Datasheet

Updated DNAscope LongRead Pipeline with ONT WGS: Fast, Accurate, and Cost-Efficient Long Reads Germline Variant Calling

Sentieon DNAscope LongRead

ONT

Long-read sequencing with Oxford Nanopore Technology (ONT) has opened the door to more complete genome characterization, enabling the detection of structural variants, repeat expansions, and complex rearrangements that short reads simply cannot resolve. But turning raw nanopore signal into accurate, clinically relevant variant calls remains a challenge. Higher per-read error rates, variable read lengths, and the sheer volume of data from modern flow cells all demand purpose-built analysis pipelines - and the computational cost of existing approaches can be a bottleneck for labs scaling up long-read sequencing.

That's the problem we set out to solve with the Sentieon DNAscope Long Read pipeline for ONT data: a fastq-to-VCF or BAM-to-VCF workflow that combines deep-learning-based variant calling with a multi-pass phasing strategy to achieve high accuracy across SNVs, indels, and structural variants while running efficiently on standard compute infrastructure.

Below, we benchmark Sentieon DNAscope LongRead using the Sentieon ONT model (version 2.3) with aligned CRAM files released by ONT during AGBT 2025.

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 trained ONT-specific model. A machine learning model (DNAModelApply) is then applied to refine call quality.

3. **Phasing**. The 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.

The Numbers

On Genome in a Bottle (GIAB) v4.2.1 benchmarks using ONT 30x-40x WGS data, the pipeline achieves:

- SNP F1 > 99.9%
- Indel F1 > 94.0%
- SV F1 >95%
- Significantly outperform Clair3 and other ONT compatible analysis pipelines

The workflow is also highly efficient. With pre-aligned whole-genome input, variant calling completes in approximately 20 min (~24 core-hours), corresponding to only less than a few dollars in on-demand compute cost.

Combined with the low cost of ONT sequencing, this enables accurate and scalable pangenome-informed analysis at true population scale—making high-quality variant calling accessible for large cohorts and routine applications.

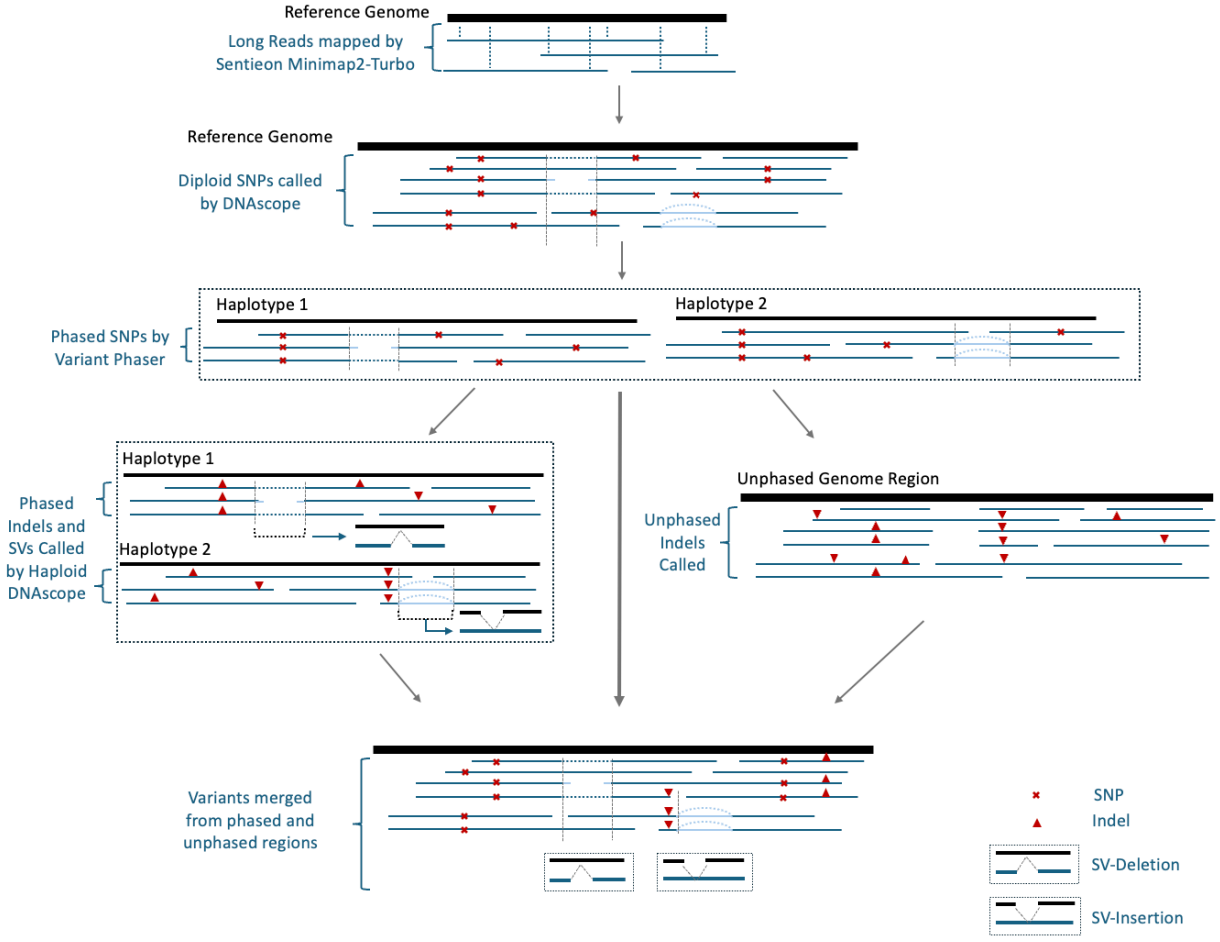

Figure 1. Sentieon DNAscope longRead pipeline Schematic Diagram. The pipeline first aligns reads using an accelerated minimap2 and performs initial variant calling. Variants and reads are then used for phasing, separating the genome into phased and unphased regions. Next, a second-pass haplotype-aware 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.

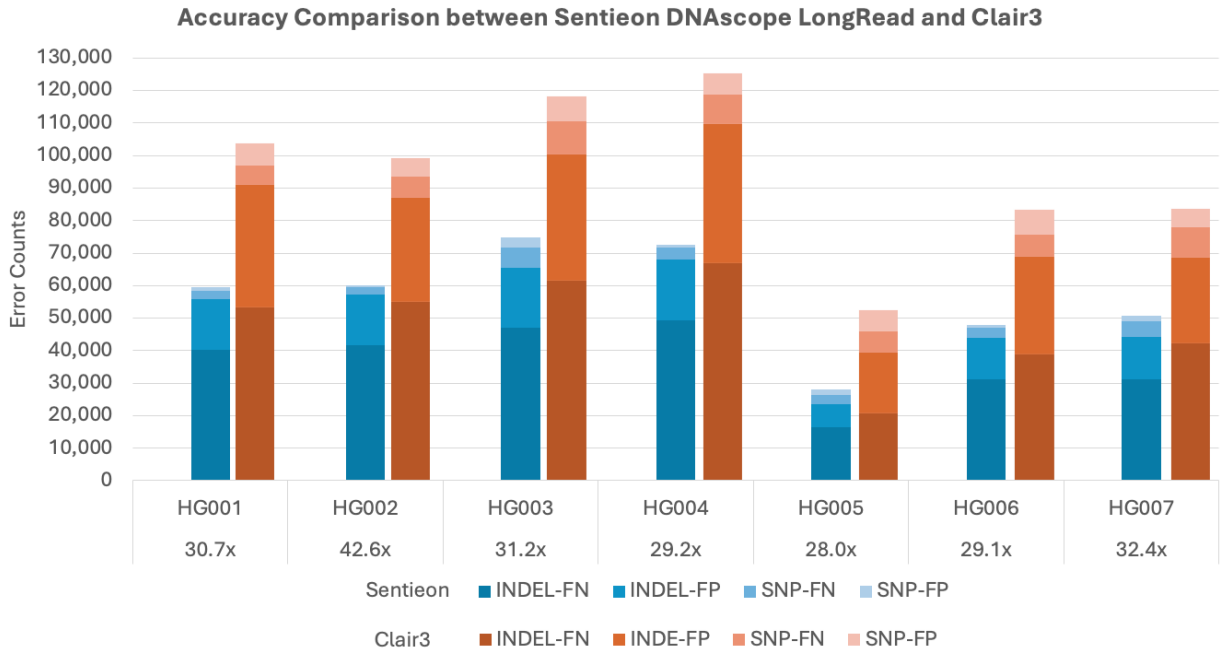

Figure 2. Error counts for DNAscope LongRead across ONT WGS datasets. Total errors fall between roughly 300k and 700k for the HG001–HG007 samples, with Indel making up most errors. Notably, sample HG003 was held out during model training, so its performance reflects true out-of-sample accuracy. In all comparisons, DNAscope LongRead has significant fewer errors than Clair3 in all error categories in all datasets.

Dataset	Depth	Platform	Pipeline	INDEL-FN	INDEL-FP	SNP-FN	SNP-FP	INDEL-F1	SNP-F1
HG001	30.7x	ONT	DNAscope LongRead	40194	15649	2487	1137	0.9423	0.9994
HG002	42.6x	ONT	DNAscope LongRead	41545	15640	2469	559	0.9477	0.9996
HG003	31.2x	ONT	DNAscope LongRead	47173	18423	6199	2919	0.9373	0.9986
HG004	29.2x	ONT	DNAscope LongRead	49322	18798	3612	842	0.9357	0.9993
HG005	28.0x	ONT	DNAscope LongRead	16532	6924	2961	1551	0.9729	0.9993
HG006	29.1x	ONT	DNAscope LongRead	31320	12689	3103	801	0.9501	0.9994
HG007	32.4x	ONT	DNAscope LongRead	31111	13074	4752	1941	0.9502	0.9990

Table 1. Accuracy at a glance — DNAscope LongRead on ONT WGS datasets at other depths.

AWS Instance Type	vCPU	Runtime	Runtime (h)	Core Hours	Peak Memory
c8i.8xlarge	32	36:58.8	0.62	19.72	15.28
c8i.12xlarge	48	29:33.0	0.49	23.64	18.88
c8i.16xlarge	64	21:18.1	0.36	22.72	22.98
c8i.24xlarge	96	16:33.9	0.28	26.50	29.99
c8i.32xlarge	128	14:27.9	0.24	30.86	34.30
c8i.48xlarge	192	13:20.1	0.22	42.67	40.42

Table 2. How fast and how cheap? Efficiency metrics for DNAscope LongRead on ONT WGS data (from CRAM to VCF).

Read More

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

Preprint with older model: https://www.biorxiv.org/content/10.1101/2025.11.14.688541v1