Campoy *et al. Genome Biology* (2020) 21:306
https://doi.org/10.1186/s13059-020-02235-5

Genome Biology

METHOD Open Access

Gamete binning: 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], Nan Wang[4], Manuel Rubio[5], Chang Liu[4,6], Christian Kukat[3], David Ruiz[5], Bruno Huettel[7] and Korbinian Schneeberger[1,2*]

* Correspondence: schneeberger@mpipz.mpg.de
† José A. Campoy and Hequan Sun contributed equally to this work.
[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
Full list of author information is available at the end of the article

Abstract

Generating chromosome-level, haplotype-resolved assemblies of heterozygous genomes remains challenging. To address this, we developed gamete binning, a method based on single-cell sequencing of haploid gametes enabling separation of the whole-genome sequencing reads into haplotype-specific reads sets. After assembling the reads of each haplotype, the contigs are scaffolded to chromosome level using a genetic map derived from the gametes. We assemble the two genomes of a diploid apricot tree based on whole-genome sequencing of 445 individual pollen grains. The two haplotype assemblies (N50: 25.5 and 25.8 Mb) feature a haplotyping precision of greater than 99% and are accurately scaffolded to chromosome-level.

Keywords: Single-cell sequencing, Haplotype-resolved assembly, Haplotyping, Phasing, De novo assembly

Introduction

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.

Recent alternatives allowing for the assembly of both haplotypes include chromosome sorting [3], Strand-seq [4–6], and high-throughput chromosome conformation capture (Hi-C) [7–13] sequencing. Chromosome sorting separates individual chromosomes before sequencing and thus enables the sequencing and assembly of individual haplotypes. However, sorting of particular chromosomes may not always be possible if they cannot be discriminated based on their fluorescence intensity or light scatter [14] and may need tedious generation of specific lines for sorting [15]. The more recent method Strand-seq is a single-cell technique that requires neither parents nor gametes which can be potentially used to cluster long sequencing reads by chromosome, phase haplotypes, and scaffold using genetic map techniques; however, the difficulty for generating Strand-seq data has limited its application to a narrow number of model species. In contrast, the analysis of the chromosome conformation, including Hi-C technologies which enable the detection of chromatin interactions at an unprecedented scale, has been successfully applied for haplotype phasing and genome scaffolding for a wide range of species [7, 9–11, 13, 16]. However, despite its simple application, Hi-C-based phasing can be error prone due to some weaknesses in defining the alleles that distinguish haplotypes, which in turn can lead to haplotype switch errors [10] and result in mis-scaffolding of small contigs due to the lack of sufficient informative connections to other contigs [8, 11, 12]. Also the reconstruction of whole chromosomes' structures can be error-prone as already one local mis-scaffolding is sufficient to introduce severe mis-assemblies like falsely joining chromosome arms [9]. It is therefore necessary to carefully inspect assemblies that rely on Hi-C for phasing or scaffolding to identify errors, which in turn require correction based on additional evidence including, for example, the integration of genetic maps [8, 9].

An elegant alternative for haplotype phasing, called 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 are unknown [17]. 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 [18–21].

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 [22, 23]. While recent improvements in long DNA molecule sequencing [24] and as mentioned above in Hi-C data generation promise the assembly of telomere-to-telomere contigs, genetic maps can reliably help to resolve mis-assemblies and guide chromosome-level scaffolding [9]. The generation of genetic maps, however, relies on a substantial amount of meiotic recombination which usually implies the genotyping of hundreds of recombinant genomes [25, 26]. Creating and genotyping sufficiently large populations is not possible in some species (like many of the mammals including humans), and for those species for which it is possible it can be time-consuming and costly and may post great challenges if the individuals show long juvenility or sterility [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

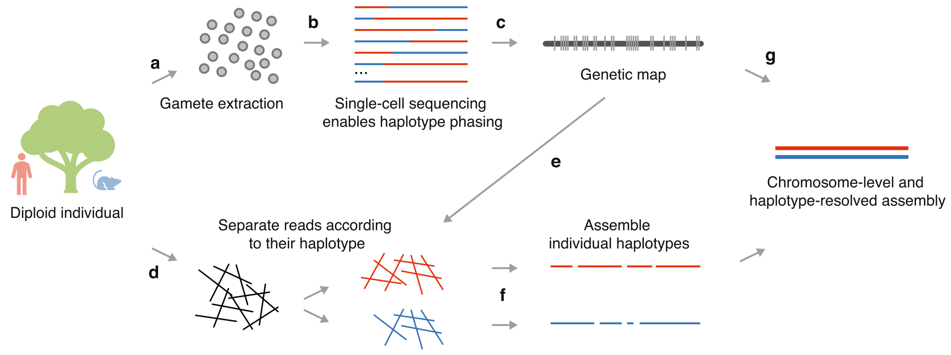

Fig. 1 Overview of gamete binning. (a) Extraction of gamete nuclei. (b) Single-cell genome sequencing of haploid gametes and haplotype phasing. (c) Genetic map construction based on the recombination patterns in the gamete genomes. (d) Long-read sequencing of somatic material. (e) Separation of long 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

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.

Results

Preliminary diploid-genome assembly

We used gamete binning to assemble the two haploid genomes of a specific, diploid apricot tree (*Prunus armeniaca*; cultivar "Rojo Pasión" [27]), which grows in Murcia, southeastern Spain (Additional file 1: Fig. S1). We first performed a preliminary de novo genome assembly using *Canu* [28] with 19.9-Gb long reads (PacBio, Additional file 1: Fig. S2) derived from DNA extracted from fruits and corresponding to 82x genome coverage according to a genome size of ~ 242.5 Mb estimated by *findGSE* [29] (see the "Online methods" section; Additional file 1: Fig. S3). 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 (see the "Online methods" section).

High-throughput single-cell sequencing of pollen

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 [30] (Fig. 1a; see the "Online methods" section). 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 6300 nuclei (Additional file 1: Figs. S1a-d; Additional file 1: Fig. S4; see the "Online methods" section). 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 5000 read pairs (see the "Online methods" section; Fig. 2a).

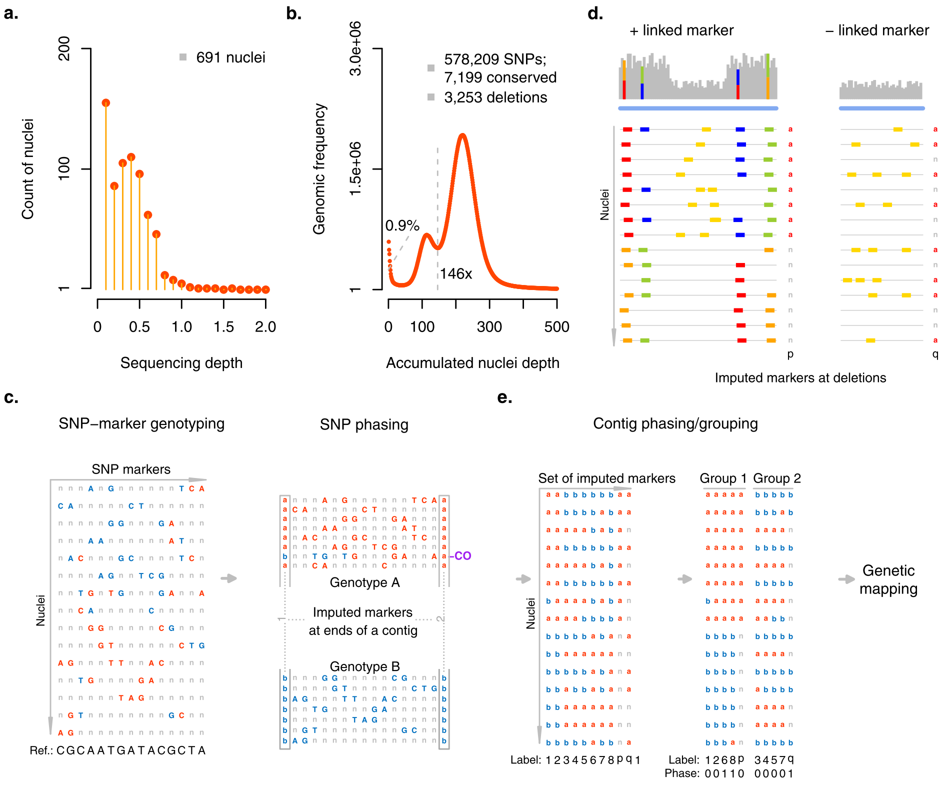

Fig. 2 Single-pollen nuclei sequencing, variant phasing and genetic mapping. **a** Sequencing depths of 691 pollen nuclei. **b** Sequencing depth histogram of pooled pollen short reads. The left-most peak revealed 0.9% of the genome that were not well covered in the pollen read sets (i.e., ≤ 5x). The 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 3253 deletions and 7199 conserved regions (see the "Online methods" section). **c** SNP phasing along contigs. Genotyping was first performed for each individual nucleus 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 of 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 4.0*) led to linkage group-wise genetic maps

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 [31] (Additional file 1: Fig. S5a-c; see the "Online methods" section). Thus, we selected a 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; Additional file 1: Fig. S5d).

Haplotype phasing and genetic mapping

With short-read alignments, we identified 578,209 heterozygous SNPs on 702 contigs with a total length of 218.0 Mb (Fig. 2b; see the "Online methods" section). Even though this implied one 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 two 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 (see the "Online methods" section). While 7199 regions (74.5 Mb) were shared between the haplotypes (and were labeled as conserved), we found that 3253 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* [32] (Figs. 2e and 3a) (see the "Online methods" section).

Haplotype-specific long-read separation and chromosome-level assembly

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; see the "Online methods" section). 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* [33], which led to 16 haplotype-specific chromosome assemblies with average N50 values ranging from 662.3 to 664.6 kb (Table 1; see the "Online methods" section). 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; see the "Online methods" section).

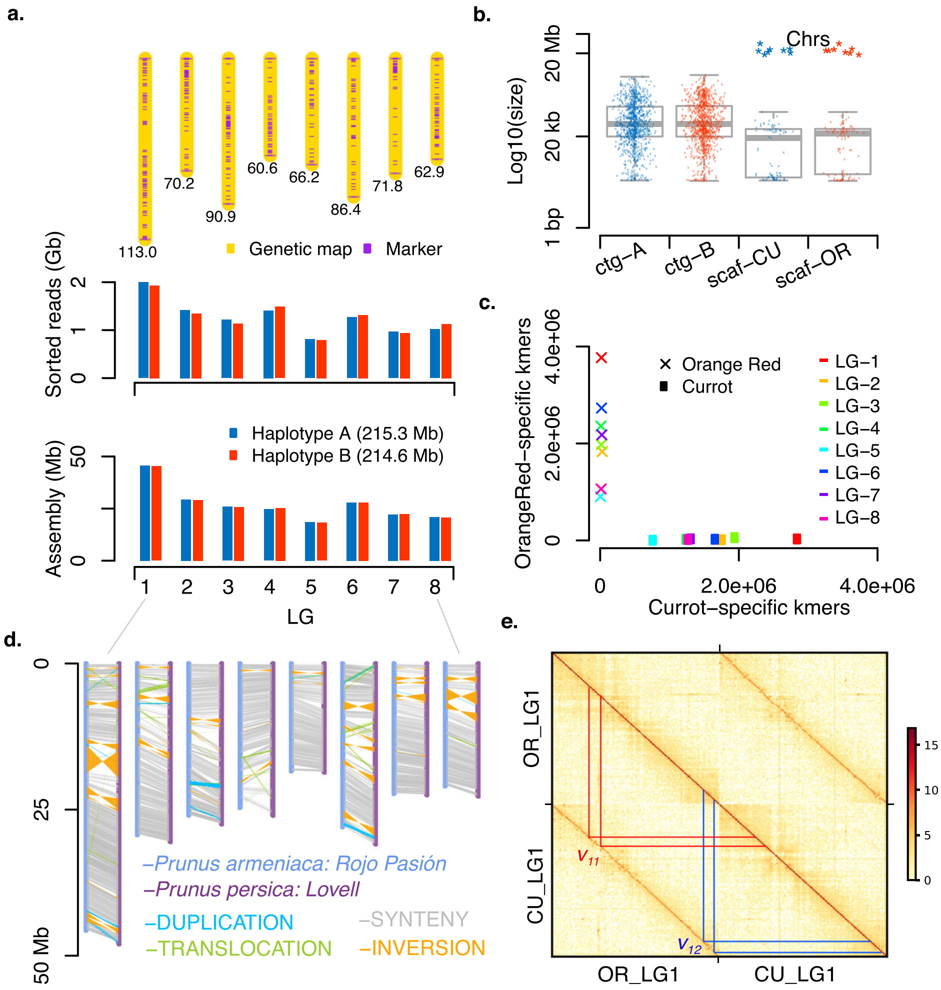

Fig. 3 Genetic mapping, haplotype-specific assembly and validation. **a** Top: Genetic map with a total genetic length of 622.0 cM (see the "Online methods" section). Middle: up to 2-Gb reads were assigned to one of the two haplotypes of each linkage group. Bottom: a combination of haplotype-A/B linkage 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; Additional file 1: Fig. S8) revealed high levels of synteny and thus implies high accuracy of the genetic map and chromosome-level scaffolding. **e** Hi-C contact map based on the assemblies of the two haplotypes of chromosome 1 (Currot (CU) and Orange Red (OR)) at a resolution of 300 kb. The contact signal showed a high contiguity within the haplotypes (main diagonal line) and confirmed two large inversions (v_{11} and v_{12}) which we observed in the assembled sequence of the two haplotypes. (See Additional file 1: Figs. S9–15 for chromosomes 2–8)

To assess haplotype accuracy, we additionally whole-genome sequenced the parental cultivars of "Rojo Pasión" 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

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

Haploid assemblies of "Rojo Pasión"	Genome-specific *k*-mers common with parental WGS		Precision in haplotyping	Size [Mb]	Chromosome scaffolds	Contig N50 [Mb]	N50 [Mb]	Protein-coding genes (total genes)
	"Currot"	"Orange Red"						
"Currot" haplotype	12,983,934	129,874	99.1%	216.0	8	0.662	25.8	30,661 (52,472)
"Orange Red" haplotype	81,422	16,807,958	99.5%	215.2	8	0.664	25.5	30,378 (51,701)

The eight main chromosome-level scaffolds of each haplotype made up ~ 99% of the respective assembly

parental genome illustrating that almost all of the variation was correctly assigned to haplotypes (Fig. 3c; Table 1; see the "Online methods" section). Having proved the haplotype accuracy, the assemblies were polished resulting in final haplotype assemblies. The final haplotype assembly sizes were 216.0 and 215.2 Mb for "Currot"-genotype (8 scaffolds, N50: 25.8 Mb) and "Orange Red"-genotype (8 scaffolds, N50: 25.5 Mb), respectively (Table 1).

We estimated the overall assembly quality by comparing the *k*-mer distributions of the assemblies and the Illumina short-read sets of the focal and parental using KAT [34] and *Merqury* [35]. Both haplotype genome assembly showed very high quality values ($QV > 36$) and the absence of allelic duplications between the haplotypes, though a fraction of ~ 7% of the heterozygous *k*-mers in the reads was missing in the assemblies (Additional file 1: Figs. S6, 7).

To further assess the overall structures of the assembled chromosomes, we compared them to recently assembled chromosomes of very closely related species such as the heterozygous "Chuanzhihong" apricot (*Prunus armeniaca*) [36], the Japanese apricot (*Prunus mume*) [37], and a more distantly related species, peach (*Prunus persica:* doubled-haploid genome) [38] using *SyRI* [22] (a tool designed for the comparison of chromosome-level assemblies). Our assemblies showed high consistency in the synteny to these assemblies across entire chromosomes, reflecting the reliability of the genetic map and the assembled genome structures (Fig. 3d; Additional file 1: Fig. S8).

As yet another way to assess the quality of the genome, we generated two Hi-C libraries from DNA extracted from leaves of Rojo Pasión and sequenced them totaling in 191.2 million read pairs (or ~ 240x haploid genome coverage). We created Hi-C contact maps using each of the homologous chromosome pairs separately as well as using the entire genome (Figs. 3e and 4a; Additional file 1: Figs. S9–15). In general, the contiguity of contact signals surrounding the main diagonal of the map again demonstrated the high quality of the structure of the assemblies.

Comparing gamete binning with Hi-C-based phasing and genome scaffolding

However, the perhaps more interesting way to use the Hi-C data is its application for genome phasing and scaffolding and the comparison of its assembly performance to that of gamete binning.

Applying *ALLHiC* [8] to the Hi-C reads sets generated 16 scaffolds (representing the 16 haploid chromosomes), with sizes ranging from 11.2 to 51.1 Mb (see the "Online methods" section). (Using a different Hi-C-based phasing and scaffolding tool, *SALSA2*

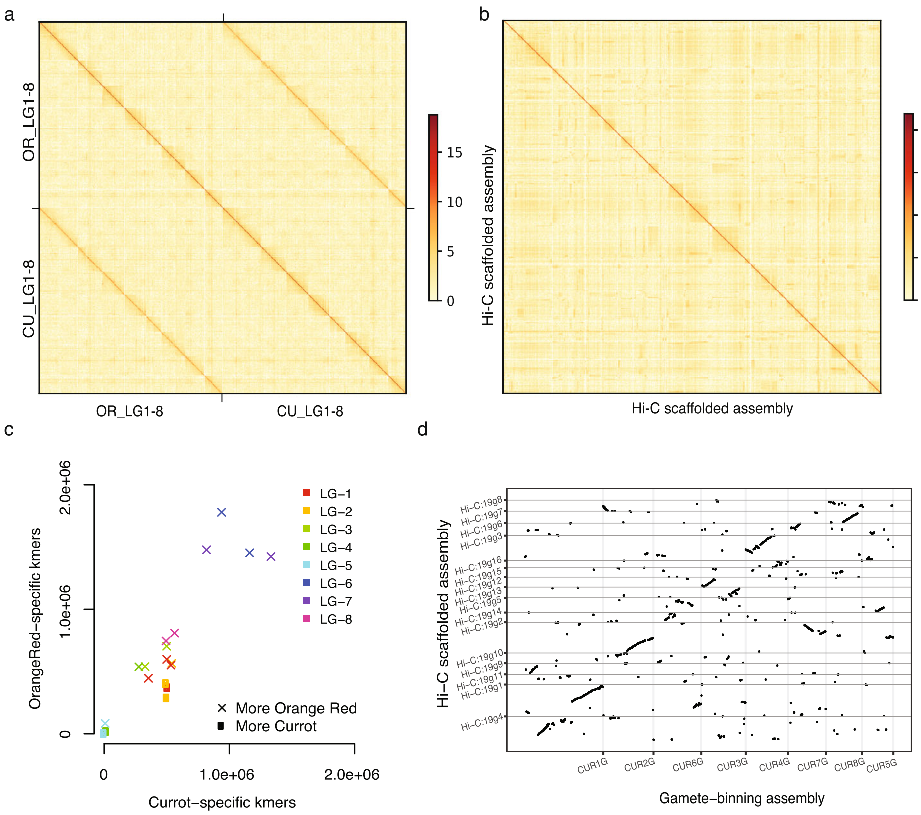

Fig. 4 Comparison of Hi-C based phasing and scaffolding with gamete binning. **a** Hi-C contact map based on all 16 haplotype assemblies generated with gamete binning (Currot (CU) and Orange Red (OR)). **b** Hi-C contact map based on all 16 haplotype assemblies generated with Hi-C data (Currot (CU) and Orange Red (OR)). Note the contact signal along the main diagonal was much weaker as compared to the signal based on the gamete binning assembly, and virtually no contact signals between two different haplotypes could be identified. **c** Haplotype validation of each linkage group (LG-1-8) of the Hi-C assembly using parent-specific k-mers. Almost all haplotype-specific assemblies included k-mers specific to both of the parental alleles indicating severe errors in the phasing. **d** Alignment of the Hi-C based assembly to genetic map derived assembly (i.e., gamete binning derived assembly) revealed mis-joining or splitting of linkage groups within the Hi-C assembly. For example, CUR1G was split into Hi-C:19 g1, 19 g4; on the other hand, alignments of 19 g2 to CUR2G, CUR5G, and CUR7G revealed mis-joins of sequences from independent linkage groups

[39], did not lead to comparable results, thus not compared further.). For comparison, we also generated Hi-C contact maps for *ALLHiC*-based assemblies (Fig. 4b). Interestingly, the contact maps of the gamete binning and *ALLHiC*-based assemblies were strikingly different. Only the gamete binning assembly showed (beside the contact within the haplotypes) the expected contact signals between two different haplotypes, which also were reported for other species [8, 40]. The absence of these signals in the Hi-C-based assembly suggests that the assembly was falsely merging sequences from different haplotypes and the contigs were likely to be scaffolded in the wrong order.

To test if the Hi-C-based assemblies were truly a mixture of the two haplotypes, we checked the presence of parental-specific *k*-mers within each of the 16 haplotype-specific chromosome-level assemblies (Fig. 4c). This revealed that the majority of the haplotype-specific assemblies were in fact mixtures of the two haplotypes, which is in great contrast with the high haplotyping accuracy of gamete binning. Finally, a whole-genome alignment of the Hi-C-based assembly to the genetic map-based assembly of gamete binning revealed many ambiguities between the genetic maps and the Hi-C-based assembly within essentially all haplotype-specific chromosome assemblies (Fig. 4d).

Taken together, besides its broad application, Hi-C-based phasing and scaffolding was far from being error-free. Some of the errors combined large pieces from different haplotypes, which resulted in falsely arranged chromosomes and severe phasing errors. Though, gamete-binning might be more tedious in its experimental requirements, the improved assembly quality might justify the additional effort.

Haplotype diversity and (non-allelic) meiotic recombination

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 Pasión" using *SyRI* [22] 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, 1132 translocation/transpositions, and 2477 distal duplications) between the haplotypes making up more than 15% of the assembled sequence (38.3 and 46.2 Mb in each of assemblies; Additional file 2: Table S1). Using the Hi-C contact maps (Fig. 3e; Additional file 1: Figs. S9–15), we validated the 17 largest rearrangements (> 500 kb) between the haplotype assemblies. Using a comprehensive RNA-seq dataset sequenced from multiple tissues of "Rojo Pasión" including reproductive buds, vegetative buds, flowers, leaves, fruits (seeds removed), and barks as well as a published apricot RNA-seq dataset [36], 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 [41] 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 (see the "Online methods" section; Additional file 2: Tables S2–3). 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 2638 meiotic crossover (CO) events (see the "Online methods" section). To improve the resolution of the predicted CO events (6.1 kb), we selected 2236 CO events detected in 369 nuclei with a sequencing depth above 0.1x genome coverage (Additional file 2: Table S4). 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. 5; see the "Online methods" section). By investigating the fine-scale pattern of short-read alignment of each nucleus, we

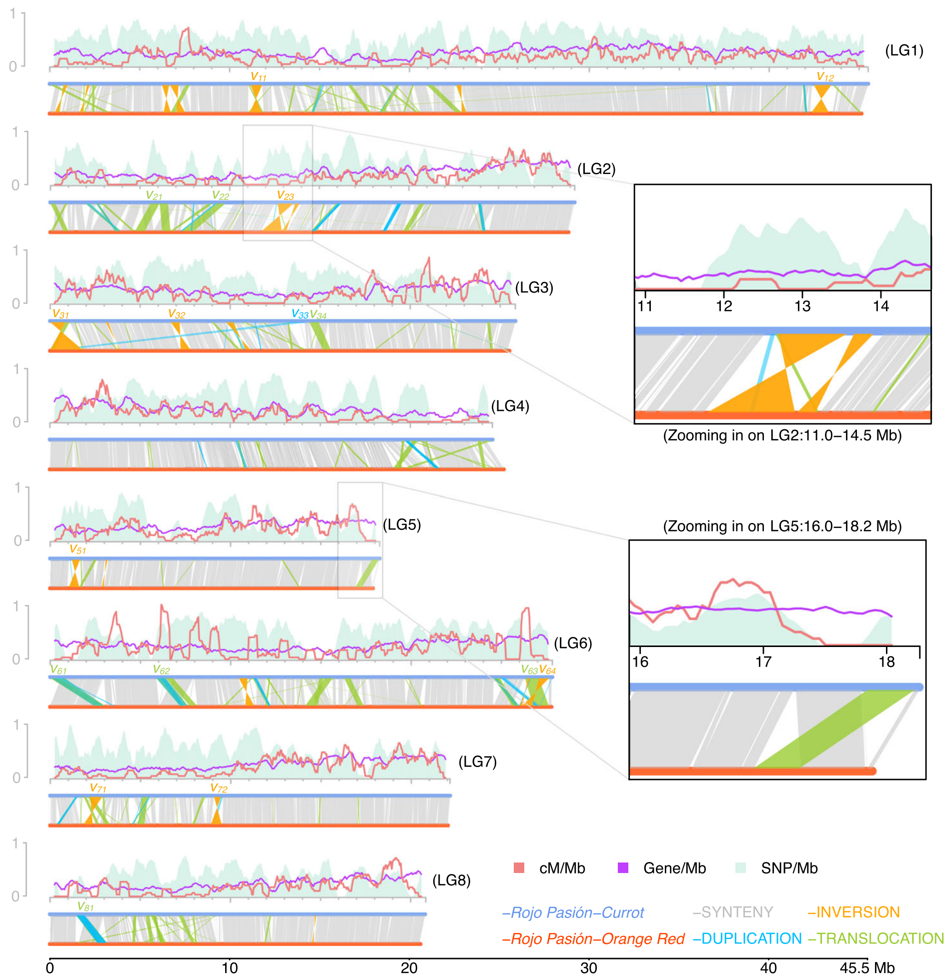

Fig. 5 Structural genome variations and meiotic recombination. Top: recombination landscape created with sliding windows of 500 kb at a step of 50 kb with COs detected in all single pollen nuclei (with coverage over 0.1x), coupled with SNP density and gene density. For *x*-axis, coordinates were based 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*), 7410 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). Variants spanning over 500 kb are labeled as v_{xy}, where *x* denotes the chromosome number and *y* the identifier of the variant in the chromosome. All these large variants were confirmed within Hi-C contact maps (Fig. 3e, Additional file 1: Figs. S9–16)

identified six CO events located in rearranged regions (0.3% of 2236 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. 6). 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.

Conclusion

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

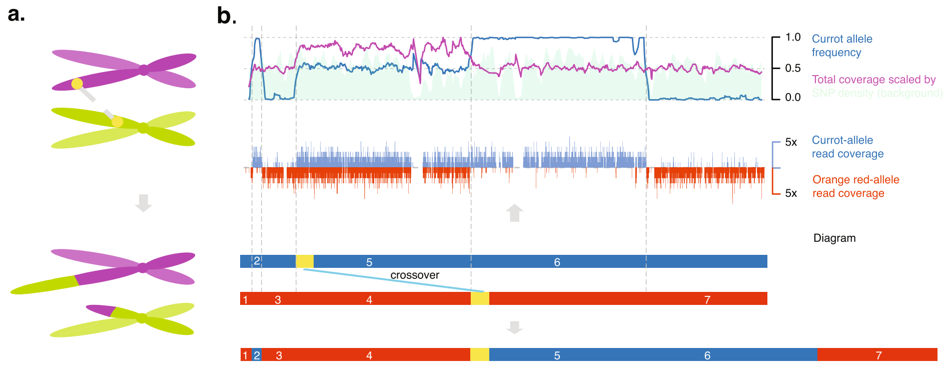

Fig. 6 Non-allelic crossovers and its consequences. **a** Illustration of a non-allelic crossover which results in a chromosomal anomaly. **b** Analysis of a single-pollen nuclei, which revealed a non-allelic CO resulting in the duplication of a large chromosomal segment. The short-read alignments of a haploid nucleus 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 labeled with "4" and "5" are now part of the same newly formed chromosome

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 Pasión" 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™ 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™.

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 Pasión" (PacBio; Additional file 1: Fig. S2) 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 Pasión" 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 Pasión" [27] flowers were extracted with forceps and submerged in woody pollen buffer (WPB) [42]. 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 [30]. Isolated pollen was prefiltered (100 μm) and bursted (30 μm) using Celltrics™ 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™ 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 (Additional file 1: Fig. S4). 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™ Chromium controller in two batches with 6300 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 2 × 151 bp paired-end mode, totaling 95 and 124 million read pairs, respectively (61.7 Gb).

Hi-C library preparation and sequencing

Approximately 0.5 g of flash-frozen leaf samples of "Rojo Pasión," which were collected from the field, were thawed and fixed with 1% formaldehyde for 30 min at room temperature under vacuum. Subsequently, the in situ Hi-C library preparation was performed according to a protocol established for rice seedlings [43]. The libraries were sequenced on an Illumina HiSeq3000 instrument; in total, around 191.2 million pair-end reads were obtained.

RNA extraction and sequencing

Fruits tissue was collected in the same way for the PacBio sequencing library. Tissue from reproductive buds, vegetative buds, flowers, leaves, and bark tissues were collected from the same shoots used for pollen nuclei isolation. RNA was extracted from these tissues using the NucleoSpin® RNA Plant of Macherey-Nagel™ to prepare Illumina libraries.

All libraries were sequenced with Illumina HiSeq 3000 at Max Planck Genome-centre Cologne (MP-GC) in the 150 bp single-end mode, which respectively led to 32.8 (reproductive buds), 28.9 (vegetative buds), 30.2 (flowers), 23.8 (leaves), 18.6 (fruit), and 26.1 (bark) million reads, totaling 24.1 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* [44]. The *k*-mer histogram was provided to *findGSE* [29] to estimate the size of the

“Rojo Pasión” genome under the heterozygous mode (with “*exp_hom*=200”; Additional file 1: Fig. S3).

Preliminary diploid-genome assembly and curation

With the 19.9-Gb raw PacBio reads of “Rojo Pasión” (Additional file 1: Fig. S2), a preliminary diploid assembly was constructed using *Canu* [28] (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* [45]. Then, the *purge haplotigs* pipeline was 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 [46]. 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 selection

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 *bowtie*2 [45]. With 87.2% reads aligned, 989,132 raw SNPs were called with *samtools and bcftools* [47]. Three criteria were used to select potential allelic SNPs (578,209), including (i) the alternative allele frequency must be between 0.38 and 0.62, (ii) the alternative allele must be carried by 60–140 reads, and (iii) the total sequencing depth at a SNP must be between 120 and 280x (as compared with genome-wide mode depth of 208x; Fig. 2b).

Deletion marker selection and genotyping

The assemblies included 10,452 regions of over 2 kb without SNP marker (total: 110.9 Mb). If the average sequencing depth of such regions was less than or equal to 146x (i.e., the value at the valley between middle and right-most peaks in the sequencing depth distribution; Fig. 2b), it was selected as a deletion-like marker. This revealed 3253 deletion markers (36.4 Mb), including 237 on contigs without a single SNP marker. The remaining 7199 regions (74.5 Mb) were defined as conserved (homozygous regions) between two haplotypes (Fig. 2b). For each deletion marker in each gamete genome, we assessed the normalized read count (*RPKM* value) within the deletion using *bedtools* [48]. The genotype at such a deletion marker was initialized as *a* or *n*, where *a* refers to the presence of reads (and therefore relates to the haplotype without the deletion) and *n* refers to the absence of reads (either the deletion haplotype or not having enough information).

Haplotype phasing and CO identification

Barcodes in the raw reads were corrected using *cellranger*, 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* [45], bases were called using *bcftools* [49], and a simple bi-marker majority voting strategy was applied to phase the SNPs along each contig (Fig. 2c). After phasing, we identified COs as consistent switches between the haplotypes.

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 (Additional file 1: Fig. S5c). 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 nucleus. 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 *JoinMap 4.0* [32] (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 [32]. 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 separation

PacBio reads (19.9 Gb) were classified based on three major cases after being aligned to the curated assembly using *minimap2* [50]. 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.

Haplotype assembly and chromosome-level scaffolding

Independent assemblies were performed with 16 sets of reads, i.e., for every two haplotypes in each of the eight linkage groups using *Flye* [33] with the default settings.

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* [51], leading to chromosome-level assemblies (i.e., those labeled with "scaf" in Fig. 3b).

Haplotype evaluation

The genotypes of the 16 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 "Currot"-genotype and "Orange Red"-genotype PacBio reads using *apollo* [52], 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 Pasión" 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 $r_O \cap p_O$, which were used to calculate average haplotyping accuracy. All *k*-mer processing (counting, intersecting and difference finding) were performed with *KMC* [53]. After haplotype validation, the assemblies were further polished with the respective parental short-read alignment using *pilon* [54] (with options "--fix bases --mindepth 0.85") generating v1.0 of the assemblies. Manual correction of the v.1.0 assemblies was performed according to focal and parental reads to generate assembly v1.1. Finally, *k*-mer-based assembly validation was performed with *KAT* [34] and *Merqury* [35].

Genome annotation

We annotated protein-coding genes for each haplotype assembly (v1.0) by integrating evidences from ab initio gene predictions (using three tools *Augustus* [55], *GlimmerHMM* [56], and *SNAP* [57]), 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* [58] 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 [36], using *HISAT* [59], and we assembled the transcripts using *StringTie* [60]. Finally, we used the tool *EvidenceModeler* [61] 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* [62] 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* [63] 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* [64] 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* [41] v4 tool based on 2326 eudicots single-copy orthologs from OrthoDB v10 [65]. A similar process was used to filter for haplotype-specific genes (Additional file 2: Tables S2–3). Finally, a genome annotation lift-over was performed from v1.0 to v1.1 using *liftoff* [66] with default parameters.

Genome assembly comparison

All genome assemblies, including "Rojo Pasión" haplotypes, "Chuanzhihong" apricot (*Prunus armeniaca*) [36], Japanese apricot (*Prunus mume*) [37], and "Lovell" peach (*Prunus persica*) [38], were aligned to each other using *nucmer* from the *MUMmer4* [67] 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* [22]. To follow the nomenclature of the Prunus community, the "Rojo Pasión" chromosomes were numbered according to the numbering in 'Lovell' peach [38].

Hi-C data analysis

We used *ALLHiC* [8] and *SALSA2* [39] for phasing and scaffolding. All 191.2 million Hi-C read pairs were aligned (using *BWA* version 0.7.15-r1140) to the haplotype-resolved unitigs assembled by *Canu*. Only uniquely mapped read pairs were selected using *filterBAM_forHiC.pl* from the *ALLHiC* package. The selected alignments were used as input for *ALLHiC_partition* ("ALLHiC_partition -b clean.bam -r unitigs.fa -e GATC -k 19") and *SALSA2* ("python run_pipeline.py -a unitigs.fa -l unitigs.fa.fai -g unitigs.gfa -m yes -b alignment.bed -e GATC -o SALSA2_out -i 8", where the file alignment.bed was generated and sorted from clean.bam using *bedtools bamtobed* (version v2.29.0) and unitigs.gfa was collected from the *Canu* output). For *ALLHiC*, we had to set group number as 19 to get 16 linkage groups (of chromosome-level size), and 3 smaller groups below 2.5 Mb, which were not considered further. We continued with *ALLHiC* pipeline as *SALSA2* did not achieve chromosome-level scaffolds. The subsequent pipeline of *ALLHiC* were run by default except for using "-RE GATC" in the "allhic extract" command. For comparison, we also aligned all raw Hi-C reads to haploid assemblies generated by gamete binning, and selected the uniquely mapped read pairs as described above. Hi-C maps were visualized using *ALLHiC_plot* at 300–500 kb resolution. Alignments of *ALLHiC* and gamete binning-based assemblies were obtained using *minimap2* and dot plot was drawn with script *pafCoordsDotPlotly.R* at *https://github.com/tpoorten/dotPlotly*.

Crossover identification

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* [45] and bases were called using *bcftools* [49]. Finally, *TIGER* [68] 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. 5), 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.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-020-02235-5.

Additional file 1: Fig. S1. Flowchart of gamete binning in haploid assembly of apricot (*Prunus armeniaca* cultivar 'Rojo Pasión'). **Fig. S2.** Size distribution of long reads (PacBio). **Fig. S3.** Genome size estimation with *k*-mers of pooled Illumina reads from pollen nuclei ($k = 21$). **Fig. S4.** Illustration of flow cytometry sorting of pollen nuclei. **Fig. S5.** Characterization and selection of pollen nuclei sequencings. **Fig. S6.** *k*-mer comparison plot for genome assembly evaluation by *KAT* [34]. **Fig. S7.** Assembly spectrum plots for evaluating *k*-mer completeness by *Merqury*. **Fig. S8.** High-scaffolding accuracy reflected by synteny to closely-related species: 'Chuanzhihong' apricot (*Prunus armeniaca*) and Japanese apricot (*Prunus mume*). **Fig. S9.** Hi-C contact along gamete binning based assemblies for chromosome 2 related to both haplotypes of Currot (CU) and Orange Red (OR) (with bin size or resolution of 300 kb). **Fig. S10.** Hi-C contact along gamete binning based assemblies for chromosome 3 related to both haplotypes of Currot (CU) and Orange Red (OR) (with bin size or resolution of 300 kb). **Fig. S11.** Hi-C contact along gamete binning based assemblies for chromosome 4 related to both haplotypes of Currot (CU) and Orange Red (OR) (with bin size or resolution of 300 kb). **Fig. S12.** Hi-C contact along gamete binning based assemblies for chromosome 5 related to both haplotypes of Currot (CU) and Orange Red (OR) (with bin size or resolution of 300 kb). **Fig. S13.** Hi-C contact along gamete binning based assemblies for chromosome 6 related to both haplotypes of Currot (CU) and Orange Red (OR) (with bin size or resolution of 300 kb). **Fig. S14.** Hi-C contact along gamete binning based assemblies for chromosome 7 related to both haplotypes of Currot (CU) and Orange Red (OR) (with bin size or resolution of 300 kb). **Fig. S15.** Hi-C contact along gamete binning based assemblies for chromosome 8 related to both haplotypes of Currot (CU) and Orange Red (OR) (with bin size or resolution of 300 kb).

Additional file 2: Table S1. Structural variation and synteny between Currot- and Orange Red-haplotype genome assemblies. **Table S2:** Haplotype-specific genes in Currot versus Orange Red haplotype. **Table S3:** Haplotype-specific genes in Orange Red versus Currot haplotype. **Table S4:** 2236 crossovers found in 369 nuclei (Currot-haplotype genome coordinate).

Additional file 3. Review history.

Acknowledgements

The authors would like to thank Antonio Molina and José Egea for kindly providing plant material, Saurabh Pophaly for help in transferring the read data to public servers, Detlef Weigel, Klaus Palme, Patrick Schaub and Florian Wüst for supportive guidance, and Kristin Krause and Vidya Oruganti for useful discussions and comments for improving the manuscript.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 3.

Authors' contributions

J.A.C, H.S. and K.S. designed the project. J.A.C., B.H., K. F.-D., C.K., D.R., M.R., N.W., and C.L. performed 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.

Funding

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 (K.S.), and the European Research Council (ERC) Grant "INTERACT" (802629) (K.S.). C.K. acknowledges the ISAC SRL Emerging Leaders Program. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

Data supporting the findings of this work are available within the paper and its additional files (Supplementary Information). Read data sequenced from two 10x sc-CNV libraries, two Hi-C libraries, one PacBio library from "Rojo Pasión," two Illumina libraries for "Currot" and "Orange Red" that support the work in this study, and the haploid assemblies and annotations generated are available in European Nucleotide Archive (ENA) under accession number "PRJEB37669" [69]. Data was uploaded to ENA using EMBLmyGFF [70]. Customs scripts supporting this work are available at *github.com/schneeberger-lab/GameteBinning* or Zenodo under MIT license [71]. All other relevant data are available upon request.

Ethics approval and consent to participate

Not applicable in the manuscript.

Consent for publication

Not applicable in the manuscript.

Competing interests

The authors declare no competing interests.

Author details

[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]Center for Plant Molecular Biology (ZMBP), University of Tübingen, Auf der Morgenstelle 32, 72076 Tübingen, Germany. [5]Departament of Plant Breeding, CEBAS-CSIC, PO Box 164, E-30100 Espinardo, Murcia, Spain. [6]Institute of Biology, University of Hohenheim, Garbenstraße 30, 70599 Stuttgart, Germany. [7]Max Planck-Genome-center Cologne, Carl-von-Linné-Weg 10, 50829 Cologne, Germany.

Received: 15 May 2020 Accepted: 11 December 2020
Published online: 29 December 2020

References

1. Korlach J, Gedman G, Kingan SB, Chin CS, Howard JT, Audet JN, 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. 2017;6(10):1–16.
2. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018;36(12):1174–82.
3. Yang H, Chen X, Wong WH. Completely phased genome sequencing through chromosome sorting. Proc Natl Acad Sci U S A. 2011;108(1):12–7.
4. Falconer E, Lansdorp PM. Strand-seq: a unifying tool for studies of chromosome segregation. Semin Cell Dev Biol. 2013;24(8–9):643–52.
5. Hills M, Falconer E, O'Neil K, Sanders AD, Howe K, Guryev V, et al. Construction of whole genomes from scaffolds using single cell strand-seq data. bioRxiv. 2018. https://www.biorxiv.org/content/10.1101/271510v1. Accessed 20 Jan 2020.
6. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10(1):1–16.
7. Selvaraj S, Dixon JR, Bansal V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat Biotechnol. 2013;31(12):1111–8.

8. 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. 2019;5(8):833–45.
9. Linsmith G, Rombauts S, Montanari S, Deng CH, Celton JM, Guérif P, et al. Pseudo-chromosome-length genome assembly of a double haploid "Bartlett" pear (Pyrus communis L.). Gigascience. 2019;8(12):1–17.
10. Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, MR MK, et al. Origin and evolution of the octoploid strawberry genome. Nat Genet. 2019;51(3):541–7.
11. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49(4):643–50.
12. Wallberg A, Bunikis I, Pettersson OV, Mosbech MB, Childers AK, Evans JD, et al. A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. BMC Genomics. 2019;20(1):1–19.
13. Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, et al. Accurate chromosome-scale haplotype-resolved assembly of human genomes. bioRxiv. 2020. https://www.biorxiv.org/content/101101/810341v2. Accessed 23 Nov 2020.
14. Doležel J, Vrána J, Cápal P, Kubaláková M, Burešová V, Šimková H. Advances in plant chromosome genomics. Biotechnol Adv. 2014;32(1):122–36.
15. International Wheat Genome Sequencing Consortium (IWGSC). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science. 2014;345(6194):1251788.
16. Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, et al. Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. Nat Genet. 2018;50(11):1565–73.
17. Zhang X, Wu R, Wang Y, Yu J, Tang H. Unzipping haplotypes in diploid and polyploid genomes. Comput Struct Biotechnol J. 2020;18:66–72.
18. Li R, Qu H, Chen J, Wang S, Chater JM, Zhang L, et al. Inference of chromosome-length haplotypes using genomic data of three or a few more single gametes. Mol Biol Evol. 2020;37(12):3684–98.
19. Kirkness EF, Grindberg RV, Yee-Greenbaum J, Marshall CR, Scherer SW, Lasken RS, et al. Sequencing of isolated sperm cells for direct haplotyping of a human genome. Genome Res. 2013;23(5):826–32.
20. Shi D, Wu J, Tang H, Yin H, Wang H, Wang R, et al. Single-pollen-cell sequencing for gamete-based phased diploid genome assembly in plants. Genome Res. 2019:1–11.
21. Wu J, Wang ZW, Shi ZB, Zhang S, Ming R, Zhu SL, et al. The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res. 2013;23(2):396–408.
22. Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019;20(1):1–13.
23. Jiao WB, Schneeberger K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun. 2020;11(1):1–10.
24. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1):1–16.
25. Sun H, Rowan BA, Flood PJ, Brandt R, Fuss J, Hancock AM, et al. Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination. Nat Commun. 2019;10(1):1–9.
26. Dréau A, Venu V, Avdievich E, Gaspar L, Jones FC. Genome-wide recombination map construction from single individuals using linked-read sequencing. Nat Commun. 2019; 10(1). https://www.nature.com/articles/s41467-019-12210-9.ris.
27. Egea J, Dicenta F, Burgos L. "Rojo Pasión" apricot. Hortscience. 2004;39(6):1490–1.
28. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.
29. Sun H, Ding J, Piednoël M, Schneeberger K. FindGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. Bioinformatics. 2018;34(4):550–7.
30. Kron P, Husband BC. Using flow cytometry to estimate pollen DNA content: improved methodology and applications. Ann Bot. 2012;110(5):1067–78.
31. Julian C, Rodrigo J, Herrero M. Stamen development and winter dormancy in apricot (*Prunus armeniaca*). Ann Bot. 2011; 108(4):617–25.
32. van Ooijen JW. JoinMap ® 4, Software for the calculation of genetic linkage maps in experimental populations. 2006. p. Wageningen: Kyazma B.V.
33. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6.
34. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 2017;33(4):574–6.
35. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21(1):1–27.
36. Jiang F, Zhang J, Wang S, Yang L, Luo Y, Gao S, et al. The apricot (*Prunus armeniaca* L.) genome elucidates Rosaceae evolution and beta-carotenoid synthesis. Hortic Res. 2019;6(1):1–12.
37. Zhang Q, Chen W, Sun L, Zhao F, Huang B, Wang J, et al. The genome of *Prunus mume*. Nat Commun. 2012;3:1–8.
38. Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, et al. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet. 2013;45(5):487–94.
39. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019;15(8):1–19.
40. Chen H, Zeng Y, Yang Y, Huang L, Tang B, Zhang H, et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat Commun. 2020;11(1). https://www.nature.com/articles/s41467-020-16338-x.ris.
41. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
42. Loureiro J, Rodriguez E, Dolezel J, Santos C. Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. Ann Bot. 2007;100(4):875–88.
43. Liu C, Cheng YJ, Wang JW, Weigel D. Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. Nat Plants. 2017;3(9):742–8.

44. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70.
45. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):1–10.
46. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):1–10.
47. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
48. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
49. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.
50. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
51. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 2019;20(1):1–17.
52. Firtina C, Kim JS, Alser M, Cali DS, Cicek AE, Alkan C, et al. Apollo: a sequencing-technology-independent, scalable, and accurate assembly polishing algorithm. Bioinformatics. 2020;36(12):1–10.
53. Kokot M, Dlugosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. Bioinformatics. 2017;33(17):2759–61.
54. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):1–14.
55. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: Ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34(WEB. SERV. ISS):435–9.
56. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004;20(16):2878–9.
57. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, De Bakker PIW. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008;24(24):2938–9.
58. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:1–11.
59. Kim D, Langmead B. Salzberg1 SL. HISAT: a fast spliced aligner with low memory requirements Daehwan HHS Public Access. Nat Methods. 2015;12(4):357–60.
60. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5.
61. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9(1):1–22.
62. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):1–14.
63. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
64. 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. 2008;9:1–12.
65. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, 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. 2019;47(D1):D807–11.
66. Shumate A, Salzberg SL. Liftoff: an accurate gene annotation mapping tool. bioRxiv. 2020. https://www.biorxiv.org/content/101101/20200624169680v1. Accessed 17 Aug 2020.
67. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol. 2018;14(1):1–14.
68. Rowan BA, 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. 2015; 5(3):385–398.
69. Campoy JA, Sun H, Goel M, Jiao W-B, Folz-Donahue K, Wang N, et al. Haplotype resolved chromosome level assembly of Apricot generated by application of gamete binning on single cell sequencing data of gametes. Datasets used in Gamete binning (Version 1.0). PRJEB37669. Eur Nucleotide Arch. https://www.ebi.ac.uk/ena/browser/view/PRJEB37669 (2020). Accessed 18 Dec 2020.
70. Norling M, Jareborg N, Dainat J. EMBLmyGFF3: a converter facilitating genome annotation submission to European Nucleotide Archive. BMC Res Notes. 2018;11(1):1–5.
71. Sun H, Campoy JA, Schneeberger K. Gamete binning. zenodo. https://zenodo.org/record/4287161 (2020). Accessed 18 Dec 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.