Single Molecule High-Fidelity (HiFi) Sequencing with >10 kb Libraries

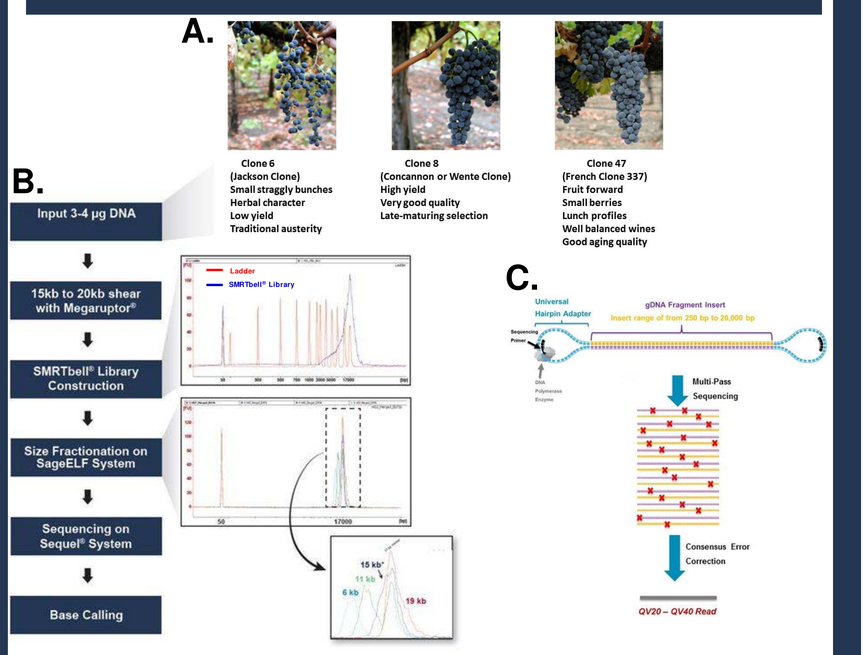
Paul Peluso[1], David Rank[1], Richard Hall[1], William Rowell[1], Greg Concepcion[1], Aaron Wenger[1], Arkarachai Fungtammasan[2], Chen-Shan Chin[2], Pi-Chuan Chang[3], Alexey Kolesnikov[3], Andrew Carroll[3], Jue Ruan[4], Sergey Koren[5], Jana Ebler[6], Tobias Mafschall[6], Andrea Minio[7], Rosa Figeroa-Balderas[7], and Dario Cantu[7]

1. Pacific Biosciences, Menlo Park CA, USA
2. DNAnexus, Mountain View CA, USA
3. Google, Inc., Mountain View CA, USA
4. Agricultural Genomics Institute, Chinese Academy of Agriculture Sciences, Shenzen, China
5. Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda MD, USA
6. Max Planck Institute for Informatics, Saarland Informatics Campus E1.4, Saarbrüken, Germany
7. Department of Viticulture & Enology, University of California Davis, Davis CA, USA

Abstract

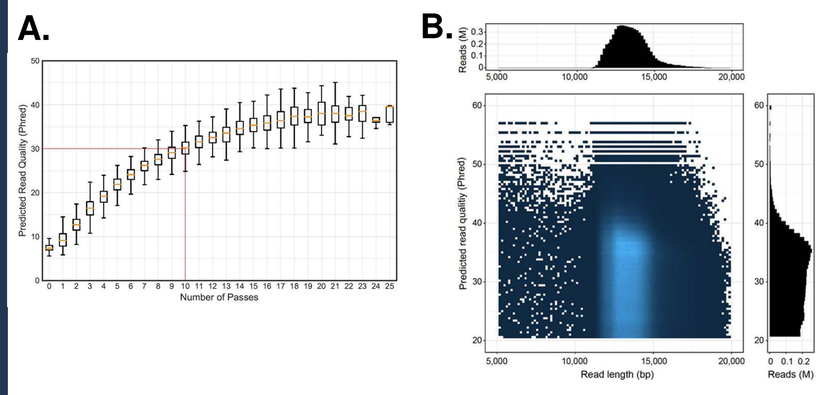
Recent improvements in sequencing chemistry and instrument performance combine to create a new PacBio data type, Single Molecule High-Fidelity reads (HiFi reads). Increased read length and improvement in library construction enables average read lengths of 10-20 kb with average sequence identity greater than 99% from raw single molecule reads. The resulting reads have the accuracy comparable to short read NGS but with 50-100 times longer read length. Here we benchmark the performance of this data type by sequencing and genotyping the Genome in a Bottle (GIAB) HG0002 human reference sample from the National Institute of Standards and Technology (NIST). We further demonstrate the general utility of HiFi reads by analyzing multiple clones of Cabernet Sauvignon. Three different clones were sequenced and *de novo* assembled with the CANU assembly algorithm, generating draft assemblies of very high contiguity equal to or better than earlier assembly efforts using PacBio long reads. Using the Cabernet Sauvignon Clone 8 assembly as a reference, we mapped the HiFi reads generated from Clone 6 and Clone 47 to identify single nucleotide polymorphisms (SNPs) and structural variants (SVs) that are specific to each of the three samples.

General Sequencing Approach



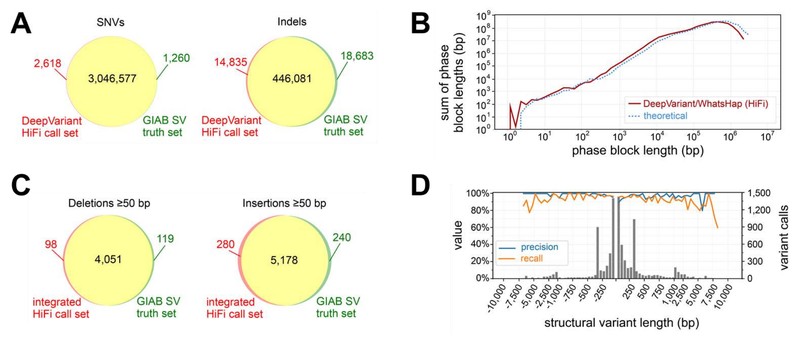
Sequencing three Clones of Caberent Sauvignon withHiFi reads. A) The three clones (6, 8, and 47) chosen for this study and the phentoyptic characteristics associated with each clone. **B)** Experimental design flowchart and size distribution QC of resulting SMRTbell templates. **C)** Schematic depicting the derivation of a single molecule HiFi read with a read quality ranging between QV20 and QV40 from the multiple reads across both forward and reverse strands of a SMRTbell molecule.

HiFi Read Characterization with NIST GIAB Human HG002 Standard



Basic characteristics of HG002 HiFi reads. A) Predicted read quality (Phred) vs. number of passes for a given SMRTbell template. Predicted QV30 is attainable by 10 passes. **B)** Read length and predicted read quality distributions for HG002 HiFi reads ≥ QV20. **C)** Empirical vs predicted QV distribution for HiFi reads ≥ QV20. Empirical estimates are only overestimating read quality by approximately 1 QV point on average. **D)** Coverage by GC fraction bin. Coverage of HiFi reads is very uniform across the human genome.

Variant Detection in HG002 using HiFi Reads



Variant calling and phasing with HiFi reads. A) Agreement of DeepVariant (HiFi) SNV and indel calls with Genome in a Bottle benchmark. B) Phasing of heterozygous DeepVariant (HiFi) variant calls with WhatsHap. C) Agreement of integrated HiFi structural variant calls with Genome in a Bottle benchmark, D) by variant size.

Summary Table

			SNVs			Indels		
Platform	Coverage	Variant Caller (training model)	Precision	Recall	F1 ^	Precision	Recall	F1
Illumina (NovaSeq)	30-fold	DeepVariant (Illumina model)	99.960%	99.940%	99.950%	99.633%	99.413%	99.523%
PacBio (HiFi)	28-fold	DeepVariant (HiFi model)	99.914%	99.959%	99.936%	96.901%	95.980%	96.438%
PacBio (HiFi)	28-fold	DeepVariant (haplotype-sorted HiFi model)	99.904%	99.963%	99.934%	97.835%	97.141%	97.486%
PacBio (HiFi)	28-fold	DeepVariant (HiFi model) + WhatsHap	99.884%	99.882%	99.883%	94.551%	86.465%	90.327%
Illumina (NovaSeq)	30-fold	GATK HaplotypeCaller (no filter)	99.852%	99.910%	99.881%	99.371%	99.156%	99.264%
PacBio (HiFi)	28-fold	GATK HaplotypeCaller (hard filter) + WhatsHap	99.693%	99.792%	99.742%	81.102%	83.818%	82.438%
PacBio (HiFi)	28-fold	GATK HaplotypeCaller (hard filter)	99.468%	99.559%	99.513%	78.977%	81.248%	80.097%

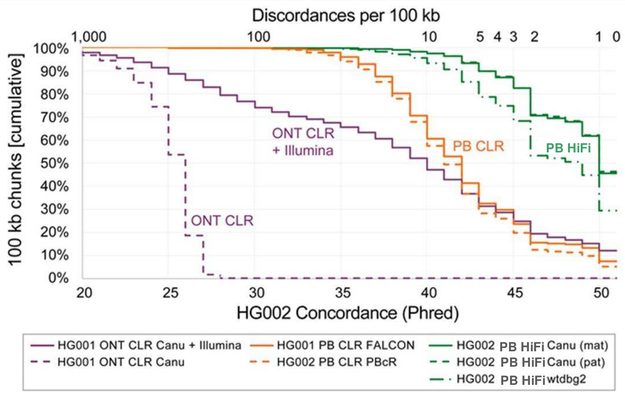
Performance of small variant calling with HiFi reads. Precision, recall, and F1 of small variant calling measured in the GIAB high-confidence regions using hap.py. **Bold values** indicates the highest value in each category column. *Values in italics* indicate that the value is higher than results obtained with the GATK HaplotypeCaller run on Illumina NovaSeq reads covering the human genome at a 30-fold depth.

HG002 *de novo* Assembly Results using HiFi Reads

Haplotype	Assembler	Total Size (Gb)	Contigs	N50 (Mb)	NG50 (Mb)	Max (Mb)	E-size[46] (Mb)	HG002 Concordance (Phred)	BUSCO Genes	RefSeq Genes
Mixed	Canu	3.42	18,006	22.78	25.02	108.46	30.16	31.1	92.3%	93.2%
Mixed	FALCON	2.91	2,541	28.95	24.51	110.21	38.04	25.8	87.6%	97.6%
Mixed	wtdbg2	2.79	1,554	15.43	12.62	84.67	22.61	44.6	94.2%	96.1%
Maternal	Canu*	3.04	5,854	18.02	17.04	48.81	19.78	47.2	94.1%	98.1%
Maternal	FALCON*	2.80	924	19.99	15.54	74.33	24.07	43.5	95.1%	97.8%
Maternal	wtdbg2	2.75	2,637	12.10	9.29	66.34	16.55	43.5	93.8%	95.6%
Paternal	Canu*	2.96	6,868	16.14	14.90	64.83	20.19	47.7	93.4%	98.2%
Paternal	FALCON*	2.70	1,489	16.40	14.06	95.34	25.61	43.5	93.6%	97.7%
Paternal	wtdbg2	2.67	1,444	13.96	10.86	50.51	15.36	42.1	92.6%	95.3%

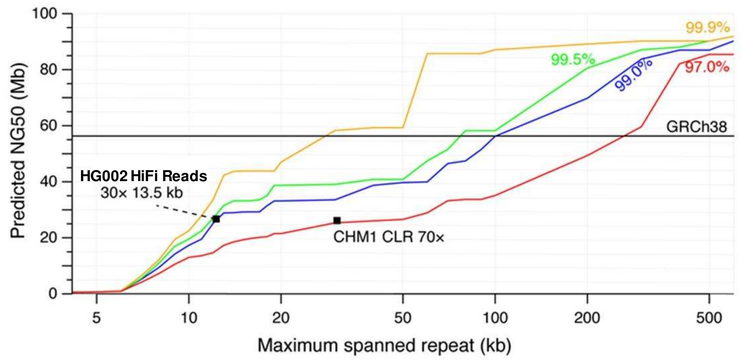
Statistics for de novo assembly using HiFi reads with three assemblers and three different haplotypes. The Ensembl "mixed" haplotype assemblies use all reads. The "maternal" and "paternal" assemblies use parent-specific reads from trio binning[3] plus unassigned reads. HG002 concordance is measured at non-variant positions in GIAB high-confidence regions. BUSCO gene completeness uses the Mammalia ODB9 gene set. RefSeq genes is the percentage of genes from R94 that are full-length, single-copy in assembly relative to the full-length, single-copy count for GRCh38. Contigs shorter than 13 kb were excluded from genome size and contiguity estimates; contigs shorter than 100 kb were excluded from the concordance measurement. "*" indicates polishing with Arrow.

HiFi Read Based Assembly QV



Impact of read accuracy on *de novo* assembly. The concordance of six assemblies to the reference genome at non-variant positions in GIAB high-confidence regions. Solid lines indicate assemblies from HiFi reads. Long-dash lines indicate accessioned assemblies from PacBio continuous long reads (CLR). Short-dash lines indicate an accessioned assembly from Oxford Nanopore (ONT) reads. Contigs longer than 100 kb were segmented into 100 kb chunks and aligned to GRCh37. Concordance was measured per chunk.

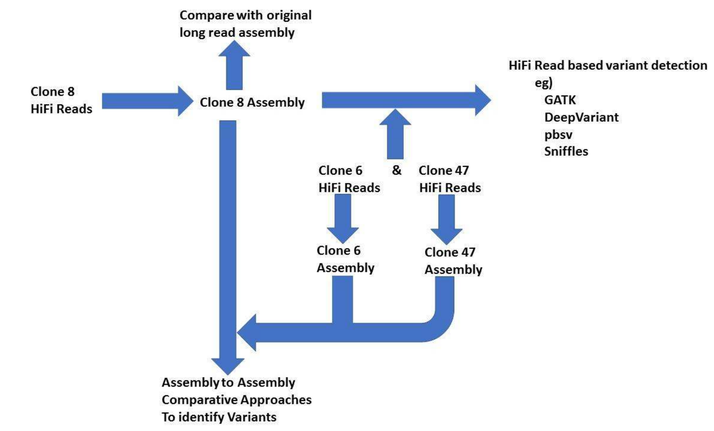
Modeling Human Assembly Contiguity Read Length vs Accuracy



Modeling human assembly contiguity. Model of the contiguity for a human assembly based on ability to resolve repeats of different length and percent identities. As a reference point, the contiguity of the current Human Ref GRCh38 is indicated as a solid line. Modeling predicts that reads with higher QV can assemble human genomes at high contiguity as compared to longer, noisier reads (compare **HG002 HiFi Reads** with **CHM1 CLR**). Moreover, 15 kb HiFi reads with read quality of ≥ QV30 can provide assemblies with Contig N50s approaching 40 Mb.

HiFi Read Data for Cabernet Clones & Analysis Strategy

	Clone 6	Clone 8	Clone 47
# of Reads	1,733,862	1,950,141	2,314,545
Yield (bases)	24,041,188,906	28,001,288,771	34,083,761,598
Average Read Length (bases)	13,865	14,358	14,725
Average Read Quality	.998	.998	.998



HiFi Read Assembly Results for Cabernet Sauvignon Clone 8

Assembly Stats	Vitis vinifera Cultivar PN0024[1]	Cabernet Sauvignon Clone8 PBI-CLR FalconUnzip Assembly[2]	Cabernet Sauvignon Clone8 PBI HiFi Read CANU[3] Assembly
Assembly Size (Mb)	486	959	988
# of Contigs	14,634	2,755	4,224
Max Contig Length (Mb)	n.a.	14.1	12.1
Contig N_{50} Length (Mb)	0.1	2.2	1.9
# of Scaffolds	2,061	N/A	N/A
Scaffold N_{50} Length (Mb)	3.4	N/A	N/A
Complete BUSCOs	93%	90%	92.4%
Duplicated BUSCOs	1.7%	69%	68%
Fragmented BUSCOs	2%	2%	1%
Missing BUSCOs	5%	8%	6%

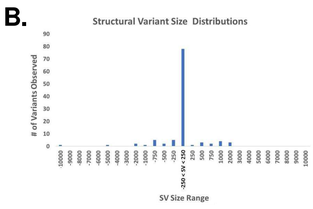
HiFi Read Based Variant Level Estimates

A.

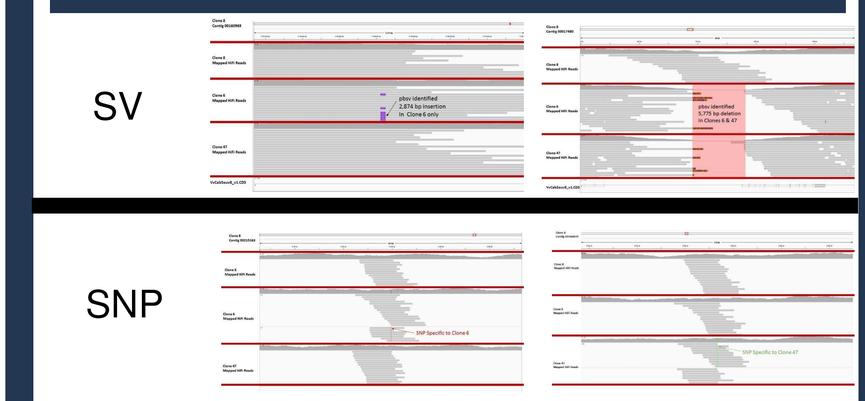
	Conservative Estimate Based on Clone6 & Clone47 HiFi Read Mapping
SNP Frequency	1 SNP in every 400,000 bases
Structural Variant Frequency	1 SV in every 9,000,000 bases

Estimating variant frequencies between Cabernet Sauvignon Clones 6, 8, and 47. A) Table of variant frequencies conservatively estimated by mapping HiFi reads from Clones 6 & 47 to the Clone 8 *de novo* assembly. Initially, SVs were called using pbsv and manually curated for half the genome. The final estimate was extrapolated to the entire genome size and reduced by one half to account for the use of two clones to determine the SV numbers. SNP calls were curated in a similar fashion manually but for only 3% of the genome **B)** Size distributions for the validated SVs from the manual curation of the data.

B.



Examples of Variant calls



Conclusion

We demonstrate

- Benchmarking studies with GIAB HG002 sample demonstrate that HiFi Reads can be used to characterize SNVs (SNPS) producing results rivaling or exceeding state of the art short read approaches
- HiFi Reads can be used to generate *de novo* assemblies as shown here for HG002 and Cabernet Sauvignon Clone 8 producing highly contiguous assemblies with extremely high-level accuracy
- HiFi Reads can be utilized to detect variants (SV and SNP) in Cabernet Sauvignon clones

Further work in progress

- *De novo* assemble Cabernet Sauvignon Clones 6 & 47
- Run assembly-based approaches to detect SVs and SNVs
- Run GATK and DeepVariant pipelines to further characterize SNVs
- Develop high confidence data regions for the variant calling among the Cabernet Sauvignon clones

References

1 https://www.ncbi.nlm.nih.gov/assembly/GCF_000003745.3/

2 Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR. Phased diploid genome assembly with single-molecule real-time sequencing. Nature methods. 2016 Dec;13(12):1050.

3 Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TP, Phillippy AM. De novo assembly of haplotype-resolved genomes with trio binning. Nature biotechnology. 2018 Dec;36(12):1174.

Acknowledgements
The authors would like to thank Adam Phillippy, Michael Schatz, Mark DePristo, Gene Meyers, Heng Li, Fritz Sedlazeck, and Justin Zook for helpful discussions and input.