Application brief

# COMPREHENSIVE HUMAN GENOMIC VARIANT DETECTION WITH HIFI LONG-READ SEQUENCING

PacBio® HiFi sequencing provides industry-leading comprehensive variant detection, unlocking access to more of the human genome than other sequencing technologies and bringing valuable insights at any sequencing depth.

## The benefits of HiFi sequencing

**Mappability**

Access >99% of the genome, including regions missed by other technologies

**Phasing**

Resolve variants by haplotype for better biological insights

**Accuracy**

Identify all classes of variants with minimal errors for a clearer view of the genome

**Methylation**

Gain single-molecule insights into 5mC methylation for epigenetic studies

## Highlights

**Unmatched access to the genome.** Map the full catalog of variation across >99% of the human genome, capturing tens of megabases of sequence missed by Illumina and Oxford Nanopore (ONT) sequencing.

**Superior structural variant (SV) detection.** 20× HiFi coverage provides industry-leading SV detection, detecting thousands of SVs missed by Illumina.

**Excellent small variant performance.** 20× HiFi coverage outperforms 60× ONT (R10.4.1 SUP), with tens of thousands fewer errors for insertions and deletions (indels).

**Phasing and DNA methylation included.** Resolve variants by haplotype and gain single-molecule 5mC methylation insights without additional workflows.

## See more of the genome with HiFi

The advent of Telomere-to-Telomere (T2T) genomes and pangenomes provides a full view of human genomic variation. Gaining that view in your own samples requires the right technology – PacBio HiFi sequencing. Illumina short reads and lower accuracy ONT long reads both fail to confidently map to many of these new reference regions, resulting in missed variation. Nurk et al. (2022) report that 200 Mb (6.5%) of the human genome – the approximate size of human chromosome 3 – is not accessible by Illumina reads and 110 Mb (3.5%) is not accessible with ONT reads. In contrast, more than 99% of the human genome can be confidently analyzed with PacBio HiFi reads (Figure 1).

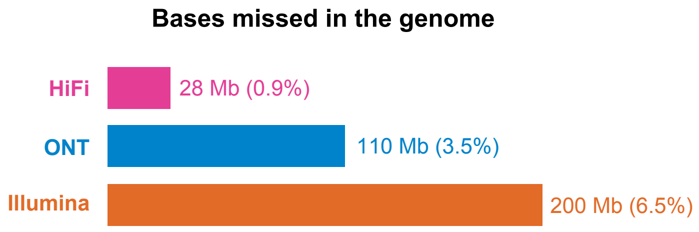

**Figure 1.** Total number of base pairs of the CHM13v1.0 T2T genome estimated to be missed by each respective sequencing technology. Read length used for analysis was 250 bp for Illumina, 25 kb for HiFi, and 100 kb for ONT (Nurk et al. 2022, Table S14).[1]

## HiFi sequencing provides the most comprehensive view of genomic variation

The superior variant calling performance of PacBio HiFi is highlighted in a comparison against recent Illumina and ONT datasets (Figure 2). All three sequencing technologies show similar SNV performance but differentiate more significantly on indels with HiFi significantly better than ONT (60×) and gaining on Illumina (35×). HiFi sequencing demonstrates industry-leading SV performance where 20× HiFi surpasses 60× ONT and far outperforms Illumina by 39% (Figure 2B).

Summing the total true sequence variation called across all variant types shows that structural variation affects more bases of the genome and that a HiFi genome (20×) identifies approximately 0.87 Mb and 8.4 Mb more true positive variation than ONT and Illumina, respectively (Figure 3).

Titrating the HiFi coverage demonstrates that a 20× HiFi genome achieves over 99% of the 30× F1 score for SNVs and SVs and over 98% of the 30× F1 score for indels (see Figure 2). The value of a 20× HiFi genome is further supported by an independent study showing that 20× HiFi coverage recalls 96.2% of the difficult, clinically-relevant germline variants identified at 30×.[2] Because an average depth of 20× identifies nearly all the variation at 30×, and more than other technologies, we recommend a 20× HiFi genome to optimize the accuracy and cost for most reference-based applications.

A single sequencing SMRT® Cell using the Revio® SPRQ™ chemistry[3] delivers 120 Gb, equivalent to one 40× genome, two 20× genomes, or four 10× genomes. The Vega™ system delivers 60 Gb per SMRT Cell[4], equal to one 20× genome or two 10× genomes.

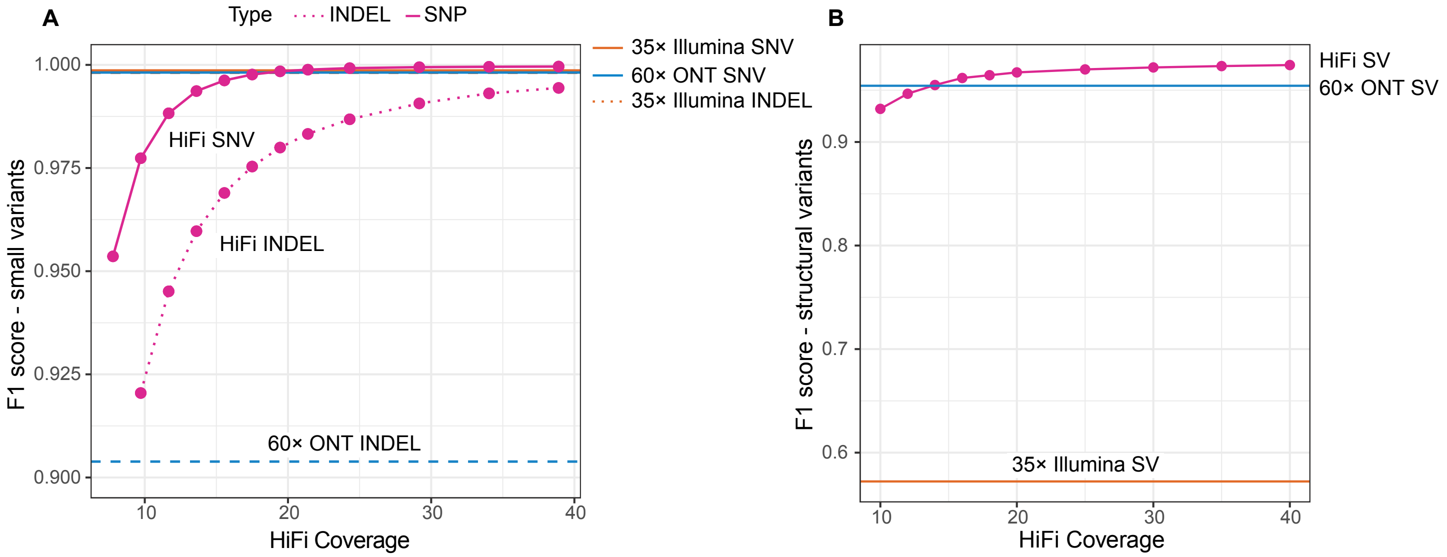

**Figure 2 HG002 variant calling performance by technology**
**A.** SNV and indel (< 50 bp) accuracy measured by F1-score against the GiaB v4.2.1 benchmark.[5] HiFi performance, using DeepVariant, is shown in magenta at various coverage levels, with each downsampled point marked by a circle. For comparison, 35× Illumina DRAGEN 4.2.1 and 60× ONT R10.4.1 SUP performance is shown by the horizontal orange and blue lines, respectively. **B.** Structural variant (SV) calling performance measured against the GiaB HG002 Q100 benchmark[6]. HiFi performance, using Sawfish[7], is shown in magenta at different coverage levels, with downsampled points marked by circles. For comparison, 35× Illumina DRAGEN 4.2.4 and 60× ONT R10.4.1 SUP performance is shown by the horizontal orange and blue lines, respectively.

For HiFi sequencing, HG002 cell-line DNA was prepared using the automated HiFi prep kit 96 workflow[8] and sequenced on a single Revio SMRT Cell using SPRQ chemistry. Sequencing generated 146 Gb of mapped HiFi data, which was downsampled to coverage levels ranging from 8× to 40×. For Illumina sequencing, publicly available data from Behera, S., et al. 2024 were used.[9] Data was collected from EPI2ME for ONT sequencing.[10] 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.[11]

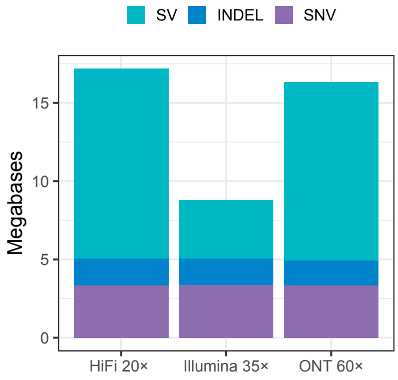

**Figure 3.** Total true positive variation identified by each technology. SV true positives were labeled by Truvari, and SNVs and indels true positives were labeled by hap.py. HiFi was measured at 20×, Illumina was measured at 35×, and ONT was measured at 60× coverage.

## KEY REFERENCES

1. Nurk, S. et al. (2022) The complete sequence of a human genome. *Science* 376:44 – 53 (Supp. Table S14)
2. Hop W., et al. (2024) HiFi long-read genomes for difficult-to-detect clinically relevant variants. *medRxiv*. https://doi.org/10.1101/2024.09.17.24313798
3. https://www.pacb.com/revio/
4. https://www.pacb.com/vega/
5. Wagner, J., et al. (2022). Benchmarking challenging small variants with linked and long reads. *Cell Genomics*, 2(5), 100128.
6. See: https://github.com/marbl/HG002 (access September 2024)
7. Saunders, C. et al. (2024) Sawfish: Improving long-read structural variant discovery and genotyping with local haplotype modeling. *bioRxiv* https://doi.org/10.1101/2024.08.19.608674
8. Protocol: https://www.pacb.com/wp-content/uploads/Guide-overview-Automated-HiFi-prep-96-and-HiFi-ABC-for-the-Hamilton-NGS-STAR-MOA-system.pdf
9. 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
10. EPI2ME: https://labs.epi2me.io/giab-2023.05/ (accessed September 2024)
11. https://github.com/PacificBiosciences/pb-benchmarks/tree/main/SPRQ-NOV-2024