Assembly of complete diploid phased chromosomes from draft genome sequences

Andrea Minio\textsuperscript{1}, Noé Cochetel\textsuperscript{1}, Amanda Vondras\textsuperscript{1}, Mélanie Massonnet\textsuperscript{1}, and Dario Cantu\textsuperscript{*}.
\textsuperscript{1}Department of Viticulture and Enology, University of California Davis, Davis, CA 95616

De novo genome assembly is essential for genomic research. High-quality genomes assembled into phased pseudomolecules are challenging to produce and often contain assembly errors because of repeats, heterozygosity, or the chosen assembly strategy. Although algorithms that produce partially phased assemblies exist, haploid draft assemblies that may lack biological information remain favored because they are easier to generate and use. We developed HaploSync, a suite of tools that produces fully phased, chromosome-scale diploid genome assemblies, and performs extensive quality control to limit assembly artifacts. HaploSync scaffolds sequences from a draft diploid assembly into phased pseudomolecules guided by a genetic map and/or the genome of a closely related species. HaploSync generates a report that visualizes the relationships between current and legacy sequences, for both haplotypes, and displays their gene and marker content. This quality control helps the user identify misassemblies and guides Haplosync’s correction of scaffolding errors. Finally, HaploSync fills assembly gaps with unplaced sequences and resolves collapsed homozygous regions. In a series of plant, fungal, and animal kingdom case studies, we demonstrate that HaploSync efficiently increases the assembly contiguity of phased chromosomes, improves completeness by filling gaps, corrects scaffolding, and correctly phases highly heterozygous, complex regions.

haplotype phasing; diploid genomes; assembly error correction; hybrid genome assembly; chromosome anchoring

Correspondence: dacantu@ucdavis.edu

Introduction

Affordable high-throughput DNA sequencing and novel assembly tools have made high-quality genome assemblies and genome research attainable and abundant. Long-read DNA sequencing technologies, like those developed by Oxford Nanopore Technologies and Pacific Biosciences, are now the preferred methods for reference genome sequencing. The assemblies produced using these technologies are more contiguous and complete than assemblies constructed using short sequencing reads and better represent repetitive content (1–4). Another important advantage of long-read sequencing is the ability to generate phased diploid assemblies. Previously, genome complexity due to heterozygosity was typically handled by generating a haploid representation of a diploid genome either by collapsing heterozygous sites into a consensus sequence or by including only one allele’s sequence (5–10).

Partially phased assemblies have revealed genomic complexities that were inaccessible in previous haploid representations, such as haplotype-specific structural variation events, trait-associated alleles, and allele-specific gene expression and methylation (11–16). However, phasing of diploid assemblies remains challenging for complex genomes. High heterozygosity and repetitive content often prevent phasing in diploid regions. This inflates the primary assembly (17, 18) and can impair scaffolding procedures that use the primary assembly as input.

Hybrid approaches that integrate additional independent data, such as optical maps or chromatin structure, help scaffold draft genome assemblies up to full-length chromosomes (19–22). Several genetic map-based and reference-guided scaffolding tools have been developed (23–26). However, these tools do not use the relationship between haplotypes to aid the assembly process. Consequently, constructing chromosome-scale pseudomolecules using these tools relies on the phasing accuracy of the draft genome, the density of genetic map markers, or similarity to a related species’ genome (24, 26, 27). Though quality control is an integral part of the assembly procedure, the relationship between haplotypes is never included in quality control processes.

Here, we present HaploSync, an open-source package that scaffolds, refines, and fully phases diploid and chromosome-anchored genomes. HaploSync leverages the relationship between haplotypes to improve the quality and accuracy of assemblies, separates haplotypes while reconstructing chromosome-scale pseudomolecule sequences, and recovers a location for genomic regions that cannot be placed during other assembly steps. Quality controls are implemented at each step to check for and correct assembly errors. HaploSync was benchmarked using five diploid species with different levels of heterozygosity from the plant, animal, and fungal kingdoms. For each species, HaploSync delivered a completely phased, chromosome-scaled genome with a quality comparable to the assemblies considered as references for each species. HaploSync, its manual, and tutorials for its use are freely available at https://github.com/andreaminio/haplosync.

Materials and methods

HaploSync has six modules: HaploSplit, HaploDup, HaploBreak, HaploFill, HaploMake, and HaploMap. The overall HaploSync workflow is summarized in Fig. 1. HaploSync accepts draft genome sequences or assembled pseudomolecules as input, preferably with minimally collapsed heterozygous sequences and no haploid consensus sequences. Allele phasing is unnecessary a priori. The tool is applicable to conventional haploid and diploid-aware assemblies.
Fig. 1: The HaploSync pipeline builds and refines haploid and diploid genome assemblies. The diploid-aware pipeline can deliver fully phased diploid pseudomolecules using a draft diploid assembly or diploid pseudomolecules as input. If draft sequences are used, HaploSplit first separates the haplotypes into two pseudomolecule sets. Pseudomolecules provided by the user or reconstructed with HaploSplit, then undergo quality control with HaploDup. If errors are found, input sequences can be edited with HaploBreak prior to rebuilding the pseudomolecules with HaploSplit. If no errors are detected and there are unplaced sequences, the pseudomolecule undergoes gap-filling with HaploFill. After each filling iteration, quality control can be performed with HaploDup. Finally, HaploMap can be used to identify colinear regions between pseudomolecules.

**A. HaploSplit.** HaploSplit uses external information to associate draft assembly sequences with original chromosones, then sorts and orients them in pseudomolecules using directed adjacency networks. Alternative sequences are detected and segregated in two different haplotypes and, if the external information relates to a chromosome, HaploSplit delivers chromosome-scale scaffolds.

External information can be a genetic map composed of sorted unique genomic markers (Fig. 2) and/or the genome assembly of a closely related species (Supplemental figure 1). When both types of information are used in hybrid mode, the genetic map is used as primary information to generate draft diploid pseudomolecules. The guide genome is used subsequently when marker information is insufficient. Phasing information between the alternative alleles is not needed a priori: HaploSplit will detect the existing relationship between haplotypes and phase them. The tool is capable of handling diploid assemblies lacking phasing information as well as diploid assemblies with inflated primary assemblies due to erroneous phasing. However, if the relationship between input sequences is known, it can be supplied to HaploSplit as a constraint to guide the reconstruction. For example, allelic information can be given to avoid placing primary contigs and haplotigs in the same haplotype.

If a genetic map is given as external evidence, HaploSplit first assesses markers’ uniqueness and congruence in the assembly. Markers present at three or more locations in draft sequences and markers present twice in the same draft sequence, are considered unreliable and are excluded from further analysis. For each sequence containing an unreliable marker, HaploSplit produces a report containing layered interactive plots (Supplemental figure 2), including the sequence’s self-alignment, the position of reliable and duplicated genetic markers, and the copy number of annotated genes if gene annotation is available. If the input draft sequence is a scaffold, its composition in terms of legacy contigs is also included. These plots can be used to investigate the source of marker duplication within a draft sequence and to correct it using either HaploBreak (see below) or a constraint file. After identifying the genetic markers that are reliable for scaffolding, HaploSplit assigns draft sequences to a chromosome based on their largest set of consecutive markers (Fig. 2B), with their orientation based on markers’ order (Fig. 2C). If marker order does not adequately define sequence orientation (e.g., only one marker is present), the sequence is aligned and oriented based on the alternative haplotype (i.e. the sequence sharing the same marker). Once each draft sequence is assigned unambiguously to a chromosome, a directed, weighted adjacency network is created for each chromosome (Fig. 2D). Directed edges are created for each draft sequence with a weight based on the number of markers composing the sequence. Directed edges with zero weight are created to connect sequences without any common genetic marker ranges. Then, two haplotypes for each genomic region are split into different network paths. The tiling path that maximizes the number of genetic markers is used to scaffold the first haplotype and its draft sequences are removed from the adjacency network (Fig. 2E). The second-best tiling path is selected from the remaining sequences in the network and is scaffolded into the second haplotype (Fig. 2F).

If a genome is used to guide scaffolding (Supplemental figure 1), draft sequences are aligned on all guide genome sequences with Minimap2 (28). Local alignments are used to generate a directed weighted adjacency network for the query draft sequence and each guide genome sequence. Each draft sequence is associated with the guide sequence with which it shares the highest identity. Directed edges are created for each draft sequence with a significant alignment on the guide sequence. Directed edges between non-overlapping hits are added to the network and connected with a weight of zero. For each adjacency network, the tiling path maximizing the number of matching bases between the draft sequences and the guide sequence is used to build the first haplotype. The second haplotype is then scaffolded using the second best path.

When a genetic map and a guide genome are used in hybrid mode, the genetic map is used as primary information to generate draft diploid (Supplemental figure 1). The draft pseudomolecules and unplaced draft sequences are aligned to the guide genome. Then, an adjacency network is created for each guide sequence using the draft sequences composing each draft pseudomolecule and the unplaced draft sequences that do not significantly overlap the alignment of the draft pseudomolecules. The two tiling paths with the highest identity with the guide sequence are used for scaffolding the two haplotypes.

HaploSplit permits diverse, user-defined relationships between sequences to constrain and/or fine-tune scaffolding. For example, the relationship between the haplotigs and primary sequence defined by a sequence assembler like Falcon Unzip can be used to maintain consistency across alternative sequences. Similarly, a list of sequences in specific linkage groups can be given to guide their placement in pseudomolecule scaffolds.
To identify misassemblies and help plan a correction strategy, HaploDup compares pseudomolecule sequences, integrates structural (e.g., contigs and scaffolds) and feature (e.g., markers and genes) information, and produces interactive plots (Fig. 3). Two kinds of plots are generated. The first compares alternative haplotypes (Fig. 3 A). The second visualizes unplaced sequences with sufficient information to be placed but are currently unplaced among scaffolds because of incompatibility with other sequences; these are compared to the two alternative pseudomolecules (Fig. 3 B).

**B.1. Alternative haplotype comparison.** HaploDup produces a report for each alternative haplotype of each linkage group (Fig. 3A). The report includes layered plots: i) alignment of the two alternative haplotypes on the target haplotype; ii) the target sequence structure, with two lines of sequences at most (if available); iii) marker position and duplication status (if available).

The dotplot is essential for visualizing colinear regions within and between pseudomolecules. Duplications, deletions, and translocations can be spotted by overlaying both haplotypes’ alignments. If this information is intersected with the structure of input contigs or scaffolds, then it is possible to determine whether these peculiarities are real or are technical errors. For example, a region duplicated in one haplotype and deleted in the other may indicate that both alleles were placed in the same scaffold instead of one placed in each haplotype (red box in Fig. 3). Genetic markers and genes’ positions also help identify assembly errors. Genetic markers that are duplicated within the genome assembly are indicative of misplaced alleles. When a gene annotation is available, HaploDup counts significant alignments (>80% coverage and identity) of each CDS on its pseudomolecule of origin and on the alternative haplotype. This is useful for spotting fused haplotypes when the whole genome dotplot lacks resolution. An unbalanced number of gene copies between haplotypes in a given region can indicate a deficit of information or a duplication error. With these plots, the user can identify misassembled regions. Misassemblies can be solved by providing either a list of the breakpoint coordinates of the misplaced sequences to HaploBreak or a constraint file to HaploSplit.

**B.2. Comparison of unplaced sequences with the two haplotypes of each pseudomolecule.** HaploDup uses external information to compare unplaced sequences to related pseudomolecules (Fig. 3B). The plot reports: i) a comparison of associated pseudomolecules structures in terms of markers and sequence content. Structure is reported on two levels (scaffolds input to HaploSplit and their composition in terms of legacy contigs) when the requisite information is available; ii) a comparison of the unplaced sequence to the associated pseudomolecules in terms of markers and sequence content at two levels (scaffolds input to HaploSplit and their composition in terms of legacy contigs); iii) a comparison of the ranges of markers covered by the unplaced sequence and the ranges covered by the draft sequences composing the pseudomolecules.

Markers and their relationship to sequences can be visual-
ized. Markers can be color-coded based on order. This plot helps resolve conflicts that prevent sequence placement into linkage groups. In Fig. 3B and Supplemental figure 6, for example, a distal marker is incorrectly ordered inside an unplaced sequence. This triggered its exclusion from any of the pseudomolecules. Once fixed, the sequence will be placed.

**C. HaploBreak.** HaploBreak (Supplemental figure 7) automatically searches for and breaks sequences at the nearest known junction or at the nearest gap. The coordinates of breakpoint pairs are given by the user to estimate where sequences should be broken to correct scaffolding errors. If a scaffolding structure is supplied by the user, these junctions are prioritized to be broken. If a pair of breakpoints leads to two distinct scaffolding junctions, the original sequences reported between the two junctions will be excluded from the tiling path. If either breakpoint in a pair is associated with a sequence instead of a junction, the corresponding original sequence is broken on the nearest gap (i.e. stretch of “N” characters between two contigs). For each pair of breakpoint coordinates queried by the user, HaploBreak will do the following procedure: (i) search for scaffolding junctions closest to the two coordinates. If a junction is found within the defined search limits, it is associated with the breakpoint, else the original sequences are searched for the closest gap (i.e. a region of “N” characters), (ii) break the sequence. If the pair of coordinates is associated with two distinct scaffolding junctions (or one junction and the end of an input sequence), the original sequence between them is classified as misplaced (i.e. “unwanted” in that tiling path). If one or both breakpoints is associated with a gap in the original sequence, the sequence is broken at the gap position.

**D. HaploFill.** A reference-independent approach, HaploFill (Supplemental figure 8) uses the relationship between homologous pseudomolecule scaffolds to improve the assembly’s completeness by integrating unplaced sequences where scaffolding gaps occur. Gaps are created during scaffolding procedures when adjacent regions in the pseudomolecule are assembled in separate sequences and lack sufficient information to connect them. Instead, a gap (i.e. stretches of “N” characters) is inserted as placeholder. When multiple scaffolding procedures are performed, gaps defined in previous iterations are inherited in the subsequent steps. HaploFill uses several reference-independent strategies to identify the specific kind of gap and the correct filler sequences. A gap in a scaffold may occur when there is insufficient reliable information to identify the correct sequence for the region. This can happen when there are a lack of digestion sites in optical maps, a shortage of markers for HaploSplit, or when multiple alternative sequences are linked with proximity ligation data (e.g., mate-pair library, HiC libraries). A gap may also occur in a scaffold when the sequence is unavailable for placement. This can occur if it was not assembled or if one consensus sequence was produced from multiple genomic loci (e.g., repeats). This might also happen in diploid assemblies at homozygous regions where no alternative sequence is produced.

HaploFill is designed to recover gap information by comparing the gap region to the sequence present in the alternative haplotype. First, unplaced sequences are searched for the missing constituent. If no suitable candidate is found, the gap is filled using the alternative allele’s sequence. HaploFill does the following steps. First, HaploFill will try to determine the ploidy of each region using sequencing coverage information: (i) align long or short sequencing reads onto each haplotype separately and calculate the base coverage along each pseudomolecule using Bedtools (29); (ii) calculate the expected haploid depth of coverage with a Savitzky–Golay filter for each pseudomolecule, excluding annotated repetitive regions; (iii) classify each region of the genome as uncovered, haploid, diploid, and repetitive based on the ratio between the depth of coverage and the expected haploid depth of coverage. Thresholds can be defined by the user. For each gap, HaploFill extracts the region upstream and downstream of the gap and the corresponding regions on the alternative haplotype to build support sequences that will assist the search for filler. If the alternative region is reliably diploid (i.e. neither repetitive nor extensively gapped on the opposite haplotype) HaploFill will (i) create a hybrid support sequence made of the regions flanking the gap and the regions corresponding to the gap on the alternative haplotype, (ii) create an alternative support region made of the regions that correspond to and flank the gap on the alternative haplotype. If the region that corresponds to the gap on the alternative haplotype is highly repetitive or gapped, HaploFill will create two alternative support sequences made of the regions flanking the gap on the two haplotypes. HaploFill will then search for gap filler among the unplaced sequences. To do this, HaploFill will first map unplaced sequences onto all the support regions with Nucmer (30). Unplaced sequences are assigned globally in a 1-to-1 relationship to supporting sequences. Pairings are ranked based on the bases that match non-repetitive portions of the support sequence and the whole support sequence. Then, the best filler is assigned to the gap. Filler priority is given to the hybrid support region filler, followed by the alternative support region, and then to the gapped support regions. If no filler can be validated to cover the gap but the corresponding region is classified as diploid based on sequencing coverage, the region is assumed to be homozygous. In this scenario, the region on the alternative haplotype corresponding to the gap is used as a filler. Like HaploSplit, HaploFill allows a wide range of user defined relationships between sequences to fine tune the filler selection procedure. For example, the relationship between the primary and haplotigs can be used to consistently place alternative sequences.

**E. HaploMake.** HaploMake automates the conversion of sequences and annotations between different assembly versions. As input, it accepts the FASTA of the genome and a structural file (e.g., AGP files, BED, and HaploFill output files) that describes the new sequence configuration. If a gene annotation, markers, or contig structures are given, HaploMake will automatically translate their coordinates rel-
Fig. 3: Example of HaploDup’s interactive reports. The figure reports two static screenshots exemplifying HaploDup interactive output. A) Assembly quality control of *M. rotundifolia* chromosome 12 Haplotype 1: whole-sequence alignment of both alternative haplotypes on Haplotype 1, legacy contig and hybrid scaffold composition of Haplotype 1, position of the genetic markers and the duplicated markers in Haplotype 1, number of significant alignment(s) per gene of Haplotype 1 in each alternative haplotype. In this example, the composition in legacy contigs and position of duplicated markers indicate that both alleles (primary contig and haplotig) and both marker copies were placed in a hybrid scaffold (red box). B) Unplaced sequence quality control: Marker content is compared between pseudomolecules and unplaced sequences to evaluate conditions that prevent the inclusion of a specific unplaced sequence. Color-coding is used to show the existing relationships between the composing sequences (e.g., primary to haplotig relationships). In this example, the presence of a marker (green box, the dark violet marker on the left of the contig) in the unplaced sequences far from its expected position on the map extends the expected coverage of the map to the end of the linkage group and prevents placement in any haplotype scaffold.
ative to the new sequence. The ends of adjacent regions in the structure files can be checked for overlaps with Nucmer (30). The coordinates of adjacent regions can be corrected by adjusting junction positions. This avoids duplicating genomic content in the final sequence and can be done without altering the gene annotation (Supplemental figure 9).

F. HaploMap. HaploMap (Supplemental figure 10) performs a pairwise comparison between haplotypes and delivers a pairwise tiling map of colinear, non-overlapping, and non-repetitive regions between different haplotypes. Like HaploSplit, local alignments between each pair of sequences are performed with Minimap2 (28) or Nucmer (30). Hits are used to create a weighted adjacency graph for identifying a bidirectional tiling path that maximizes the identity between the two sequences. The coordinates of the colinear regions that form the bidirectional tiling path are listed in a pairwise, phased map of matching sequences.

G. Testing datasets. HaploSync performance was tested using a wide range of species and assembly protocols (Table 1). The diploid Candida albicans draft assembly (31), built using PacBio reads and FalconUnzip (17), was anchored to chromosomes using the genetic map generated by (32). A diploid genome assembly of Arabidopsis thaliana Columbia-0 (Col-0) X Cape Verde Islands (Cvi-0) (17) was anchored using a genetic map from (33). The Bos taurus Angus x Brahma genome from (34) was assembled using FalconUnzip, anchored to chromosomes using the genetic map from (12), and integrated with sex chromosome information from the Integrated Bovine Map from Btau_4.0 release available from https://www.hgsc.bcm.edu/other-mammals/bovine-genome-project. To support the assembly and quality control of pseudomolecule reconstruction, the locations of unique genes from the respective reference annotations (C. albicans SC5314_A22, A. thaliana TAIR10, B. Taurus Btau_ARS-UCD1.2) were identified by mapping CDS sequences on primary and haplotype sequences using GMAP (ver. 2019.09.12 (35)). Unique gene models were defined by mapping CDS sequences from the reference genomes annotations on the respective reference genome sequences using GMAP (ver. 2019.09.12 (35)). All CDS mapping on multiple locations in the haploid genome sequences were not represented in the pseudomolecules produced by HaploSplit. This number was reduced to 182 (3.0%) in a single iteration of HaploFill. BUSCO analysis confirmed the nearly complete separation of alternative alleles with only five complete gene models found in multiple copies in Haplotype 1 (3 genes) and Haplotype 2 (2 genes) (Supplemental table 1). With 18.6 ± 0.6 markers/Mbp, the genetic map of B. taurus autosomal chromosomes was the most dense out of the species used in this study. HaploSplit produced pseudomolecules almost identical in size to the ARS-UCD1.2 genome assembly (39), with Haplotype 1 pseudomolecules deviating by 0.7 ± 0.7% and Haplotype 2 by 6.7 ± 3.0% (Supplemental table 2, Supplemental figure 11). HaploFill inserted 151 Mbp, mostly in Haplotype 2 pseudomolecules, reducing missing information in Haplotype 2 pseudomolecules to 1.4 ± 1.9% of ARS-UCD1.2 chromosome sizes. For sex chromosomes, only a genetic map of the X chromosome with low marker density was available (2.1 markers/Mbp, assembly ver. Btau_4.0 available at https://www.hgsc.bcm.edu/other-mammals/bovine-genome-project). As a consequence, HaploSplit’s performance dropped. HaploSplit retrieved 79.8% of the expected 139 Mbp X chromosome. However, HaploFill reduced missing information to 9.7% (Supplemental table 2, Supplemental figure 11). Without markers available, the length of the Y chromosome was only 11% of its expected size (4.5 Mbp). The gene space was more complete in terms of single copy reference genes. Only 7 of 57,974 single copy CDSs mapping on the assembled sequences were not placed in the initial pseudomolecules produced with HaploSplit. This was reduced to 5 by HaploFill.

Results and discussion

To evaluate HaploSync’s performance, five diploid species from three different kingdoms were selected. This included a Muscadina rotundifolia, Vitis vinifera, and an F1 progeny of Arabidopsis thaliana (Col-0 x Cvi-0) (17), the bull B. taurus Angus x Brahma (12, 34), and pathogenic yeast Candida albicans. These species are diverse and vary in genome size, chromosome number, repeat content, and amount of heterozygosity. Long sequencing reads, genetic maps, and public reference genomes are available for those species.

H. HaploSync adaptability to different species. HaploSync produced high-quality genomes for all five species (Table 1). The resulting assemblies were nearly twice the size of their original haploid assemblies, with 1.87X to 2.03X their gene space represented (Supplemental table 1). This indicates that most of both haplotypes were assembled separately. High-density genetic maps and highly contiguous draft assemblies enabled HaploSplit to produce high-quality pseudomolecules that differed 5.8 - 17.8% from their expected chromosome sizes. In one iteration, HaploFill increased assembly completeness and reduced the difference in length between haplotypes. For C. albicans, the limited number of markers were used to anchor 2.4 Mbp of sequences to pseudomolecules in HaploSplit. The assembly had the highest share of unplaced sequences (5.2Mbp), but HaploFill recovered 17.9% of the missing genomic content in one iteration. The final pseudomolecules were up to 97.9% complete. Only 231 (3.8%) of 6,079 single copy genes in the reference annotation mapping on the assembled sequences were not represented in the pseudomolecules produced by HaploSplit. This number was reduced to 182 (3.0%) in a single iteration of HaploFill. BUSCO analysis confirmed the nearly complete separation of alternative alleles with only five complete gene models found in multiple copies in Haplotype 1 (3 genes) and Haplotype 2 (2 genes) (Supplemental table 1). With 18.6 ± 0.6 markers/Mbp, the genetic map of B. taurus autosomal chromosomes was the most dense out of the species used in this study. HaploSplit produced pseudomolecules almost identical in size to the ARS-UCD1.2 genome assembly (39), with Haplotype 1 pseudomolecules deviating by 0.7 ± 0.7% and Haplotype 2 by 6.7 ± 3.0% (Supplemental table 2, Supplemental figure 11). HaploFill inserted 151 Mbp, mostly in Haplotype 2 pseudomolecules, reducing missing information in Haplotype 2 pseudomolecules to 1.4 ± 1.9% of ARS-UCD1.2 chromosome sizes. For sex chromosomes, only a genetic map of the X chromosome with low marker density was available (2.1 markers/Mbp, assembly ver. Btau_4.0 available at https://www.hgsc.bcm.edu/other-mammals/bovine-genome-project). As a consequence, HaploSplit’s performance dropped. HaploSplit retrieved 79.8% of the expected 139 Mbp X chromosome. However, HaploFill reduced missing information to 9.7% (Supplemental table 2, Supplemental figure 11). Without markers available, the length of the Y chromosome was only 11% of its expected size (4.5 Mbp). The gene space was more complete in terms of single copy reference genes. Only 7 of 57,974 single copy CDSs mapping on the assembled sequences were not placed in the initial pseudomolecules produced with HaploSplit. This was reduced to 5 by HaploFill.
BUSCO analysis confirmed the completeness and the separation of the alleles, with 92.5% complete gene models found in Haplotype 1 (1.3% in multiple copies) and 86.8% in Haplotype 2 (1.2% multiple copies) (Supplemental table 1).

In plants, the high level of polymorphism and structural variation between haplotypes make assembly and phasing challenging (17).

The high level of heterozygosity in the A. thaliana accession used to test HaploSync is caused by sequence variation between its parents, Col-0 and Cvi-0. This led to a primary assembly 17% longer and haplotigs 11.8% shorter than the haploid reference genome. After Haplosync, the two sets of pseudomolecules differed by 3.6% and 6.3% from the haploid reference genome size. This supports the tool’s ability to phase duplicated primary content between haplotypes. When gene space completeness was estimated using single-copy genes in the reference annotation, similar results were obtained. The amount of single copy CDSs mapping on the assembled sequences represent the 99.7% of the entire dataset (34,741 out of 34,854). After HaploSplit, unplaced sequences included 1,966 putative loci (5.7%). Of these, 261 (1.1%) were missing from the pseudomolecules. HaploFill further increased the completeness of the pseudomolecules to include 98.1% and 95.8% of the gene space in the two haplotypes. This reduced the putative, single copy CDS loci among unplaced sequences to only 123. Over 97% complete BUSCO gene models were complete in Haplotype 1 and the Haplotype 2, with only 1.3% and 1.5% in multiple copies, respectively (Supplemental table 1).

Vitis species can be 12% heterozygous (40). Assemblies of the species can exhibit extensive loss of phase between primary sequences and associated haplotigs (17, 18, 41–43). In Cabernet Franc, for example, the primary assembly is inflated by 18.8% and haplotigs are 40.7% shorter than the expected haploid genome size. HaploSync was able to overcome these limitations for both species and placed over 93.0% of the sequences in phased pseudomolecules that were no more than 9.8% different in size. HaploSplit also automatically placed and correctly phased the grape sex determining region (13) in Muscadinia and Vitis species. Using the unique CDS sequences from PN40024 as a reference for Vitis gene space, 1,233 (6.2%) of genes could not be placed in Cabernet Franc pseudomolecules with HaploSplit and 223 (1.4%) of genes could not be placed in M. rotundifolia pseudomolecules. This fraction of gene coding sequences could not be placed because of high fragmentation and low, uneven marker density that negatively affected pseudomolecule reconstruction performance. Several iterations of HaploFill reduced the number of unplaced CDSs to 0.3% for both genomes. This included 91 and 46 unique genes among unplaced sequences for Cabernet Franc and Trayshed, respectively. This highlights HaploFill’s ability to recover gene space information. Completeness and phasing of both Haplotypes was confirmed with BUSCO: 93% complete models in Haplotype 1 and 83% in Haplotype 2.

I. HaploSync performance adaptability to different assembly procedures. HaploSync was applied to two grapes, M. rotundifolia cv. Trayshed (37) and V. vinifera cv. Cabernet Franc (36), to assess its adaptability to genomes assemblies produced using different strategies. Although contigs were produced with PacBio data and FalconUnzip for both draft assemblies, Trayshed and Cabernet Franc were scaffolded with different technologies. M. rotundifolia underwent hybrid scaffolding with PacBio and a NGM map, which matches optical fingerprints of DNA molecules with assembled sequences digested in silico with the same enzyme. Gaps were introduced where there was a low density of digestion sites. Systematic errors were observed at highly repetitive and heterozygous regions, including the RNU1/RPV1 locus on chromosome 12 (Supplemental figure 2). The differential expansion of TIR-NBS-LRR genes between haplotypes (37) may have caused their fusion in the same scaffold. These issues affected 50 hybrid scaffolds (326.2 Mbp), required correction, and were easily found with HaploDup. For Cabernet Franc, scaffolding was performed using HiC data that produced chimeric scaffolds due to the presence of diploid information in the primary assembly. Both haplotypes of 108 of scaffolds (449 Mbp) were included in the same assembled sequence (Supplemental figure 4). After scaffold correction, both genome assemblies were anchored to chromosomes using a Vitis consensus genetic map (38). Low specificity and marker density (3.5 markers/Mbp) affected the construction of pseudomolecules by HaploSplit and negatively affected HaploSync’s performance. Cabernet Franc was most affected, with only 350.8 Mbp and 263.4 Mbp placed on Haplotype 1 and Haplotype 2, respectively (i.e. 75% and 55% of the reference haploid genome). Unplaceable sequences were nearly half of Cabernet Franc’s expected haploid genome size (240Mbp). Trayshed’s assembly was more complete; Haplotype 1 and Haplotype 2 assemblies were 374.3 Mbp and 338.8 Mbp long, respectively.

Three iterations of HaploFill were performed on Cabernet Franc’s assembly. Each iteration reduced unplaced sequences by nearly one half (Supplemental figure 12). The final Cabernet Franc pseudomolecules were 456 Mbp (Haplotype 1) and a 411 Mbp (Haplotype 2). Afterwards, 47 Mbp (5.4%) of sequences remained unplaced. In contrast, only two iterations of HaploFill were sufficient to leave just 8% of Trayshed sequences unplaced. Haplotype 1 and Haplotype 2 of Trayshed’s pseudomolecules were 400 Mbp and 370 Mbp, respectively. The total sizes of both haplotypes in both chromosome-scale assemblies were similar to their expected haploid reference genome sizes (44, 45) and to Cabernet Sauvignon’s haplotypes (459 Mbp and 449 Mbp, respectively) (13).

J. HaploSync performance assessment. The performance of different HaploSync tools, in terms of result quality and processing time, are influenced by multiple factors. Unsurprisingly, the genome size and the number of linkage groups affect all assembly phases and the duration of alignment procedures. For HaploDup, HaploFill, HaploMap, HaploBreak, and HaploMake, genome size determines the size of the output and how long alignments take to complete, which can constitute over 90% of the computational
time. The number of linkage groups exponentially increase the number of comparisons and plots needed. For example, HaploDup required 40 hours to process B. taurus, which has a 2.6 Gbp haploid genome size in 30 linkage groups and is the largest dataset used in this study. Nearly 15 of these hours were consumed for alignments between sequences while using 24 cores. C. albicans is the smallest dataset, with 14 Mbp in 8 linkage groups. In contrast to B. taurus, the same procedure required 75 minutes, with only 5 minutes dedicated to mapping.

HaploFill performance is also affected by the number of phased genomic sequences in the pseudomolecules. Alternative pseudomolecules are the backbone that enable the algorithm to retrieve gap filling information. The completeness of the pseudomolecules directly affects the amount of information usable as support for sequence placement. Unplaced sequences are information that might be recovered. The workflows adopted for A. thaliana and for the Vitis genotypes were selected based on pseudomolecule completeness. The A. thaliana assembly had relatively low sequence fragmentation and a high density map. The pseudomolecules created for A. thaliana with HaploSplit were fairly complete after a single filling procedure. HaploSplit was less effective for Cabernet Franc and Trayshed because their assemblies were more fragmented and their maps were less dense. The workflow used for the grape genomes included several iterations of HaploFill to achieve highly complete pseudomolecules (Supplemental figure 12).

HaploSplit is fast. It takes between a few seconds and one minute to build the adjacency graph, traverse it, find the two best tiling paths, and report the structure of the phased pseudomolecules. In contrast, the input quality control and the alignment between the draft sequences and the guide genome in preparation for the graph creation can be time-consuming. HaploSplit result quality is affected by several factors. The disparity and incomplete representation of both alternative alleles affects the completeness of the diploid pseudomolecules produced and necessitates filling. A. thaliana and B. taurus are F1 progeny. Their considerable structural variability is captured by the FalconUnzip assembler, which reconstructs the alleles fully and separately. In contrast, Cabernet Franc and Trayshed have several homozygous regions that were assembled in a single copy and highly heterozygous regions that increased the fragmentation of the contigs by foiling the assembler into overassembling the primary sequences. This difference is reflected in HaploSplit’s results. HaploSplit was able to separate alleles and deliver a nearly complete diploid assembly of A. thaliana and B. taurus. Vitis required a more extensive filling procedure to recover the missing information.

HaploSplit can use a genetic map and/or a guide-genome as information to facilitate scaffolding. We tested how HaploSplit performs given different scaffolding information using Vitis vinifera cv. Cabernet franc cl. 04 (36). Reference genomes of closely related accessions, PN40024 (45) and Cabernet Sauvignon (13), are available. Cabernet Franc contigs were scaffolded (i) with the genetic map of (38), (ii) using the PN40024 V2 assembly or the first haplotype of Cabernet Sauvignon as guides, or (iii) using both the genetic map and a guide genome. The reference-based approach incorporated more sequences into pseudomolecules than when only a genetic map was used. As expected, the best results were obtained using Cabernet Sauvignon as a reference, which shares one allele with Cabernet Franc. This approach, however, led to overfitting of the scaffolding results to the guide. Small structural variants in long draft sequences (Supplemental figure 13 A boxes) can find a proper representation thanks to neighbouring colinear regions. Larger structural variants that encompass multiple sequences may fail to be reported correctly together. Each draft sequence location is identified independently from the others based on colinearity with guide genome, so placement is based on the structure of guide sequences rather than their actual order (Supplemental figure 13 B boxes). Moreover, gaps or the lack of information in the guide genome may impede the recovery of novel information. Only draft sequences that partially anchor within present information can be placed (Supplemental figure 13 C boxes). As a consequence, fragmented draft assemblies and the second haplotype are prone to be artificially similar to the guide genome. The hybrid approach performs better. The reconstruction of both haplotypes is more complete than the map-based approach. with the second haplotype benefiting most from this strategy (Fig. 4 A). Though no overassembly was observed, the mapping phase duplicated some alleles. Both copies of several markers occurred in the same pseudomolecule scaffold when Cabernet Sauvignon (5 markers) and PN40024 (4 markers) were used as guides.

The effect of the number of reliable genetic markers on the performance of HaploSplit was tested on the genome of B. taurus Angus x Brahma (12, 34). The same diploid genome underwent chromosome scale reconstruction using a randomly selected subset of 479 markers (1%, 0.2 markers/Mbp) out of its available genetic map (46,323 markers, 17.6 markers/Mbp; Fig. 4 B and C, Supplemental figure 14). Unsurprisingly, the number of unplaced sequences increased to 37% of the total assembly length given lower marker density. HaploSplit found a location in pseudomolecules for 99.4% to 100% of the sequences with markers. The performance of the algorithm, in terms of completeness of the delivered pseudomolecules, is primarily influenced by input assembly fragmentation and the genetic map’s marker density. This limits the number and the sizes of the sequences with markers. The primary assembly is composed of extremely long sequences that likely contain markers and are placed even when map density is low. In contrast, Haplontigs are more fragmented and require high marker density for comparable coverage. As a result, the first haplotype assembly is more complete even with fewer markers present (Fig. 4 B and C).

In summary, the type and quality of the external guide information have a large effect on the quality of the final assembly. A guide genome aids assembly via local sequence alignments; lack of homology between sequences and repetitive regions can cause segregation errors (Supplemental figure 6),
HaploSync suite of tools can be used to address some of the remaining impediments to genome reconstruction and improve assembly quality by taking advantage of diploid information that is readily available. HaploSync correctly and completely phases diploid genomes, reconstructs pseudomolecules by recovering missing information, and exerts quality control over the results.

Web resources

HaploSync is freely available for download at GitHub https://github.com/andreaminio/haplosync. Instructions for installation, a full list of dependencies, a description of each tool, and tutorials are available in HaploSync’s manual (https://github.com/andreaminio/HaploSync/tree/master/manual).

Data availability

The data used in this study are summarized in Supplemental table 3. Pseudomolecule reconstructions of *Candida albicans* NCYC4145 (31), *Arabidopsis thaliana* Columbia-0 (Col-0) X Cape Verde Islands (Cvi-0) (17), and *Bos taurus* Angus x Brahma (34) are available at Zenodo (https://zenodo.org/record/3987518, DOI 10.5281/zenodo.3987518). *Vitis vinifera* cv. Cabernet Franc cl. 04 (36) and *Muscadinia rotundifolia* cv. Trayshed (37) pseudomolecule assemblies are available at www.grapegenomics.com.

Funding

This work was funded by the National Science Foundation (NSF) award #1741627 and the US Department of Agriculture (USDA)-National Institute of Food and Agriculture (NIFA) Specialty Crop Research Initiative award #2017-51181-26829. It was also partially supported by funds from E.&J. Gallo Winery and the Louis P. Martini Endowment in Viticulture.

Authors’ contributions

AM and DC conceptualized the project. AM developed the methodology and software. AM and NC performed bioinformatic analyses and tested the software. AM, NC, and AMV wrote the software manual and pipeline walk-through. AM, AMV, MM, and DC wrote the manuscript. DC secured the funding and supervised the project. All authors have read and approved the manuscript.

Bibliography

1. Anthony Rhoads and Kin Fai Au. PacBio Sequencing and its applications. Genomics, Proteomics and Bioinformatics, 13(5):278–289, 2015. ISSN 22103244. doi: 10.1016/j.gpb.2015.08.002.
2. Philipp Bongartz and Siegfried Schloissnig. Deep repeat resolution—the assembly of the *Drosophila* Histone Complex. Nucleic Acids Research, 47(3):e18–e18, feb 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1194.
3. Huilong Du and Chengzhi Liang. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. Nature Communications, 10(1):1–10. 2019. ISSN 20411723. doi: 10.1038/s41467-019-13555-3.
4. Prita Paajanen, George Kettleborough, Elena López-Girona, Michael Giolai, Darren Heav- 

5. Kerrin S. Small, Michael Brudno, Matthew M. Hill, and Arend Sidow. A haplome align-

6. Hua Ying, David C. Hayward, Ira Cooke, Weiwen Wang, Aurelie Moya, Kirby R. Siemering, 

7. Alex Di Genova, Andrea M. Almeida, Claudia Muñoz-Espinoza, Paula Vizoso, Dante Trav-

8. Hideki Hirakawa, Kenta Shinsawa, Shunaki Kosugi, Kosuke Tashiro, Shinobu Nakayama, 

9. Lorenzo Barchi, Marco Pietrella, Luca Venturini, Andrea Minio, Laura Toppino, Alberto 

10. Hua Ying, David C. Hayward, Ira Cooke, Weiwen Wang, Aurelie Moya, Kirby R. Siemering, 

11. Jin Kim, Karthik N. Saih, Hideki Noguchi, Mikaela Ayers-Toyoda, Yoshitoshi Ogura, Miki 

12. Pirita Paajanen, George Kettleborough, Elena López-Girona, Michael Giolai, Darren Heav-

13. Kerrin S. Small, Michael Brudno, Matthew M. Hill, and Arend Sidow. A haplome align-

14. Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P. Smith, and Adam M. 

15. Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P. Smith, and Adam M. 

16. Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P. Smith, and Adam M. 

17. Alex Di Genova, Andrea M. Almeida, Claudia Muñoz-Espinoza, Paula Vizoso, Dante Trav-

18. Kerrin S. Small, Michael Brudno, Matthew M. Hill, and Arend Sidow. A haplome align-

19. Lorenzo Barchi, Marco Pietrella, Luca Venturini, Andrea Minio, Laura Toppino, Alberto 

20. Stefano Nardelli, Giuseppe Ferrari, Anna Volpato, Antonio Giavedoni, Andrea Minio, 

21. Andrea Walberg, Ignas Bunkis, Olga Vinniere, Mia-Britt Mosbach, Anna K. 

22. Karen H. Miga, Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrea 

23. Arthur Akopyan, Wei-Wei Yang, Hailin Li, Yulan Yang, Tao Zhang, Pasi Sainio, Yehuda 

24. Pirita Paajanen, George Kettleborough, Elena López-Girona, Michael Giolai, Darren Heav-

25. Patrick S. Schnable, Eric Lyons, and Jianguo Lu. ALLMAPS: Robust scaffold ordering 

26. Patrick S. Schnable, Eric Lyons, and Jianguo Lu. ALLMAPS: Robust scaffold ordering 

27. Paarth S. Joshi, Qi Sun, and Lance Cadle-Davidson. Haplotyping the Vitis collinear core 

28. Patrick S. Schnable, Eric Lyons, and Jianguo Lu. ALLMAPS: Robust scaffold ordering 

29. Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing 

30. Anja Ferchle, B. M. Magee, and Georgiana May. Genome-wide single nucleotide polymorphism map for Candida albicans. Eur. Sacchar. J. 33(3):705–14, 2004. ISSN 1535-9757. 

31. Jennafer A. P. Hamlin, Guilherme B. Dias, Casey M. Bergman, and Douda BhensCan. 

32. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

33. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

34. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

35. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

36. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

37. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

38. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

39. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

40. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

41. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

42. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

43. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

44. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

45. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

46. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

47. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

48. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

49. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

50. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

51. Tanja Singer, Yiping Fan, Hua-Hua Peng, Long-Tzu, Samuel P. Williams, and Steven P. 

52. tanja.singer@ucdenver.edu
Jinna Hoffman, Brenda M. Murdoch, Warren M. Snelling, Tara G. McDanel, John A. Hammon, John C. Schwartz, Wilson Nandolo, Darren E. Hagen, Christian Dreischer, Sebastian J. Schultheiss, Steven G. Schroeder, Adam M. Philippy, John B. Cole, Curtis P Van Tassell, George Liu, Timothy P.L. Smith, and Juan F. Medrano. De novo assembly of the cattle reference genome with single-molecule sequencing. GigaScience, 9(3):1–9, mar 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa021.

40. Riccardo Velasco, Andrey Zharkikh, Michela Troggio, Dustin A. Cartwright, Alessandro Cestaro, Dmitry Prusak, Massimo Pinto, Lisa M Fitzgerald, Silvia Vezzulli, Julia Reid, Giulia Malacarne, Diana Ilevu, Giuseppina Coppola, Bryan Wardell, Diego Micheletti, Teresita Macalma, Marco Faczi, Jeff T. Mitchell, Michele Perazzoli, Glenn Eldredge, Pamela Gatto, Rozan Ozyerski, Marco Moretto, Natalia Gutin, Marco Stefani, Yang Chen, Cinzia Segala, Christine Davenport, Lorenzo Dematte, Amy Miraz, Juri Battilana, Keith Stormo,Fabrizio Costa, Guanhuo Tao, Azeddine Si-Ammour, Tim Harris, Angie Lackey, Ciclilde Perbost, Bruce Talloin, Alessandra Stella, Victor Soleoyev, Jeffrey A. Fawcett, Lieven Sterck, Klaas Vandepoele, Stellia M. Grandi, Stefano Toppo, Claudio Moser, Jerry Lanchbury, Robert Bigden, Mark Skolnick, Vittorio Sgardamella, Salath K. Bhanagar, Paolo Fontana, Alexander Gutin, Yves Van de Peer, Francesco Salamini, and Roberto Viola. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS one, 2(12): e1326, dec 2007. ISSN 1932-6203. doi: 10.1371/journal.pone.0001326.

41. Michael J. Roach, Simon A. Schmidt, and Anthony R. Borneman. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics, 19(1): 460, dec 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2485-7.

42. Amanda M. Vondras, Andrea Minio, Barbara Blanco-Ulate, Rosa Figueroa-Balderas, Michael A. Penn, Yongteng Zhou, Danielle Seymour, Zirou Ye, Dongren Liang, Lucero K. Espinoa, Michael M. Anderson, M. Andrew Walker, Brandon Gaut, and Dario Cantu. The genomic diversification of grapevine clones. BMC genomics, 20(1):972, 2019. ISSN 14712164. doi: 10.1186/s12864-019-6211-2.

43. Yongteng Zhou, Andrea Minio, Melanie Massonnet, Edwin Solares, Yuanda Lv, Tengiz Beridze, Dario Cantu, and Brandon S. Gaut. The genomic diversification of grapevine clones. Nature Plants, 5(9):965–979, 2019. ISSN 20550278. doi: 10.1038/s41477-019-0507-8.

44. Olivier Jaillon, Jean Marc Aury, Benjamin Noel, Alberto Policriti, Christian Clepet, Alberto Casagrande, Nathalie Choisne, Sébastien Aubourg, Nicola Viuto, Claire Jablin, Alessandro Vezzi, Fabrice Legaai, Philippe Hugueney, Corinne Dasliva, David Horner, Erica Mica, Delphine Jubitot, Julie Poulain, Claire-Emmanuelle Bruley, Alain Billault, Béatrice Seguere, Michel Gouyvenoux, Edgardo Ugarti, Federica Cattonani, Véronique Anthouard, Virginie Vico, Cristian Del Fabbro, Michaeli Aiaux, Gabriele Di Gaspero, Vincent Dumas, Nicoletta Felice, Sophie Paillard, Irena Juman, Marco Moroldo, Simone Scalabrin, Aurélie Canaguier, Isabelle Le Clainche, Giorgio Malacrida, Eleonore Durand, Graziano Pesole, Valérie Laucou, Philippe Chatelot, Didier Merdngou, Massimo Delledonne, Mario Pezzotti, Alan Lechamry, Claude Scarpell, François Antguenave, M. Enrico Pè, Giorgio Valle, Michelle Mogant, Michel Cabcio, Anne Françoise Adam-Blondon, Jean-Weissenbach, Francis Qultier, and Patrick Wincker. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 449(7161):463–467, sep 2007. ISSN 14764687. doi: 10.1038/nature06148.

45. Aurélie Canaguier, Jérôme Grimplet, Gabriele Di Gaspero, Simone Scalabrin, Éric Duchêne, Nathalie Choisne, Nacer Mohellibi, Cécile Guichard, Stéphane Rombauts, Isabelle Le Clainche, Aurélie Bérard, Aurélie Chauveau, Rémi Bouron, Camille Rustenholz, Michele Morgante, Marie Christine Le Paslier, Dominique Brunel, and Anne-Francoise Adam-Blondon. A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). Genomics data, 14(July):56–62, dec 2017. ISSN 2213-5960. doi: 10.1016/j.gdata.2017.09.002.

46. Sergey Koren, Arang Rhee, Brian P. Walenz, Alexander T. Dillhey, Derek M. Bickhart, Sarah B. Kinger, Stefan Hiendleder, John L. Williams, Timothy P.L. Smith, and Adam M. Philippy. De novo assembly of haplotype-resolved genomes with trio binning. Nature Biotechnology, 2018. ISSN 15461696. doi: 10.1038/nbt.4277.
Table 1: Assembly statistics

| Genotype            | Kingdom | Haploid Size | Chromosomes | Technology | Markers (per Mbp) | Input sequences | Results | HaploSplit | HaploFill |
|---------------------|---------|--------------|-------------|------------|-------------------|-----------------|---------|------------|-----------|
|                     |         |              |             |            |                   |                 |         |            |           |
| C. albicans         | Fungi   | 14 Mbp       | 7 + R       | PacBio     | 116 (8.3)         | Primary: 15.5 Mbp | HaploSplit | 11.6 Mbp  | 12.9 Mbp  |
|                     |         |              |             |            |                   | Haplotigs Total: 13.8 Mbp | 29.2 Mbp  |           |           |
|                     |         |              |             |            |                   | Primary: 140.0 Mbp | HaploFill | 109.0 Mbp | 114.7 Mbp |
|                     |         |              |             |            |                   | Haplotigs Total: 104.9 Mbp | 245.0 Mbp |           |           |
| A. thaliana         | Plantae | 119 Mbp      | 5           | PacBio     | 676 (5.7)         | Primary: 2.7 Gbp | HaploSplit | 2.6 Gbp   | 2.6 Gbp   |
|                     |         |              |             |            |                   | Haplotigs Total: 2.5 Gbp | 5.2 Gbp   |           |           |
| B. taurus           | Animalia| 2.62 Gbp (29+X) | 29 + XY   | PacBio     | 46,325 (17.6)    | Primary: 2.7 Gbp | HaploSplit | 2.6 Gbp   | 2.6 Gbp   |
|                     |         | 2.49 Gbp (29 + Y) |         |            |                   | Haplotigs Total: 5.2 Gbp |           |           |           |
| V. vinifera cv. Cabernet Franc | Plantae | 487 Mbp 557 Mbp | 19 | PacBio + Dovetail HiC | 1,661 (3.5) | Primary: 570.2 Mbp | HaploSplit | 350.8 Mbp  | 455.6 Mbp  |
|                     |         |              |             |            |                   | Haplotigs Total: 284.7 Mbp | 410.9 Mbp |           |           |
| M. rotundifolia     | Plantae | 483 Mbp      | 20          | PacBio + BioNano | 1,661 (3.5) | Primary: 459.5 Mbp | HaploSplit | 374.3 Mbp  | 400.5 Mbp  |
|                     |         |              |             |            |                   | Haplotigs Total: 364.8 Mbp | 370.0 Mbp |           |           |
|                     |         |              |             |            |                   | Total: 896.0 Mbp | HaploFill | 165.5 Mbp | 63.0 Mbp   |

1. FalconUnzip Hamlinet al. (2019) (31)
2. Forche et al. (2004) (32)
3. FalconUnzip, Chinet al. (2016) (17)
4. Singer et al. (2006) (33)
5. FalconUnzip, Koren et al. (2018) (46)
6. Low et al. (2020) using the Integrated Bovine Map of sex chromosome (ver. Btau_4.0, https://www.hgsc.bcm.edu/other-mammals/bovine-genome-project) (12)
7. Haploid genome size estimated to be in the range 487-557 Mbp as reported for PN40024 in (44) and Cabernet Sauvignon in (37), FalconUnzip + SSPACE + HiRise, Vondras et al. (2021)
8. Zou et al. (2020) (38)
9. FalconUnzip + Hybrid Scaffolder, Cochetel et al. (2021) (37)
10. Zou et al. (2020) (38)
11. Reported for FalconUnzip assembly as haplotype separation is lost during Hybrid Scaffolding
12. Reported for FalconUnzip assembly as haplotype separation is lost during Hybrid Scaffolding
Supplemental figure 1: HaploSplit workflow diagrams for genome guided pseudomolecule reconstruction and hybrid reconstruction with both map and guide genome information. HaploSplit reconstructs pseudomolecules using colinearity with the guide-genome to sort sequences into different haplotypes. i) Query draft sequences are aligned onto the guide genome and used to build a directed adjacency network of non-overlapping hits. ii) The tiling path that maximizes the number of identical bases between the query and guide sequence is selected. iii) A directed adjacency network of all query sequences associated with a guide sequence is created. iv) The tiling path with the highest number of identical bases with the guide sequence is used to generate the first haplotype. v) Sequences belonging to the first haplotype are removed from the adjacency network and the second-best tiling path is used to scaffold the second haplotype. In hybrid mode, after performing the reconstruction with the genomic map, the intermediate pseudomolecules and unplaced sequences are mapped on the genome are mapped on the guide genome; ii) location of each draft input sequence composing the pseudomolecule is translated from whole sequence local alignments; iii) an adjacency graph graph of all mapped draft sequences is created; iv) the two best tiling paths are identified and final pseudomolecules are delivered.
Supplemental figure 2: HaploSplit input QC interactive plot. Example of interactive plot generated by HaploSplit to perform a quality control of the input sequences as an intra-sequence duplication of genetic markers has been detected. The self-alignment of the sequence is performed by 1 Kbp windows to increase the resolution of duplicated regions identification. Markers present in the sequence and duplicated ones, are also reported and, when available, the sequence structure and the gene copy count tracks are layered to support the duplication analysis. The plot evidence and example of systematic errors observed in correspondence to highly repetitive and heterozygous regions when FalconUnzip produced contigs undergo Hybrid Scaffolding with BioNano NGM maps.

In the example, RUN1/RPV1 locus on chromosome 12 (same as Figure 3). Different expansion of TIR-NBS-LRR genes between haplotypes is the probable cause the fusion of both haplotypes of the region in the same scaffold. These issues affected 50 hybrid scaffolds (326.2Mbp) requiring correction.
Supplemental figure 3: Schematic representation of HaploDup procedure. Diagram of HaploDup data usage to create Quality Control interactive plots.
Supplemental figure 4: Example of assembly error due to scaffolding of primary sequences with Dovetail HiC data. Scaffolding Falcon\Unzip produced primary contigs with HiC data produced chimeric scaffolds. This is due to the presence of both haplotype information in the overassembled primary set of contigs. As a result, one haplotype is split to accommodate the alternative one inside. For Cabernet Franc, correction was required by 108 of scaffolds encompassing 449Mbp.
Supplemental figure 5: Example of errors due to lack of overlap. Lack of collinearity information in the sequences extremities in HaploSplit may not allow a proper identification of homologous sequences. The example here reported shows an example. Plot was made using Nucmer and reporting all alignment hits without filtering. It is evident that the inner extremity of the two inserted sequences (in grey) do actually represent the same genomic region (in green). Due to the alignment with Minimap on the genome, however, the projection of the sequences on the guide genome did not report the conflicting overlap, thus allowing the placement of both sequences in the same pseudomolecule scaffold.
Supplemental figure 6: Example of errors due to marker sorting errors. Erroneous sorting order of markers may not allow placement of a sequence because of conflicting map coverage. The presence of the marker at position #123 of the map in the contig caused HaploSync to overestimate the extension of the range covered markers, making it impossible to place the draft sequence in any haplotype.

Supplemental figure 7: Schematic representation of HaploBreak procedure. HaploBreak selects appropriate breakpoints in mis-assembled regions by either breaking sequences at the closest scaffolding junction or gap. i) Given a pair of breakpoint coordinates, HaploBreak searches for the closest junction. If a scaffolding junction is found, it is associated with the breakpoint. Otherwise, the breakpoint is associated with the closest gap. ii) Then, HaploBreak breaks the assembly. If a pair of coordinates is associated with two different scaffolding junctions, the sequence between the coordinates is classified as misplaced. Otherwise, the sequence is broken at the gap.
Supplemental figure 8: Schematic of the HaploFill procedure. HaploFill procedure: i) calculates sequencing coverage along each pseudomolecule, ii) evaluates the ploidy and repetitiveness along each pseudomolecule using the sequencing depth information and the repeat annotation, respectively, and iii) classifies each region of the genome based on the expected haploid depth of coverage; iv) for each gap, HaploFill identifies the corresponding region on the alternative haplotype and generates support sequences, and maps unplaced sequences on support sequences; v) the best filling information is assigned to the gap, prioritizing (in order of importance) the hybrid support region filler, the alternative support region filler, then the gap support region filler; vi) if the gap is diploid but cannot be filled, the region on the alternative haplotype corresponding to the gap is used as filler.
Supplemental figure 9: Schematic of the HaploMake procedure. HaploMake does an overlap-free sequence reconstruction without detrimentally affecting the gene annotation. When overlaps between sequences occur, the coordinates are corrected to avoid duplicating genomic content.

Supplemental figure 10: Schematic of the HaploMap procedure. HaploMap reports a map of the relationship between the two haplotypes. i) HaploMap performs local alignments between each pair of sequences, ii) the tiling path that maximizes the identity between both sequences is found, and iii) the coordinates of colinear regions in the bi-dimensional tiling path are reported.
Supplemental figure 11: Results of B. taurus pseudomolecules reconstruction after HaploSplit and after HaploFill. Reconstructed pseudomolecules are compared in size to the ARS-UCD1.2 genome chromosome sequences. Each plots report a bar graph of the actual sizes and in terms of difference percentage from ARS-UCD1.2 genome chromosomes the expected genome size. In HaploSplit Haplotype 1 pseudomolecules deviate only by 0.7±0.7% from the expected size, Haplotype 2 by 6.7±3.0%. HaploFill further reduces the divergence, mostly in Haplotype 2 pseudomolecules where it goes down to 1.4±1.9%.
Supplemental figure 12: Results of HaploSync reconstruction of *M. rotundifolia* cv. Trayshed and *V. vinifera* cv. Cabernet Franc. Overview of pseudomolecule reconstruction results for the two Vitis species. The graphs report both pseudomolecule sequence size and gene content at each step of HaploSync pipeline.
Supplemental figure 13: Guide genome overfitting

Dotplots comparing chromosome 12 and chromosome 13 of PN40024, Cabernet Sauvignon and Cabernet Franc, reconstructed with HaploSplit using PN40024 or Cabernet Sauvignon as guide genome, to the sequences of PN40024 and Cabernet Sauvignon. A boxes (green): Structural variants present in Cabernet Franc are reported in the results when part of long draft sequences. B boxes (red): Draft sequence location an orientation are placed accordance to the guide genome structure. With higher fragmentation (ex. sequences used for the second haplotype) increases also the overfit to the guide genome. C boxes (blue): Lack of information in the guide genome (gaps in Cabernet Sauvignon sequence) do not allow to insert place information inside the missing region unless the draft sequences do not anchor to the known part of the pseudomolecule.
Supplemental figure 14: Performance of HaploSplit on markers subsampling. Overview of the HaploSplit performance in reconstructing B. taurus Angus x Brahma (12, 34) pseudomolecules using a subset of the full genetic map in terms of the input draft sequences placed. The results evidence the interrelated effect of map density and assembly fragmentation on the final assembly completeness, with only longer sequences used with the lowest amount of markers and shorter sequences requiring a higher density map to find a location.

A) Cumulative length of placed and unplaced Primary sequences and Haplotigs with each marker dataset; B) Percentage of Primary sequences and Haplotigs finding a placement in the assembled pseudomolecules with each marker dataset; c) distribution of Primary sequences and Haplotigs lengths, placed and unplaced with each marker dataset.
**Supplemental table 1: Gene content representation in the different assemblies.** Table reporting the full statistics of the genomes assembled as testing dataset for HaploSync. C. albicans, A. thaliana and B. taurus genomes gene counts were obtained by mapping unique CDS sequences from the respective reference genome annotations (C. albicans SC5314_A22, A. thaliana TAIR10, B. Taurus Blau_ARS-UCD1.2) using GMAP (35). The unique gene mapping datasets were obtained by mapping the CDS sequences of reference genome annotations to the respective reference genome, CDSs mapping in multiple locations in the haploid genome were removed from the dataset. Gene counts reported for V. vinifera cv Cabernet Franc and M. rotundifolia cv. Trayshed were obtained after performing the whole genome annotation as reported in (37) and (36) respectively.

| Chromosome | Minio A. et al. | HaploSync | Minio A. et al. | HaploSync | Minio A. et al. | HaploSync |
|------------|-----------------|-----------|-----------------|-----------|-----------------|-----------|
| 1          | 25              | 25        | 25              | 25        | 25              | 25        |
| 2          | 25              | 25        | 25              | 25        | 25              | 25        |
| 3          | 25              | 25        | 25              | 25        | 25              | 25        |
| 4          | 25              | 25        | 25              | 25        | 25              | 25        |
| 5          | 25              | 25        | 25              | 25        | 25              | 25        |
| 6          | 25              | 25        | 25              | 25        | 25              | 25        |
| 7          | 25              | 25        | 25              | 25        | 25              | 25        |
| 8          | 25              | 25        | 25              | 25        | 25              | 25        |
| 9          | 25              | 25        | 25              | 25        | 25              | 25        |
| 10         | 25              | 25        | 25              | 25        | 25              | 25        |
| 11         | 25              | 25        | 25              | 25        | 25              | 25        |
| 12         | 25              | 25        | 25              | 25        | 25              | 25        |
| 13         | 25              | 25        | 25              | 25        | 25              | 25        |
| 14         | 25              | 25        | 25              | 25        | 25              | 25        |
| 15         | 25              | 25        | 25              | 25        | 25              | 25        |
| 16         | 25              | 25        | 25              | 25        | 25              | 25        |
| 17         | 25              | 25        | 25              | 25        | 25              | 25        |
| 18         | 25              | 25        | 25              | 25        | 25              | 25        |
| 19         | 25              | 25        | 25              | 25        | 25              | 25        |
| 20         | 25              | 25        | 25              | 25        | 25              | 25        |
| 21         | 25              | 25        | 25              | 25        | 25              | 25        |
| 22         | 25              | 25        | 25              | 25        | 25              | 25        |

The copyright holder for this preprint (which this version posted May 2, 2022) is the author(s), who has/have granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.
Supplemental Table 2: Results of B. taurus pseudomolecules reconstruction after HaploSplit and after HaploFill. Reconstructed pseudomolecule are compared in size to the ARS-UCD1.2 genome chromosome sequences. Each plots report a bar graph of the actual sizes and in terms of difference percentage from ARS-UCD1.2 genome chromosomes the expected genome size. In HaploSplit Haplotype 1 pseudomolecules deviate only by 0.7±0.7% from the expected size, Haplotype 2 by 6.7±3.0%. HaploFill further reduces the divergence, mostly in Haplotype 2 pseudomolecules where it goes down to 1.4±1.9%.

Supplemental Table 3: Data availability for input data and assembly results for testing datasets. Data availability for input genomic sequences, input maps used for each of testing datasets, and availability of the final assembly results.

Supplemental Table 4: Data availability for input data and assembly results for testing datasets. Data availability for input genomic sequences, input maps used for each of testing datasets, and availability of the final assembly results.

Supplemental Table 5: Data availability for input data and assembly results for testing datasets. Data availability for input genomic sequences, input maps used for each of testing datasets, and availability of the final assembly results.