Datasheet

DNAscope LongRead Pipeline for PacBio HiFi: Illuminating the "Dark Matter" of the Genome

Introduction: A New Standard for Long-Read Accuracy

PacBio HiFi sequencing provides 99.9% single-molecule read accuracy (Q30+) with the best homopolymer resolution of any long-read sequencing technology. With read lengths greater than 10kb, HiFi sequencing can illuminate variation hidden in dark regions of the genomes that are inaccessible with traditional short-read sequencing approaches, providing a solid sequencing platform for whole-genome characterization.

Today, we are introducing an updated **Sentieon DNAscope LongRead** pipeline, with a model specifically optimized for PacBio HiFi data. This pipeline doesn't just maintain Sentieon's signature speed; it provides an optimized workflow that is specifically designed to provide high variant calling accuracy for the most difficult variation in the genome's most challenging regions.

Examples of regions that are difficult to sequence using traditional approaches:

1. **Segmental Duplications (SegDup):** Highly identical sequences in the genome cause ambiguity during read alignment and variant calling for both short and standard long-read alignments.
2. **Pseudogenes and gene duplications:** Variation in gene families where active or inactive genes are present in multiple copies provide a unique challenge for variant calling.
3. **Homopolymers and other long repeats:** Long homopolymer sequences are inherently difficult to resolve as these regions are prone to sequencing error.

Linear reference genomes (like GRCh38) cannot capture the structural diversity of human populations. The DNAscope LongRead pipeline solves this by combining highly accurate small variant calling with long-read structural variant calling for accurate assessment of all classes of genomic variation.

How It Works

The pipeline takes aligned or unaligned reads as input and produces phased, filtered variant calls through a series of coordinated steps:

1. **Alignment**. If starting from FASTQ or unaligned BAM/CRAM, reads are aligned to the reference genome using Sentieon's accelerated minimap2 implementation, then coordinate-sorted into analysis-ready alignments.

2. **First-pass diploid calling**. Sentieon DNAscope performs an initial round of variant calling across diploid regions of the genome using a PacBio-specific model. A machine learning model (DNAModelApply) is then applied to refine call quality.

3. **Phasing**. VariantPhaser uses heterozygous variants and read-level haplotype information to partition the genome into phased and unphased regions, producing a phased VCF and a set of phase assignments for each read.

4. **Second-pass, haplotype-aware calling**. Within phased regions, reads are split by haplotype and variant calling is performed independently on each haplotype to improve indel accuracy. A patching step reconciles the per-haplotype calls with the phased diploid variants from the first pass.

5. **Unphased region calling**. Regions that could not be phased are called separately, then patched against the first-pass calls.

6. **Merging**. Phased haplotype calls and unphased diploid calls are merged into a single, cohesive output VCF. Optional haploid calling can be performed on user-specified regions (e.g., chrX in males). gVCF output is available for joint genotyping workflows.

7. **Structural variant calling**. In parallel, Sentieon LongReadSV identifies structural variants using a dedicated model trained for long-read data, producing a separate SV VCF.

The result is a comprehensive set of variant calls: SNVs, indels, and SVs, from a single pipeline invocation.

Performance Breakthrough: Accuracy Where It Matters Most

Below, we benchmark Sentieon DNAscope LongRead using the Sentieon PacBio model (version 2.3) with Fastq files sequenced by Revio platform (non SPRQ chemistry for HG001, SPRQ for

HG002-HG004), released by PacBio during 2024Q4. The DNAscope LongRead pipeline demonstrates industry-leading performance on the PacBio platform.

Whole Genome Region Performance:

- F1-Score of SNP > 0.999, total error counts <~8000
- F1-Score of Indel > 0.996, total error counts <~4000

- **SegDups Region Performance:**
  - Total error counts <~1500.
  - 50% fewer errors than Illumina.

- **CMRG Performance**
  - Total error counts <~450.
  - 30% fewer errors than Illumina.
  - These metrics represent a significant leap in resolving variants that were previously hidden in "dark" duplicated regions.

Speed and Efficiency

Efficiency remains at the core of Sentieon. Compared to standard open-source pipelines, DNAscope LongRead provides a massive boost in processing speed for PacBio data, drastically reducing compute costs and turnaround times.

- **Rapid Turnaround:** Process a 30x HiFi WGS sample from Fastq to VCF in 40 minutes, 3-5x less runtime required by traditional tools.
- **Resource Optimized:** Lower memory footprint, designed for on-premise HPC environments or the cloud.

Conclusion

PacBio HiFi technology provides an unprecedented view of the genome, and **Sentieon DNAscope LongRead** provides a powerful lens to view the “Dark Matter” of the genome clearly. By combining high-accuracy sequencing with accurate and performant algorithms, we are pushing the boundaries of what is possible in clinical WGS.

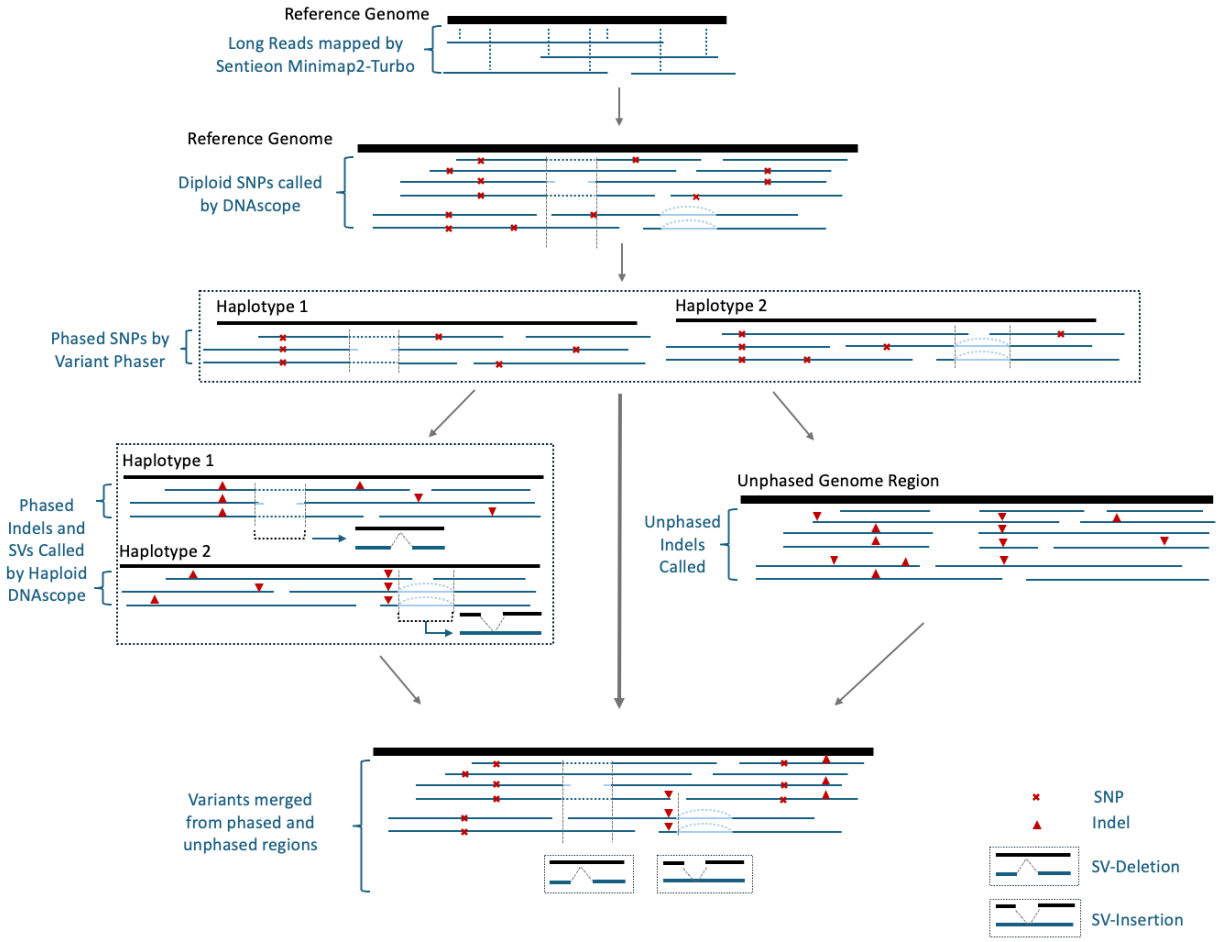

Figure 1. Sentieon DNAscope LongRead pipeline Schematic Diagram. The pipeline first aligns reads using an accelerated minimap2 and performs initial variant calling. Variants and reads are then used for phasing, separating the genome into phased and unphased regions. Next, a second-pass haplotype-aware calling is performed in phased regions. In parallel, unphased regions are called separately, and all results are merged into a single cohesive output VCF. Sentieon LongReadSV module runs independently to detect structural variants, resulting in a comprehensive set of SNPs, Indels, and SVs.

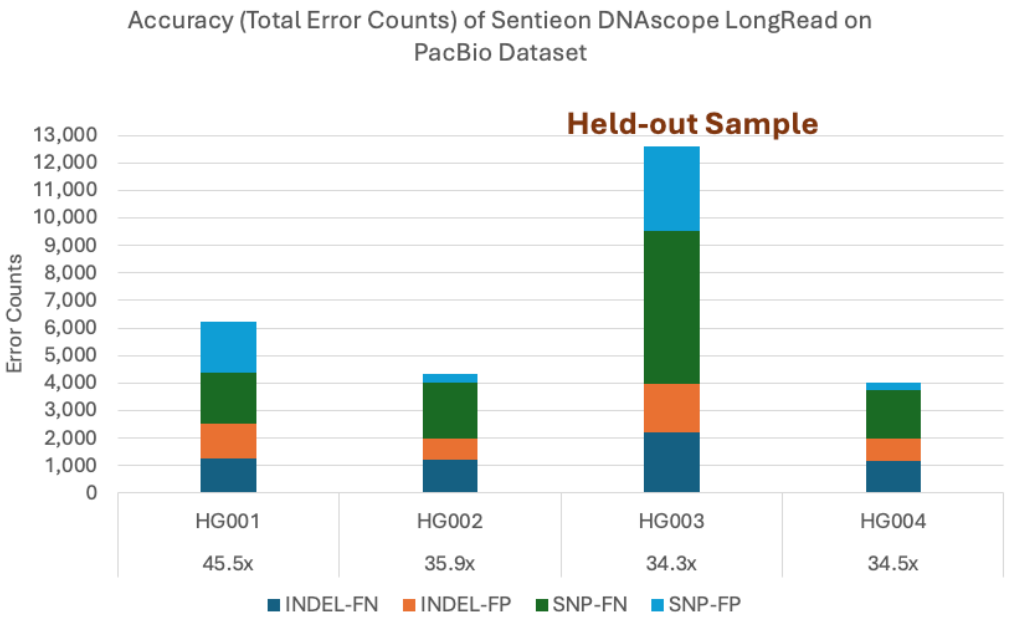

Figure 2. Error counts for DNAscope LongRead across PacBio WGS datasets. Total errors fall between roughly 4k and 12k for the HG001–HG004 samples. Notably, sample HG003 was held out during model training, so its performance reflects true out-of-sample accuracy.

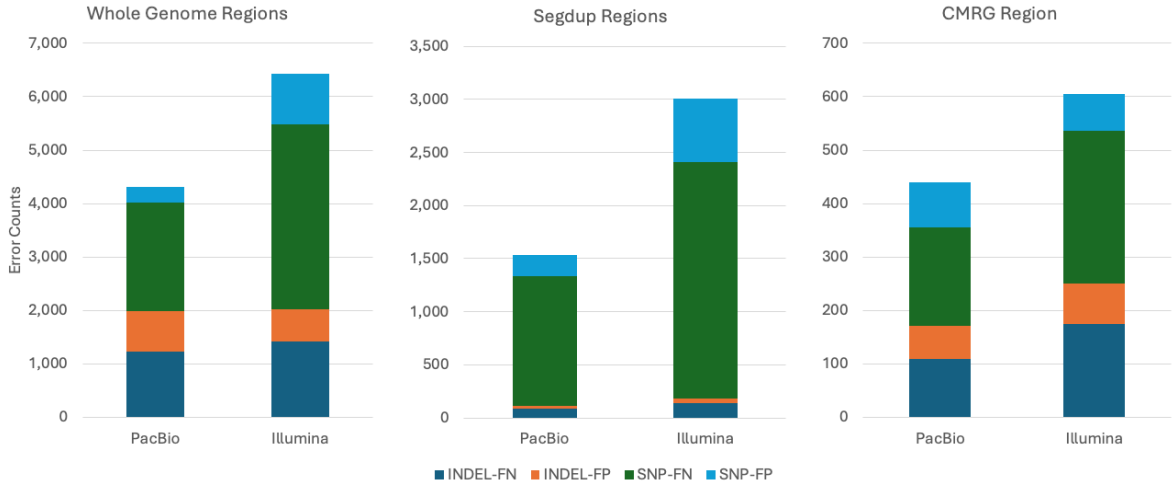

Figure 3. Error counts comparison between PacBio and Illumina in difficult genome regions. There are approximately 30-50% error reductions of PacBio accuracy comparing to Illumina platform.

Dataset	Depth	Platform	Pipeline	INDEL-FN	INDEL-FP	SNP-FN	SNP-FP	INDEL-F1	SNP-F1
HG001	45.5x	PacBio	DNAscope LongRead	1,248	1,262	1,874	1,840	0.9975	0.9994
HG002	35.9x	PacBio	DNAscope LongRead	1,223	761	2,030	303	0.9982	0.9997
HG003	34.3x	PacBio	DNAscope LongRead	2,216	1,762	5,553	3,068	0.9963	0.9987
HG004	34.5x	PacBio	DNAscope LongRead	1,190	793	1,762	262	0.9982	0.9997

Table 1. Accuracy at a glance — DNAscope LongRead on PacBio WGS datasets.

Dataset	Depth	vCPUs	Runtime (min)	Core Hours
HG001	45.5x	64	46:39.4	49.8
HG002	35.9x	64	39:57.8	42.6
HG003	34.3x	64	38:27.8	41.0
HG004	34.5x	64	41:24.8	44.2

Table 2. How fast and how cheap? Efficiency metrics for DNAscope LongRead on PacBio WGS data (on c8i.16xlarge, from Fastq to VCF).

Read More

Manual: https://support.sentieon.com/docs/sentieon_cli/#dnascope-longread

Preprint with older model: https://www.biorxiv.org/content/10.1101/2022.06.01.494452v1