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

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

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

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

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

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

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

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

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

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

In contrast to conventional diploid genome assemblies where the two haplotypes are merged into one artificial consensus sequence, separate haploid assemblies allow for the analysis of haplotype diversity. Comparing the two haplotype assemblies of ‘Rojo Pasion’ using *SyRI*[12] allowed us to gain first insights into the haplotype diversity within an individual apricot tree. Despite high levels of synteny, the two assemblies revealed large-scale rearrangements (23 inversions, 1,132 translocation/transpositions and 2,477 distal duplications) between the haplotypes making up more than 15% of the assembled sequence (38.3 and 46.2 Mb in each of assemblies; Supplementary Table 1). Using a comprehensive RNA-seq dataset sequenced from multiple tissues of ‘Rojo Pasion’ including reproductive buds, vegetative buds, flowers, leaves, fruits (seeds removed) and barks as well as a published apricot RNA-seq dataset[23], we predicted 30,378 and 30,661 protein-coding genes within each of the haplotypes (with an annotation completeness of 96.4% according to a BUSCO[27] analysis). Mirroring the huge differences in the sequences, we found the vast amount of 942 and 865 expressed, haplotype-specific genes in each of the haplotypes (Methods; Supplementary Tables 2-3). Such deep insights into the differences between the haplotypes, which are only enabled by chromosome-level and haplotype-resolved assemblies, will generally be of high value for the analysis of agronomically relevant variation.

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

Online Methods

DNA extraction, Illumina/PacBio library preparation and sequencing

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

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

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

Dormant shoots of ‘Rojo Pasion’ bearing developed flower buds were collected in Murcia, Spain. Then, the shoots were shipped at 4 °C to MPIPZ (Cologne, Germany) and were grown in long-day conditions in the greenhouse. Flowers at the pre-anthesis stage were frozen in liquid nitrogen. Anthers from ten ‘Rojo Pasion’[17] flowers were extracted with forceps and submerged in woody pollen buffer (WPB)[28]. Around 500,000 pollen grains were extracted from anthers by vortexing them in WPB. The nuclei were isolated from the pollen using a modified bursting method[20]. Isolated pollen was prefiltered (100µm) and bursted (30um) using Celltrics^{TM} sieves and woody pollen buffer. The nuclei were then stained with propidium iodide (PI) at 50 µg/mL just before sorting and counting by flow cytometry to remove pollen grain debris using a BD FACSAria Fusion^{TM} with high-speed sort settings (70 µm nozzle and 70 PSI sheath pressure) and 0.9% NaCl as sheath fluid. The nuclei were identified by PI fluorescence, light scattering, and autofluorescence characteristics (Supplementary Figure 4). A total of 12,600 nuclei were counted and collected in a solution of 4.2 μL phosphate-buffered saline with 0.1% bovine serum albumin.

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

Genome size estimation

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

Initial diploid-genome assembly and curation

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

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

SNP marker identification

After trimming off 10x barcodes and hexamers, all pooled Illumina reads from the 10x sc-CNV libraries (61.7 Gb) were re-aligned to the curated haploid assembly using *bowtie2*[30]. With 87.2% reads aligned, 989,132 raw SNPs were called with *samtools and bcftools*[32]. Three criteria were used to select potential allelic SNPs (578,209), including i) the alternative allele frequency must be between 0.38 to 0.62; ii) the alternative allele must be carried by 60-140 reads; iii) the total sequencing depth at a SNP must be between 120-280x (as compared with genome-wide mode depth of 208x; Fig. 2b).

Deletion marker identification and genotyping

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

Variant phasing and CO identification

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

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

Ploidy evaluation of single-cell sequencing

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

Imputation of virtual markers at ends of contigs

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

Linkage grouping and genetic mapping

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

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

Haplotype-specific PacBio read classification

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

Haplotype-genome assembly and scaffolding

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

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

Haplotype evaluation on the two haploid assemblies

The genotypes of the sixteen assemblies were firstly identified by comparing *k*-mers in each assembly with Illumina WGS of the parental cultivar (*k*=21; Fig. 3c). Although evaluation can always be performed in each linkage group, we combined the eight linkage-group-wise assemblies for 'Currot'-genotype and the other eight for 'Orange Red'-genotype, respectively.

After polishing the assemblies respectively with the classified 'Currot'-genotype and 'Orange Red'-genotype PacBio reads using *apollo*[37], we built up two sets of haplotype-specific *k*-mers from the assemblies, r_C and r_O. Correspondingly, a set of 'Currot'-specific *k*-mers (with coverage from 10 to 60x), p_C, was selected from the parental Illumina WGS that did not exist in 'Orange Red' short reads (coverage over 1x) but in 'Rojo Pasion' pollen short reads (coverage from 10 to 300x); similarly, a set of 'Orange Red'-specific *k*-mers, p_O, was also collected. Then we intersected r_C and r_O with p_C and p_O respectively, leading to four subsets $r_C \cap p_C$, $r_C \cap p_O$, $r_O \cap p_C$, and $r_O \cap p_O$. This calculation gave an average haplotyping accuracy of 99.1% (Table 1). All *k*-mer processing (counting, intersecting and difference finding) were performed with *KMC*[38]. After haplotype validation, the assemblies were further polished with the respective parental short read alignment using *pilon*[39] (with options '--fix bases --mindepth 0.85'). The final haplotype assembly sizes were 216.0 and 215.2 Mb for 'Currot'-genotype (93 scaffolds, N50: 25.8 Mb) and 'Orange Red'-genotype (104 scaffolds, N50: 25.5 Mb), respectively (Table 1). Note, the eight main chromosome-level scaffolds of each haplotype made up ~99% of the respective assembly.

Genome annotation

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

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

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

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

Genome assembly comparison

All genome assemblies, including 'Rojo Pasion' haplotypes, 'Chuanzhihong' apricot (*Prunus armeniaca*)[23], Japanese apricot (*Prunus mume*)[25] and 'Lovell' peach (*Prunus *persica*)[26], were aligned to each other using *nucmer* from the *MUMmer4*[51] toolbox with parameters '-max -l 40 -g 90 -b 100 -c 200'. The alignments were further filtered for alignment length (>100 bp) and identity (>90%), with which structural rearrangements and local variations were identified using *SyRI*[12]. To follow the nomenclature of the Prunus community, the 'Rojo Pasion' chromosomes were numbered according to the numbering in 'Lovell' peach[26].

Crossover identification and landscape creation

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

REFERENCES

1. Korlach, J. *et al.* De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *Gigascience* **6**, 1–16 (2017).

2. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).

3. Yang, H., Chen, X. & Wong, W. H. Completely phased genome sequencing through chromosome sorting. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 12–17 (2011).

4. Selvaraj, S., Dixon, J. R., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118 (2013).

5. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).

6. Falconer, E. & Lansdorp, P. M. *Strand-seq*: A unifying tool for studies of chromosome segregation. *Semin. Cell Dev. Biol.* **24**, 643–652 (2013).

7. Zhang, X., Wu, R., Wang, Y., Yu, J. & Tang, H. Unzipping haplotypes in diploid and polyploid genomes. *Comput. Struct. Biotechnol. J.* **18**, 66–72 (2020).

8. Li, R. *et al.* Inference of Chromosome-length Haplotypes Using Genomic Data of Three to Five Single Gametes. *bioRxiv* 361873 (2018). doi:10.1101/361873

9. Kirkness, E. F. *et al.* Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome Res.* **23**, 826–832 (2013).

10. Shi, D. *et al.* Single-pollen-cell sequencing for gamete-based phased diploid genome assembly in plants. *Genome Res.* 1–11 (2019).

11. Wu, J. *et al.* The genome of the pear (Pyrus bretschneideri Rehd.). *Genome Res.* **23**, 396–408 (2013).

12. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. *SyRI*: finding genomic rearrangements and local sequence differences from whole-genome assemblies.

Genome Biol. **20**, 1–13 (2019).

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

25. Zhang, Q. *et al.* The genome of Prunus mume. *Nat. Commun.* **3**, 1–8 (2012).

26. Verde, I. *et al.* The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* **45**, 487–494 (2013).

27. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

28. Loureiro, J. Two new nuclear isolation buffers for plant DNA flow cytometry: A test with 37 species. *Ann. Bot.* 875–888 (2007).

29. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

30. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, 1–10 (2009).

31. Roach, M. J., Schmidt, S. a. & Borneman, A. R. *Purge Haplotigs*: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 1–10 (2018).

32. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

33. Quinlan, A. R. & Hall, I. M. *BEDTools*: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

34. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

35. Li, H. *Minimap2*: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

36. Alonge, M. *et al. RaGOO*: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 1–17 (2019).

37. Firtina, C. *et al.* Apollo: A Sequencing-Technology-Independent, Scalable, and Accurate Assembly Polishing Algorithm. *Bioinformatics* 1–10 (2020).

38. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).

39. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, 1–14 (2014).

40. Stanke, M. *et al.* AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, 435–439 (2006).

41. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).

42. Johnson, A. D. *et al.* SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).

43. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 1–11 (2005).

44. Kim, D., Langmead, B. & Salzberg1, S. L. HISAT: a fast spliced aligner with low memory requirements Daehwan HHS Public Access. *Nat. Methods* **12**, 357–360 (2015).

45. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

46. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22 (2008).

47. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).

48. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

49. Keller, O., Odronitz, F., Stanke, M., Kollmar, M. & Waack, S. Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs 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).

Figure legends

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

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

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

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

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

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

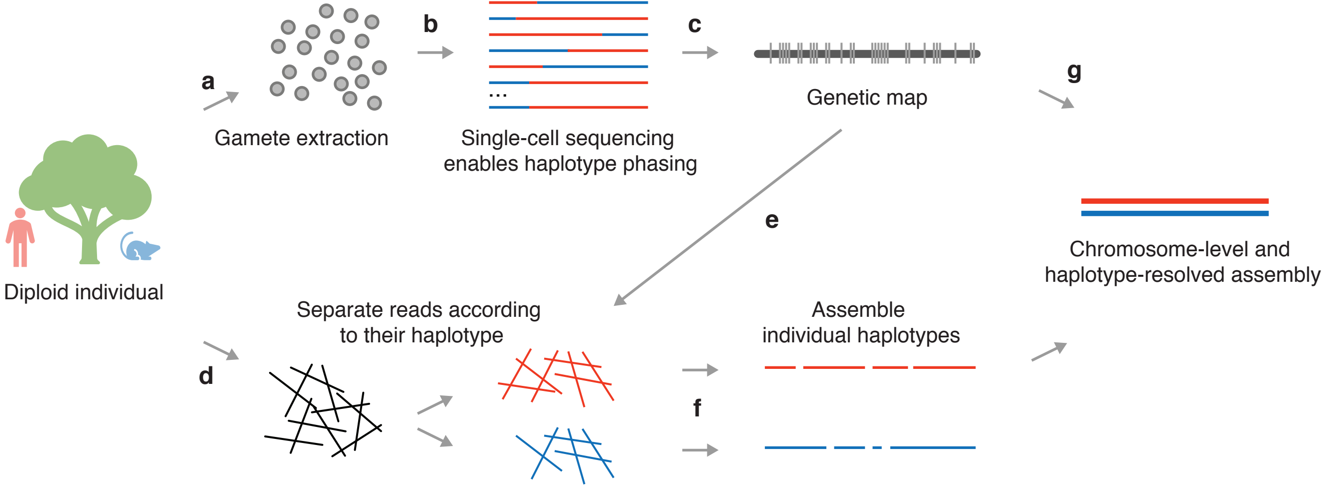
Figure 5. Non-allelic crossovers and its consequences. a. Illustration of a non-allelic crossover which results in a chromosomal anomaly. **b.** Analysis of a single-pollen nuclei, which revealed a non-allelic CO resulting in the duplication of a large chromosomal segment. The short-read alignments of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

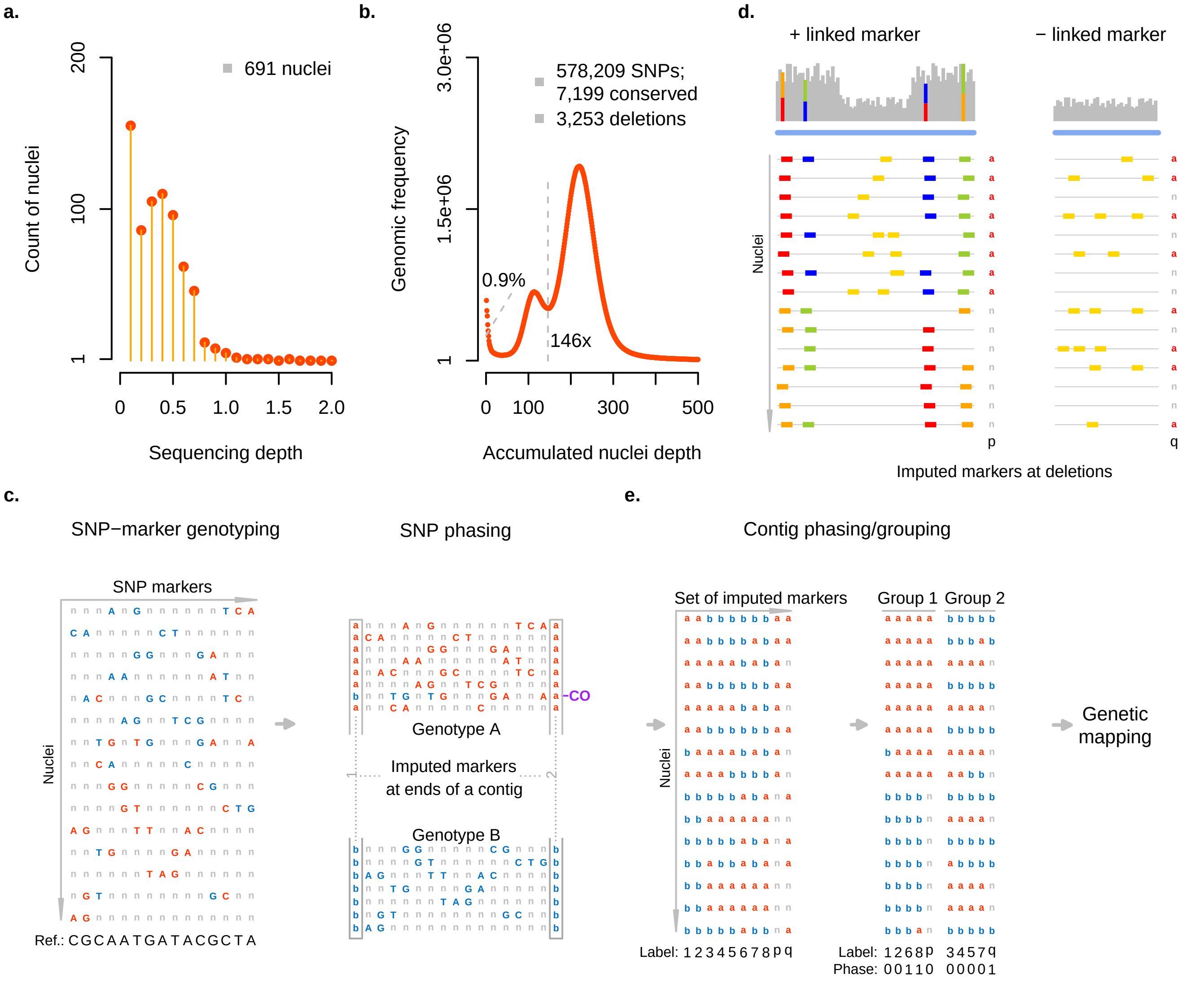
homologous chromosomal regions labelled with “4” and “5” are now part of the same newly formed chromosome.

Tables

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)


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



```
p q
a a
a a
a n
a a
a n
a a
a n
a n
n a
n n
n a
n a
n n
n n
n a
```

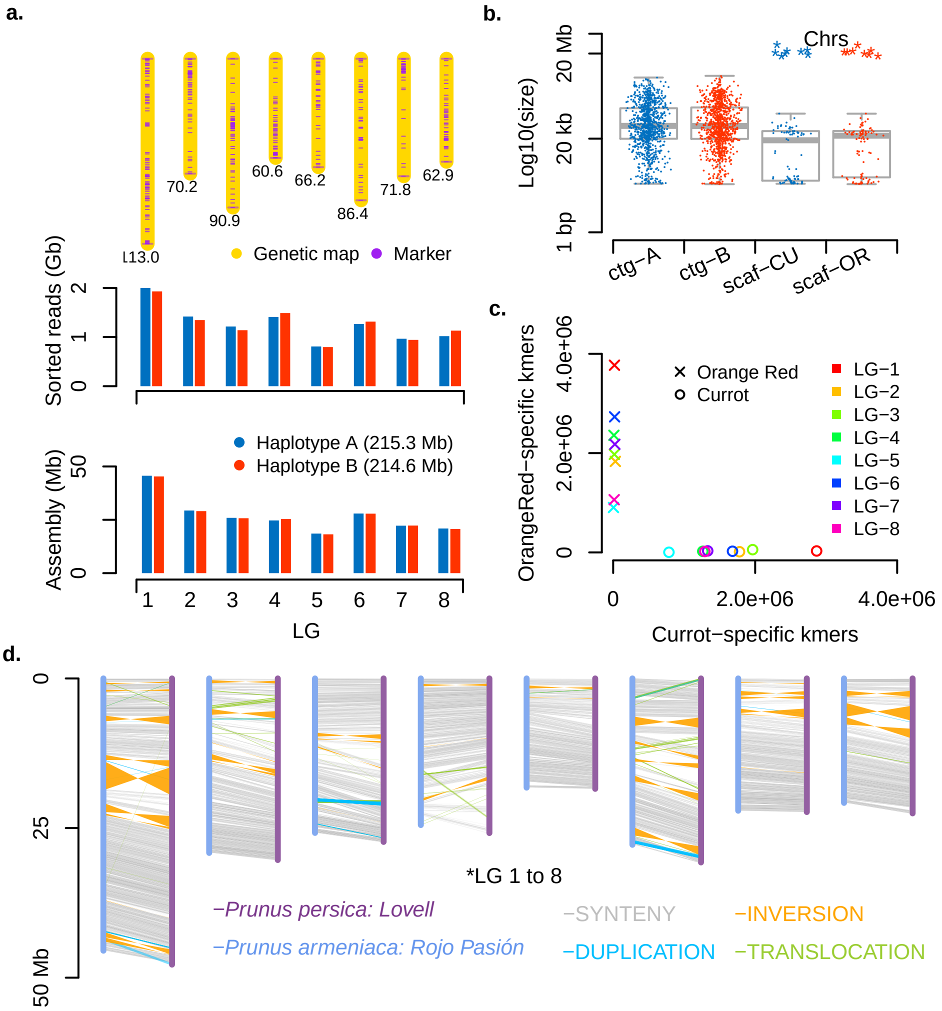
```
n n n A n G n n n n n n T C A
C A n n n n n C T n n n n n n
n n n n n G G n n n G A n n n
n n n A A n n n n n n A T n n
n A C n n n G C n n n n T C n
n n n n A G n n T C G n n n n
n n T G n T G n n n G A n n A
n n C A n n n n n C n n n n n
n n n G G n n n n n C G n n n
n n n n G T n n n n n n C T G
A G n n n T T n n A C n n n n
n n T G n n n n G A n n n n n
n n n n n n T A G n n n n n n
n G T n n n n n n n n G C n n
A G n n n n n n n n n n n n n
Ref.: C G C A A T G A T A C G C T A
```

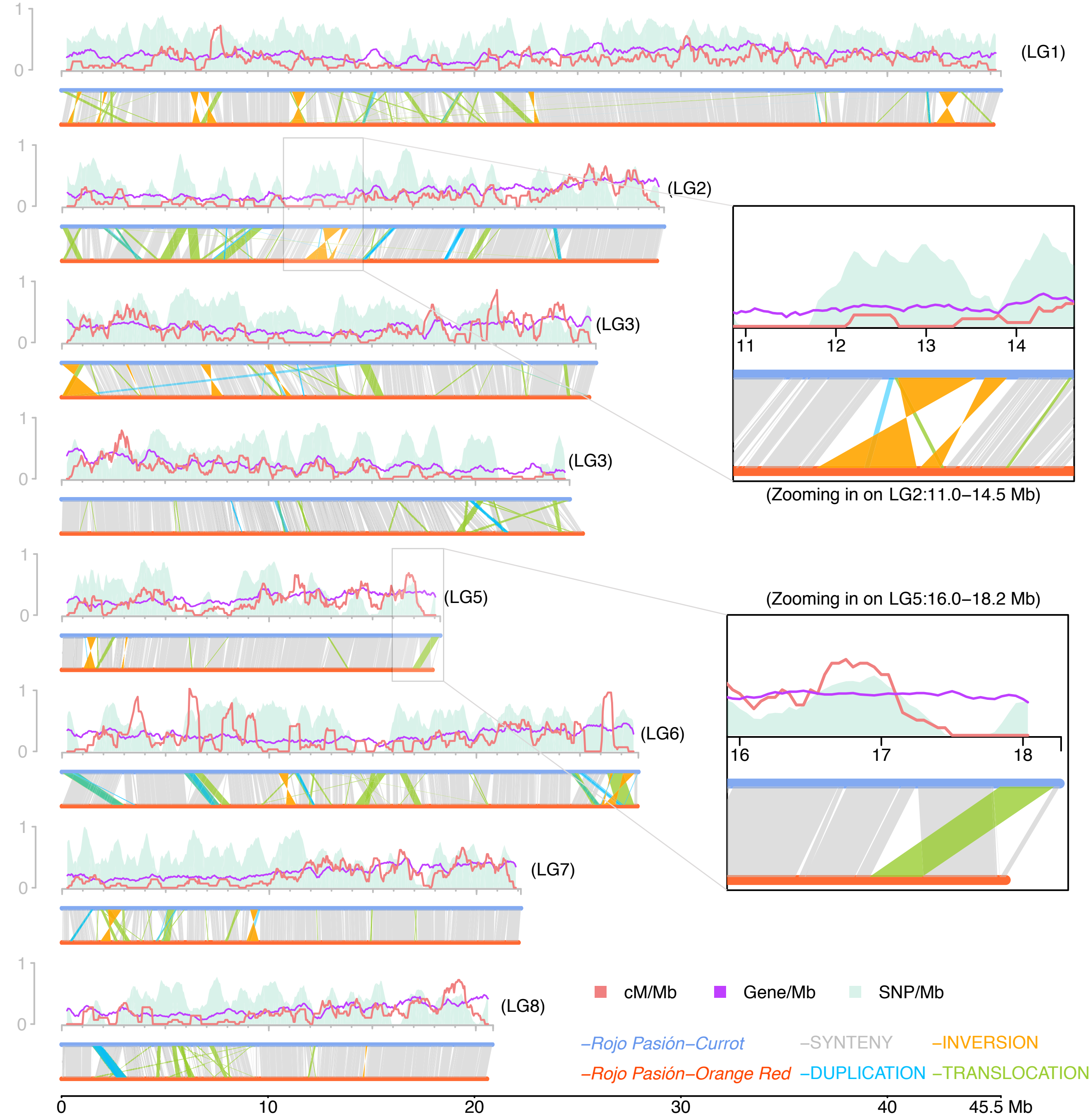
```
a n n n A n G n n n n n n T C A a
a C A n n n n n C T n n n n n n a
a n n n n n G G n n n G A n n n a
a n n n A A n n n n n n A T n n a
a n A C n n n G C n n n n T C n a
a n n n n A G n n T C G n n n n a
b n n T G n T G n n n G A n n A a
a n n C A n n n n n C n n n n n a
```

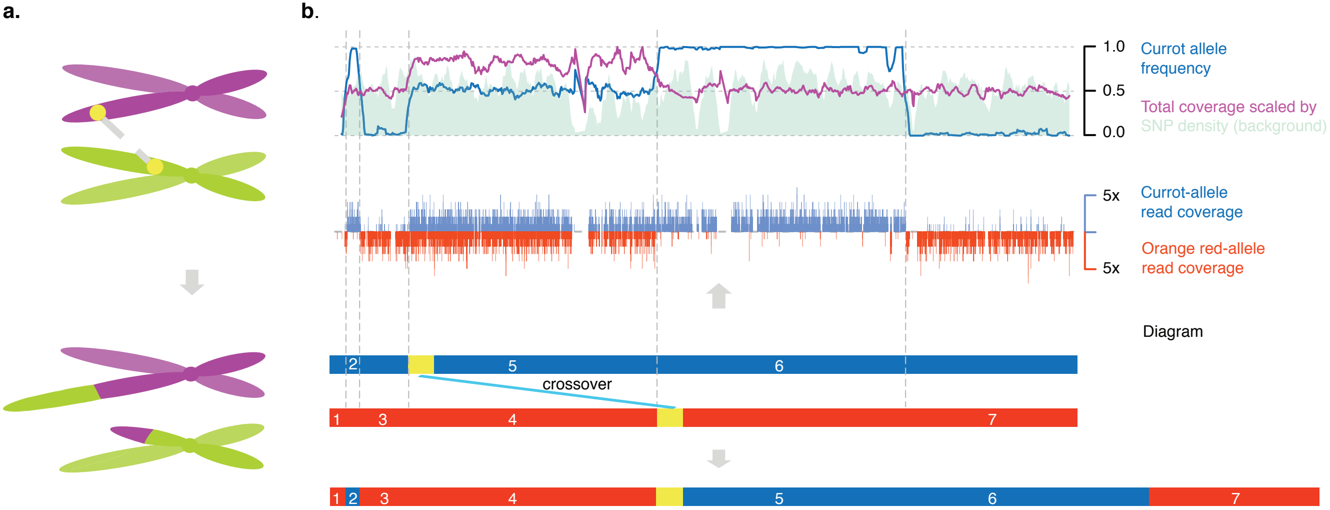
```
b n n n G G n n n n n C G n n n b
b n n n n G T n n n n n n C T G b
b A G n n n T T n n A C n n n n b
b n n T G n n n n G A n n n n n b
b n n n n n n T A G n n n n n n b
b n G T n n n n n n n n G C n n b
b A G n n n n n n n n n n n n n b
```

```
a a b b b b b b a a
a a b b b b a b a a
a a a a a b a b a n
a a b b b b b b a a
a a a a a b a b a n
a a b b b b b b a a
b a a a a b a b a n
a a a a b b b b a n
b b b b b a b a n a
b b a a a a a a n n
b b b b b a b a n a
b b a b b a b a n a
b b a a a a a a n n
b b a a a a a a n n
b b b b b a b b n a
Label: 1 2 3 4 5 6 7 8 p q
```

```
a a a a a   b b b b b
a a a a a   b b b a b
a a a a a   a a a a n
a a a a a   b b b b b
a a a a a   a a a a n
a a a a a   b b b b b
b a a a a   a a a a n
a a a a a   a a b b n
b b b b n   b b b b b
b b b b n   a a a a n
b b b b n   b b b b b
b b b b n   a b b b b
b b b b n   a a a a n
b b b b n   a a a a n
b b b a n   b b b b b
Label: 1 2 6 8 p   3 4 5 7 q
Phase: 0 0 1 1 0   0 0 0 0 1
```


a.
113.0
70.2
90.9
60.6
66.2
86.4
71.8
62.9
● Genetic map ● Marker
Sorted reads (Gb)
0
1
2
● Haplotype A (215.3 Mb)
● Haplotype B (214.6 Mb)
Assembly (Mb)
0
50
1
2
3
4
5
6
7
8
LG
b.
Chrs
Log10(size)
1 bp
20 kb
20 Mb
ctg−A
ctg−B
scaf−CU
scaf−OR
c.
OrangeRed−specific kmers
0
2.0e+06
4.0e+06
× Orange Red
○ Currot
🟥 LG−1
🟧 LG−2
🟩 LG−3
🟩 LG−4
🟦 LG−5
🟦 LG−6
🟪 LG−7
🟪 LG−8
0
2.0e+06
4.0e+06
Currot−specific kmers
d.
0
25
50 Mb
*LG 1 to 8
−Prunus persica: Lovell
−Prunus armeniaca: Rojo Pasión
−SYNTENY
−DUPLICATION
−INVERSION
−TRANSLOCATION


1
0
(LG1)
(LG2)
(LG3)
(LG3)
(LG5)
(LG6)
(LG7)
(LG8)
11
12
13
14
(Zooming in on LG2:11.0–14.5 Mb)
(Zooming in on LG5:16.0–18.2 Mb)
16
17
18
cM/Mb
Gene/Mb
SNP/Mb
–Rojo Pasión–Currot
–Rojo Pasión–Orange Red
–SYNTENY
–DUPLICATION
–INVERSION
–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
5x
Currot-allele read coverage
Orange red-allele read coverage
Diagram
2
5
6
crossover
1
3
4
7
1
2
3
4
5
6
7