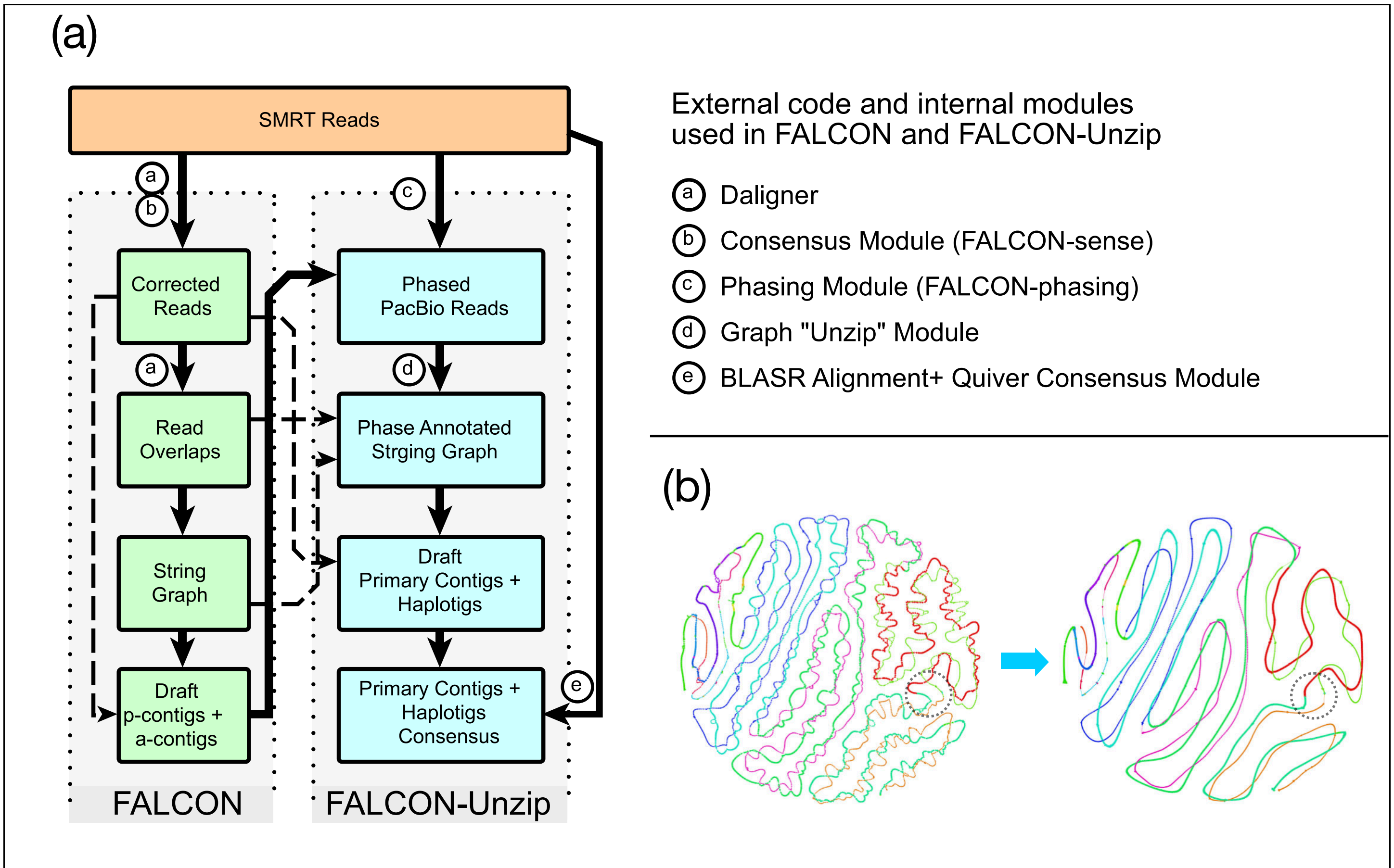
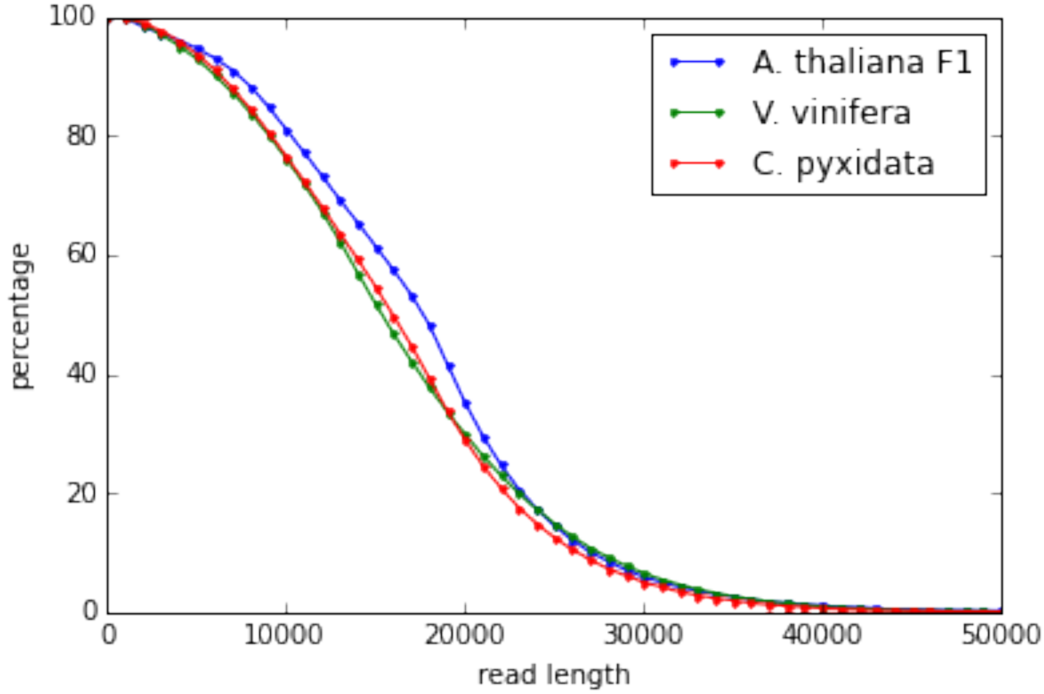


Supplementary Figure 1

Schematics of the software and data process modules and the FACLON-Unzip assembly graph process for resolving haplotypes.

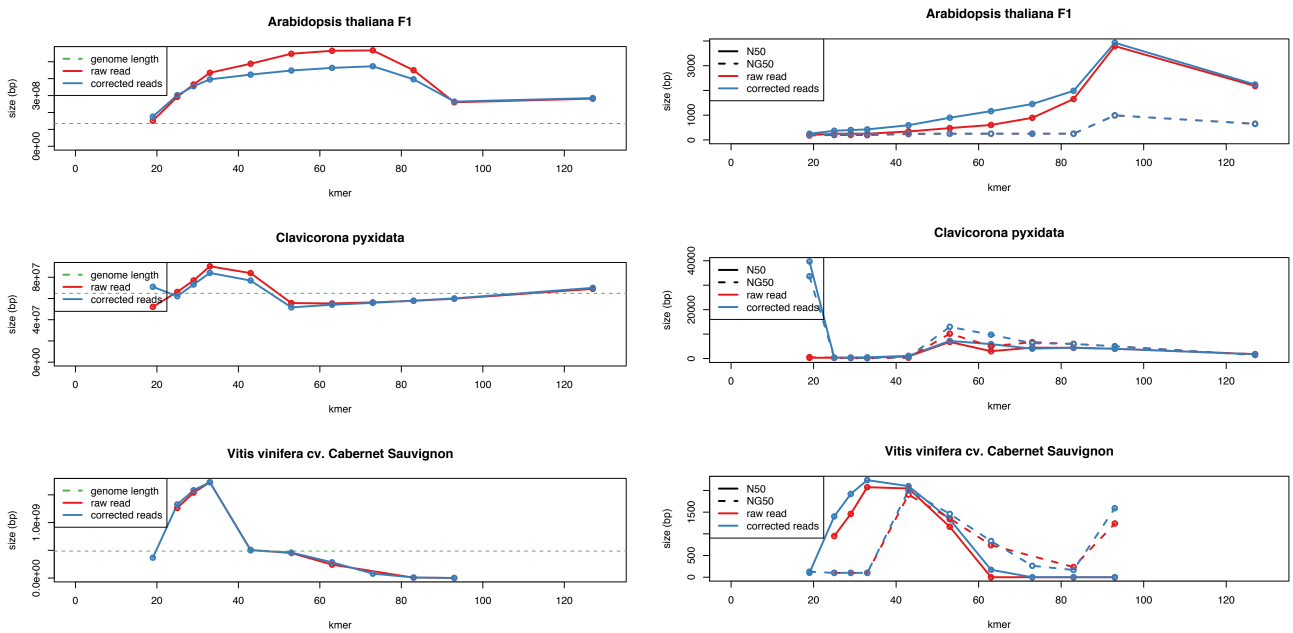
(**a**) Data dependence flow and software modules inside FALCON and FALCON-Unzip

(**b**) Left: Initial assembly graph of a contig in the *Arabidopsis* F1 hybrid assembly. The different colors represent different haplotype blocks and phases. Right: The assembly graph after "unzipping". Conceptually, the unzipping step identifies the heterozygous SNPs and uses them to remove overlaps between reads from different haplotypes. After removing such overlaps, nodes from the different haplotypes in the assembly graph will no longer have edges between them. This allows FALCON-Unzip to identify long haplotype specific paths and construct haplotigs of them. The dashed circle region indicates haplotype blocks that can be extended through a bubble region.



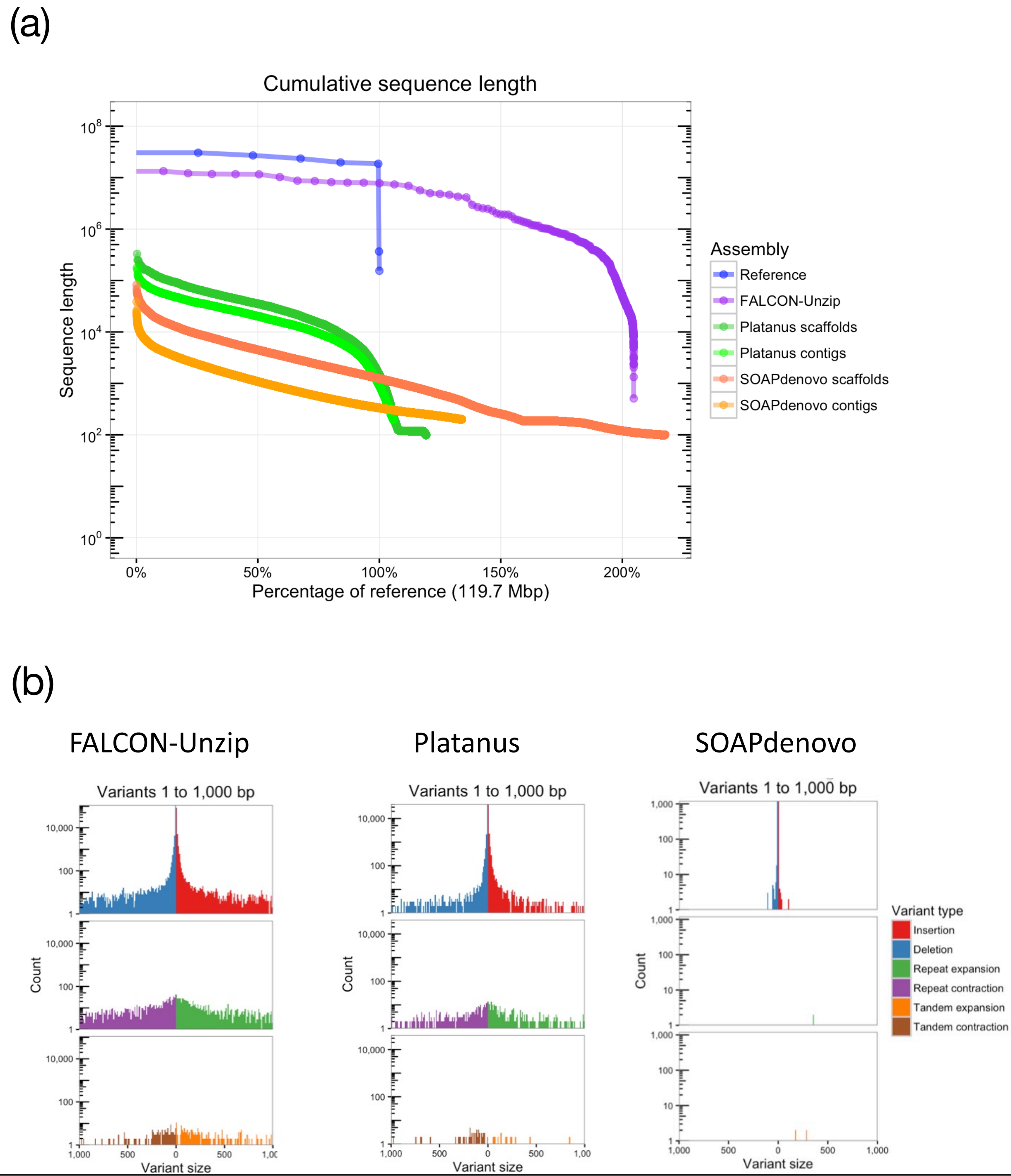
Supplementary Figure 2

Reverse accumulative read length distribution of the three diploid genome datasets



Supplementary Figure 3

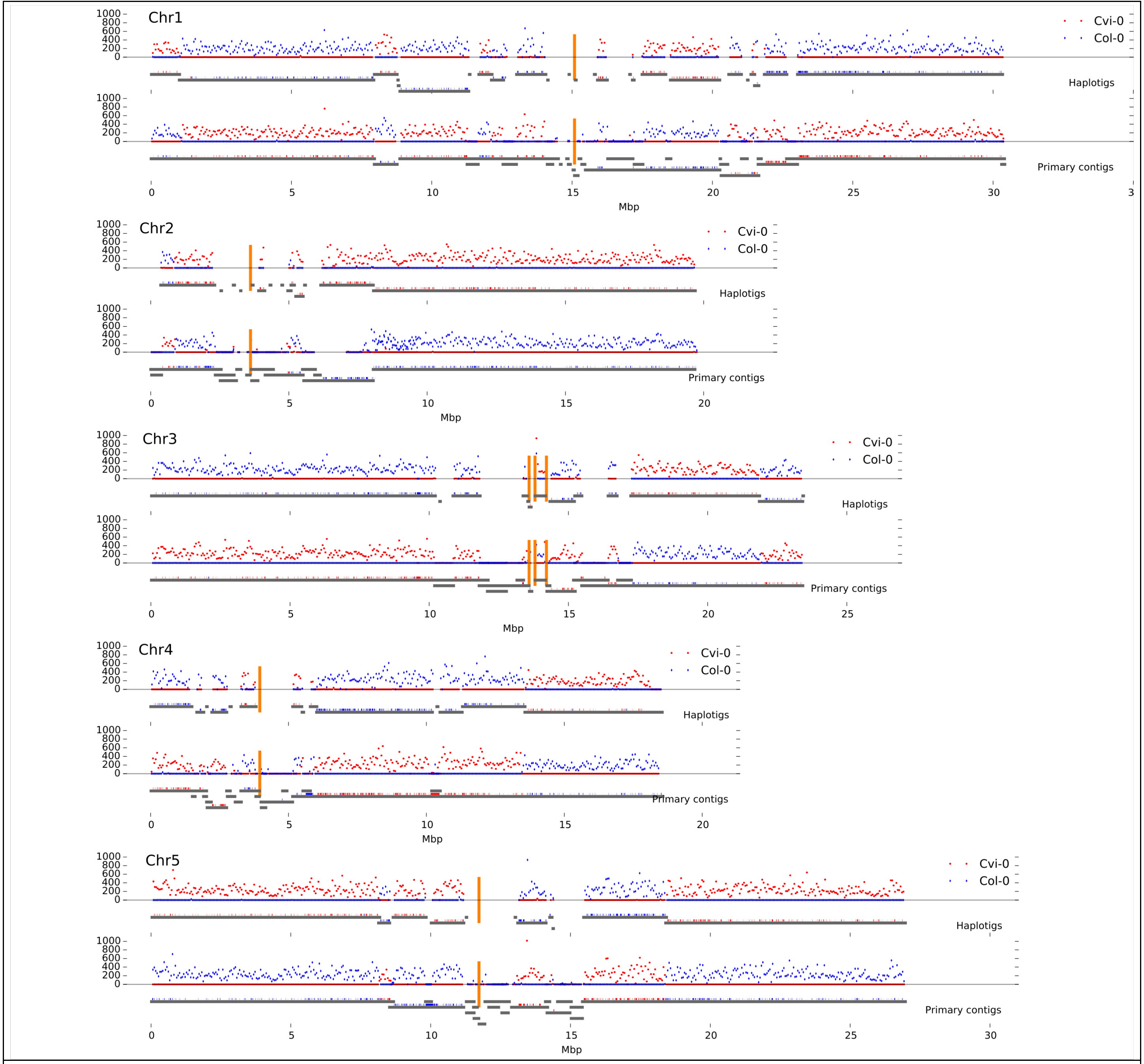
SOAPdenovo assembly sizes and N50 and NG50 sizes of the 3 genomes using different values of k using the raw reads and corrected by Lighter.



Supplementary Figure 4

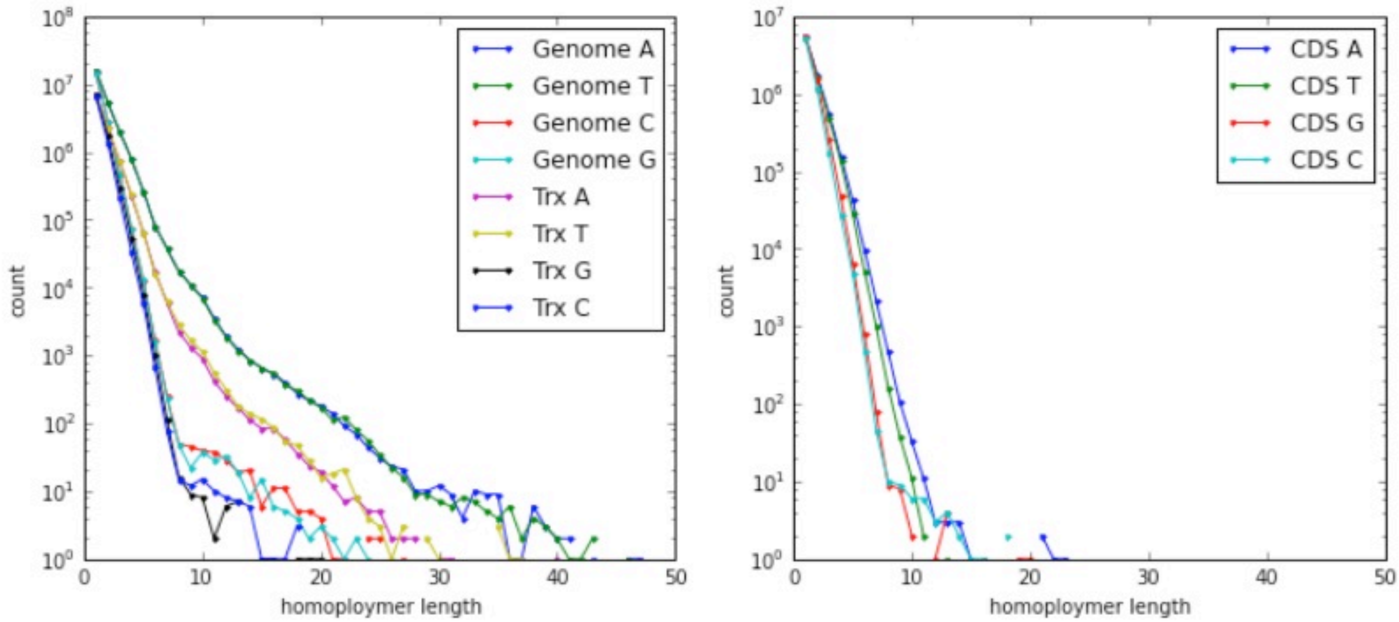
Assemblytic analysis comparison of the *Arabidopsis* F1 assemblies from FALCON-Unzip, Platanus, and SOAPdenovo.

(**a**) Cumulative sequence length of three *Arabidopsis* F1 assemblies created by FALCON-Unzip, Platanus, and SOAPdenovo compared to the TAIR10 reference. (**b**) Variants called using Assemblytics from three *Arabidopsis* F1 assemblies created by FALCON-Unzip ,Platanus, and SOAPdenovo.



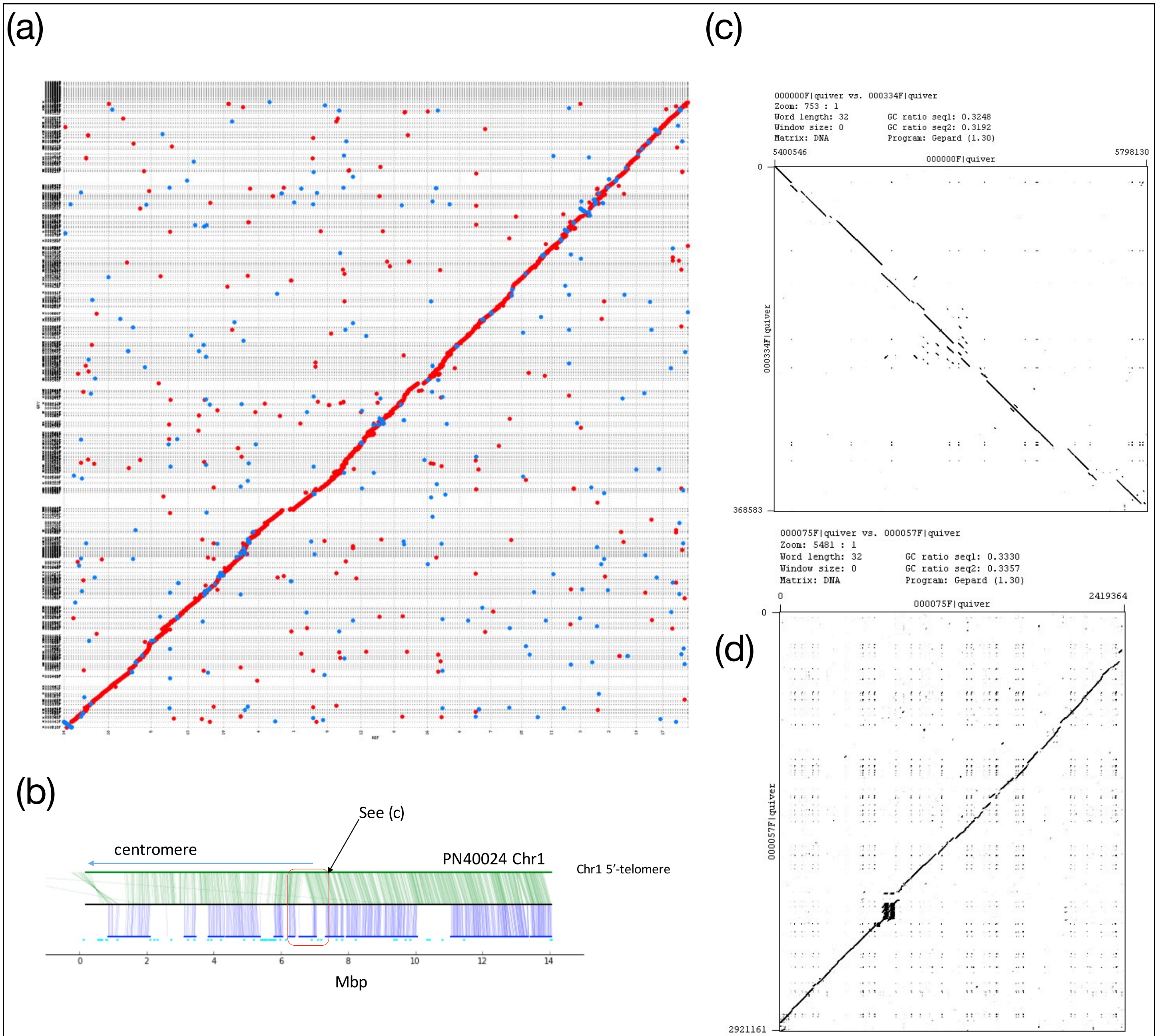
Supplementary Figure 5

Variation comparison between the inbred line assemblies and the F1-hybrid for all *Arabidopsis* chromosome along with TAIR10 references.



Supplementary Figure 6

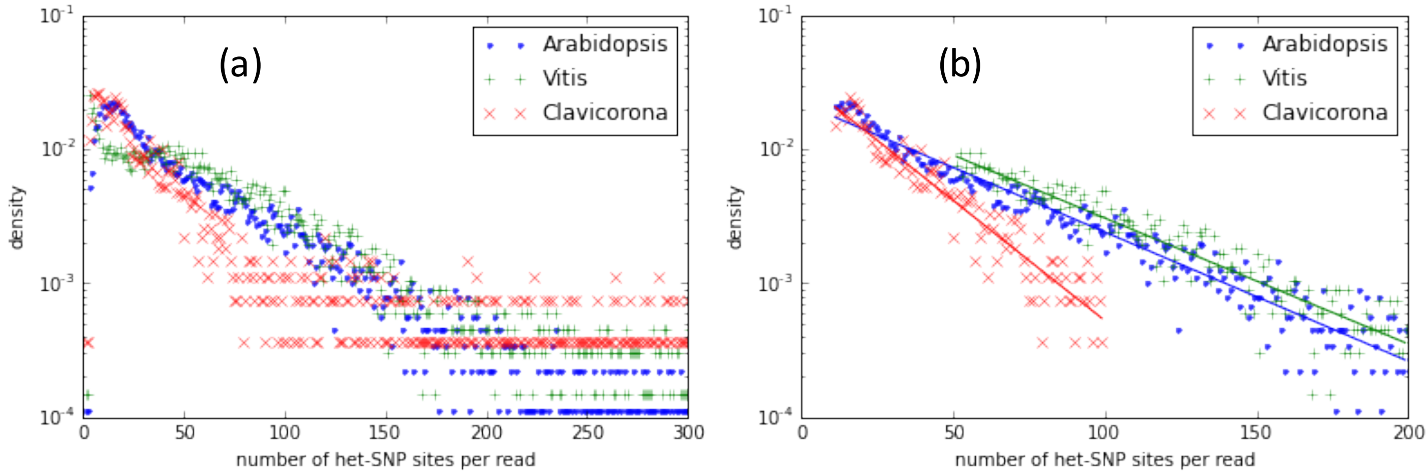
Homopolymer length and frequency in the TAIR10 Assembly.



Supplementary Figure 7

Assembly comparison: FALCON-Unzip V. vinifera cv. Cabernet Sauvignon assembly versus V. vinifera reference genome

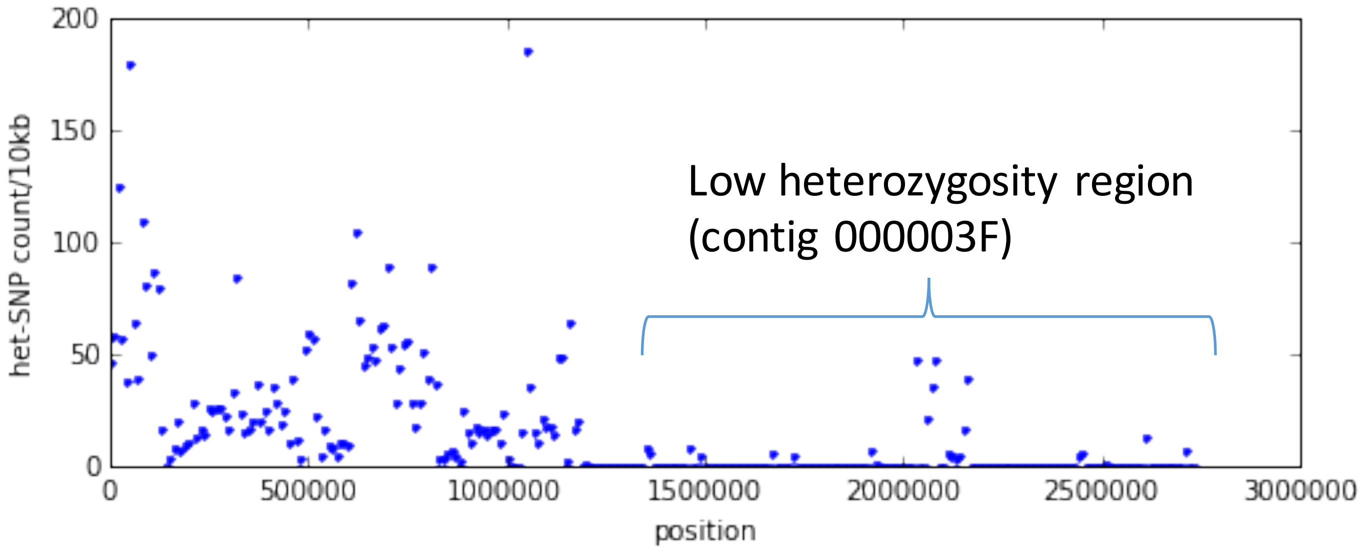
(**a**) MUMmerplot of FALCON-Unzip V. vinifera cv. Cabernet Sauvignon assembly versus V. vinifera reference genome. For clarity only alignments >= 10,000 bp long to the primary chromosomes are displayed. (**b**) The synteny between PN40024 Chr1 from 5'- telomere to centromere (green line) to the longest contig 000000F (black line) and its associated haplotigs (blue lines). The vertical green and blue lines indicated homologous coding sequences between the sequences. The cyan lines in the bottom indicate the synteny between the primary contig and other primary contigs. (**c**) Synteny alignment between two primary contigs 000334F vs. 000000F (**d**) Synteny alignment between two primary contigs 000057F vs 000075F



Supplementary Figure 8

Comparison of the distribution the het-SNP site density of the three genomes

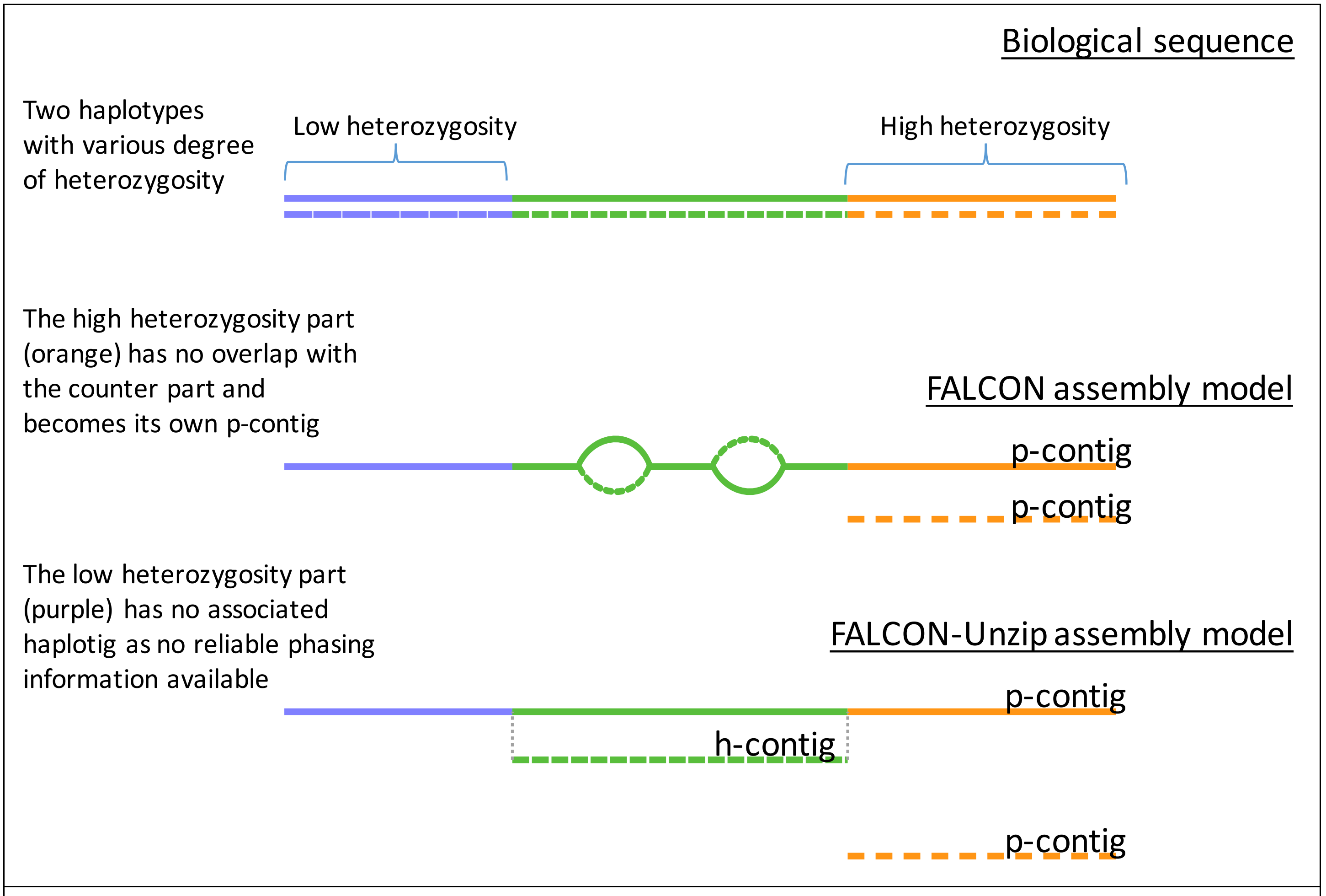
(**a**) The distribution of number of het-SNPs observed of the reads used for phasing of the longest contig of each genome in semi-log plot . (**b**) Fitting the distributions with a exponential function (density ~ c * exp(-a * het-SNP count)). We pick het-SNP count range of 10 to 200 for Arabidopsis, 50 to 200 for *Vitis*, and 10 to 100 for *Clavicorona* to catch the exponential decay part. The fitted parameter a = -0.0222, 0.0216, 0.0412 for *Arabidopsis*, *Vitis* and *Clavicorona* respectively. The fastest decay rate for Clavicorona indicates it has the least variation between the haplotypes among the three genomes. From this fitting, we expect to see about 45 (*Arabidopsis)*, 46 (*Vitis*), and 24 (*Clavicorona)* per 10kb in the regions of interests.



Supplementary Figure 9

Example of a low heterozygosity region observed in *Clavicorona* genome.

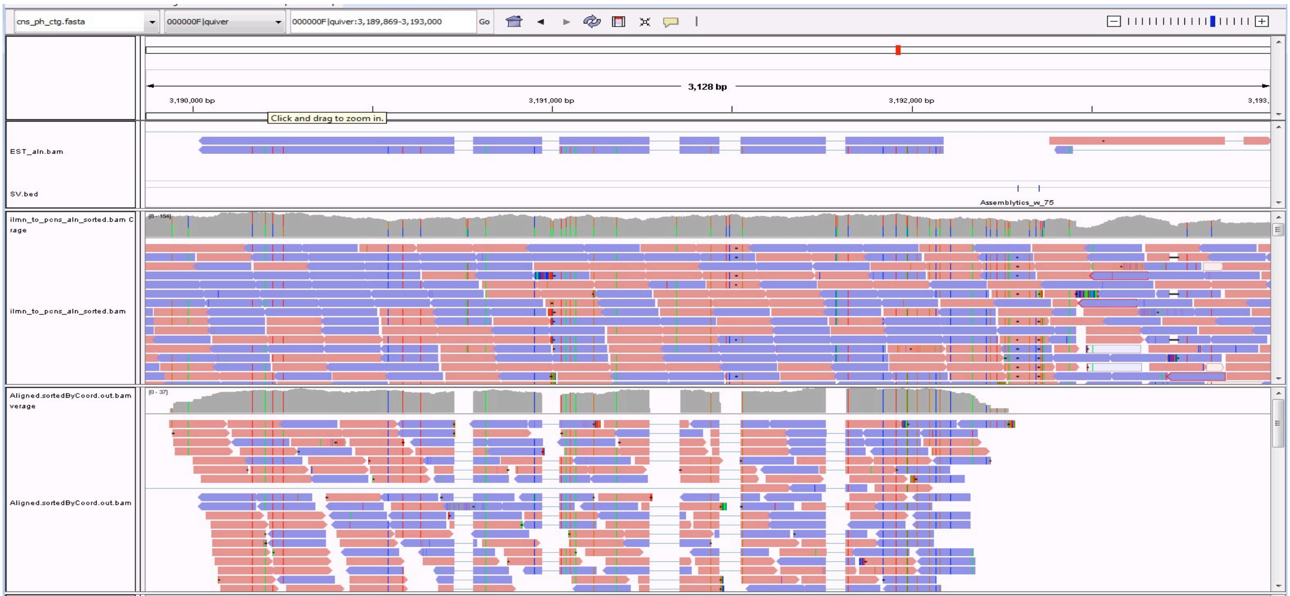
The het-SNPs are called with FreeBayes on the alignments of the short read data to only the primary contigs. The contig 00003F has a low heterozygosity region from ~1.2Mb to ~2.7Mb.



Supplementary Figure 10

General schematic about how different levels of heterozygosity can affect the contig layout.

(a)

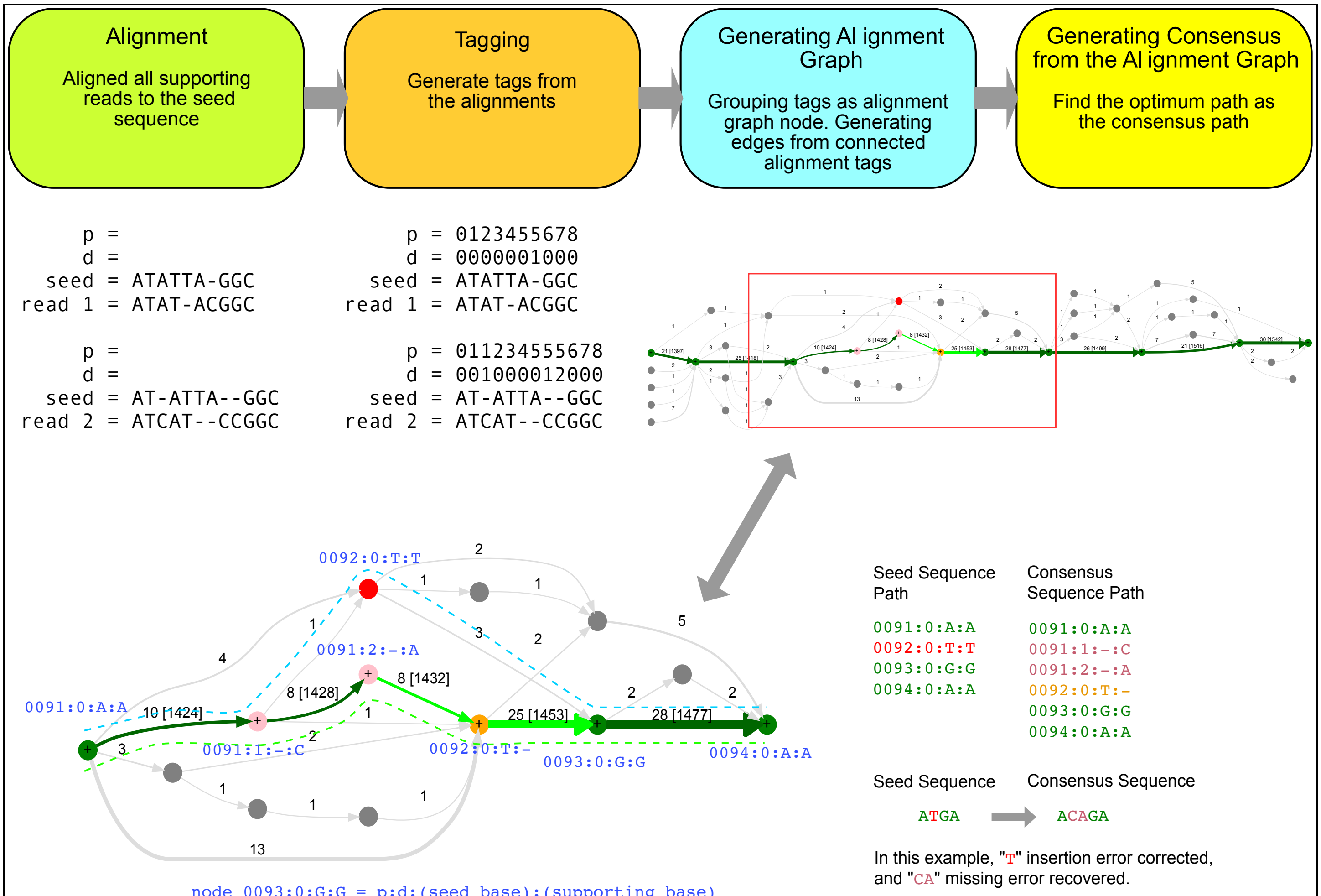


(b)

Supplementary Figure 11

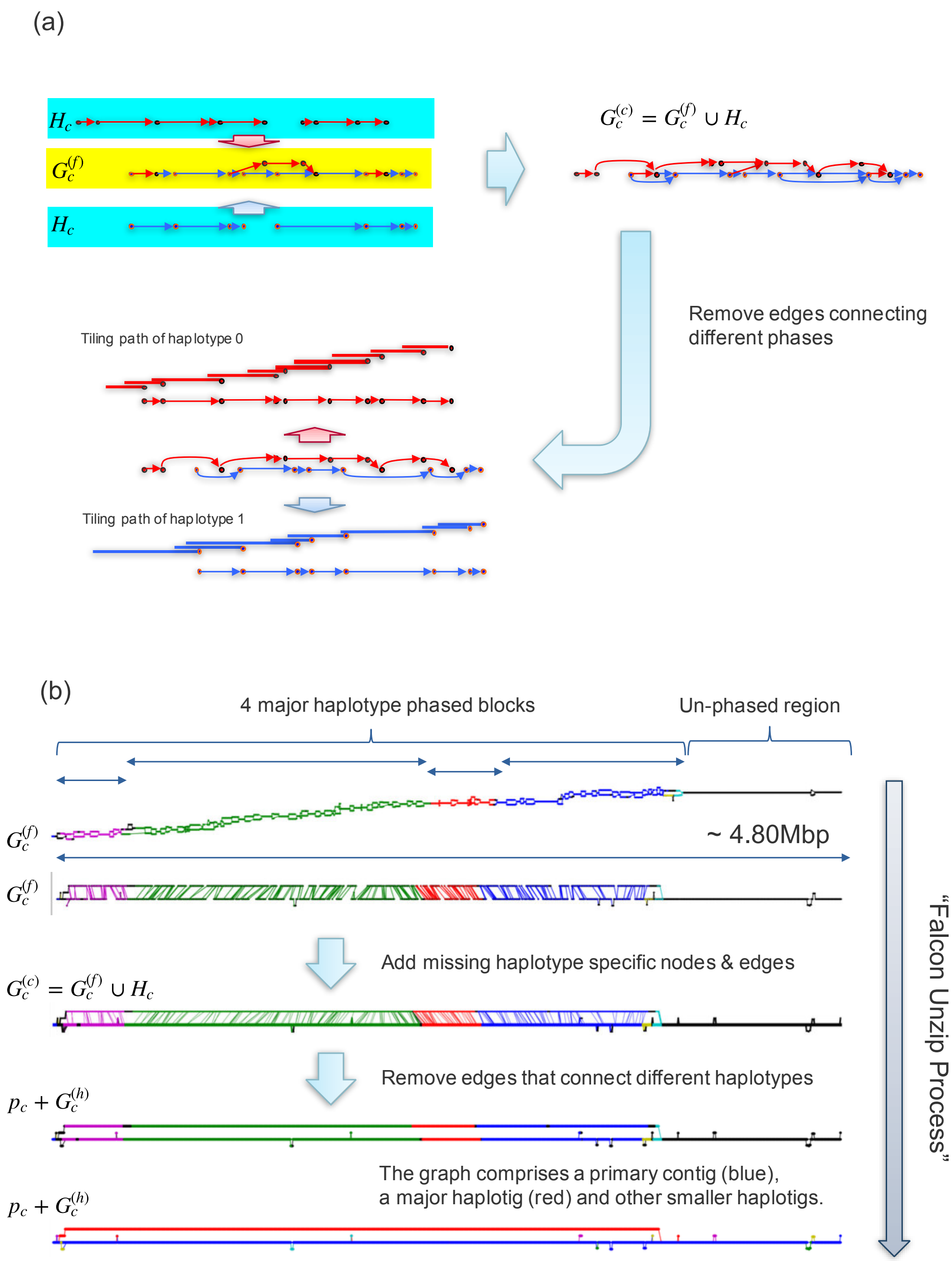
Candidates for differentially expressed alleles from RNA-seq data.

(**a**)(**b**)We mapped both genomic reads (middle panel) and cDNA reads (lower panel) to the primary contigs from our *Clavicorona pyxidata* assembly. We also shows curated CDS sequences mapped to the contig (top panel). The genomic reads shows both alleles mapped while we only observe on major allele in the transcript reads.



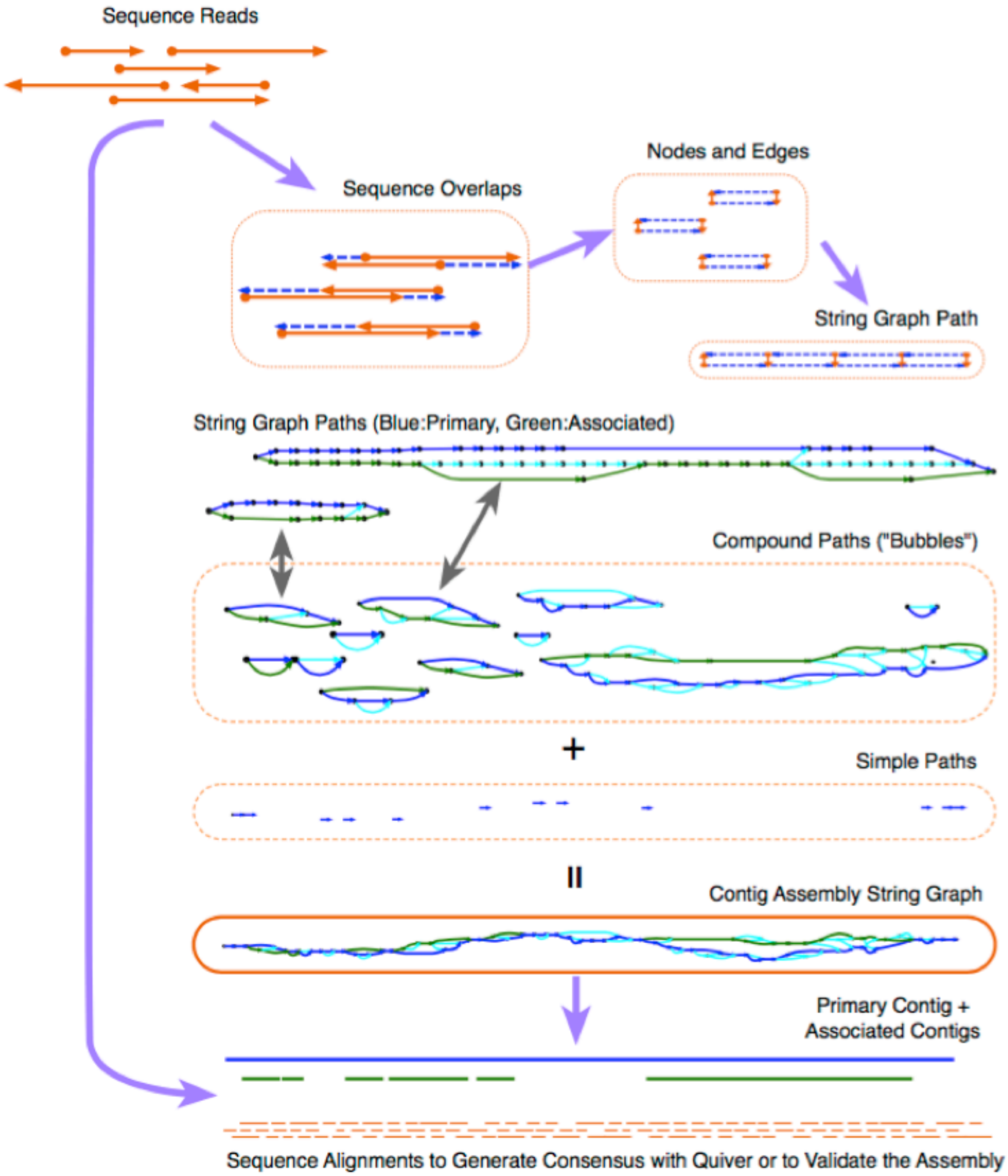
Supplementary Figure 12

An Example of how the FALCON-sense algorithm generates consensus sequence.



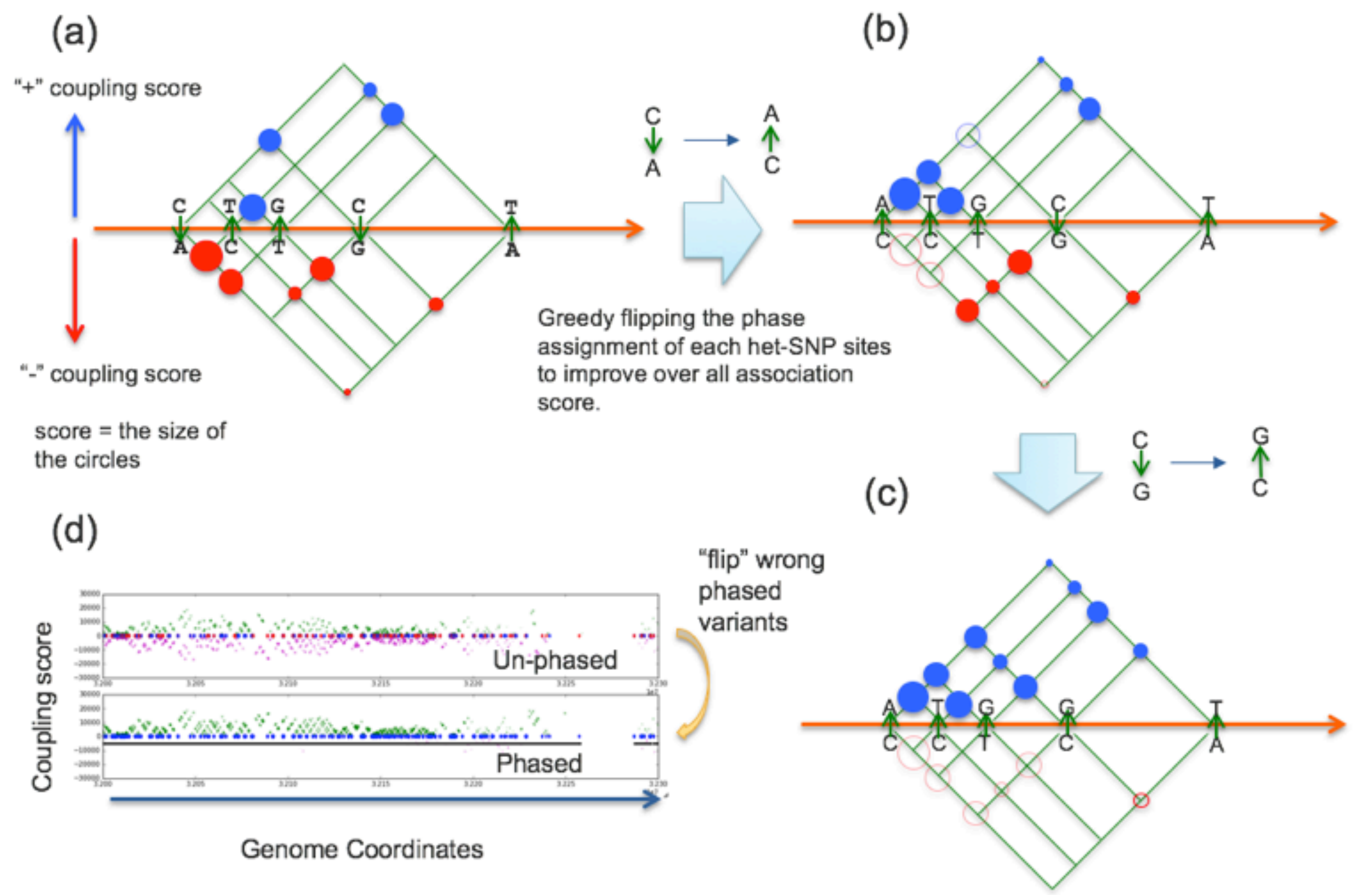
Supplementary Figure 13

(**a**) Summary of the graph reduction from sequence overlaps to contigs. (**b**) Example on constructing haplotigs in the Clavicorona pyxidata assembly



Supplementary Figure 14

Summary of the graph reduction from sequence overlaps to contigs



Supplementary Figure 15

Summary of the greedy SNP phasing algorithm

(**a**) All pairs of het-SNPs that are covered by multiple reads are evaluation. A "coupling score" is calculation from the number reads that support current haplotype assignment of the het-SNPs. (**b**)(**c**) We linearly scan through the het-SNP positions. If the total score is improved by flipping the haplotype assigned at one location, then we flip the assignment. (**d**) An example showing the "coupling score" before the flipping process (un-phased het-SNPs assignment) and afterward (phased het-SNP assignment).