Optimizing coverage for long-read variant detection in rare disease genomics

Abstract # 216577
Nina Gonzaludo, Gregory Young, Zev Kronenberg, Aaron M Wenger, Michael Eberle
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

Selecting the appropriate genomic coverage for reference-based variant detection requires balancing sequencing cost with accuracy. While 30x coverage is the standard for comprehensive variant detection with short-read technologies, it remains unclear whether this benchmark applies to long-read sequencing, which offers improved coverage uniformity, access to challenging genomic regions, and haplotype resolution.

To address this, we analyzed a 40x HiFi dataset of HG002 generated on a single Revio SMRT Cell using PacBio's new SPRQ chemistry. We downsampled the data to simulate varying coverage levels and assessed small and structural variant calling accuracy against Genome in a Bottle (GIAB) benchmarks.

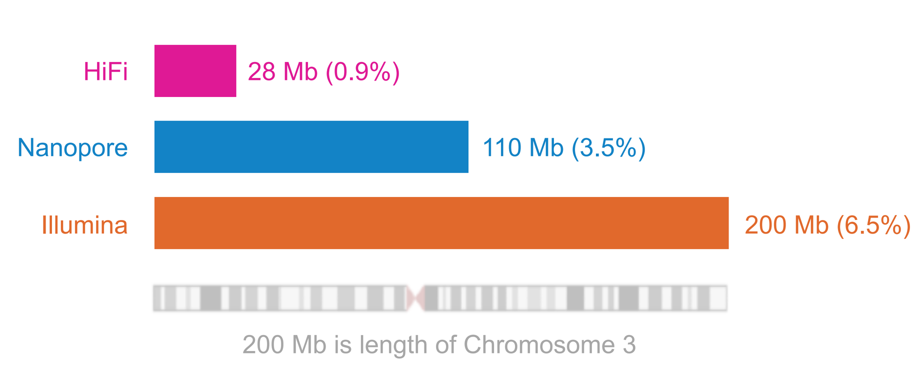


Figure 1. Total number of base pairs of the CHM13v1.0 T2T genome estimated to be missed by each respective sequencing technology. Read length used for analysis was 250 bp for Illumina, 25 kb for HiFi, and 100 kb for ONT (adapted from Nurk et al. 2022, Table S14).[1]

The advent of telomere-to-telomere (T2T) genomes and pangenomes provide a more complete view of human variation. Being able to access this variation, especially in the challenging regions of the genome that also tend to be the most polymorphic, is critical to better understanding the genetic basis of many rare diseases. More than 99% of the genome can be confidently analyzed with PacBio HiFi reads.

Methods

PacBio dataset

- HG002 cell-line DNA prepared with HiFi prep kit 96 following standard protocol
- SPRQ sequencing chemistry on single Revio SMRT Cell
- 146 Gb of mapped HiFi data
- Downsampled from 8- to 40-fold aligned depth

Revio +SPRQ

Alternative technology datasets

- Illumina dataset from Behera, S., et al. 2024 [2]
- ONT data was collected from EPI2ME [3]

Please see the "SPRQ Nov 2024" benchmark at github.com/PacificBiosciences/pb-benchmarks for technical details on the analysis and links to each of the datasets used.

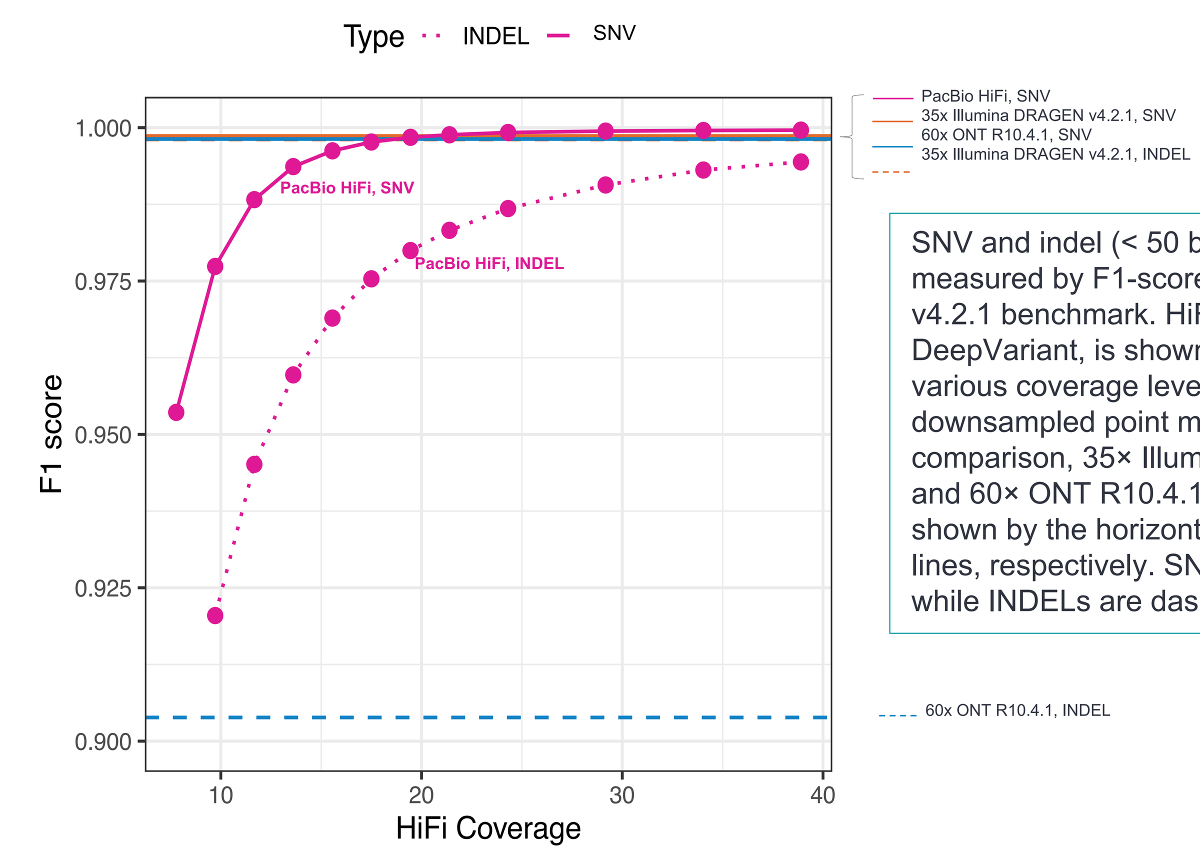
Results

Small variant calling performance

Table 1. SNV and INDEL calling performance against the GiaB v4.2.1 benchmark.

		SNV performance			INDEL performance		
Tech	Coverage	Recall SNV	Precision SNV	F1-score SNV	Recall INDEL	Precision INDEL	F1-score INDEL
HiFi	20x	0.997527	0.999330	**0.998428**	0.977192	0.98275	**0.979963**
HiFi	30x	0.999218	0.999658	**0.999438**	0.989809	0.991506	**0.990657**
HiFi	40x	0.999441	0.999747	**0.999594**	0.993939	0.994882	**0.994410**
Illumina	35x	0.997838	0.999447	**0.998642**	0.997473	0.998683	**0.998077**
ONT	60x	0.997757	0.998536	**0.998147**	0.862416	0.949482	**0.903857**

Figure 2 HG002 small variant calling performance by technology



SNV and indel (< 50 bp) accuracy measured by F1-score against the GiaB v4.2.1 benchmark. HiFi performance, using DeepVariant, is shown in magenta at various coverage levels, with each downsampled point marked by a circle. For comparison, 35× Illumina DRAGEN 4.2.1 and 60× ONT R10.4.1 SUP performance is shown by the horizontal orange and blue lines, respectively. SNV are solid lines, while INDELs are dashed.

20x HiFi coverage matches F1-score SNV performance of 35x Illumina and 60x ONT

30x HiFi coverage outperforms SNV performance of 35x Illumina and 60x ONT datasets. Moving from **30x** to **40x** HiFi coverage shows minimal gains in F1-score.

20x HiFi coverage shows superior F1-score INDEL performance to 60x ONT but lags behind 35x Illumina. Increasing coverage shows continued gains in INDEL calling performance (e.g., 40x at 99.44% F1-score).

Structural variant (SV) calling performance

Table 2. SV calling performance against the GiaB HG002 Q100 benchmark.

Tech	Coverage	Recall SV	Precision SV	F1-score SV
HiFi	10x	0.8804	0.9900	**0.9320**
HiFi	20x	0.9463	0.9891	**0.9672**
HiFi	30x	0.9560	0.9885	**0.9720**
HiFi	40x	0.9608	0.9882	**0.9743**
Illumina	35x	0.4055	0.9716	**0.5722**
ONT	60x	0.9237	0.9871	**0.9543**

Structural variant (SV) calling performance

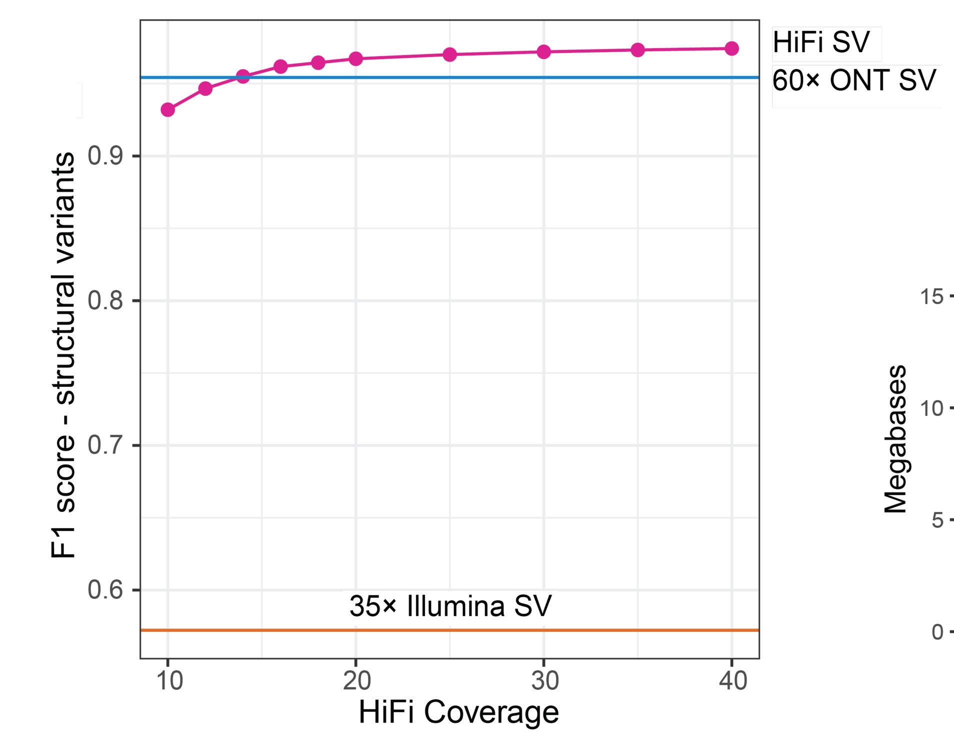


Figure 3. HG002 structural variant calling performance by technology.

Performance measured against the GIAB HG002 Q100 benchmark[4]. HiFi performance, using Sawfish[5], is shown in magenta at different coverage levels, with downsampled points marked by circles. For comparison, 35x Illumina DRAGEN 4.2.4 and 60x ONT R10.4.1 SUP performance is shown by the horizontal orange and blue lines, respectively.

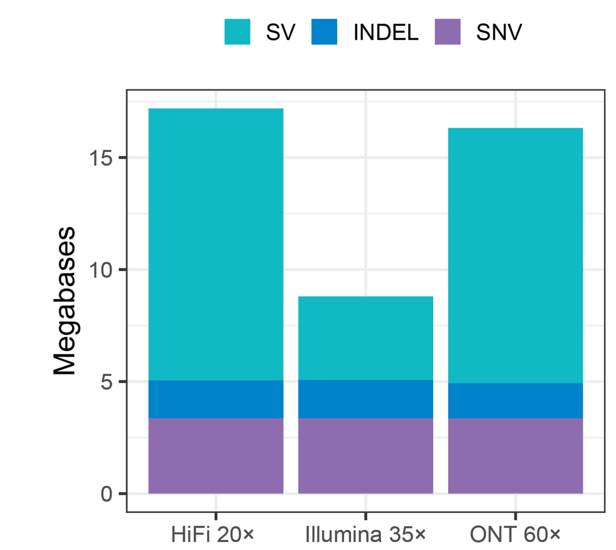


Figure 4. Total true positive variation identified by each technology.

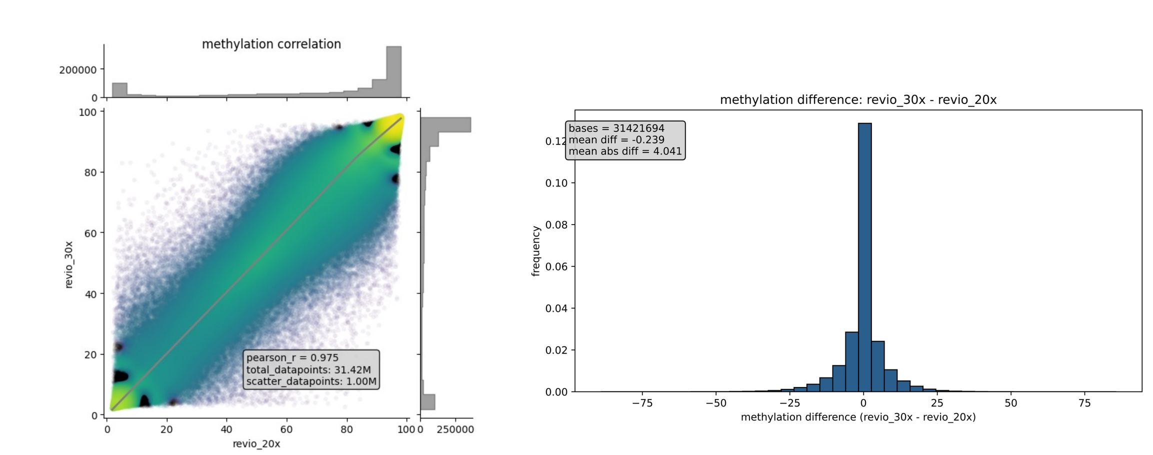
SV true positives were labeled by Truvari, and SNVs and indel true positives were labeled by hap.py. HiFi was measured at 20x, Illumina at 35x, and ONT at 60x coverage.

20x HiFi coverage outperforms other technologies at higher coverages for SV detection

HiFi genome identifies more true variation at lower coverage levels than other technologies

5mC concordance (20x vs 30x)

Methylation at CpG sites is detected during sequencing without the need for any additional prep or sequencing. This enables additional epigenetics insights with standard PacBio WGS



5mC calls at 20x remain highly concordant with 30x depth

Conclusion

These results indicate that optimal coverage should be tailored to study objectives and technology. For population-scale studies, a 20x HiFi genome offers the best balance of cost and accuracy, while coverage >20x may be warranted in individual cases demanding higher sensitivity. The value of a 20x HiFi genome is further supported by an independent study showing that 20x coverage recalls 96.2% of the difficult, clinically-relevant germline variants identified at 30x.[6] These results may change as algorithms and technologies continue to improve.

References

1. Nurk, S. et al. (2022) The complete sequence of a human genome. *Science* 376:44 – 53 (Supp. Table S14)
2. Behera, S., Catreux, S., Rossi, M. et al. (2024) Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol.* https://doi.org/10.1038/s41587-024-02382-1
3. EPI2ME: https://labs.epi2me.io/giab-2023.05/ (accessed September 2024)
4. See: https://github.com/marbl/HG002 (access September 2024)
5. Saunders, C. et al. (2025) Sawfish: Improving long-read structural variant discovery and genotyping with local haplotype modeling. *Bioinformatics,* 41(4)
6. Hop W., et al. (2024) HiFi long-read genomes for difficult-to-detect clinically relevant variants. *American Journal of Human Genetics*, 112(2)

Research use only. Not for use in diagnostic procedures. © 2025 Pacific Biosciences of California, Inc. ("PacBio"). All rights reserved. Information in this document is subject to change without notice. PacBio assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of PacBio products and/or third-party products. Refer to the applicable PacBio terms and conditions of sale and to the applicable license terms at pacb.com/license. Pacific Biosciences, the PacBio logo, PacBio, Circulomics, Omniome, SMRT, SMRTbell, Iso-Seq, Sequel, Nanobind, SBB, Revio, Onso, Apton, Kinnex, PureTarget, SPRQ, and Vega are trademarks of PacBio.