Chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes

José A. Campoy[1,∀], Hequan Sun[1,2,∀], Manish Goel[1], Wen-Biao Jiao[1], Kat Folz-Donahue[3], Christian Kukat[3], Manuel Rubio[4], David Ruiz[4], Bruno Huettel[5] and Korbinian Schneeberger[1,2,*]

[∀]These authors contributed equally.

[1]Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany [2]Faculty of Biology, LMU Munich, Großhaderner Str. 2, 82152 Planegg-Martinsried, Germany [3]FACS & Imaging Core Facility, Max Planck Institute for Biology of Ageing, 50931 Cologne, Germany [4]Departament of Plant Breeding, CEBAS-CSIC, PO Box 164, E-30100, Espinardo, Murcia, Spain [5]Max Planck-Genome-center Cologne, Carl-von-Linné-Weg 10, 50829 Cologne, Germany

*Correspondence: Korbinian Schneeberger (schneeberger@mpipz.mpg.de)

Key words: single-cell sequencing, haplotype-resolved assembly, haplotyping, phasing, *de novo* assembly

Generating haplotype-resolved, chromosome-level assemblies of heterozygous genomes remains challenging. To address this, we developed gamete binning, a method based on single-cell sequencing of hundreds of haploid gamete genomes, which enables the separation of conventional long sequencing reads into two haplotype-specific read sets. After independently assembling the reads of each haplotype, the contigs are scaffolded to chromosome-level using a genetic map derived from the recombination patterns within the same gamete genomes. As a proof-of-concept, we assembled the two genomes of a diploid apricot tree supported by the analysis of 445 pollen genomes. Both assemblies (N50: 25.5 and 25.8 Mb) featured a haplotyping precision of >99% and were accurately scaffolded to chromosome-level as reflected by high levels of synteny to closely-related species. These two assemblies allowed for first insights into haplotype diversity of apricot and enabled the identification of non-allelic crossover events introducing severe chromosomal anomalies in 1.6% of the pollen genomes.

Currently, most diploid genome assemblies ignore the differences between the homologous chromosomes and assemble the genomes into one pseudo-haploid sequence, which is an artificial consensus of both haplotypes. Such an artificial consensus can result in imprecise gene annotation and misleading biological interpretation[1,2]. To avoid these problems, it is a common strategy to inbreed or to generate double-haploid genotypes to enable the assembly of homozygous genomes. More recent alternatives include chromosome sorting[3], Hi-C[4,5] and Strand-seq[6] to either separate the chromosomes before sequencing or to generate additional information that discriminates between the two haplotypes and thereby reconstructs the sequence of both haplotypes separately. Another elegant method, trio binning, is based on the separation of whole-genome sequencing reads into haplotype-specific read sets before assembly using the genomic differences between the parental genomes[2]. While this is a powerful method, it can be limiting if the parents are not available or unknown[7]. A solution for this is the sequencing of a few gamete genomes (derived from the focal

individual), which is sufficient for the inference of genome-wide haplotypes, but relies on existing long-contiguity reference sequences[8,9,10,11].

In addition to resolving haplotypes, the generation of chromosome-level assemblies, which are necessary to understand the full complexity of genomic differences including all kinds of structural rearrangements, is similarly challenging[12,13]. While recent improvements in long DNA molecule sequencing[14] promise the assembly of telomere-to-telomere contigs, genetic maps can reliably help to resolve mis-assemblies as well as guide chromosome-level scaffolding. The generation of genetic maps, however, relies on a substantial amount of meiotic recombination which usually implies the genotyping of hundreds of recombinant genomes. Creating and genotyping sufficiently large populations can be time-consuming and costly and posts great challenges in species with long juvenile periods[15,16].

To address all these challenges, we present gamete binning, a method for chromosome-level, haplotype-resolved genome assembly - independent of parental genomes or recombinant progenies (Fig. 1). The method starts by isolating gamete nuclei from the focal individual followed by high-throughput single-cell sequencing of hundreds of the haploid gamete genomes. (For clarification, we collectively refer to both gametophytes in plants and gametes in animals collectively as gametes, as both have haploid genomes.) The segregation of sequence variation in the gamete genomes enables a straightforward phasing of all variants into two haplotypes, which subsequently allows for genetic mapping and sorting of whole-genome sequencing reads into distinct read sets - each representing a different haplotype. Assembling these independent read sets leads to haplotype-resolved genome assemblies, which can be scaffolded to chromosome-level using a gamete genome-derived genetic map.

We used gamete binning to assemble the two haploid genomes of a specific, diploid apricot tree (*Prunus armeniaca*; cultivar 'Rojo Pasion'[17]), which grows in Murcia, southeastern Spain (Supplementary Figure 1). We first performed a preliminary *de novo* genome assembly using *Canu*[18] with 19.9 Gb long reads (PacBio, Supplementary Figure 2) derived from DNA extracted from fruits and corresponding to 82x genome coverage according to a genome size of ~242.5 Mb estimated by *findGSE*[19] (Methods; Supplementary Figure 3). After purging

77 haplotype-specific contigs, the curated assembly consisted of 939 contigs with a combined
78 length of 230.9 Mb and an N50 of 563.8 kb, which represents a haploid, but mosaic assembly
79 of the apricot genome (Methods).

80 To advance this assembly, we isolated pollen grains from ten closed flowers (to avoid
81 contamination of foreign pollen) and released their nuclei following a protocol based on pre-
82 filtering followed by bursting[20] (Fig. 1a; Methods). The nuclei mixture was cleaned up using
83 propidium iodide staining plus sorting by flow cytometry, leading to a solution with 12,600 nuclei
84 that were loaded into a 10x Chromium Controller in two batches - each with 6,300 nuclei
85 (Supplementary Figures 1a-d; Supplementary Figure 4; Methods). With this we generated two
86 10x single-cell genome (CNV) sequencing libraries, which were sequenced with 95 and 124
87 million 151 bp paired-end reads (Illumina). By exploring the *cellranger*-corrected cell barcodes
88 within the read data of both libraries, we extracted 691 read sets - each with a minimum of
89 5,000 read pairs (Methods; Fig. 2a).

90 Aligning the reads of each pollen genome to the curated assembly, we found that the
91 reads of 246 sets featured high similarity to thrip genomes or included more than one haploid
92 genome, possibly due to random attachment of multiple nuclei during 10x Genomics library
93 preparation or the uncompleted separation of pollen nuclei during pollen maturation[21]
94 (Supplementary Figure 5a-c; Methods). Thus, we selected the set of 445 haploid pollen
95 genomes. In general, the short read alignments did not show any biases or preferences for
96 specific regions of the genome as reported for some of the single-cell genome amplification
97 kits, but covered nearly all regions (99.1%) of the curated assembly (Fig. 2b; Supplementary
98 Figure 5d).

99 With short read alignments, we identified 578,209 heterozygous SNPs on 702 contigs
100 with a total length of 218.0 Mb (Fig. 2b; Methods). Even though this implied 1 SNP marker
101 every 377 bp on average, we observed that the distances between some of the SNP markers
102 were larger than the usual long reads, which would hamper the haplotype assignment of reads
103 whenever they aligned to such regions. Overall, we observed 10,452 regions larger than 2 kb
104 without markers (110.9 Mb) including 237 regions (12.5 Mb), that spanned entire contigs.

105 Regions without markers occur if the two haplotypes are identical (which is a common
106 phenomenon in domesticated genomes) or if a region exists only in one of the haplotypes (e.g.
107 a large indel). We distinguished these two cases using the short-read coverage of the
108 combined pollen read sets, assuming that the regions that are only present in one haplotype
109 are supported by only approximately half of the reads (Methods). While 7,199 regions (74.5
110 Mb) were shared between the haplotypes (and were labelled as conserved), we found that
111 3,253 regions (36.4 Mb) were specific to one of the haplotypes (i.e. deletions; Fig. 2b). Such
112 regions (i.e. deletions) which are specific to one haplotype can also be used as markers. If
113 such deletions were linked to nearby SNP markers, we phased them according to their linked
114 alleles. For deletions on contigs without additional markers, we used the absence and
115 presence of read alignments in the pollen to assign genotypes.

116 The haploid nature of the 445 selected individual pollen genomes allowed us to phase
117 all SNP and deletion markers into two haplotypes simply by using the linkage within the pollen
118 genomes (Fig. 2c-d). To phase the haplotypes across the contigs, we generated two virtual
119 markers for each contig representing the (imputed) alleles at both ends of the contig. The
120 markers were grouped into a genetic map with eight linkage groups (corresponding to the eight
121 homologous chromosome pairs) including 891 contigs with a total length of 228.0 Mb
122 (corresponding to about 99% of the complete assembly) using *JoinMap* 4.0[22] (Fig. 2e; Fig. 3a)
123 (Methods).

124 After this, we aligned the PacBio reads to the curated assembly. Using the phased
125 alleles (of the SNP and deletion markers) within each of the individual PacBio read alignments,
126 we separated 93.4% of the reads into one of 16 haplotype-specific clusters representing the
127 two haplotypes of each of the eight linkage groups. Reads that aligned in regions that were
128 conserved between the two haplotypes were randomly assigned to one of the two haplotype-
129 specific clusters (Fig. 3a; Methods). Similarity analyses revealed that most of the remaining
130 6.6% reads were related to organellar genomes or repetitive sequences.

131 The 16 haplotype-specific read sets were independently assembled using *Flye*[24], which
132 led to 16 haplotype-specific chromosome assemblies with average N50 values ranging from

133 662.3 kb to 664.6 (Methods). Using the genetic map, we combined the contigs of each
134 assembly into a pseudo-molecule. This led to two haplotype-resolved chromosome-level
135 assemblies, both with N50 above 25.0 Mb (Fig. 3a-b; Methods).

136 To assess haplotype accuracy, we additionally whole-genome sequenced the parental
137 cultivars of 'Rojo Pasion' known as 'Currot' and 'Orange Red'. Using Illumina sequencing
138 technology, we generated 15.7 and 16.2 Gb short reads of each of the diploid parental
139 genomes, respectively. Overall, we found that ~99.1% of the *k*-mers that were specific to one
140 of the haplotype assemblies could be found in the corresponding parental genome illustrating
141 that almost all of the variation was correctly assigned to haplotypes (Fig. 3c; Table 1; Methods).
142 Having proved the haplotype accuracy, the assemblies were polished resulting in final
143 assemblies (N50: 25.5 Mb and 25.8 Mb; Table 1; Methods). To further assess the quality of
144 the scaffolded chromosome structure, we compared our assemblies with recently assembled
145 genomes, including those of very closely-related species such as the heterozygous
146 'Chuanzhihong' apricot (*Prunus armeniaca*)[23] and the Japanese apricot (*Prunus mume*)[25], and
147 a more distantly-related species, peach (*Prunus persica:* doubled-haploid genome)[26], using
148 *SyRI*[12] (a tool designed for the comparison of chromosome-level assemblies). Our assemblies
149 showed high consistency in the synteny to these assemblies, reflecting the reliability of the
150 genetic map and the assembled genome structure (Fig. 3d; Supplementary Figure 6).

151 In contrast to conventional diploid genome assemblies where the two haplotypes are
152 merged into one artificial consensus sequence, separate haploid assemblies allow for the
153 analysis of haplotype diversity. Comparing the two haplotype assemblies of 'Rojo Pasion' using
154 *SyRI*[12] allowed us to gain first insights into the haplotype diversity within an individual apricot
155 tree. Despite high levels of synteny, the two assemblies revealed large-scale rearrangements
156 (23 inversions, 1,132 translocation/transpositions and 2,477 distal duplications) between the
157 haplotypes making up more than 15% of the assembled sequence (38.3 and 46.2 Mb in each
158 of assemblies; Supplementary Table 1). Using a comprehensive RNA-seq dataset sequenced
159 from multiple tissues of 'Rojo Pasion' including reproductive buds, vegetative buds, flowers,
160 leaves, fruits (seeds removed) and barks as well as a published apricot RNA-seq dataset[23],

161 we predicted 30,378 and 30,661 protein-coding genes within each of the haplotypes (with an
162 annotation completeness of 96.4% according to a BUSCO[27] analysis). Mirroring the huge
163 differences in the sequences, we found the vast amount of 942 and 865 expressed, haplotype-
164 specific genes in each of the haplotypes (Methods; Supplementary Tables 2-3). Such deep
165 insights into the differences between the haplotypes, which are only enabled by chromosome-
166 level and haplotype-resolved assemblies, will generally be of high value for the analysis of
167 agronomically relevant variation.

168 Moreover, the chromosome-level assemblies also allow for fine-grained analyses of the
169 haploid pollen genomes, which have already undergone recombination during meiosis. Meiotic
170 recombination is the major mechanism to generate novel variation in offspring genomes.
171 During meiosis new haplotypes are formed by sequence exchanges between two homologous
172 chromosomes. To keep chromosome structures intact during such exchanges, it is essential
173 that recombination only occurs in syntenic regions as otherwise large parts of the chromosome
174 can be lost or duplicated in the newly formed molecules. Re-analyzing the 445 pollen nuclei
175 genomes using one of the chromosome-level assemblies as reference, we detected 2,638
176 meiotic crossover (CO) events (Methods). To improve the resolution of the predicted CO
177 events (6.1 kb), we selected 2,236 CO events detected in 369 nuclei with a sequencing depth
178 above 0.1x genome coverage (Supplementary Table 4). Along the chromosomes, CO events
179 were broadly and positively correlated with the density of protein-coding genes and were
180 almost completely absent in rearranged regions as expected (Fig. 4; Methods). By
181 investigating the fine-scale pattern of short read alignment of each nuclei, we identified six CO
182 events located in rearranged regions (0.3% of 2,236 CO events found in 1.6% of the pollen
183 genomes), which led to stark chromosomal rearrangements. In each of the six chromosomes
184 we found duplicated read coverage and pseudo-heterozygous variation in the regions that
185 were involved in the chromosome rearrangements as induced by the non-allelic CO (Fig. 5).
186 This evidences the existence of non-allelic recombination in pollen genomes and might open
187 up a more detailed view on the actual meiotic recombination patterns as compared to what
188 could be observed in offspring individuals.

Taken together, following the elegant rationale of haplotype-based read separation before genome assembly introduced by trio binning[2], we present gamete binning. In contrast to trio binning, gamete binning does not rely on paternal genomes, but instead uses the genomes of individual gametes to resolve haplotypes. In addition, the recombination patterns in these gamete genomes can be used to calculate a genetic map, which in turn enables the generation of chromosome-level assemblies. High-throughput analysis of gamete genomes avoids tedious generation of offspring progeny and allows to sample the required material in its ecological context, which makes it possible to analyze meiotic recombination as it occurs in natural environments. As a result, gamete binning can efficiently and effectively enable haplotype-resolved and chromosome-level genome assembly of any heterozygous individual with accessible gametes.

200 # Online Methods

201 ## DNA extraction, Illumina/PacBio library preparation and sequencing

202 Fresh developing fruits of 'Rojo Pasion' were frozen in liquid nitrogen immediately after
203 being sampled in Murcia, Spain. After being shipped to the Max Planck Institute for Plant
204 Breeding Research (MPIPZ, Cologne, Germany), DNA was extracted from the mesocarp and
205 exocarp of the fruits using the Plant DNA Kit of Macherey-Nagel™ to create a PacBio
206 sequencing library. Meanwhile, fresh leaves were sampled from the parental cultivars ('Currot'
207 and 'Orange Red') at the experimental field of CEBAS-CSIC in Murcia, Spain, and Illumina
208 short read libraries were prepared after DNA extraction using the Plant DNA Kit of Macherey-
209 Nagel™.

210 All libraries were sequenced with the respective sequencing machines (Illumina HiSeq
211 3000 and PacBio Sequel I) at Max Planck Genome-centre Cologne (MP-GC), which led to
212 19.9 Gb long reads for 'Rojo Pasion' (PacBio; Supplementary Figure 2) and 15.7 and 16.2 Gb
213 short reads for the parental cultivars (Illumina). Note that the parental WGS data were only
214 used for haplotype validation and for sorting the individual chromosome assemblies to two sets
215 of eight chromosomes to match the inheritance of the chromosomes.

216 ## Pollen nuclei DNA extraction, 10x sc-CNV library preparation and sequencing

217 Dormant shoots of 'Rojo Pasion' bearing developed flower buds were collected in
218 Murcia, Spain. Then, the shoots were shipped at 4 °C to MPIPZ (Cologne, Germany) and were
219 grown in long-day conditions in the greenhouse. Flowers at the pre-anthesis stage were frozen
220 in liquid nitrogen. Anthers from ten 'Rojo Pasion'[17] flowers were extracted with forceps and
221 submerged in woody pollen buffer (WPB)[28]. Around 500,000 pollen grains were extracted from
222 anthers by vortexing them in WPB. The nuclei were isolated from the pollen using a modified
223 bursting method[20]. Isolated pollen was prefiltered (100µm) and bursted (30um) using
224 Celltrics™ sieves and woody pollen buffer. The nuclei were then stained with propidium iodide
225 (PI) at 50 µg/mL just before sorting and counting by flow cytometry to remove pollen grain
226 debris using a BD FACSAria Fusion™ with high-speed sort settings (70 µm nozzle and 70 PSI

227 sheath pressure) and 0.9% NaCl as sheath fluid. The nuclei were identified by PI fluorescence,
228 light scattering, and autofluorescence characteristics (Supplementary Figure 4). A total of
229 12,600 nuclei were counted and collected in a solution of 4.2 μL phosphate-buffered saline
230 with 0.1% bovine serum albumin.

231 According to manufacturer's instructions, the nuclei were loaded into a 10x™ Chromium
232 controller in two batches with 6,300 nuclei each, i.e., two 10x sc-CNV libraries were prepared.
233 In each library, DNA fragments from the same nucleus were ligated with a unique 16-bp
234 barcode sequence (of A/C/G/T). Both libraries were sequenced using Illumina HiSeq3000 in
235 the 2x151 bp paired-end mode, totaling 95 and 124 million read pairs, respectively (61.7 Gb).

236 **Genome size estimation**

237 After trimming off 10x Genomics barcodes and hexamers from the 61.7 Gb reads of
238 the two 10x sc-CNV libraries, *k*-mer counting (*k*=21) was performed with *Jellyfish*[29]. The *k*-mer
239 histogram was provided to *findGSE*[19] to estimate the size of the 'Rojo Pasion' genome under
240 the heterozygous mode (with '*exp_hom*=200'; Supplementary Figure 3).

241 **Initial diploid-genome assembly and curation**

242 With the 19.9 Gb raw PacBio reads of 'Rojo Pasion' (Supplementary Figure 2), a
243 preliminary diploid assembly was constructed using *canu*[18] (with options
244 'genomeSize=242500000 corMhapSensitivity=high corMinCoverage=0 corOutCoverage=100
245 correctedErrorRate=0.105').

246 All raw Illumina reads from the 10x libraries were firstly aligned to the initial assembly
247 using *bowtie2*[30]. Then the *purge haplotigs* pipeline was then used to remove haplotigs (i.e.,
248 haplotype-specific contigs inflating the true haploid genome) based on statistical analysis of
249 sequencing depth, and identify primary contigs to build up a curated haploid assembly[31]. To
250 reduce the false positive rate in defining haplotigs, each haplotig was blasted to the curated
251 assembly; if over 50% of the haplotig could not be covered by any primary contigs, it was re-
252 collected as a primary contig.

253 **SNP marker identification**

254 After trimming off 10x barcodes and hexamers, all pooled Illumina reads from the 10x
255 sc-CNV libraries (61.7 Gb) were re-aligned to the curated haploid assembly using *bowtie*2[30].
256 With 87.2% reads aligned, 989,132 raw SNPs were called with *samtools and bcftools*[32]. Three
257 criteria were used to select potential allelic SNPs (578,209), including i) the alternative allele
258 frequency must be between 0.38 to 0.62; ii) the alternative allele must be carried by 60-140
259 reads; iii) the total sequencing depth at a SNP must be between 120-280x (as compared with
260 genome-wide mode depth of 208x; Fig. 2b).

261 **Deletion marker identification and genotyping**

262 There were 10,452 regions of over 2 kb but without a single SNP marker defined (total:
263 110.9 Mb). If the average sequencing depth of such a region was less than or equal to 146x
264 (i.e., the value at the valley between middle and right-most peaks in sequencing depth
265 distribution; Fig. 2b), it was selected as a deletion-like marker. This led to a list of 3,253 large-
266 scale deletion markers (36.4 Mb), among which 237 contigs (12.5 Mb) did not have a single
267 SNP marker. The remaining 7,199 regions (74.5 Mb) were defined as conserved between two
268 haplotypes (Fig. 2b). For a deletion marker, raw reads of each nucleus were counted within
269 the deletion with *bedtools*[33] and were further normalized as *reads per kilobase per million*
270 *mapped reads* (*RPKM*) to reduce the effect of sequencing depth and deletion size. The
271 genotype at such a deletion marker was initialized as *a* or *n*, where *a* means the presence of
272 reads (or non-deletion, which might be changed to *b* during later linkage grouping and
273 mapping) and *n* means an absence of reads (either deletion or not available; Fig. 2d).

274 **Variant phasing and CO identification**

275 Barcode in the raw reads were corrected using *cellranger* from 10x Genomics, with
276 which 182.1 million read pairs (51.0 Gb) were clustered into 691 read sets. Reads of each read
277 set were aligned to the curated assembly using *bowtie2*[30], bases were called using *bcftools*[34],
278 and a simple bi-marker majority voting strategy was applied in phasing SNPs along each contig

279 (Fig. 2c). After phasing, we could identify COs within contigs to facilitate later genetic mapping,
280 for example, there was a CO for the nucleus with "***nnTGnTGnnnGAnnA*".**

281 **Ploidy evaluation of single-cell sequencing**

282 For each nucleus, with short read alignment and base calling to the curated assembly,
283 we counted the number of inter-genotype transitions (genotype *a* to *b* and *b* to *a*) at phased
284 SNP markers over all contigs. Correlating this to the number of covered markers revealed two
285 clusters of nuclei (Supplementary Figure 5c). One cluster with 217 nuclei showed that inter-
286 genotype transitions increased linearly with the number of covered markers (while there were
287 high ratios of more than 5 transitions in every 100 markers), which indicated the sequencing
288 data were mixed from more than one nuclei. The other cluster of 445 nuclei (31.2 Gb with
289 111.4 million read pairs) showed a limited increase (probably due to sequencing errors or
290 markers from repetitive regions), which supported the expected haploid status.

291 **Imputation of virtual markers at ends of contigs**

292 Let *a* and *b* denote the parental genotypes. The genotype of a nucleus at both ends of
293 a contig (referred to as virtual markers) can be represented by *aa*, *bb* or *ab* (or *ba*) where *aa*/*bb*
294 indicates an identical genotype along the contig while *ab* (or *ba*) indicates a CO event in the
295 regions of contig. Then we can build up genotype sequences at the two ends of all contigs
296 (with SNP markers) by imputing at all nuclei. For example, given a contig, sequences of
297 *aaaaaa**b**abbbbbbb* (marker 1) and *aaaaaa**a**abbbbbbb* (marker 2) means there is a CO (in bold)
298 at the 7th (of 15) nuclei (Fig. 2c).

299 **Linkage grouping and genetic mapping**

300 All virtual markers (defined using SNP markers along contigs) were classified into 8
301 linkage groups (653 contigs: 212.9 Mb) after pairwise comparison of their genotype sequences
302 using *JoinMap4.0*[22] (with haploid population type: HAP; and logarithm of the odds (LOD) values
303 larger than 3.0).

304 After filtering out contigs with >10% missing nuclei information or nuclei with >10%
305 missing contigs, a high-quality genetic map consisting of 216 contigs (147.3 Mb, corresponding
306 to 622.0 cM; Fig. 3a) was first obtained using regression mapping in *JoinMap* 4.0® with the
307 following settings: LOD larger than 3.0, a "*goodness-of-fit jump"* threshold of 5.0 for removal
308 of loci and a "two rounds" mapping strategy[22]. Genotype sequences imputed at contig ends or
309 deletions (i.e., respective virtual markers) were used to integrate the remaining 723 contigs
310 into the genetic map. For example, given a deletion marker (e.g., *p* and *q* in Fig. 2c-e), if the
311 respective contig had already existed in the genetic map, phasing was only performed at the
312 deletion (according to surrounding phased SNPs); otherwise, phasing plus positioning to the
313 genetic map would be applied. Both operations were based on finding the minimum divergence
314 of the genotype sequence of the marker to that of the other contigs (in the corresponding
315 genetic map). The final genetic map was completed as 891 contigs of 228.0 Mb.

316 **Haplotype-specific PacBio read classification**

317 PacBio reads (19.9 Gb) were classified based on three major cases after being aligned
318 to the curated assembly using *minimap2*[35]. First, a read covering phased SNP markers was
319 directly clustered into the haplotype supported by the respective alleles in the read. Second, a
320 read covering no SNP markers but overlapping a deletion marker was clustered into the
321 respective genotype based on its phasing with neighboring imputed markers in genetic map.
322 Third, a read in a conserved region was assigned to one of the haplotypes randomly. Overall,
323 93.4% reads could be classified into two genotypes for eight linkage groups (Fig. 3a). Non-
324 classified reads (6.6%) were found (by blasting) to be related to organelle genomes and
325 repeats.

326 **Haplotype-genome assembly and scaffolding**

327 Independent assemblies were performed with sixteen sets of reads, i.e., for every two
328 haplotypes in each of the eight linkage groups using *flye*[24] with the default settings. As an
329 intermediate evaluation, combining eight assemblies from eight linkage groups could lead to
330 two artificial assemblies with 992-1017 contigs and N50 values of 662.3-664.6 kb.

331 Using the 891 contigs of the curated assembled that were assigned to chromosomal
332 positions with the genetic mapping, we created a pseudo reference genome, with which the
333 newly assembled contigs were scaffolded using *RAGOO*[36], leading to chromosome-level
334 assemblies (i.e., those labeled with 'scaf' in Fig. 3b).

335 **Haplotype evaluation on the two haploid assemblies**

336 The genotypes of the sixteen assemblies were firstly identified by comparing *k*-mers in
337 each assembly with Illumina WGS of the parental cultivar (*k*=21; Fig. 3c). Although evaluation
338 can always be performed in each linkage group, we combined the eight linkage-group-wise
339 assemblies for 'Currot'-genotype and the other eight for 'Orange Red'-genotype, respectively.
340 After polishing the assemblies respectively with the classified 'Currot'-genotype and
341 'Orange Red'-genotype PacBio reads using *apollo*[37], we built up two sets of haplotype-specific
342 *k*-mers from the assemblies, r_C and r_O. Correspondingly, a set of 'Currot'-specific *k*-mers (with
343 coverage from 10 to 60x), p_C, was selected from the parental Illumina WGS that did not exist
344 in 'Orange Red' short reads (coverage over 1x) but in 'Rojo Pasion' pollen short reads
345 (coverage from 10 to 300x); similarly, a set of 'Orange Red'-specific *k*-mers, p_O, was also
346 collected. Then we intersected r_C and r_O with p_C and p_O respectively, leading to four subsets
347 $r_C \cap p_C$, $r_C \cap p_O$, $r_O \cap p_C$, and $r_O \cap p_O$. This calculation gave an average haplotyping accuracy of
348 99.1% (Table 1). All *k*-mer processing (counting, intersecting and difference finding) were
349 performed with *KMC*[38]. After haplotype validation, the assemblies were further polished with
350 the respective parental short read alignment using *pilon*[39] (with options '--fix bases --mindepth
351 0.85'). The final haplotype assembly sizes were 216.0 and 215.2 Mb for 'Currot'-genotype (93
352 scaffolds, N50: 25.8 Mb) and 'Orange Red'-genotype (104 scaffolds, N50: 25.5 Mb),
353 respectively (Table 1). Note, the eight main chromosome-level scaffolds of each haplotype
354 made up ~99% of the respective assembly.

355 **Genome annotation**

356 We annotated protein-coding genes for each haplotype assembly by integrating
357 evidences from *ab initio* gene predictions (using three tools *Augustus*[40], *GlimmerHMM*[41] and

358 *SNAP*[42]), RNA-seq read assembled transcripts and homologous protein sequence alignments.
359 We aligned protein sequences from the database UniProt/Swiss-Prot, *Arabidopsis thaliana*
360 and *Prunus persica* to each haplotype assembly using the tool *Exonerate*[43] with the options "-
361 -percent 60 --minintron 10 --maxintron 60000". We mapped RNA-seq reads from reproductive
362 buds, vegetative buds, flowers, leaves, fruits (except seeds) and bark tissues, as well as a
363 published Apricot RNA-seq dataset[23], using HISAT[44], and we assembled the transcripts using
364 *StringTie*[45]. Finally, we used the tool *EvidenceModeler*[46] to integrate the above evidence in
365 order to generate consensus gene models for each haplotype assembly.

366 We annotated the transposon elements (TE) using the tools *RepeatModeler* and
367 *RepeatMasker* (http://www.repeatmasker.org). We filtered the TE related genes based on their
368 coordinates overlapping with TEs (overlapping percent > 30%), sequence alignment with TE-
369 related protein sequences and *A. thaliana* TE related gene sequences (both requiring *blastn*
370 alignment identity and coverage both larger than 30%).

371 We improved the resulting gene models using in-house scripts. Firstly, we ran a primary
372 gene family clustering using *orthoFinder*[47] based on the resulting gene models from each
373 haplotype to find haplotype-specific genes. We then aligned these specific gene sequences to
374 the other haplotype using *blastn*[48] to check whether it was specific because the ortholog was
375 unannotated in the other haplotype. For these potentially unannotated genes (blastn identity >
376 60% and blastn coverage > 60%), we checked the gene models from *ab initio* prediction around
377 the aligned regions to add the unannotated gene if both the gene model and the aligned region
378 had an overlapping rate larger than 80%. We also directly generated new gene models based
379 on the *Scipio*[49] alignment after confirming the existence of start codon, stop codon and splicing
380 site. Finally, the completeness of assembly and annotation were evaluated by the *BUSCO*[27]
381 v4 tool based on 2,326 eudicots single-copy orthologs from OrthoDB v10[50]. A similar process
382 was used to filter for haplotype-specific genes (Supplementary Tables 2-3).

383 **Genome assembly comparison**

384 All genome assemblies, including 'Rojo Pasion' haplotypes, 'Chuanzhihong' apricot
385 (*Prunus armeniaca*)[23], Japanese apricot (*Prunus mume*)[25] and 'Lovell' peach (*Prunus*

386 *persica*)[26], were aligned to each other using *nucmer* from the *MUMmer4*[51] toolbox with
387 parameters '-max -l 40 -g 90 -b 100 -c 200'. The alignments were further filtered for alignment
388 length (>100 bp) and identity (>90%), with which structural rearrangements and local variations
389 were identified using *SyRI*[12]. To follow the nomenclature of the Prunus community, the 'Rojo
390 Pasion' chromosomes were numbered according to the numbering in 'Lovell' peach[26].

391 **Crossover identification and landscape creation**

392 All 220 million pollen nuclei-derived short read pairs were pooled and aligned to the
393 'Currot'-genotype assembly, from which 739,342 SNP markers were defined with an
394 alternative allele frequency distribution of 0.38 to 0.62 and alternative allele coverage of 50 to
395 150x. Then, short reads of 445 nuclei were independently aligned to the 'Currot'-genotype
396 assembly using *bowtie2*[30] and bases were called using *bcftools*[34]. Finally, *TIGER*[52] was used
397 to identify COs. The landscape of COs from 369 nuclei with a sequencing depth over 0.1x was
398 calculated within 500 kb sliding windows along each chromosome at a step of 50 kb (Fig. 4),
399 where for each window, the recombination frequency (*cM/Mb*) was defined as *C*/*n*/(*w*/10^6)*
400 100%, where *C* is the number of recombinant nuclei in that window, *n* is the total number of
401 nuclei (369) and *w* is the window size. *SNP/Mb* and *gene/Mb* were calculated for the same
402 windows as *x/*(*w*/10^6), where *x* was the count of the feature in the respective window.

Acknowledgements

The authors would like to thank Antonio Molina and Jose Egea for kindly providing plant material, Saurabh Pophaly for help in transferring the read data to public servers and Kristin Krause and Vidya Oruganti for useful discussions and comments for improving the manuscript. This work was funded by the “Humboldt Research Fellowship for Experienced Researchers” (Alexander von Humboldt Foundation) (J.A.C.), the Marie Skłodowska-Curie Individual Fellowship PrunMut (789673) (J.A.C.), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2048/1– 390686111, and the European Research Council (ERC) Grant “INTERACT” (802629) (K.S.). C.K. acknowledges the ISAC SRL Emerging Leaders Program.

Author contributions

J.A.C, H.S. and K.S. designed the project. J.A.C., B.H., K. F.-D., C.K., D.R., and M.R. performed all wet-lab experiments. H.S., J.A.C, M.G., and W-B.J. performed all data analysis. J.A.C., H.S. and K.S. wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. Read data sequenced from two 10x sc-CNV libraries, one PacBio library from ‘Rojo Pasion’, two Illumina libraries for ‘Currot’ and ‘Orange Red’ that support the work in this study as well as the haploid assemblies and annotations generated are available in European Nucleotide Archive (ENA) under accession number “PRJEB37669”. Data was uploaded to ENA using EMBLmyGFF[53]. All other relevant data are available upon request.

Code availability

Customs scripts supporting this work are available at *github.com/schneeberger-lab/GameteBinning*.

431 **REFERENCES**

432 1. Korlach, J. *et al.* De novo PacBio long-read and phased avian genome assemblies
433 correct and add to reference genes generated with intermediate and short reads.
434 *Gigascience* **6**, 1–16 (2017).
435 2. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat.*
436 *Biotechnol.* **36**, 1174–1182 (2018).
437 3. Yang, H., Chen, X. & Wong, W. H. Completely phased genome sequencing through
438 chromosome sorting. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 12–17 (2011).
439 4. Selvaraj, S., Dixon, J. R., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction
440 using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118
441 (2013).
442 5. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware,
443 chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845
444 (2019).
445 6. Falconer, E. & Lansdorp, P. M. *Strand-seq*: A unifying tool for studies of chromosome
446 segregation. *Semin. Cell Dev. Biol.* **24**, 643–652 (2013).
447 7. Zhang, X., Wu, R., Wang, Y., Yu, J. & Tang, H. Unzipping haplotypes in diploid and
448 polyploid genomes. *Comput. Struct. Biotechnol. J.* **18**, 66–72 (2020).
449 8. Li, R. *et al.* Inference of Chromosome-length Haplotypes Using Genomic Data of Three
450 to Five Single Gametes. *bioRxiv* 361873 (2018). doi:10.1101/361873
451 9. Kirkness, E. F. *et al.* Sequencing of isolated sperm cells for direct haplotyping of a
452 human genome. *Genome Res.* **23**, 826–832 (2013).
453 10. Shi, D. *et al.* Single-pollen-cell sequencing for gamete-based phased diploid genome
454 assembly in plants. *Genome Res.* 1–11 (2019).
455 11. Wu, J. *et al.* The genome of the pear (Pyrus bretschneideri Rehd.). *Genome Res.* **23**,
456 396–408 (2013).
457 12. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. *SyRI*: finding genomic
458 rearrangements and local sequence differences from whole-genome assemblies.

459 *Genome Biol.* **20**, 1–13 (2019).
460 13. Jiao, W. B. & Schneeberger, K. Chromosome-level assemblies of multiple Arabidopsis
461 genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat.*
462 *Commun.* **11**, 1–10 (2020).
463 14. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data
464 analysis. *Genome Biol.* **21**, 1–16 (2020).
465 15. Sun, H. *et al.* Linked-read sequencing of gametes allows efficient genome-wide analysis
466 of meiotic recombination. *Nat. Commun.* **10**, 1–9 (2019).
467 16. Dréau, A., Venu, V., Avdievich, E., Gaspar, L. & Jones, F. C. Genome-wide
468 recombination map construction from single individuals using linked-read sequencing.
469 *Nat. Commun.* **10**, (2019).
470 17. Egea, J., Dicenta, F. & Burgos, L. 'Rojo Pasion' apricot. *Hortscience* **39**, 1490–1491
471 (2004).
472 18. Koren, S. *et al. Canu*: scalable and accurate long-read assembly via adaptive k-mer
473 weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
474 19. Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. *FindGSE*: Estimating genome size
475 variation within human and Arabidopsis using *k*-mer frequencies. *Bioinformatics* **34**,
476 550–557 (2018).
477 20. Kron, P. & Husband, B. C. Using flow cytometry to estimate pollen DNA content:
478 Improved methodology and applications. *Ann. Bot.* **110**, 1067–1078 (2012).
479 21. Julian, C., Rodrigo, J. & Herrero, M. Stamen development and winter dormancy in
480 apricot (Prunus armeniaca). *Ann. Bot.* **108**, 617–625 (2011).
481 22. van Ooijen, J. W. JoinMap ® 4, Software for the calculation of genetic linkage maps in
482 experimental populations. Wageningen, Netherlands: Kyazma B.V. (2006).
483 23. Jiang, F. *et al.* The apricot (Prunus armeniaca L.) genome elucidates Rosaceae
484 evolution and beta-carotenoid synthesis. *Hortic. Res.* **6**, 1–12 (2019).
485 24. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads
486 using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

487 25. Zhang, Q. *et al.* The genome of Prunus mume. *Nat. Commun.* **3**, 1–8 (2012).
488 26. Verde, I. *et al.* The high-quality draft genome of peach (Prunus persica) identifies unique
489 patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* **45**, 487–
490 494 (2013).
491 27. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M.
492 BUSCO: Assessing genome assembly and annotation completeness with single-copy
493 orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
494 28. Loureiro, J. Two new nuclear isolation buffers for plant DNA flow cytometry: A test with
495 37 species. *Ann. Bot.* 875–888 (2007).
496 29. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of
497 occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
498 30. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient
499 alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, 1–10
500 (2009).
501 31. Roach, M. J., Schmidt, S. a. & Borneman, A. R. *Purge Haplotigs*: Allelic contig
502 reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 1–10
503 (2018).
504 32. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
505 2078–2079 (2009).
506 33. Quinlan, A. R. & Hall, I. M. *BEDTools*: A flexible suite of utilities for comparing genomic
507 features. *Bioinformatics* **26**, 841–842 (2010).
508 34. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping
509 and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**,
510 2987–2993 (2011).
511 35. Li, H. *Minimap2*: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
512 3100 (2018).
513 36. Alonge, M. *et al. RaGOO*: Fast and accurate reference-guided scaffolding of draft
514 genomes. *Genome Biol.* **20**, 1–17 (2019).

515 37. Firtina, C. *et al.* Apollo: A Sequencing-Technology-Independent, Scalable, and Accurate
516 Assembly Polishing Algorithm. *Bioinformatics* 1–10 (2020).
517 38. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer
518 statistics. *Bioinformatics* **33**, 2759–2761 (2017).
519 39. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant
520 detection and genome assembly improvement. *PLoS One* **9**, 1–14 (2014).
521 40. Stanke, M. *et al.* AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic*
522 *Acids Res.* **34**, 435–439 (2006).
523 41. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: Two open
524 source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
525 42. Johnson, A. D. *et al.* SNAP: A web-based tool for identification and annotation of proxy
526 SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).
527 43. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence
528 comparison. *BMC Bioinformatics* **6**, 1–11 (2005).
529 44. Kim, D., Langmead, B. & Salzberg1, S. L. HISAT: a fast spliced aligner with low memory
530 requirements Daehwan HHS Public Access. *Nat. Methods* **12**, 357–360 (2015).
531 45. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from
532 RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
533 46. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using
534 EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**,
535 1–22 (2008).
536 47. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative
537 genomics. *Genome Biol.* **20**, 1–14 (2019).
538 48. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
539 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
540 49. Keller, O., Odronitz, F., Stanke, M., Kollmar, M. & Waack, S. Scipio: Using protein
541 sequences to determine the precise exon/intron structures of genes and their orthologs
542 in closely related species. *BMC Bioinformatics* **9**, 1–12 (2008).

50. Kriventseva, E. V. *et al.* OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).

51. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, 1–14 (2018).

52. Rowan, B. A., Patel, V., Weigel, D. & Schneeberger, K. Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. *G3 Genes, Genomes, Genet.* **5**, 385–398 (2015).

53. Norling, M., Jareborg, N. & Dainat, J. EMBLmyGFF3: A converter facilitating genome annotation submission to European Nucleotide Archive. *BMC Res. Notes* **11**, 1–5 (2018).

556 ## Figure legends

557 **Figure 1. Overview of gamete binning. a.** Extraction of gamete nuclei. **b.** Single-cell
558 genome sequencing of haploid gametes and haplotype phasing. **c.** Genetic map
559 construction based on the recombination patterns in the gamete genomes. **d.**
560 Long-read sequencing of somatic material. **e.** Separation of long reads based on
561 genetic linkage groups using phased alleles. **f.** Independent assembly of each
562 haplotype of each linkage group. **g.** Scaffolding assemblies to chromosome-level
563 using the gamete-derived genetic map.

564 **Figure 2. Single-pollen nuclei sequencing, variant phasing and genetic mapping.**
565 **a.** Sequencing depths of 691 pollen nuclei**. b.** Sequencing depth histogram of
566 pooled pollen short reads. The left-most peak revealed 0.9% of the genome that
567 were not well covered in the pollen read sets (i.e., ≤5x). The middle peak
568 indicated regions covered only by half of the genomes and present in only one of
569 the haplotypes, and the right-most peak indicated regions, which were present in
570 both haplotypes and showed the expected coverage. In regions represented in
571 both haplotypes, 578,209 SNPs were defined. Regions without SNP markers
572 were classified into 3,253 deletions and 7,199 conserved regions (Methods). **c.**
573 SNP phasing along contigs. Genotyping was first performed for each individual
574 nuclei at each SNP marker. As shown, both genotypes (in red and blue) were
575 mixed in the curated but mosaic assembly. After phasing, 8 and 7 nuclei were
576 respectively clustered for genotype A and B, and crossover could be identified.
577 With this, representative markers were imputed at ends of contigs. **d.** Imputation
578 of markers at deletions by genotyping using normalized read count. Two cases
579 were considered for phasing (and positioning) a deletion marker (in the genetic

580 map). If it was linked with surrounding SNP alleles, it could be phased
581 accordingly; otherwise, comparison its genotype sequence to genotype
582 sequences of all other markers (including SNP-derived markers at ends of
583 contigs) would be performed to find its value of phase (and positioning). **e**.
584 Linkage group and genetic map construction using the set of imputed markers
585 (SNP-derived markers labeled as 1-8 and deletion markers as *p* and *q*). For
586 example, the genotype sequences of 6, 8 and *q* needed to be flipped (i.e., phase
587 values were 1 - contig phasing). Further ordering of the markers (using *JoinMap*)
588 led to linkage group-wise genetic maps.

589 **Figure 3. Genetic mapping, haplotype-specific assembly and validation. a.** Top:
590 Genetic map with a total genetic length of 622.0 cM (Methods). Middle: up to 2
591 Gb reads were assigned to one of the two haplotypes of each linkage group.
592 Bottom: a combination of haplotype-A/B linkage groups led to two assemblies
593 with 214.6 and 215.3 Mb. **b.** Contig size distributions before (ctg-A, ctg-B) and
594 after scaffolding (scaf-CU for the assembly with sequence from 'Currot'; and scaf-
595 OR for the assembly with sequence from 'Orange Red'). After scaffolding, eight
596 chromosome-scale pseudo-molecules were obtained for each haplotype as
597 labeled by "Chrs". **c.** Haplotype validation for the two assemblies of each linkage
598 group (LG-1-8) using parent-specific *k*-mers (of 'Orange Red' and 'Currot'). With
599 each linkage group, the two assemblies could be clearly identified as either
600 'Currot'-haplotype or 'Orange Red'-haplotype using parental *k*-mers. After
601 combining the 'Currot'-related assemblies and 'Orange Red'-related assemblies
602 to genome-level, *k*-mer comparison revealed a haplotype accuracy of 99.1%. **d**.
603 Using the 'Currot'-haplotype as representative and comparing it to the assembly
604 of the double haploid Prunus ssp. reference genome (*Prunus persica*, and other

605 closely-related species; Supplementary Figure 6) revealed high levels of synteny
606 and thus implies high accuracy of the genetic map and chromosome-level
607 scaffolding.

608 **Figure 4. Structural genome variations and meiotic recombination.** Top:
609 recombination landscape created with sliding windows of 500 kb at a step of 50
610 kb with COs detected in all single pollen nuclei (with coverage over 0.1x), coupled
611 with SNP density and gene density. For x-axis, coordinates were based on the
612 haploid assembly of 'Currot'-genotype. For y-axis, all features were scaled to 1.0,
613 which stands for a maximum of 18 for recombination frequency (*cM/Mb*), 7,410
614 for SNP density and 480 for gene density. Bottom: structural variations (>50 kb)
615 identified between the two haploid assemblies. In general, crossovers are almost
616 completely absent in SVs, for example, at LG2:11.0−14.5 Mb (inversion case)
617 and LG5:16.0−18.2 Mb (translocation case).

618 **Figure 5. Non-allelic crossovers and its consequences. a.** Illustration of a non-
619 allelic crossover which results in a chromosomal anomaly. **b.** Analysis of a single-
620 pollen nuclei, which revealed a non-allelic CO resulting in the duplication of a
621 large chromosomal segment. The short-read alignments of a haploid nuclei
622 revealed a pseudo-heterozygous region with increased read coverage, which is
623 the hallmark of a long duplication specific to this genome. All other chromosomes
624 were haploid (not shown). (Top row: 'Currot' allele frequency, SNP density (in
625 sliding windows of 500 kb at a step of 50 kb), and read coverage scaled by SNP
626 density. Middle row: count of 'Currot' or 'Orange Red' alleles at SNP markers.
627 Bottom row: diagram illustrating how a non-allelic CO in transposed regions (as
628 indicated by yellow rectangles) resulted in a large duplication, i.e., the original

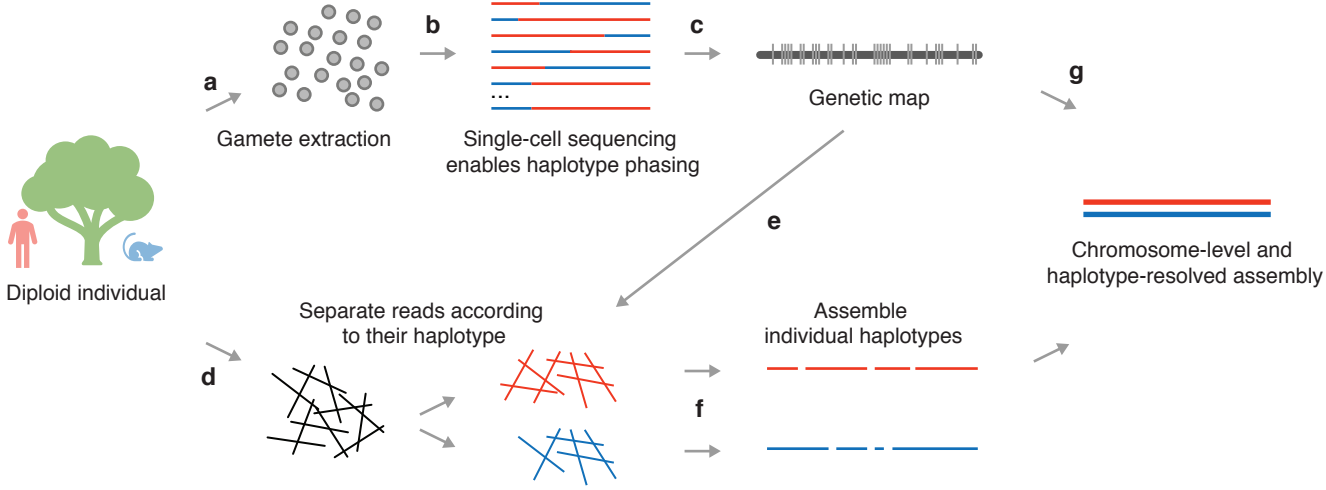
homologous chromosomal regions labelled with “4” and “5” are now part of the same newly formed chromosome.

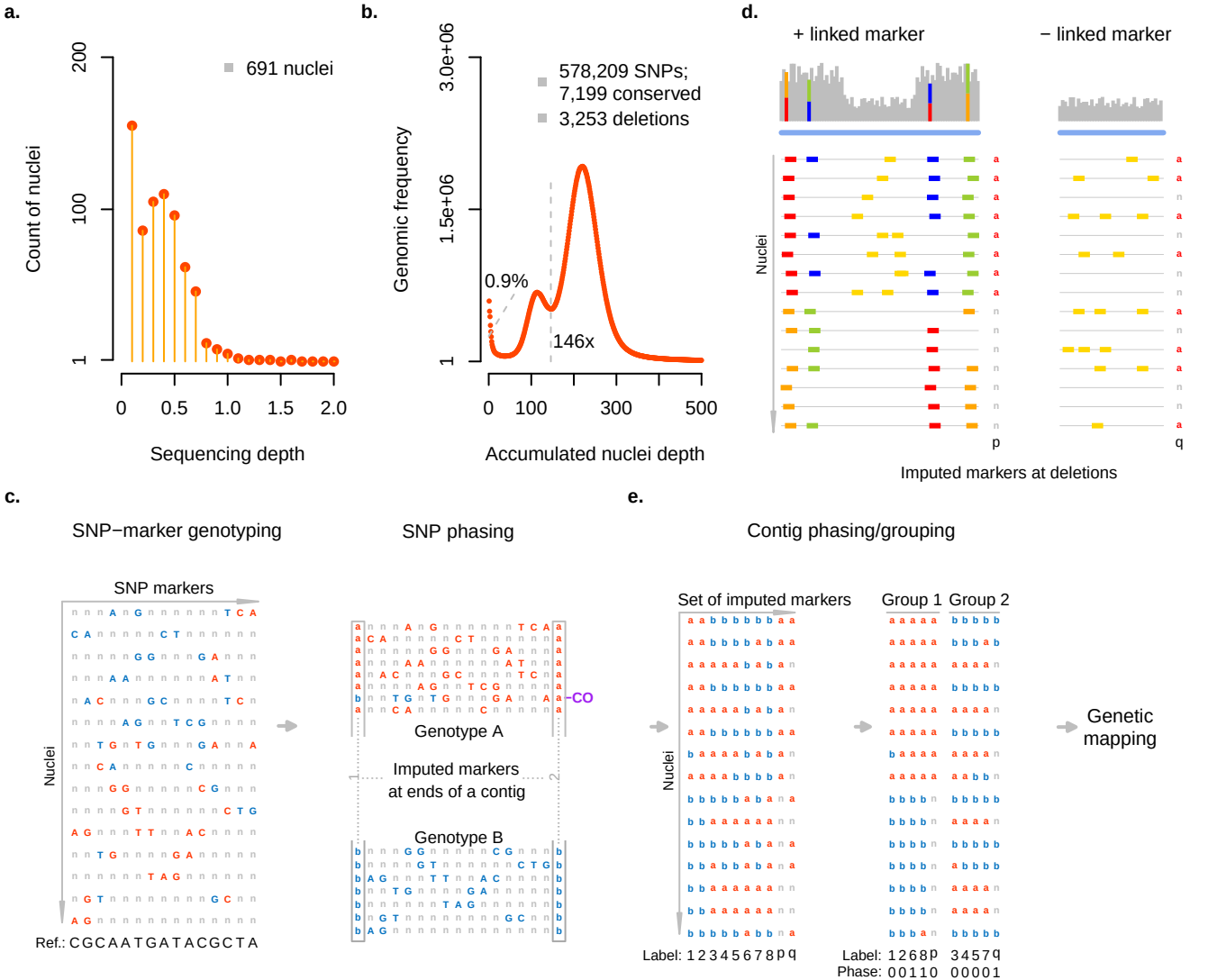
632 # Tables

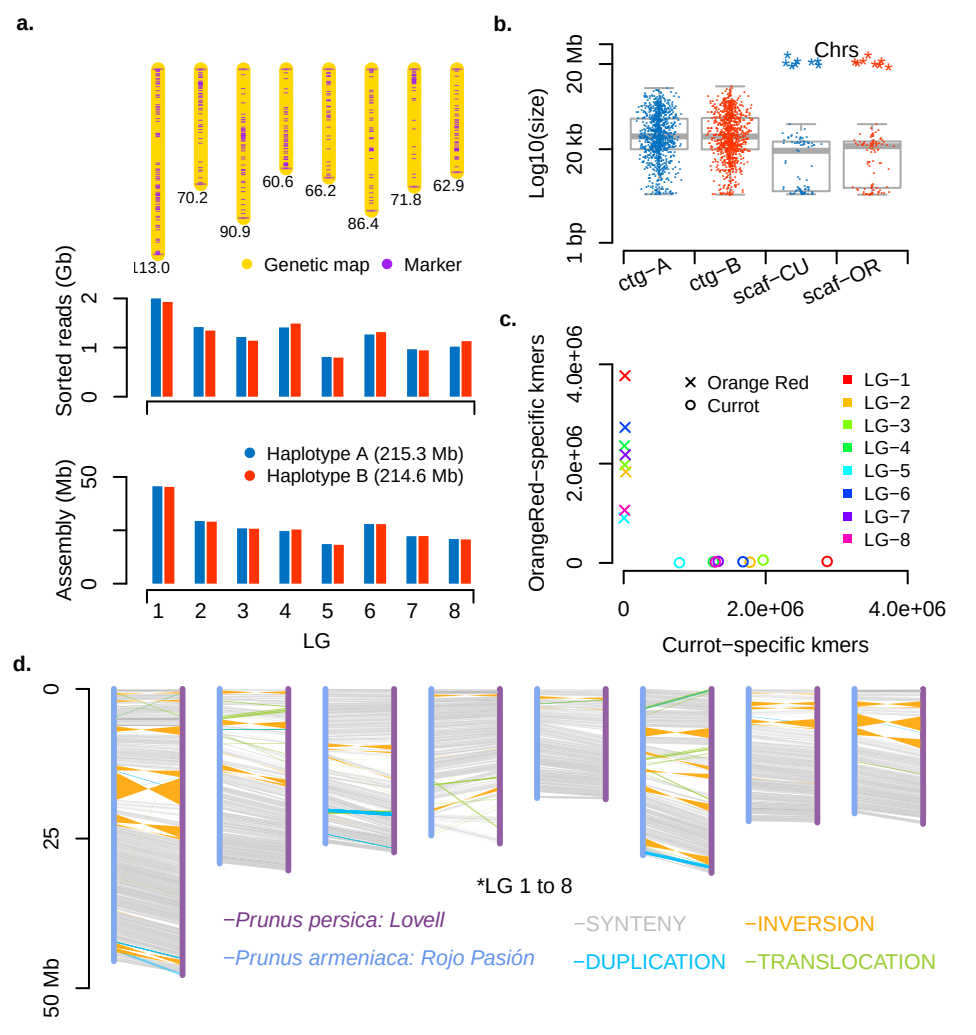
633 **Table 1 Assembly and validation statistics of two haplotype-resolved genome assemblies**

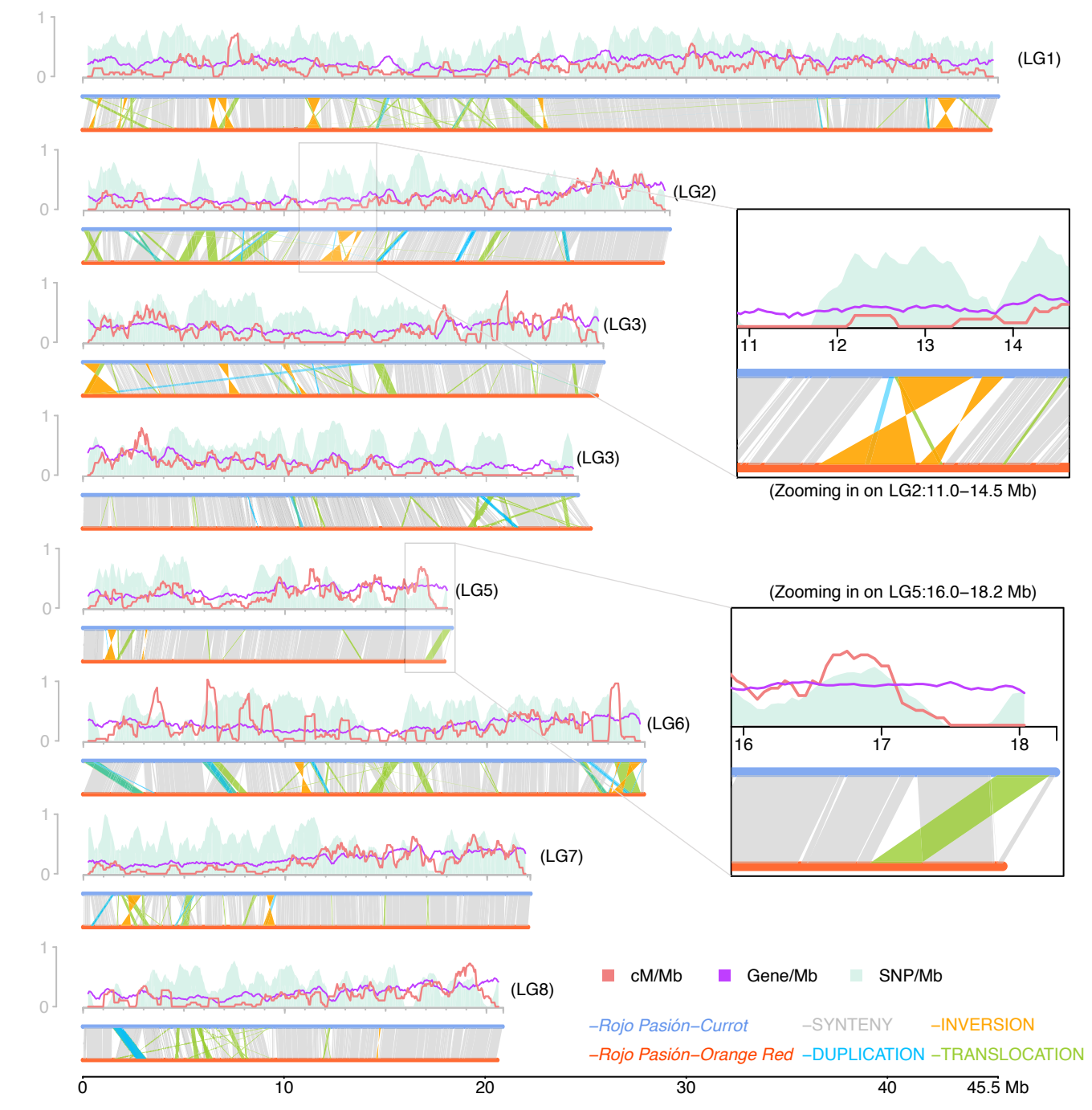
Haploid assemblies of 'Rojo Pasion'	Number of genome-specific *k*-mers common with parental WGS of		Precision in haplotyping	Size [Mb]	Chromosome scaffolds	N50 [Mb]	Protein-coding genes (Total genes)
	'Currot'	'Orange Red'					
'Currot'-haplotype	12,754,496	162,794	98.7%	216.0	8	25.8	30,661 (52,472)
'Orange Red'-haplotype	108,261	16,566,104	99.4%	215.2	8	25.5	30,378 (51,701)

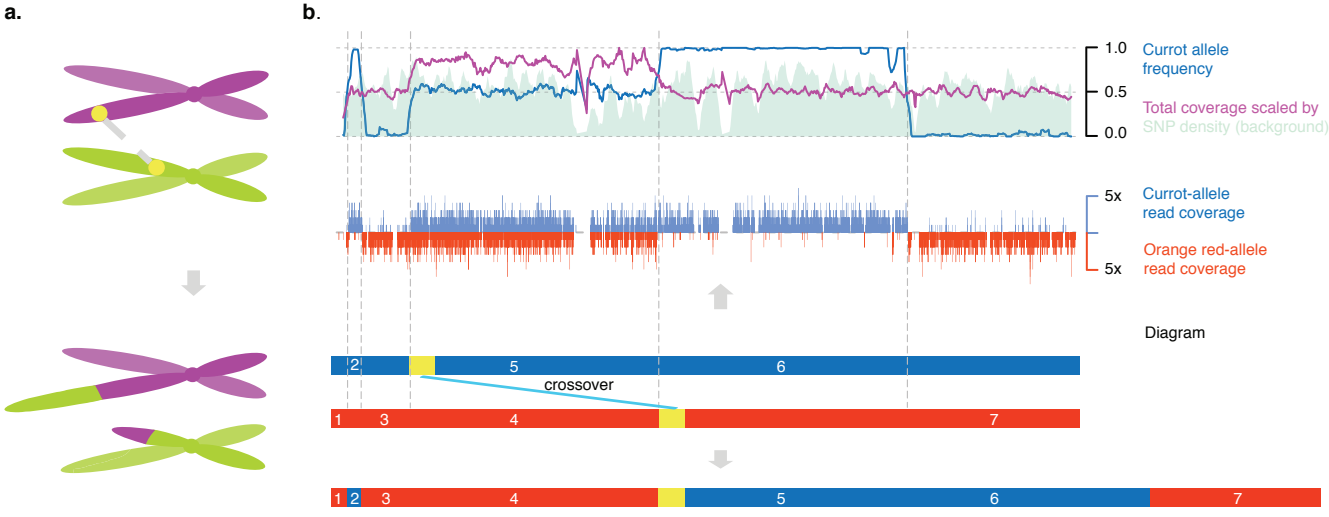
634


a
Gamete extraction
b
...
Single-cell sequencing
enables haplotype phasing
c
Genetic map
g
Chromosome-level and
haplotype-resolved assembly
Diploid individual
e
Separate reads according
to their haplotype
d
f
Assemble
individual haplotypes


a.
691 nuclei
Count of nuclei
200
100
1
0
0.5
1.0
1.5
2.0
Sequencing depth
b.
578,209 SNPs;
7,199 conserved
3,253 deletions
Genomic frequency
3.0e+06
1.5e+06
1
0.9%
146x
0
100
300
500
Accumulated nuclei depth
d.
+ linked marker
− linked marker
Nuclei
p
q
Imputed markers at deletions
c.
SNP−marker genotyping
SNP phasing
SNP markers
Nuclei
Ref.: CGCAATGATACGCTA
Genotype A
Imputed markers
at ends of a contig
Genotype B
CO
e.
Contig phasing/grouping
Set of imputed markers
Group 1
Group 2
Nuclei
Label: 12345678pq
Label: 1268p 3457q
Phase: 00110 00001
Genetic
mapping


a.
113.0
70.2
90.9
60.6
66.2
86.4
71.8
62.9
Genetic map
Marker
Sorted reads (Gb)
Haplotype A (215.3 Mb)
Haplotype B (214.6 Mb)
Assembly (Mb)
LG
b.
Chrs
Log10(size)
20 Mb
20 kb
1 bp
ctg−A
ctg−B
scaf−CU
scaf−OR
c.
Orange Red
Currot
LG−1
LG−2
LG−3
LG−4
LG−5
LG−6
LG−7
LG−8
OrangeRed−specific kmers
Currot−specific kmers
d.
*LG 1 to 8
−Prunus persica: Lovell
−Prunus armeniaca: Rojo Pasión
−SYNTENY
−INVERSION
−DUPLICATION
−TRANSLOCATION
50 Mb


(LG1)
(LG2)
(LG3)
(LG3)
(LG5)
(LG6)
(LG7)
(LG8)
(Zooming in on LG2:11.0–14.5 Mb)
(Zooming in on LG5:16.0–18.2 Mb)
11
12
13
14
16
17
18
cM/Mb
Gene/Mb
SNP/Mb
–Rojo Pasión–Currot
–SYNTENY
–INVERSION
–Rojo Pasión–Orange Red
–DUPLICATION
–TRANSLOCATION
0
10
20
30
40
45.5 Mb


a.
b.
1.0
0.5
0.0
Currot allele frequency
Total coverage scaled by SNP density (background)
5x
Currot-allele read coverage
Orange red-allele read coverage
5x
Diagram
2
5
6
crossover
1
3
4
7
1
2
3
4
5
6
7