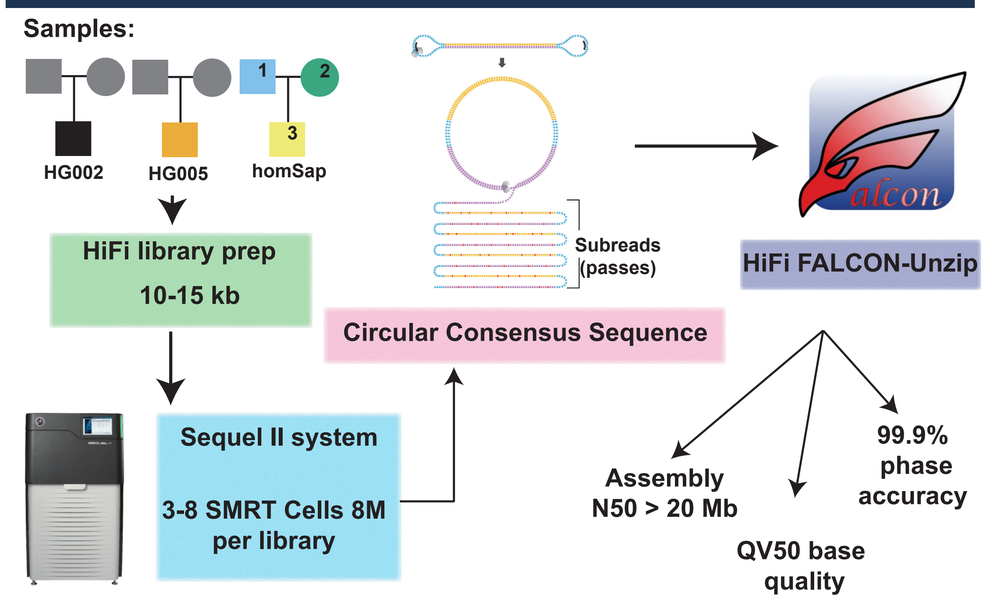
High-quality human genomes achieved through HiFi sequence data and FALCON-Unzip Assembly

Z.N. Kronenberg[1], I. Sovic[1], C. Dunn[1], S.B. Kingan[1], G.T. Concepcion[1], J. Drake[1], J. Korlach[1], P. Peluso[1], A.M. Wenger[1], J. Mountcastle[2,3], B. Haase[2,3], O. Fedrigo[2,3], E.D. Jarvis[2,3]

[1]Pacific Biosciences, Menlo Park, CA, USA; [2]Laboratory of Neurogenetics of Language, Box 54, The Rockefeller University, New York, NY 10065, USA; [3]Howard Hughes Medical Institute, Chevy Chase, MD, USA.

PACBIO® VERTEBRATE GENOMES PROJECT The Rockefeller University

Abstract



HiFi Assembly Statistics

Statistic	HG002	HG005	homSap1	homSap2	homSap3
Insert Size (kb)	15	10 &12	10	10	10
SMRT Cells 8M	3	8	4	4	4
CCS >= Q20 yield (Gb)	68	120	73	66	64
HiFi Coverage (fold)	22	40	24	22	21
Assembly Length (Gb)	2.91	2.87	2.89	2.87	2.89
Primary Ctgs.	2310	3702	3424	2986	3617
Primary N50 (Mb)	30.5	19.8	26.1	28.7	22.8
Asm. Phased	60%	56.20%	68.80%	64.50%	63.90%
Phase N50 (kb)	148	134	127	98	93

Table 1. Summary statistics for the five HiFi human de novo genome assemblies. Library preparation, depth of coverage, and chemistries varied, however, the resulting assemblies have similar statistics.

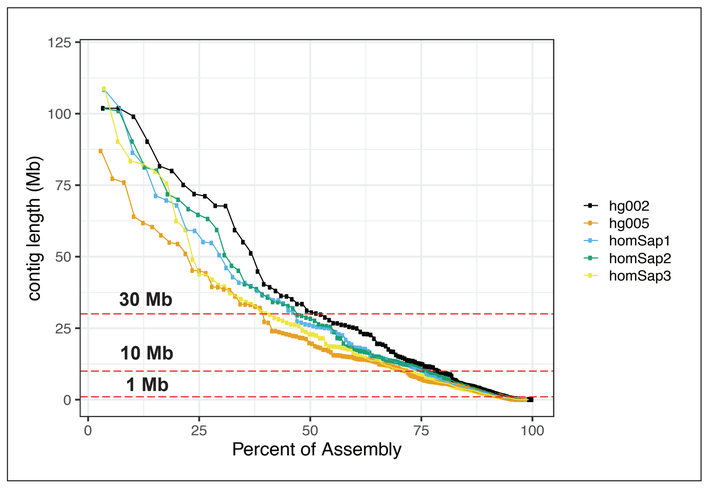


Figure 1. Primary contig sizes shown against the length of the genome assembly. Contigs are sorted from largest to smallest.

HiFi Contigs (>1 Mb) Mapped on GRCh38 Ideogram

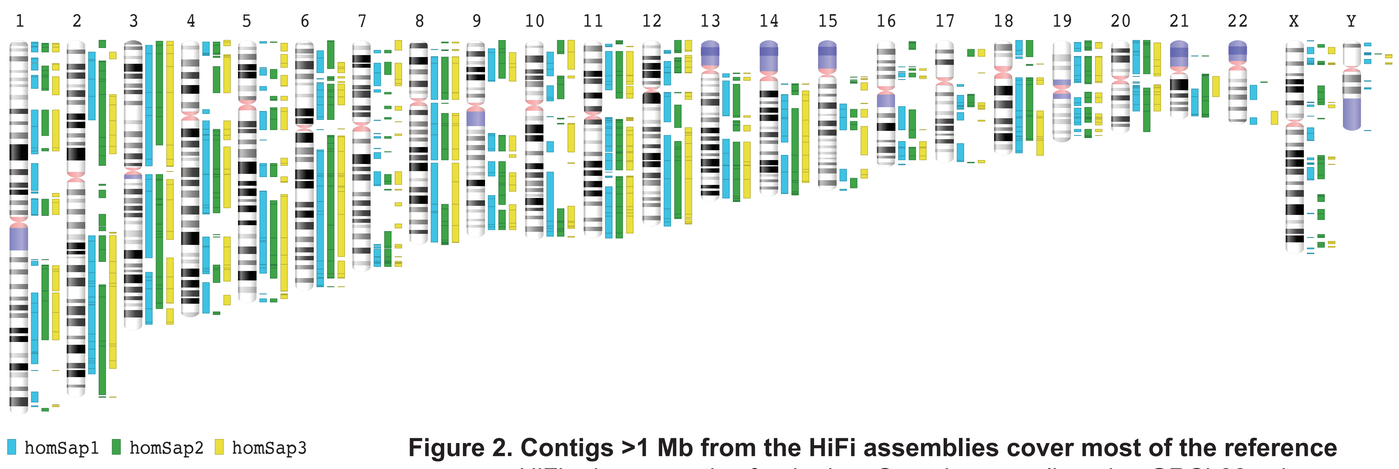


Figure 2. Contigs >1 Mb from the HiFi assemblies cover most of the reference genome. HiFi primary contigs for the homSap trio were aligned to GRCh38 using minimap2. Only contigs greater than 1Mb are shown. Breaks in the alignments do not always correspond to contig boundaries.

Basepair Accuracy of HiFi Contigs

–HiFi assembly of HG002 has higher basepair accuracy than Long Read assembly for the same sample.

–Median quality across 100 kb windows (N=25,789) is 5.5 Phred QV points, which translates to 3.5-fold fewer errors (**Figure 3**).

–Both primary and haplotigs are nearly QV50 (**Table 2**).

statistic	primary	haplotig
mean	49.29	47.7
median	50.6	49.88
QV40 windows	97.80%	92.40%

Table 2. Base Quality values for HiFi primary/haplotig contigs. The mean, median and number of 100 kb windows greater than QV40 are shown.

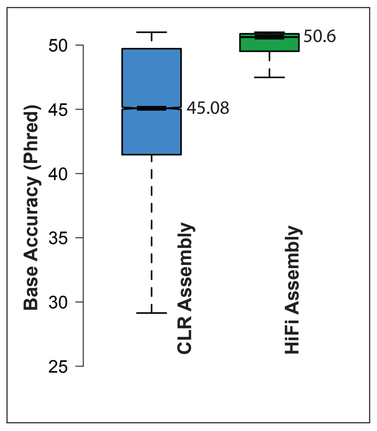


Figure 3. Comparison of Base QV between HiFi and CLR. The box and whisker diagram shows the distribution of Phred qualities in 100 kb windows across the HG002 *de novo* assembly.

Highly Accurate Phasing Achieved Through FALCON-Unzip

–FALCON-Unzip phased between 56.2-68.8% of the genomes we assembled (**Table1**).

–Average haplotig phasing accuracy was ~ 99.9% (**Figure 4**).

–Haplotigs (phase blocks) contiguity, measured as N50, ranged from 93-150 kb and increases with HiFi insert size.

–Phase blocks can span many genes. For example, in **Figure 5**, several genes are phased, including those with clinically relevant variants (OMIM; ClinVar)

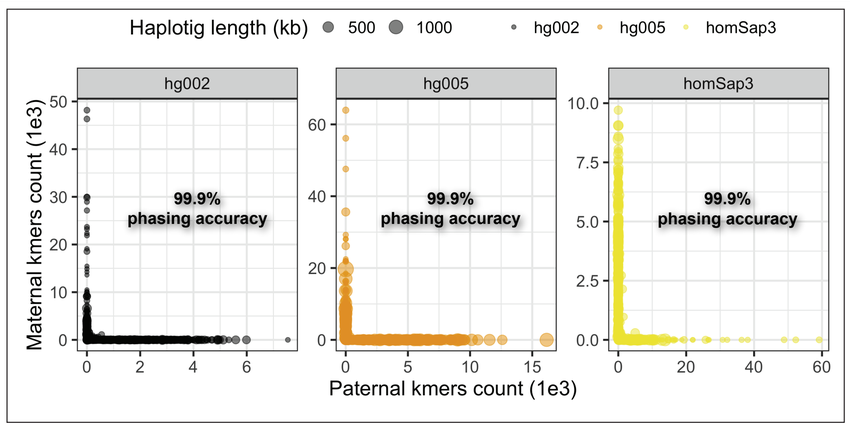


Figure 4. Haplotig Phasing Accuracy. Each point represents a haplotig, the size of the point denotes haplotig size. Points along the vertical and horizontal axis are enriched for either maternal or paternal kmers. The overall accuracy is listed within each pane.

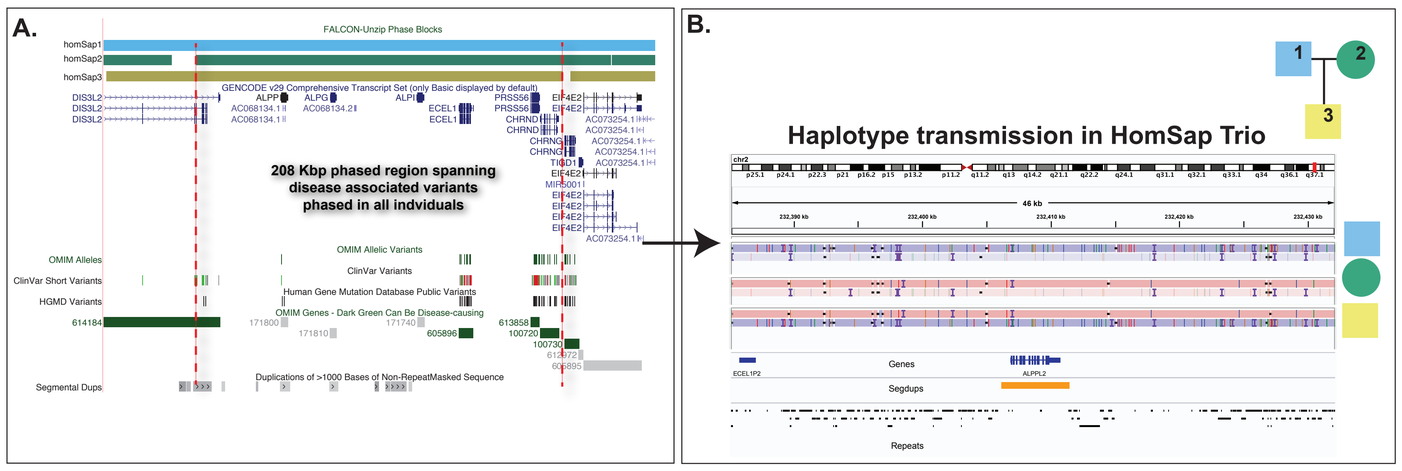
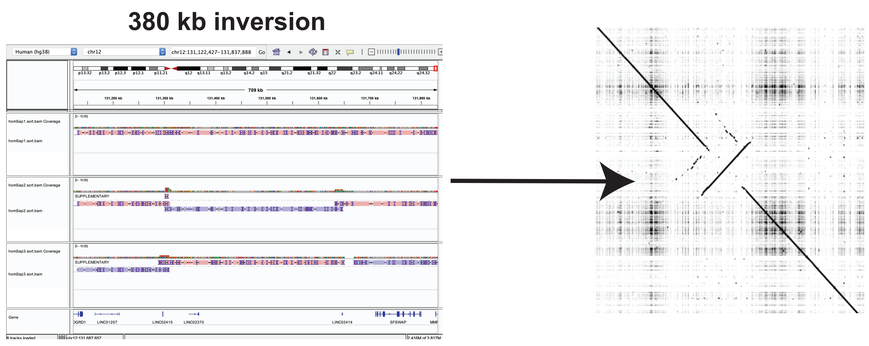


Figure 5. Example of haplotigs phasing clinically relevant genes. A) The top panel of genome browser shows the haplotigs/phased regions aligned to GRCh38. Genes tracks and clinical variants are shown below. **B)** Haplotype transmission in the HomSap trio, for the same region shown in Figure 5A.

Sequence Resolved Structural Variation



Conclusions

FALCON-Unzip produces highly accurate, well phased, human assemblies at 20 fold HiFi coverage. Here we have demonstrated the accuracy of FALCON-Unzip on five human genomes across three different trios.

–FALCON-Unzip now natively supports HiFi data.

–FALCON-Unzip is available at: https://github.com/PacificBiosciences/pbbioconda

–FALCON-Unzip HiFi user guide is available at: https://github.com/PacificBiosciences/pbbioconda/wiki/HiFi-:--FALCON-Unzip3-User-Guide

For further reading:

Wenger, Aaron M., et al. "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome." Nature biotechnology (2019): 1-8.

Vollger, Mitchell R., et al. "Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads." BioRxiv (2019): 635037.

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2019 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of Pacific Biosciences. Pacific Biosciences does not sell a kit for carrying out the overall No-Amp Targeted Sequencing method. Use of the No-Amp method may require rights to third-party owned intellectual property. BluePippin and SageELF are trademarks of Sage Science. NGS-go and NGSengine are trademarks of GenDx. FEMTO Pulse and Fragment Analyzer are trademarks of Agilent Technologies Inc. All other trademarks are the sole property of their respective owners