Datasheet

Pangenome Analysis Without the Pain: Fast, Accurate, and Affordable Germline Variant Calling with Illumina WGS Data

The human reference genome has been the backbone of genomic analysis for over two decades, serving as a "map" for understanding, referencing and interpreting genetic variation. However, the linear reference genome uses a single haploid representation, which cannot capture the full breadth of structural and sequence diversity across global populations. This leads to alignment artifacts and variant-calling errors across some regions of the genome, especially when a sample's haplotypes diverge from the linear reference.

Pangenome approaches offer a way forward. Projects like the Human Pangenome Reference Consortium (HPRC) incorporate diverse haplotypes into graph-based models, reducing alignment artifacts and variant calling errors and providing a much more complete picture of human variation. The catch is that existing pangenome workflows tend to be resource-hungry, often demanding high-end compute or specialized hardware that puts them out of reach for many labs.

That's the problem we set out to solve with the Sentieon pangenome pipeline, a fastq-to-VCF workflow designed to bring pangenome-informed analysis to standard x86 and ARM hardware without sacrificing speed or accuracy.

How It Works

The pipeline leverages pangenome data to construct a personalized reference for each sample, then applies an optimized selective alignment strategy to re-align reads across some regions of the genome. Pangenome alignments are lifted back to GRCh38/hg38, providing compatibility with the downstream tools and pipelines you already use, and performs variant calling on the linear reference genome using Sentieon DNAscope. The workflow supports the HPRC minigraph-cactus pangenome today, with readiness for the upcoming 400-sample HPRC release.

The Numbers

On GIAB v4.2.1 benchmarks, the pipeline delivers strong results: total errors drop to around 6,000, with F1 scores above 99.89% for SNPs, 99.78% for indels, and 94.77% for Structural Variants (at long reads accuracy level!). It does this fast; with pre-aligned whole-genome input,

variant calling finishes in roughly 60 minutes (about 80 core-hours), for around $4 in on-demand compute with general-purpose hardware. This price point that makes pangenome-informed analysis realistic not just for one-off studies, but at true population scale.

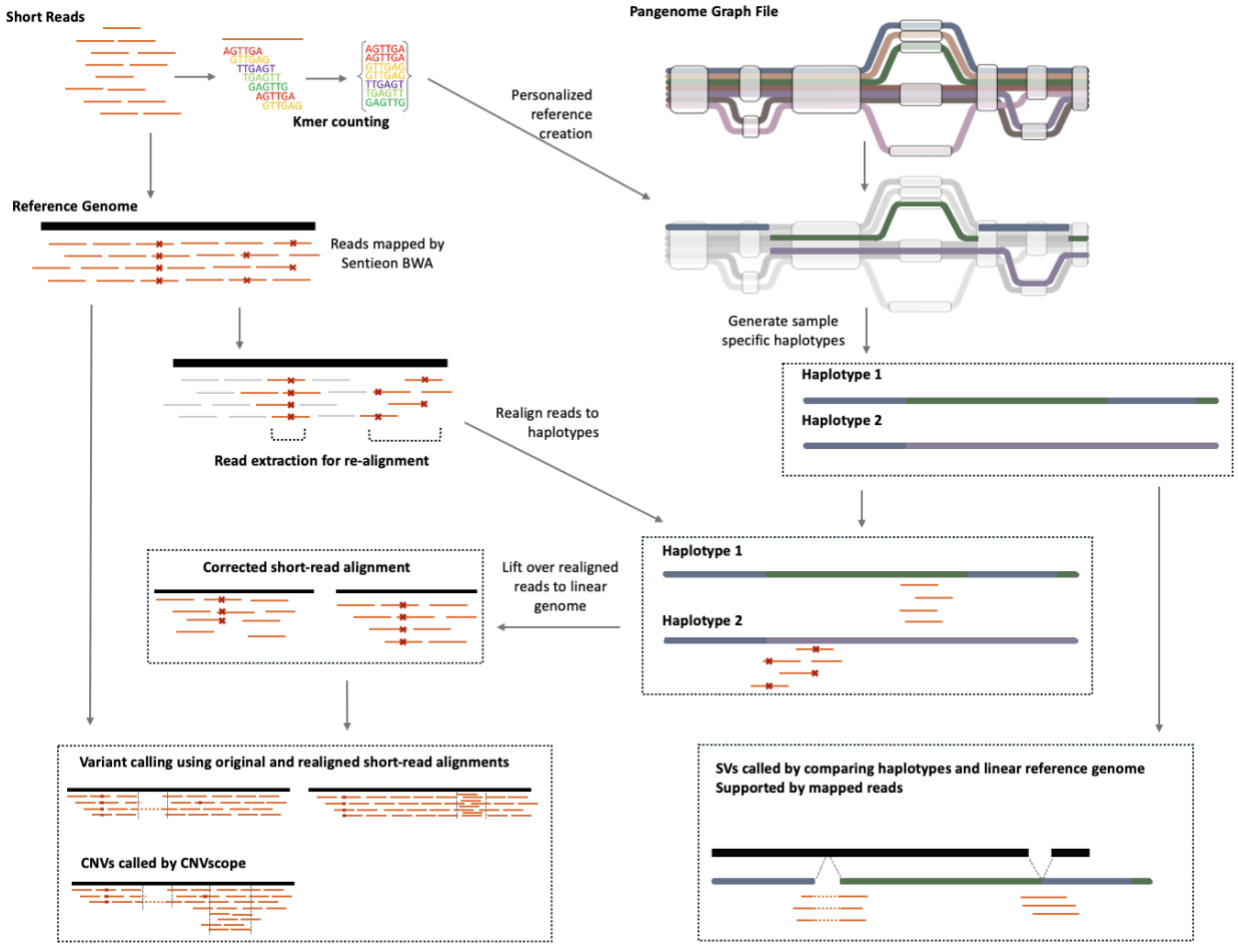

Figure 1. How the DNAscope Pangenome pipeline works. First, k-mers from the input reads are counted and matched against the full pangenome graph to identify the closest haplotypes. Reads are then aligned to the GRCh38 linear reference, and a targeted subset is pulled out and realigned to those best-matching haplotypes. 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.

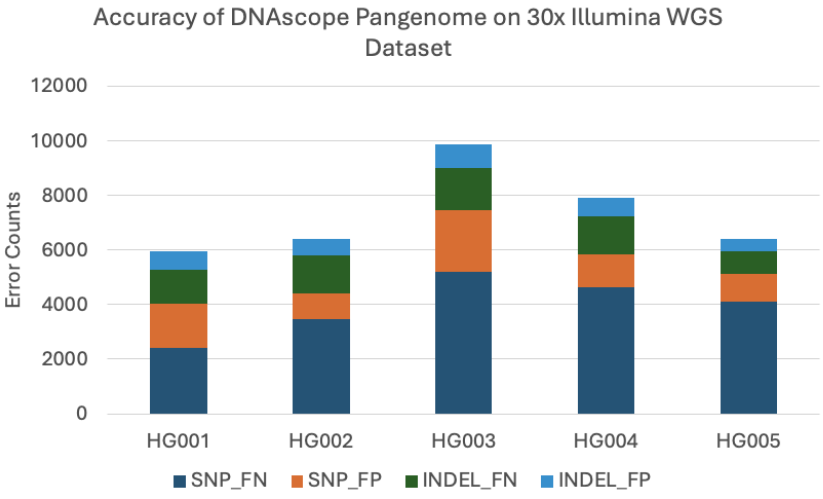

Figure 2. Error counts for DNAscope Pangenome across 30× Illumina WGS datasets. Total errors fall between roughly 6k and 10k for the HG001–HG005 samples, with SNP false-negatives making up most errors. Notably, sample HG003 was held out during model training, so its performance reflects true out-of-sample accuracy.

Dataset	Platform	Pipeline	SNP_FN	SNP_FP	INDEL_FN	INDEL_FP	SNP_F1	INDEL_F1
HG001	ILMN	DNAscope Pangenome	2416	1605	1241	681	0.9994	0.9981
HG002	ILMN	DNAscope Pangenome	3457	945	1412	613	0.9994	0.9982
HG003	ILMN	DNAscope Pangenome	5222	2259	1545	853	0.9989	0.9978
HG004	ILMN	DNAscope Pangenome	4647	1206	1377	695	0.9991	0.9981
HG005	ILMN	DNAscope Pangenome	4115	1004	822	470	0.9992	0.9985

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

AWS Instance	vCPUs	Run time (min)	Peak Memory (GB)	Core Hours	Compute Cost (On Demand)
c8i.24xlarge	96	59.0	61.5	94.3	$ 4.42
c8i.16xlarge	64	72.2	61.8	77.0	$ 3.61
c8i.12xlarge	48	87.7	60.4	70.1	$ 3.29

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