Title

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

Journal

Nature methods, 13(12)

ISSN

1548-7091

Authors

Chin, Chen-Shan
Peluso, Paul
Sedlazeck, Fritz J
et al.

Publication Date

2016-12-01

DOI

10.1038/nmeth.4035

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

Chen-Shan Chin[1,10], Paul Peluso[1,10], Fritz J Sedlazeck[2], Maria Nattestad[3], Gregory T Concepcion[1], Alicia Clum[4], Christopher Dunn[1], Ronan O'Malley[5], Rosa Figueroa-Balderas[6], Abraham Morales-Cruz[6], Grant R Cramer[7], Massimo Delledonne[8], Chongyuan Luo[5], Joseph R Ecker[5], Dario Cantu[6], David R Rank[1] & Michael C Schatz[2,3,9]

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 (https://github.com/PacificBiosciences/FALCON/) 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.

De novo genome assembly is a fundamental pursuit in genome research[1–3] that has led to the creation of high-quality reference genomes for many haploid or highly inbred species and has promoted gene discovery, comparative genomics, and other studies[4–6]. However, currently available genome assemblies rarely capture the heterozygosity present within a diploid or polyploid species[7]. Most assemblers output a mosaic genome sequence that arbitrarily alternates between parental alleles[8]. Consequently, variation between homologous chromosomes—including differences in sequence, structure, and gene presence—is undetected. Heterozygous genome assemblies are also typically more fragmented, which has limited the identification of allele-specific expression, long-range expression quantitative trait loci (eQTLs), and other haplotype-specific features[9–11]. These challenges are becoming more prominent as *de novo* sequencing projects shift toward more heterogeneous samples such as outbred, wild-type diploid, and polyploid nonmodel organisms, as well as to highly rearranged disease samples including samples from human cancers.

While the problem of assembling diploid and polymorphic genomes is not new[12,13], it lacks a universal and scalable solution. Computational methods for diploid assembly tend to generate short contigs averaging from just a few hundred bases to several kilobases[12,14,15]. Approaches such as sequencing both parents and offspring (i.e., trios)[16], haploid sex cells[17], clonal fosmid[18], and synthetic long reads[19,20] are labor intensive and costly, and they often produce assemblies with limited contiguity. Long-range scaffolding technologies such as optical mapping and chromatin assays are often inapplicable to heterozygous short-read assemblies, as they demand well-assembled contig sequences (minimal contig N50 size of 50 kbp to 100 kbp) and can leave unresolved regions (N characters) inside the scaffolds.

Single-molecule real-time (SMRT) Sequencing is commonly used to finish bacterial genomes and provide high-contiguity assemblies for mammalian-scale genomes[21,22]. The long reads (currently ~10 kbp, on average, with some approaching 100 kbp) can span many repetitive elements and help resolve complicated diploid genomes. Nonetheless, existing assemblers do not take advantage of the long reads to resolve haplotypes. In this paper, we present FALCON, a diploid-aware long-read assembler, and FALCON-Unzip, an associated haplotype-resolving tool, to assemble haplotype contigs or 'haplotigs' that represent the diploid genome with correctly phased homologous chromosomes (**Fig. 1**).

The FALCON assembler follows the design of the hierarchical genome assembly process (HGAP)[23] but uses more computationally optimized components (**Supplementary Fig. 1a**). It begins by using reads to construct a string graph that contains sets of 'haplotype-fused' contigs as well as bubbles representing divergent regions between homologous sequences[24] (**Fig. 1a**). Next, FALCON-Unzip identifies read haplotypes using phasing

[1]Pacific Biosciences, Menlo Park, California, USA. [2]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA. [3]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [4]DOE Joint Genome Institute, Walnut Creek, California, USA. [5]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, California, USA. [6]Department of Viticulture and Enology, University of California Davis, Davis, California, USA. [7]Department of Biochemistry and Molecular Biology, University of Nevada, Reno, Nevada, USA. [8]Dipartimento di Biotecnologie, Universita' degli Studi di Verona, Verona, Italy. [9]Department of Biology, Johns Hopkins University, Baltimore, Maryland, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to C.-S.C. (jchin@pacb.com) or M.C.S. (michael.schatz@gmail.com).

RECEIVED 6 JUNE; ACCEPTED 25 AUGUST; PUBLISHED ONLINE 17 OCTOBER 2016; DOI:10.1038/NMETH.4035

information from heterozygous positions that it identifies (**Fig. 1b**). Phased reads are then used to assemble haplotigs and primary contigs (backbone contigs for both haplotypes) (**Fig. 1c** and **Supplementary Fig. 1b**) that form the final diploid assembly with phased single-nucleotide polymorphisms (SNPs) and structural variants (SVs).

To evaluate the accuracy of FALCON-Unzip, we applied it to a trio of *Arabidopsis* genomes (Col-0, Cvi-0, and the hybrid Col-0–Cvi-0) and analyzed the results with respect to each other and the TAIR10 reference genome[25]. We also assessed performance on the genomes of *Vitis vinifera* cv. Cabernet Sauvignon, a highly heterozygous outcrossed grape cultivar of agricultural importance, and on a highly heterozygous wild-type diploid fungus, *Clavicorona pyxidata*, which has resisted previous short-read assembly approaches.

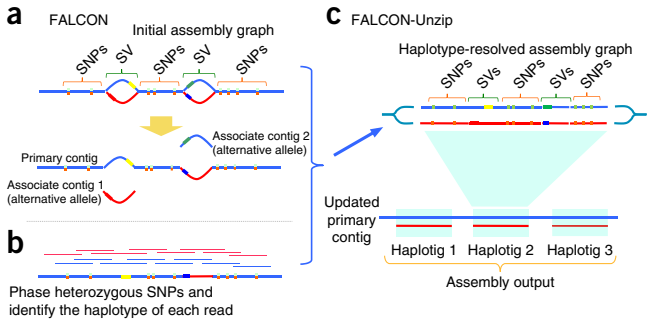

Figure 1 | Overview of FALCON and FALCON-Unzip. (**a**) An initial assembly is computed by FALCON, which error corrects the raw reads (not shown) and then assembles them using a string graph of the read overlaps. The assembled contigs are further refined by FALCON-Unzip into a final set of contigs and haplotigs. (**b**) Phase heterozygous SNPs and group reads by haplotype. (**c**) The phased reads are used to open up the haplotype-fused path and generate as output a set of primary contigs and associated haplotigs.

RESULTS

Sequencing and assembly of an *Arabidopsis* trio

We individually sequenced and assembled the inbred Col-0 and Cvi-0 genomes using FALCON (**Supplementary Table 1**). Contig N50 sizes were 7.4 Mb (Col-0) and 6.0 Mb (Cvi-0), about 10 to 100 times more contiguous than other recently published *Arabidopsis* assembly[26] (**Table 1**) and approaching the continuity of the highly curated TAIR10 assembly (10.9 Mbp), which was assembled using expensive BAC sequencing[25]. The largest FALCON contigs spanned entire chromosome arms (**Fig. 2**), creating a high-quality draft reference for Cvi-0.

When comparing our Col-0 assembly to the TAIR10 assembly, the nucleotide sequence identity was greater than 99.98% (**Supplementary Table 2**). We applied BUSCO[27] to evaluate the assembly completeness by identifying a set of highly conserved plant orthologs in the assembly (**Supplementary Table 3**). BUSCO identified 914 (95.6%) and 906 (94.8%) genes in the Col-0 and Cvi-0 assemblies, respectively, compared with 915 (95.7%) in the TAIR10 reference. The variations between Col-0 and Cvi-0 assemblies are summarized in **Table 2**.

To assess performance on heterozygous genomes, we generated and assembled short- and long-read sequencing data of the F1 progeny with four leading assembly algorithms (**Table 1**). Canu[28] was used to assemble long-read sequence data (**Table 1** and **Supplementary Fig. 2**) from the Col-0–Cvi-0 F1 hybrid sample. The total size of the assembly was 219 Mb, slightly smaller than the expected diploid size of 238 Mb. The high level of polymorphisms, including a SNP rate of ~1/200 bp and 1,051 SVs larger than 50 bp between the strains (**Table 2**), might cause fragmented assembly, as the algorithm is not currently optimized for diploid genomes. Consequently, the contiguity of the F1 assembly was substantially worse (~3-fold less) than the Canu assembly of either inbred parent alone (**Table 1**). Short-read assemblies with SOAPdenovo2 (ref. 29) and Platanus[15], which were designed to assemble heterogeneous diploid genomes, were significantly less contiguous compared with Canu; SOAPdenovo2 assembled a total of 260 Mbp with an N50 = 990 bp even after k-mer optimization and error correction (**Supplementary Fig. 3**). Contigs assembled using Platanus were marginally improved, with an N50 = 26.9 kbp and a total assembly size of 143 Mbp, which was only slightly larger than the haploid genome size.

Most assemblers generate a single set of contigs, but FALCON generates 'primary contigs' (p-contigs) and 'alternative contigs' (a-contig) that comprise the genome regions typified by SVs from the p-contigs (see Online Methods). The a-contigs, representing local alternative sequences, spanned a total of 57 Mbp (~40% of the p-contigs) with an N50 = 146 kbp. Thus, FALCON alone produced 84% of the estimated 238-Mbp diploid genome. After the initial assembly, the FALCON-Unzip algorithm used the heterozygosity information within the initial primary contigs for haplotype phasing (**Fig. 1b** and **Supplementary Note**). With phasing information from the raw reads, FALCON-Unzip generated a subsequent set of p-contigs and the final haplotig set (h-contigs) that represented more contiguous haplotype-specific sequence information than the a-contigs (**Fig. 1c**). After the 'unzipping' process, the total size of the p-contigs was 140 Mbp (N50 = 7.96 Mbp), and the total size of the haplotigs was 105 Mbp (N50 = 6.92 Mbp). FALCON-Unzip generated phased diploid genome assemblies with continuity comparable to that of the individual inbred parental genomes (**Table 1**).

Comparison of the F1 assembly of FALCON-Unzip, Platanus, and SOAPdenovo2 directly with the TAIR10 reference is detailed in the **Supplementary Note** (**Supplementary Fig. 4** and **Supplementary Table 4**). Overall, the variants from the FALCON-Unzip assembly captured 89% of the Platanus variants and 90% of the SOAP variants at a stringent requirement of the exact same variant type, size, and genomic location. However, the Platanus and SOAP assemblies captured only 37% and 1% of the FALCON-Unzip variants, respectively.

Col-0–Cvi-0 F1 haplotig phasing quality

We aligned p-contigs and haplotigs to the parental inbred assemblies to evaluate the accuracy of haplotype separations. Ideally, each haplotig should be identical to one of the parental haplotypes and show variations against the other. We observed that most of the haplotigs only showed SNPs or SVs in one of the parental genomes, indicating that the phasing approach works accurately (**Fig. 2** and **Supplementary Fig. 5**). We assessed accuracy by computing the ratio of differences (for example, SNPs) to either of the parental assemblies within each haplotig (**Supplementary Table 5**). For the largest six haplotigs spanning 50% of the genome, the minority SNP percentages were all lower than 0.2%. The small minority SNP ratio represents either a small number of (i) local phasing errors,

Table 1 | Assembly results

Species	Sample (total coverage, read length N50)	Assembler	Sequence	Assembly size (Mb)	No. contigs (scaffolds)	N50 size (Mb)	N50 no.	N90 size (Mb)	Max contig size (Mb)
A. thaliana	Inbred Col-0	Canu	Contigs	131	1,102	4.573	8	0.0069	11.186
	(130x, read N50 = 9 kbp)	FALCON	P-contigs	120	377	7.353	7	1.278	12.197
	Inbred Cvi-0	Canu	Contigs	127	676	4.817	9	0.364	12.393
	(120x, read N50 = 9 kbp)	FALCON	P-contigs	120	260	6.073	7	1.993	14.370
		Canu	Contigs	219	1,897	1.554	17	0.042	15.379
	F1 Col-0–Cvi-0	FALCON	P-contigs	143	426	7.923	6	0.387	13.386
	(120x, read N50 = 17 kbp)		A-contigs	57	551	0.146	117	0.05	0.688
		FALCON-Unzip	P-contigs	140	172	7.961	7	0.504	13.319
			Haplotigs	105	248	6.920	6	0.571	11.648
	F1 Col-0–Cvi-0 (short reads)	Platanus	Scaffolds	143	151,779	0.0269	1,290	0.00014	0.329
	(60x, 250 bp reads)	SOAPdenovo2, *k* = 93	Scaffolds	260	691,629	0.00099	43,570	0.00013	0.0825
V. vinifera		Canu	Contigs	1,066	14,489	0.139	1,778	0.03	2.211
	Cabernet Sauvignon	FALCON	P-contigs	633	1,314	2.392	72	0.362	14.114
	(140x, read N50 = 15 kbp)		A-contigs	184	1,164	0.278	220	0.073	0.804
		Falcon Unzip	P-contigs	591	718	2.173	72	0.402	14.079
			Haplotigs	368	2,037	0.779	127	0.075	3.926
	Cabernet Sauvignon (short reads)	SOAPdenovo2, *k* = 33	Scaffolds	1,728	12,879,081	0.0001	791,053	0.0001	0.0368
	(46x, 100-bp reads)	SOAPdenovo2, *k* = 43	Scaffolds	507	767,707	0.0019	63,857	0.0018	0.0310
		Canu	Contigs	60	432	0.646	16	0.045	4.390
C. pyxidata	*Clavicorona pyxidata*	Falcon	P-contigs	43	133	1.49	8	0.218	4.829
	(100x, read N50 = 16 kb)		A-contigs	12	172	0.0805	41	0.037	0.407
		Falcon-Unzip	P-contigs	42	82	1.484	8	0.252	4.778
			Haplotigs	24	93	0.872	9	0.141	2.218
	Clavicorona pyxidata (short reads)	Platanus	Scaffolds	39	26,702	0.045	225	0.0013	0.489
	(86x, 100-bp reads)	SOAPdenovo2, *k* = 19	Scaffolds	52	157,941	0.00055	15,065	0.00013	0.070

P-contigs, primary contigs; A-contigs, alternate contigs. *k*, k-mer size.

(ii) incorrect SNP calls, and/or (iii) assembly base errors; but it demonstrates that there are no significant segmental switching errors. Only nine haplotigs (~2.5% of all haplotig bases) showed a minority SNP ratio over 10%, and they were generally associated with repetitive or low heterozygous regions. Finally, we aligned the haplotigs of the FALCON-Unzip assembly to analyze its ability to incorporate SNPs. We identified 450,680 SNPs among the haplotigs, compared with 501,243 found by aligning the Col-0 and Cvi-0 assemblies. Thus, FALCON-Unzip phased 85.7% of all SNPs and 91.9% of all SVs directly from the shotgun sequence assembly.

Col-0–Cvi-0 F1 coding sequence prediction

The overall base-to-base concordance rate is about 99.99% (QV40 in Phred scale) in the F1 FALCON-Unzip assembly. The insertion and deletion (indel) concordances to the parental lines were lower (about QV40) than the SNP concordance rate (about QV50), with most residual errors concentrated in long homopolymer sequences (**Supplementary Data 1**, **Supplementary Table 6** and **Supplementary Fig. 6**). We evaluated the impact of such errors on coding sequence prediction with AUGUSTUS (**Supplementary Note** and **Supplementary Table 7**). Interestingly, AUGUSTUS[30] aligned 97% of all coding sequences (CDS) of TAIR10 to our assembly without any indels, and the vast majority of BUSCO genes (877) were even found to be phased.

Vitis vinifera sequencing and diploid assembly

V. vinifera cv. Cabernet Sauvignon is an F1 of two very distinct cultivars, Cabernet Franc and Sauvignon Blanc, and it is one of the world's most widely cultivated red wine grape varieties. Long reads (**Supplementary Table 1**) were generated and assembled using Canu, FALCON, and FALCON-Unzip (**Table 1**). FALCON-Unzip yielded the most contiguous assembly of 590 Mbp (N50 = 2.17 Mbp) and generated a total of 368 Mbp of associated haplotigs (N50 = 779 kbp). Both primary and associated contigs displayed overall high macrosynteny with the current *V. vinifera* genome reference (PN40024 (ref. 31); **Supplementary Fig. 7**). The total p-contig size was larger than the estimated genome size of *V. vinifera* (~500 Mbp[31]). This suggests that in some cases FALCON-Unzip underestimated the alternative haplotype sequences because of high heterozygosity between homologous regions. An analysis of synteny between different p-contigs to determine the extent of inclusion of redundant regions identified a total of 25 Mbp of syntenic blocks in the primary assembly (**Supplementary Data 2**,**3** and **Supplementary Note**).

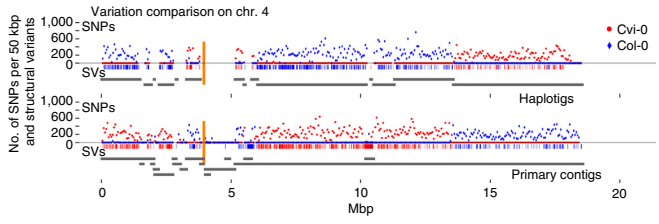

Figure 2 | SNP density and structural variation in the FALCON-Unzip F1 *Arabidopsis* assembly. The plot shows the primary contigs and haplotigs aligned to chromosome 4 (chr. 4) of the TAIR reference assembly as gray line segments. Colored dots show the number of Col-0 and Cvi-0 specific SNPs per 50-kbp region of the assembled contig. Vertical orange lines indicate centromere locations. Short vertical lines below the plot indicate structural variations against Col-0 (blue) and Cvi-0 (red) references.

Table 2 | *Arabidopsis* genome assembly comparisons

	HGAP inbreds, Col-0 versus Cvi-0		Falcon-Unzip haplotigs versus primary contigs	
Variant type	**Events**	**Affected bases**	**Events**	**Affected bases**
SNP count	501,243	1,002,486	450,680	901,360
Indel > 50 bp	1,051	882,736	966	798,438
Repeat contraction/ expansion > 50 bp	1,670	3,746,572	1,479	3,130,205
Tandem contraction/ expansion > 50 bp	73	97,319	65	85,495
Total SV > 50 bp detected	2,794	4,726,627	2,510	4,014,138
Predicted CDS	Col-0: 28,176; Cvi-0: 27,797		p: 31,679; h: 24,808	
Aligned CDS pairs	27,424		24,808	
Predicted coding sequence SNPs	183,942	367,884	147,811	295,622
Other predicted coding SVs	16,748	153,260	15,151	136,245
Local inframe variants	5,135	82,929	4,090	66,681
Local noninframe variants	11,613	70,331	11,061	69,564

p, primary contigs; *h*, haplotigs.

Compared with *Arabidopsis*, the *V. vinifera* genome has more repeats and higher heterozygosity, making it more challenging to assemble. Canu generated an assembly of 1,006 Mbp, which is roughly twice the haploid genome size with a significantly smaller N50 = 139 kbp. Even with optimized k-mer sizes (33–43 bp), SOAPdenovo2's scaffold N50 size was smaller than 2 kbp and the contig N50 < 1 kbp (**Supplementary Fig. 3**). The Platanus results were unacceptably incomplete, with less than 1% of the expected genome size reported, most likely because of the limited available coverage. Nevertheless, even with high coverage (1,577 million reads) and multiple libraries, published assemblies of different grape cultivars report contig N50 sizes of at most 41 kbp using Platanus[32].

To assess completeness of the assemblies we used BUSCO and aligned the 29,971 mRNA sequences annotated from the current *V. vinifera* genome reference PN40024. Both approaches highlighted the completeness of the gene space in the FALCON-Unzip assembly (**Supplementary Tables 3** and **8**). Overall, 80% of the 956 BUSCO genes and 16,981 of the 29,971 predicted complete genes from PN40024 were phased in the assembly. In contrast, less than 15% of the 956 BUSCO proteins were found within the most contiguous short-read assemblies, suggesting that these assemblies are not only highly fragmented, but also markedly incomplete (**Supplementary Table 3**).

Clavicorona pyxidata sequencing and assembly

To demonstrate the generality of FALCON-Unzip to wild-type heterozygous genomes, we analyzed *C. pyxidata*, a common coral fungus that grows on hardwoods across North America (haploid size, ~42 Mbp). FALCON-Unzip produced the most contiguous assembly, followed by Canu (~2-fold less contiguous) and short-read assemblies (30- to >100-fold less contiguous) (**Table 1**). In lieu of a reference, we evaluated the assemblies using BUSCO and genomic sequencing data (SRA accession SRR1800147, 86×, 150-bp reads) (**Supplementary Note** and **Supplementary Table 3**).

In contrast to the *V. vinifera* genome, the *C. pyxidata* genome has significantly skewed rates of heterozygosity, and about 50% of the genome is essentially homozygous. This suggests that naturally occurring inbreeding or other selective pressures limit variation in these regions. Different levels of heterozygosity between homologous chromosomes, seen in all three genomes, also affect the assembly sizes (**Supplementary Note** and **Supplementary Figs. 8–10**).

For evaluating phasing accuracy, we used the 150-bp paired-end short-read data and called phased SNPs relative to the primary contigs with FreeBayes[33] and HapCut[34] (**Supplementary Table 9**). Because of the insert size limit of the short-read data set, the phasing data only covered about 23% (9.72 Mbp) of the genome, but nearly all phased blocks (96% to 98%, depending on variant call quality threshold) were fully concordant with the FALCON-Unzip assembly (**Supplementary Table 9**). Comparison of homologous alleles within the genome with publicly available RNA sequencing data (SRA accession SRR1589642) identified several candidate differentially expressed alleles (**Supplementary Fig. 11**).

DISCUSSION

We have demonstrated that FALCON and FALCON-Unzip can assemble PacBio SMRT Sequencing data from heterozygous diploid genomes into highly accurate, contiguous, and correctly phased primary contigs and haplotigs. Such haplotype-specific assemblies represent the true genome and both enable and strengthen studies of haplotype structures and heterozygous variants such as SVs and SNPs between homologous chromosomes.

In all three genomes that we studied, the FALCON-Unzip assembly was two- to threefold more contiguous than alternative long-read assemblers and 30- to >100-fold more contiguous than state-of-the-art short-read assemblers. In the *Arabidopsis* F1-hybrid assembly, the haplotigs almost perfectly matched one of their parental genomes with only ~2.5% incorrectly phased sequences. In future work, we aim to improve phasing accuracy further by analyzing the local assembly graph to predict hard-to-resolve regions and potential errors in the assembly. We showed that the low frequency of residual sequencing errors (<0.1%) had almost no effect on the identification of gene sequences. In the other two assemblies, we demonstrated greatly improved diploid representations of core genes from the FALCON-Unzip assembly (for example, >90% in the *Arabidopsis* F1 genome) and accurate phasing measured using orthogonal data (**Supplementary Table 9**).

Both the raw sequencing read lengths and error rates affect haplotype and consensus accuracies. Genome complexity, especially the rate of heterozygous positions and the repetitive sequences, is also a major factor impacting performance. Most haplotype-phasing algorithms utilize heterozygous SNPs and ignore SVs. In contrast, FALCON-Unzip is designed to combine SNPs and SVs to separate haplotype information beyond what either method alone provides to construct haplotype-specific contigs. With long read lengths from SMRT Sequencing and increased levels of heterozygosity, this allows us to almost fully resolve both haplotype chromosomes for practically the entire *Arabidopsis* F1 genome with high contiguity. The other two genomes highlight additional complexities that are possible for diploid genomes. In *V. vinifera*, we found homologous regions with very high variation rates likely due to the outcrossing nature of the organism; while in *C. pyxidata* we discovered extended regions of unexpectedly low heterozygosity, suggesting increased selective pressures or complex naturally occurring inbreeding. While future read-length increase will improve the separation of the haplotypes, we can already begin to utilize the assembly output to understand and represent heterozygosity variations

within a wide range of diploid genomes (**Supplementary Table 10**). The assembly results can, in principle, also be improved with other types of data, especially long-range scaffolding data, and extend to higher ploidy genomes in the future.

The lack of haplotype resolution in mosaic genome assemblies makes it difficult to probe the impact of epigenetic and differential gene expression and can exacerbate 'reference bias' when remapping sequencing data[35]. With FALCON-Unzip, however, almost all the heterozygosity information is captured in the p-contigs and haplotigs, so the question of how haplotype-specific variations affect gene expression, methylation patterns, or other regulatory interactions can be examined further. More systematic study of phased diploid references will expose the detailed *cis*-regulatory mechanisms of differential expression in diploid genomes to improve our general understanding of the biology beyond haploid genomes. Looking forward, we expect many new opportunities for understanding diploid and polyploid genomic diversity and its impact on genome annotation, gene regulation, and evolution.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Data available at BioProject, accession codes PRJNA314706 (*Arabidopsis*), PRJNA316730 (*V. vinifera* cv. Cabernet Sauvignon), and PRJNA336540 (*Clavicorona pyxidata*). Assemblies can be downloaded from https://downloads.pacbcloud.com/public/dataset/PhasedDiploidAsmPaperData/FUNZIP-PhasedDiploidAssemblies.tgz.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The sequencing of the Cabernet Sauvignon genome was supported in part by a gift from the J. Lohr Vineyards and Wines to D.C. We would also like to thank F. Neto for providing an early-release BUSCO plant data set. *Clavicorona pyxidata* DNA was provided by L. Nagy (Institute of Biochemistry Biological Research Centre of the Hungarian Academy of Sciences). We thank J. Puglisi, F. Jupe, A. Copeland, and A. Wenger for reading and critiquing the manuscript. The project was supported in part by National Institutes of Health award (R01-HG006677 to M.C.S.) and by National Science Foundation awards (DBI-1350041 and IOS-1237880 to M.C.S.; MCB 0929402; and MCB 1122246 to J.R.E.). J.R.E. is an investigator at the Howard Hughes Medical Institute and Gordon and Betty Moore Foundation (GBMF 3034).

AUTHOR CONTRIBUTIONS

C-S.C., P.P., A.C., D.R.R., and M.C.S. conceived the idea of the FALCON–FALCON-Unzip assembler. C.-S.C, P.P., F.J.S., M.N., G.T.C., D.R.R., D.C., and M.C.S. designed the experiments and performed the analysis. P.P., D.C., D.R.R., and M.C.S. collected the sequencing data. R.O'M. C.L., and J.R.E. constructed the Col-0-Cvi-1. A.C., R.O'M. R.F.-B., A.M.-C., G.R.C., M.D., C.L., J.R.E., and D.C. collected the samples and prepared DNA for sequencing. C.-S.C., P.P., F.J.S., M.N., G.T.C., D.C., D.R.R., and M.C.S. wrote the manuscript. C.-S.C. and C.D. implemented the computer code.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563–567 (1996).
2. Myers, E.W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
3. Bonfield, J.K., Smith, Kf. & Staden, R. A new DNA sequence assembly program. *Nucleic Acids Res.* **23**, 4992–4999 (1995).
4. Mouse ENCODE Consortium. *et al.* An encyclopedia of mouse DNA elements (mouse ENCODE). *Genome Biol.* **13**, 418 (2012).
5. Celniker, S.E. *et al.* Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
6. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
7. Earl, D. *et al.* Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. *Genome Res.* **21**, 2224–2241 (2011).
8. Church, D.M. *et al.* Extending reference assembly models. *Genome Biol.* **16**, 13 (2015).
9. Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J. & Schork, N.J. The importance of phase information for human genomics. *Nat. Rev. Genet.* **12**, 215–223 (2011).
10. Henson, J., Tischler, G. & Ning, Z. Next-generation sequencing and large genome assemblies. *Pharmacogenomics* **13**, 901–915 (2012).
11. Alkan, C., Sajjadian, S. & Eichler, E.E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).
12. Vinson, J.P. *et al.* Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. *Genome Res.* **15**, 1127–1135 (2005).
13. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
14. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).
15. Kajitani, R. *et al.* Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
16. Roach, J.C. *et al.* Chromosomal haplotypes by genetic phasing of human families. *Am. J. Hum. Genet.* **89**, 382–397 (2011).
17. Kirkness, E.F. *et al.* Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome Res.* **23**, 826–832 (2013).
18. Kitzman, J.O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29**, 59–63 (2011).
19. McCoy, R.C. *et al.* Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. *PloS One* **9**, e106689 (2014).
20. Mostovoy, Y. *et al.* A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nat. Methods* **13**, 587–590 (2016).
21. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
22. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016).
23. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
24. Fasulo, D., Halpern, A., Dew, I. & Mobarry, C. Efficiently detecting polymorphisms during the fragment assembly process. *Bioinformatics* **18**, S294–S302 (2002).
25. The *Arabidopsis* Genome Initiative Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
26. Gan, X. *et al.* Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**, 419–423 (2011).
27. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
28. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R. & Phillippy, A.M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Preprint at *bioRxiv* http://dx.doi.org/10.1101/071282 (2016).
29. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18 (2012).
30. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
31. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
32. Patel, S., Swaminathan, P., Fennell, A. & Zeng, E. in *Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (eds. Huan, J. *et al.*) 1771–1773 (EEE, 2015).
33. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at arXiv:1207.3907v2 [q-bio.GN] (2012).
34. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153–i159 (2008).
35. Degner, J.F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).

ONLINE METHODS

DNA isolation and library preparation. For the *Arabidopsis* sample preparation, to minimize chloroplast DNA contamination, nuclei were isolated from leaf tissue as previous described[36]. Genomic DNA was isolated using standard purification columns and protocols (Qiagen). For grapevine DNA extraction, young leaves (~1 cm diameter) were collected from *Vitis vinifera* cv. Cabernet Sauvignon clone 08 at Foundation Plant Services (UC Davis, Davis, CA). Plant tissue (1 g) was ground to a powder in a mortar containing liquid nitrogen. 10 mL of prewarmed (65 °C) extraction buffer (300 mM Tris–HCl pH 8.0, 25 mM EDTA pH 8.0, 2 M NaCl, 2% (w/v) soluble PVP (MW 40000), 2% CTAB, 2% 2-mercaptoethanol) was added, and the suspension was homogenized by inversion and incubated (65 °C) for 30 min in a water bath, mixing by inversion (every 5 min). Plant debris was removed by centrifugation (5,000 r.p.m.) for 5 min at room temperature, and the supernatant was transferred into a new tube. Equal volume of chloroform:isoamyl alcohol (CIA, 24:1 v/v) was added and mixed by inversion for 5 min. Aqueous phase was segregated by 10 min centrifugation (5,000 r.p.m.) at room temperature and transferred gently into a new tube. RNase A was added to the sample (2 µg) and was incubated (37 °C) for 30 min. After RNAse treatment, equal volume of CIA was added and centrifuged as above. 0.1 volume of 3 M NaOAc pH 5.2 and an equal volume of isopropanol were added for DNA precipitation, and the sample was mixed by inversion and then incubated (−80 °C) for 30 min. DNA was collected by centrifugation (5,000 r.p.m.) for 30 min and the pellet was washed twice with 3 mL of 70% ethanol. After 10 min centrifugation (5,000 r.p.m.), DNA pellet was air dried at room temperature and resuspended in 500 µl of nuclease-free water. DNA quality was evaluated by pulse-gel electrophoresis, and quantity was determined using the Qubit fluorometer.

Shearing of the DNA was performed either with G-tubes (Covaris) or by passage through a small bore needle[37] to average size of 15 kbp to 40 kbp. The needle method was used during an evaluation of shearing techniques. However, both shearing methods produced libraries of comparable quality and sequencing performance. Sheared DNA was enzymatically repaired and converted into SMRTbell libraries prepared as described by the manufacturer (Pacific Biosciences). Non-SMRTbell DNA was removed by exonuclease treatment. Finally, a BluePippin preparative electrophoresis purification step was performed (Sage Sciences) on the library to select insert sizes ranging from 7 to 50 kbp or from 15 to 50 kbp depending on the sequencing experiment. These size-selected libraries were used in subsequent sequencing steps.

Sequencing methods. Sequencing was performed on the PacBio RS II instrument per the manufacturer's recommendations. The Col-0 and Cvi-0 inbred *Arabidopsis* data sets were collected using P4-C2 chemistry with 4 h movie lengths. The F1 Col-0–Cvi-0 and the *C. pyxidata* and the *V. vinifera* cv. Cabernet Sauvignon samples were run with P6 chemistry and 6 h data-collection movies.

Raw long-read error correction. All raw long-read sequences were aligned to each other using 'daligner'[38] executed by the main script of the FALCON assembler. The overlap data and raw subreads were then processed to generate consensus sequences. The consensus-calling algorithm (FALCON-sense) was designed to preserve the information from heterozygous single-nucleotide polymorphisms (SNP) and is described in detail in the **Supplementary Note** (section "Updated FALCON consensus algorithm" and **Supplementary Fig. 12**).

Initial 'haplotype-fused' assembly with a collapsed diploid-aware contig layout. After the error-correction step, FALCON identified the overlaps between all pairs of the preassembled error-corrected reads. The read overlaps were used to construct a directed (in contrast to bidirected) string graph following the Myers' algorithm[39]. For diploid genomes with high heterozygosity, the string graph typically contained linear chains of 'bubbles' (**Supplementary Fig. 1b** and **Supplementary Fig. 13**). We can decompose such linear chains into 'simple' and 'compound' paths, in which a simple path is a path where there is no internal branching node, and it also has unique source node and sink node; and a compound path is a collection of edges that represents a bubble with unique source and sink in the assembly graph. The algorithm for constructing such compound paths is described in the **Supplementary Note**. The nonbranched collection of compound paths and simple paths are further combined to create unitigs. Genome repeats, sequencing errors, or missing overlaps can introduce spurious unitigs. Empirically derived heuristic rules were applied to remove these artifacts and layout the primary contigs and the associated contigs. The graph-reduction process is detailed in **Supplementary Figure 14**. We call the final assembly graph the 'haplotype-fused assembly graph $G^{(f)}$'.

Mapping and phasing the raw reads. In the draft assembly, each contig is simply a tiling sequence from the subsequences of a set of error-corrected reads. Some of the raw reads have not yet been associated with any contigs. For example, if a read is 'contained' within other reads (overlaps completely to a substring of another read), it is not used in constructing the first draft of the contigs. There are two strategies for identifying the raw-read-to-contig associations: (i) remap all raw reads to the contigs and find the best alignments or (ii) trace the read overlapping information to find out where a raw-read is most likely to be associated. FALCON-Unzip applies strategy (ii) to avoid the time penalty for the remapping process, as the overlap information already exists. For each raw read, FALCON-Unzip examines all overlapping reads. If a read is uniquely associated with one contig, then the raw read is assigned to that contig. If there are multiple contigs associated with a read, it scores the matching contigs by the overlap lengths. In this case, a read is assigned to a target contig with the highest sum of overlap lengths.

For each primary contig, we collect all raw reads associated with the primary contig and its associated contigs. We align the raw reads to the contigs with the BLASR aligner[40] and call heterozygous SNPs (het-SNPs) by analyzing the base frequency of the detailed sequence alignments. A simple phasing algorithm was developed to identify phased SNPs (see **Supplementary Note** and **Supplementary Fig. 15**). Along each contig, the algorithm assigns phasing blocks where chained phased SNPs can be identified. Within each block, if a raw read contains a sufficient number of het-SNPs, it assigns a haplotype phase for the read unambiguously. Combined with the block and the haplotype phase information, it assigns a 'block-phase' tag for each phased read in each phasing block. Some reads might not have enough

doi:10.1038/nmeth.4035

phasing information. For example, if there are not enough het-SNP sites covered by a read, it assigns a special 'un-phased tag' for each un-phased read.

Overview of the algorithm constructing haplotype-specific contigs. The algorithm to construct the haplotype-specific contigs (haplotigs) is summarized in **Figure 1** and **Supplementary Figure 13**. Briefly, for each contig c, it constructs a haplotype-specific assembly graph from all reads that mapped to it, denoted as H_c, by ignoring the overlaps between any two reads from the same block but different phases. It then combines this graph H_c to the fused assembly subgraph $G_c^{(f)} \subset G^{(f)}$ that contains the paths of contig c to construct a complete contig subgraph $G_c^{(c)} = G_c^{(f)} \cup H_c$. Unlike the initial subgraph $G_c^{(f)}$, where some reads are masked out by reads from different phases, the complete contig subgraph $G_c^{(c)}$ rescues such masked-out reads and has complete read representation from both haplotypes.

In the fused assembly graph $G_c^{(f)}$, there is a path that is corresponding to the original contig c from node s to node t. It is desirable to generate a new locally phased contig that also starts from the same node s and ends at the same node t as new primary contig p_c. While such primary contig p_c may not be fully phased end to end, the collection of p_c of all contig c can serve as a haploid assembly representation with annotated locally phased regions. Additionally, the variations between the two haplotypes can be identified by aligning other haplotigs to the primary contigs. Once p_c is identified, the corresponding edges of p_c in $G_c^{(c)}$ are removed. It also removes all other edges connecting different phases of the same block. Namely, it constructs a subgraph $G_c^{(h)}$ of $G_c^{(c)}$ by removing edges which are already in p_c or connect distinctly phased nodes. We identify all linear paths within $G_c^{(h)}$ as the haplotigs $h_{c,i=1\ldots n}$, where n is the total number of haplotigs associated with the primary contig. Some of the haplotigs might be caused by missing overlaps or sequence errors. The haplotig sequences are aligned to the primary contig. If the alignment identity is high and no phased reads are associated with the haplotig, the haplotig will be marked as duplicated and removed. Note that a haplotig may contain multiple haplotype-phased blocks. For example, haplotype-specific SVs may affect the initial mapping such that the phasing algorithm cannot connect two neighboring blocks. However, reads from different phasing blocks might be uniquely overlapped if the SVs between the haplotypes are distinguishable. Such haplotype-specific overlaps can connect broken haplotype-phased blocks into to larger haplotigs.

Polishing partially phased primary contigs and their associated haplotigs. Conceptually, FALCON-Unzip generates one new primary contig p_c and n haplotigs $h_{c,i=1\ldots n}$ from the original assembly graph $G_c^{(i)}$ of the contig c. It uses the phasing information to decide whether a phased read belongs to the primary contig p_c or one of the haplotigs $h_{c,i=1\ldots n}$. Each unphased read may also contain structural-level variations that are the same as in a particular haplotig. In such cases, by examining the overlaps between the read to those in the haplotigs, FALCON-Unzip can find the best hit from the unphased read to one haplotig. In the end, each raw read will be augmented with the information regarding which haplotig or primary contig it belongs to and will be mapped accordingly. This ensures that the haplotig consensus is generated from the appropriate reads belonging to the correct haplotype. Finally, FALCON-Unzip uses the Quiver algorithm[23] to remove residual errors in the haplotig consensus from the haplotype-specific alignments.

FALCON-Unzip outputs a set of partially phased primary contigs (p-contigs) and the associated haplotigs (h-contigs) for each primary contig. The phased regions in the primary contig can be identified by simply aligning the associated haplotigs to the primary contig or by directly examining the assembly graph identifying the anchoring nodes from the haplotigs to the primary contig.

Software availability. FALCON and FALCON-Unzip are written in C and Python. falcon and its dependences are hosted open source on GitHub (https://github.com/PacificBiosciences/falcon). FALCON-Unzip is also hosted open source on GitHub (https://github.com/PacificBiosciences/FALCON_unzip). The specific git repositories of the various modules used for generating the assemblies presented in this paper are listed in the **Supplementary Note**. We have also prepared an Amazon Web Services EBS volume that contains all of the preconfigured software and an example *C. pyxidata* data set (see **Supplementary Data 4** and **Supplementary Note** for a walkthrough).

36. Liu, Y.-G. & Whittier, R.F. Rapid preparation of megabase plant DNA from nuclei in agarose plugs and microbeads. *Nucleic Acids Res.* **22**, 2168–2169 (1994).
37. Hayward, G.S. Unique double-stranded fragments of bacteriophage T5 DNA resulting from preferential shear-induced breakage at nicks. *Proc. Natl. Acad. Sci. USA* **71**, 2108–2112 (1974).
38. Myers, G. *Algorithms in Bioinformatics* (eds. Brown, D. & Morgenstern, B.) 52–67 (Springer, 2014).
39. Myers, E.W. The fragment assembly string graph. *Bioinformatics* **21**, ii79–ii85 (2005).
40. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).