences, University of California, Davis