Datasheet

DNAscope Pangenome Analysis with Ultima WGS: Fast, Accurate, and Cost-Efficient Germline Variant Calling

The human reference genome has long served as the foundation for genomic analysis, but its linear, haploid structure cannot fully represent the diversity of human genomes. This limitation leads to alignment artifacts and variant-calling errors, particularly in regions where sample haplotypes diverge from the reference. These challenges are further compounded in sequencing data with context-dependent error profiles.

Pangenome approaches address these limitations by incorporating diverse haplotypes into graph-based references, improving alignment and variant detection in complex regions. However, existing pangenome workflows are often computationally intensive and difficult to scale.

To improve variant calling accuracy and runtime with Ultima data, we developed the Sentieon pangenome pipeline, a fastq-to-VCF workflow optimized specifically for Ultima whole-genome sequencing data. The pipeline delivers high accuracy and efficiency on standard x86 and ARM CPUs, enabling scalable analysis without specialized hardware.

Below, we benchmark Sentieon DNAscope Pangenome using the Sentieon Ultima model (version 1.2) with aligned CRAM files released by Ultima during AGBT 2025.

How It Works

The pipeline leverages pangenome data to construct a personalized reference for each sample, improving alignment in regions where the linear reference is insufficient. It applies an optimized selective realignment strategy designed to handle context-specific sequencing errors, including those associated with homopolymers.

Reads are first aligned to the linear reference, and candidate regions are selectively realigned against sample-specific haplotypes derived from the pangenome. The resulting alignments are then projected back onto GRCh38/hg38 coordinates, ensuring compatibility with standard downstream tools and workflows.

Variant discovery and genotyping are performed using Sentieon DNAscope, enabling robust detection of SNPs and indels even in challenging genomic contexts. The workflow currently supports the HPRC minigraph-cactus pangenome and is ready for expanded future releases.

The Numbers

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

- SNP F1 > 99.8%
- Indel F1 > 95.0%
- Significantly outperform other Ultima compatible analysis pipelines

Indels in long homopolymer regions (>7 bp) account for the majority of errors. Simply excluding the “AllHomopolymers_ge7bp_imperfectge11bp_slop5” regions (~10% of truth variants) reduces total errors from ~80k to <15k (Indel F1 > 99%)—surpassing the error profile of Illumina linear-genome analysis.

The workflow is also highly efficient. With pre-aligned whole-genome input, variant calling completes in approximately 100 min (~110 core-hours), corresponding to only a few dollars in on-demand compute cost.

Combined with the low cost of Ultima 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.

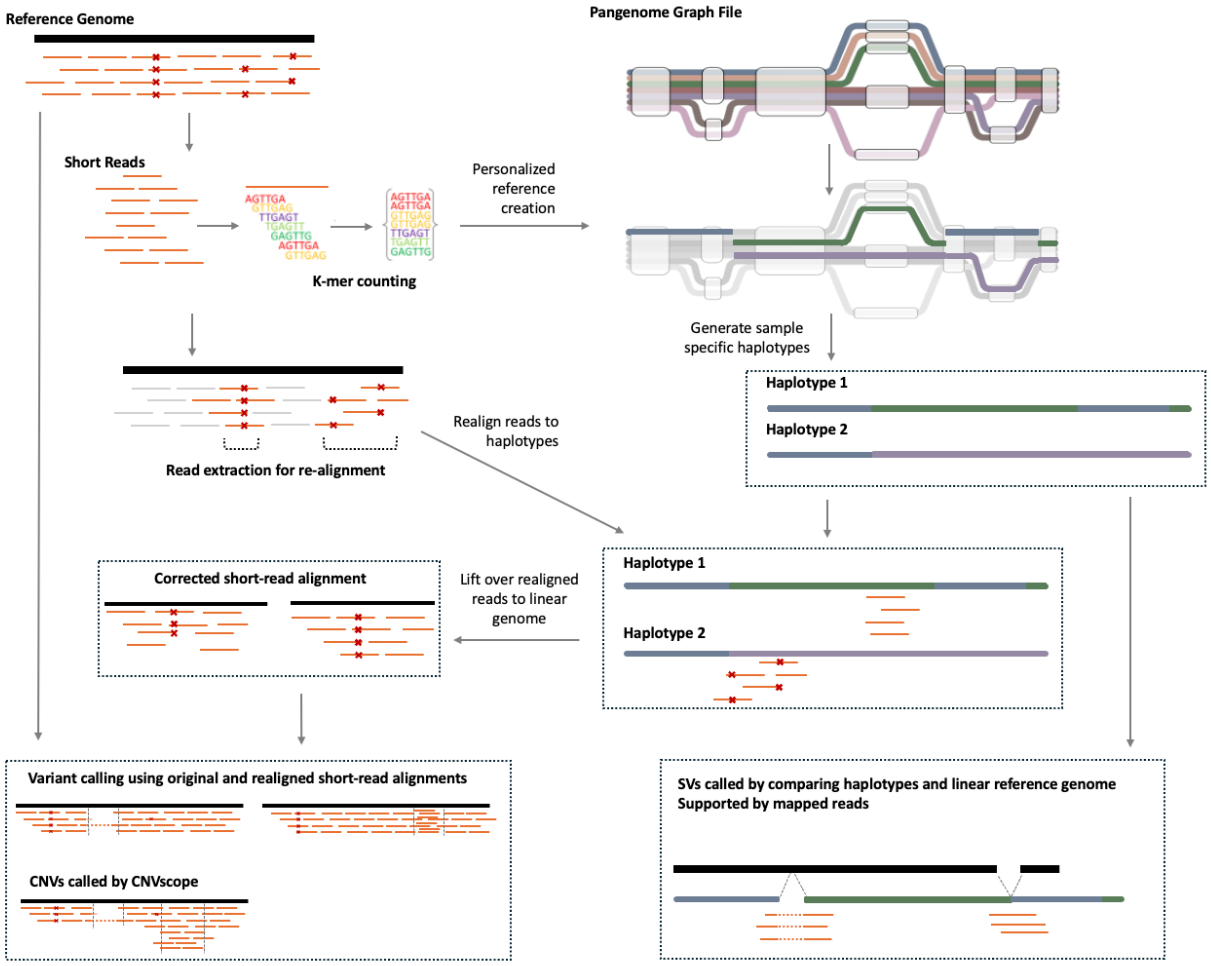

Figure 1. How the DNAscope Pangenome pipeline works. First, k-mers from the input data are counted and matched against the full pangenome graph to identify the closest haplotypes. A targeted subset of reads are extracted and realigned to those best-matching haplotypes from the pangenome. The updated alignments are lifted back to the linear reference, and both sets of alignments are used in small variant calling. For structural variants, the pipeline compares the sample's haplotypes directly against the linear reference and uses the read alignments for additional support and validation.

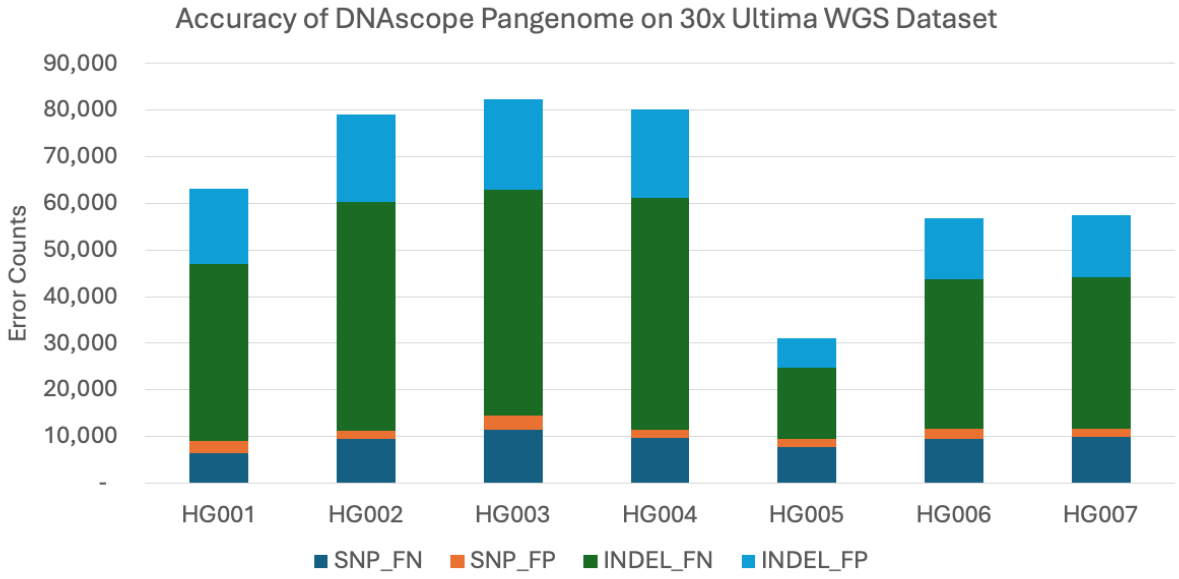

Figure 2. Error counts for DNAscope Pangenome across 30× Ultima WGS datasets. Total errors fall between roughly 30k and 80k for the HG001–HG007 samples, with Indel false-negatives making up most errors. Notably, sample HG003 was held out during model training, so its performance reflects true out-of-sample accuracy.

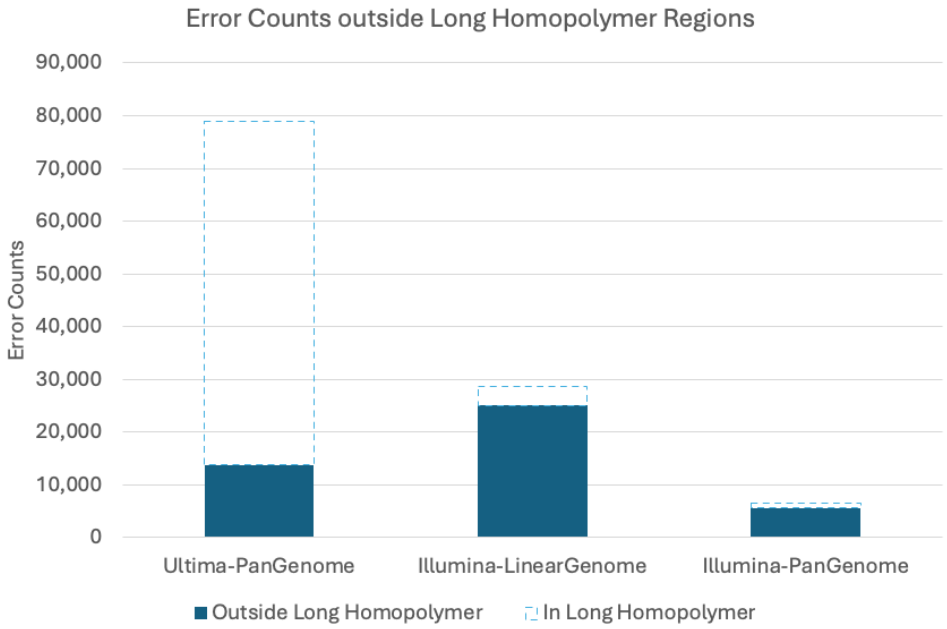

Figure 3. Error counts for the HG002 dataset across three workflows: the Ultima pangenome pipeline (DNAscope Pangenome), the Illumina linear-genome pipeline (DNAscope), and the Illumina pangenome pipeline. Solid color bars represent errors outside long homopolymer regions (AllHomopolymers_ge7bp_imperfectge11bp_slop5), and dashed boxs represent errors within these regions. Excluding long homopolymer regions markedly reduces errors and brings Ultima performance in line with Illumina.

Dataset	Platform	Pipeline	SNP_FN	SNP_FP	INDEL_FN	INDEL_FP	SNP_F1	INDEL_F1
HG001	Ultima	DNAscope Pangenome	6,480	2,474	38,109	16,161	0.9986	0.9439
HG002	Ultima	DNAscope Pangenome	9,406	1,757	49,191	18,658	0.9983	0.9375
HG003	Ultima	DNAscope Pangenome	11,424	3,133	48,409	19,272	0.9978	0.9351
HG004	Ultima	DNAscope Pangenome	9,694	1,763	49,616	19,177	0.9983	0.9348
HG005	Ultima	DNAscope Pangenome	7,691	1,687	15,267	6,416	0.9986	0.9749
HG006	Ultima	DNAscope Pangenome	9,415	2,138	32,201	13,113	0.9982	0.9485
HG007	Ultima	DNAscope Pangenome	9,873	1,878	32,361	13,260	0.9982	0.9484

Table 1. Accuracy at a glance — DNAscope Pangenome on 30× Ultima WGS data.

Dataset	Depth	vCPUs	Runtime	Core Hours
HG001	30x	64	1:43:17	110.17
HG002	30x	64	1:46:04	113.14
HG003	30x	64	1:39:35	106.22
HG004	30x	64	1:39:33	106.19
HG005	30x	64	1:44:02	110.97
HG006	30x	64	1:51:23	118.81
HG007	30x	64	1:46:06	113.17

Table 2. How fast and how cheap? Efficiency metrics for DNAscope Pangenome on 30× Ultima WGS data.

Read More

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

©Sentieon Inc.
160 E Tasman Dr #208, San Jose, CA 95134
www.sentieon.com