Phased diploid genome assembly with single-molecule real-time sequencing

CS Chin[1], P Peluso[1], FJ Sedlazeck[2], M Nattestad[4], ***GT Concepcion[1]**, A Clum[5], C Dunn[1], R O'Malley[6], R Figueroa-Balderas[7], A Morales-Cruz[7], GR Cramer[7], M Delledonne[9], Chongyuan Luo[6], JR Ecker[6], D Cantu[7], DR Rank[1], MC Schatz[2,3,4]

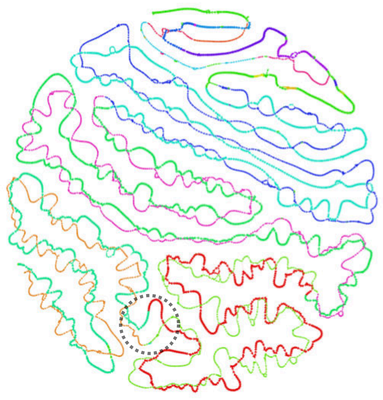
[1]PacBio, 1380 Willow Road, Menlo Park, CA 940251Pacific Biosciences, Menlo Park, CA; [2]Deptartment of Comp Sci, Johns Hopkins University, Baltimore, MD; [3]Dept of Biology, Johns Hopkins University, Baltimore, MD; [4]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; [5]DOE Joint Genome Institute, Walnut Creek, CA; [6]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA; [7]Department of Viticulture and Enology, University of California, Davis, CA; [8]Department of Biochemistry and Molecular Biology, University of Nevada, Reno, NV; [9]Dipartimento di Biotecnologie, Universita' degli Studi di Verona, Verona, Italy

Abstract

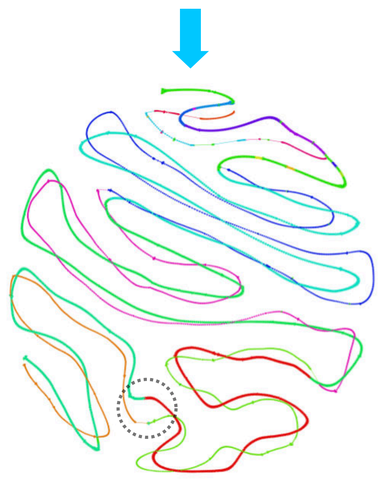
While genome assembly projects have been successful in many haploid and inbred species, the assembly of noninbred or rearranged heterozygous genomes remains a major challenge. To address this challenge, we introduce the open-source FALCON and FALCON-Unzip algorithms to assemble long-read sequencing data into highly accurate, contiguous, and correctly phased diploid genomes. We generate new reference sequences for heterozygous samples including an F1 hybrid of *Arabidopsis thaliana*, the widely cultivated *Vitis vinifera* cv. Cabernet Sauvignon, and the coral fungus *Clavicorona pyxidata*, samples that have challenged short-read assembly approaches. The FALCON-based assemblies are substantially more contiguous and complete than alternate short- or long-read approaches. The phased diploid assembly enabled the study of haplotype structure and heterozygosities between homologous chromosomes, including the identification of widespread heterozygous structural variation within coding sequences.

Resolving Phases

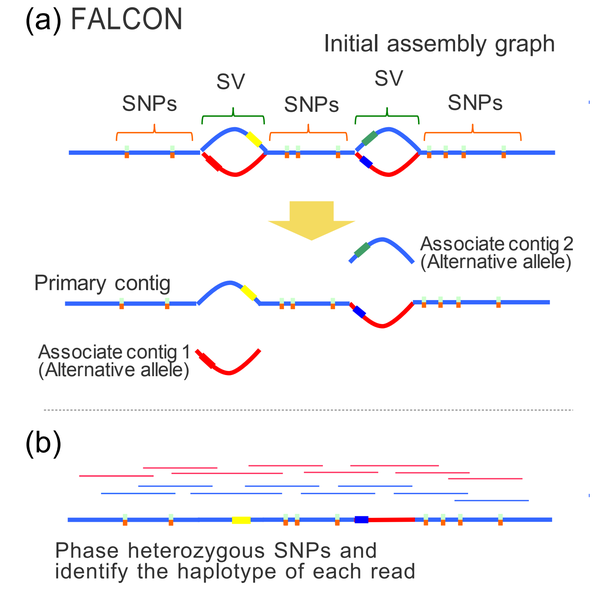
(a)Initial assembly graph of a contig in the *Arabidopsis* F1 hybrid assembly. The different colors represents different haplotype blocks and phases.



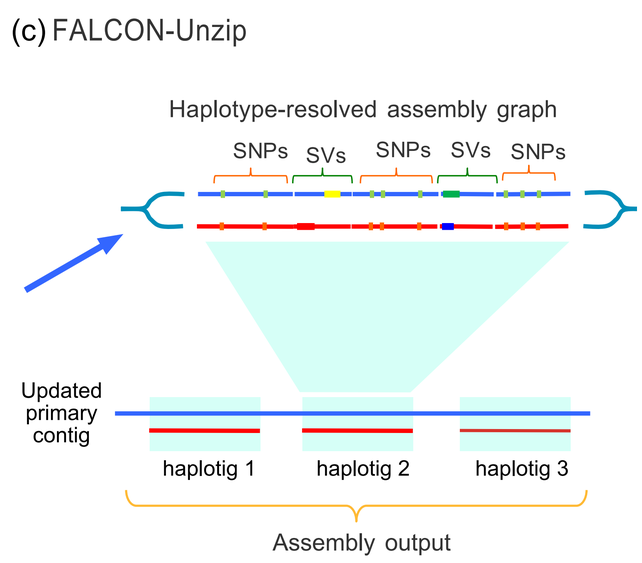
(b)Assembly graph after "unzipping." Conceptually, the unzipping step identifies the heterozygous SNPs and uses them to remove overlaps between reads from different haplotypes. After removing such overlaps, nodes from the different haplotypes in the assembly graph will no longer have edges between them. 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.



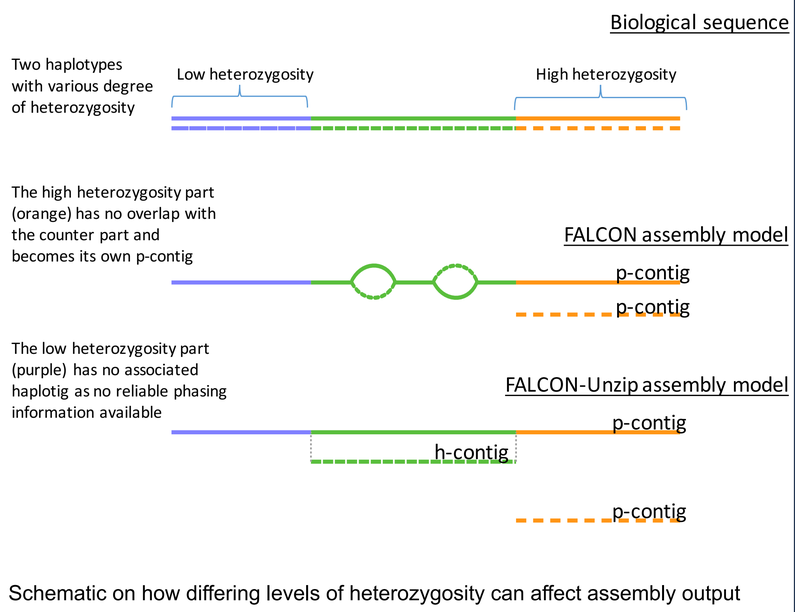
Methods



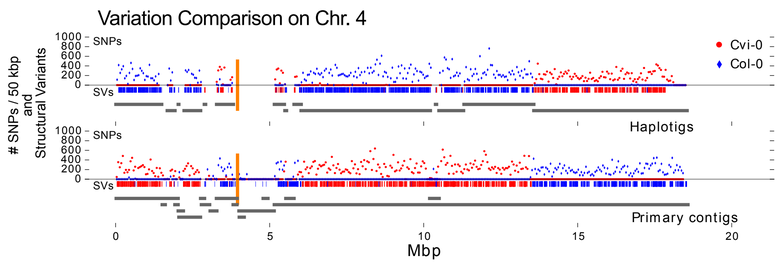
a) The initial assembly is computed by FALCON, which error corrects the raw reads (not shown) and then assembles the reads using a string graph of the read overlaps. The assembled contigs are further refined by FALCON-Unzip into a final set of contigs and haplotigs. **b)** Phase heterozygous SNPs and group reads by haplotype.



c) Phased reads are used to open up the haplotype-fused path and subsequently generate a set of primary contigs and associated haplotigs.



Schematic on how differing levels of heterozygosity can affect assembly output



This plot shows the primary contigs and haplotigs aligned to chromosome 4 of the *Arabidopsis* TAIR reference assembly as grey line segments. Blue and Red colored dots show the number of Col-0 and Cvi-0 specific SNPs, respectively, per 50 kbp region of the assembled contig. The vertical orange lines indicate the centromere locations. The short vertical tick marks above the grey lines indicate the structural variations against Col-0 (blue) and Cvi-0 (red).

Assembly Results

Species	Sample Stat. Total coverage, Read length N50	Assembler	Sequence	Assembly Size (Mb)	# contigs *scaffolds*	N50 size (Mb)	Max Contig Size (Mb)
A. thaliana	Inbred Col-0 130-fold read N50 = 9 kb	Canu	contigs	131	1,102	4.573	11.186
		FALCON	p-contigs	120	377	7.353	12.197
	Inbred Cvi-0 120-fold read N50 = 9 kb	Canu	contigs	127	676	4.817	12.393
		FALCON	p-contigs	120	260	6.073	14.370
	F1 Col-0 x CVI-0 120-fold, read N50 = 17 kb	Canu	contigs	219	1,897	1.554	15.379
		FALCON	p-contigs	143	426	7.923	13.386
			a-contigs	57	551	0.146	0.688
		FALCON-Unzip	p-contigs	140	172	7.961	13.319
			haplotigs	105	248	6.920	11.648
	F1 Col-0 x CVI-0 short reads 60-fold 250 bp reads	Platanus	scaffolds	143	151,779	0.0269	0.329
		SOAPdenovo k=93	scaffolds	260	691,629	0.00099	0.0825
V. vinifera	140-fold read N50 = 15 kb	Canu	contigs	1066	14,489	0.139	2.211
		FALCON	p-contigs	633	1,314	2.392	14.114
			a-contigs	184	1,164	0.278	0.804
		FALCON-Unzip	p-contigs	591	718	2.173	14.079
			haplotigs	368	2,037	0.779	3.926
	short reads 46-fold 100 bp reads	SOAPdenovo k=33	scaffolds	1728	12,879,081	0.0001	0.0368
		SOAPdenovo k=43	scaffolds	507	767,707	0.0019	0.0310
C. pyxidata	100-fold read N50 = 16 kb	Canu	contigs	60	432	0.646	4.390
		FALCON	p-contigs	43	133	1.49	4.829
			a-contigs	12	172	0.0805	0.407
		FALCON-Unzip	p-contigs	42	82	1.484	4.778
			haplotigs	24	93	0.872	2.218
	short reads 86-fold 100 bp reads	Platanus	scaffolds	39	26,702	0.045	0.489
		SOAPdenovo k=19	scaffolds	52	157,941	0.00055	0.070

Conclusions

The highly fragmented polyploid genome output typical of contemporary assembly algorithms makes it difficult to probe for haplotype specific variation. With FALCON-Unzip, however, this heterozygosity information can be captured in the primary contigs and associated haplotigs, so the question of how haplotype specific variation can affect gene expression, methylation patterns and other regulatory interactions can be further examined. Looking forward, we expect SMRT Sequencing technology will allow us to further probe diploid and polyploid genomic diversity and understand its impact on genomic annotation, gene regulation and evolution.

References

1. Chin et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Meth. (13)12, 1050-1054.
2. GitHub: https://github.com/PacificBiosciences/FALCON

Acknowledgements

J. Lohr Vineyards and Wines, Felipe Simao Neto, L. Nagy, Joseph D. Puglisi, Florian Jupe, Alex Copeland and Various PacBio contributors. The project was supported in part by NIH award (R01-HG006677) and NSF awards (DBI-1350041 and IOS-1237880 to MCS; MCB 0929402 and MCB 1122246 to J.R.E). J.R.E is an investigator of the HHMI and Gordon and Betty Moore Foundation (GBMF 3034).