Salus 赛陆医疗
innovation in biomed

Sentieon

High Accuracy and Affordable Human Genome and Exome sequencing and analysis for individual and populations

Introduction

Whole genome sequencing (WGS) and whole exome sequencing (WES) have become a pivotal tool in various applications, including the study of genetic diseases, heritable risk assessment, and the reconstruction of human population history[1]. High-throughput paired-end (PE) 150bp+ sequencing represents a major milestone for WGS, generating sequence reads from both ends of longer genomic fragments. This approach effectively bridges most genomic repeats with relatively short reads, enhancing sequencing accuracy and genome assembly.

The Salus sequencing platform introduces an affordable and reliable solution for next-generation sequencing (NGS) data generation. By employing sequencing-by-synthesis (SBS) principles, Salus integrates several proprietary innovations[2]:

1. Wide-Field Imaging: Expands the field of view by over 100% compared to conventional lenses, reducing imaging time by 50%.

2. 3D Chips: High-density 3D covalent bond-modified primers significantly enhance throughput, reduce costs, and improve both robustness and adaptability.

3. High-Efficiency Sequencing Enzymes: Proprietary enzymes extend read length from PE150 to PE300 or SE400.

4. Ultra-bright Fluorescent Dyes: Optimized usage of synthesized dyes minimizes reagent costs.

5. Rapid Chemistry Reagents: Ultra-fast sequencing mode reduces SE50+8+8 sequencing time to just 4.8 hours.

The Salus Pro instrument, a mid-throughput sequencer in the Salus product line, Salus Pro is Chinese NMPA Class III Approved and CE-IVDR certified and generates up to 300 GB of data in 45 hours, equivalent to three WGS or 20 WES datasets[3].

WGS projects routinely produce terabytes of data, posing challenges in both data infrastructure and method development for downstream analysis. Fortunately, a robust ecosystem of cloud-based infrastructure and evolving bioinformatics tools now enables fast and reliable data analysis. These tools are critical for challenging clinical applications, including rapid small-variant (SNPs and InDels) calling with high sensitivity and precision.

Sentieon DNAscope is such an advanced solution for accurate and efficient germline small-variant calling. It combines established methods from haplotype-based variant callers with machine learning to achieve improved accuracy. As a successor to GATK, DNAscope retains its logical architecture while enhancing active region detection and local assembly for better sensitivity, particularly in high-complexity genomic regions. Moreover, platform-specific machine learning models further enhance variant calling accuracy[4].

Study Overview

This study demonstrates the integration of WGS and WES data generated on the Salus sequencing platform with Sentieon DNAscope pipeline. To validate the performance of this combination, multiple replicates of the well-characterized human genomes HG001–HG007 were sequenced. These genomes were selected due to the availability of high-quality variant truth sets provided by NIST, which facilitated accurate measurement of SNP and InDel sensitivity and precision. Seven samples (HG001–HG007) were sequenced at approximately 40X WGS and 250X whole-exome sequencing (WES) coverage using both the Salus Pro sequencer and the Illumina NovaSeq platform for direct comparison.

The DNAscope model was trained using the Sentieon software package (202308.03). Reference datasets from HG001, HG002, HG003, HG005, and HG007 generated on the Salus Pro platform were used for training. Data were randomly split, with 20% reserved for validation, while chromosome 20 was held out for testing. All datasets were mapped to the hg38 reference genome using Sentieon BWA-Turbo, followed by quality checks of the generated BAM files.

For model training, the aligned datasets were downsampled to create multiple WGS training sets with depths ranging from 15X to 40X and WES sets from 50X to 250X to enhance depth tolerance. A gradient boosting decision tree (GBDT) was built using candidate variants from DNAscope's highly sensitive mode.

The HG004 and HG006 datasets were reserved for validation and downsampled to typical WGS (30X) and WES (120X) depths. Variants were compared to the GIAB v4.2.1 benchmark VCF using hap.py v0.3.10 with RTGtools vcfeval v3.9.2 for accuracy calculations.

Pipeline Implementation

The DNAscope pipeline was executed via the Sentieon CLI interface as described in the documentation. The command line used was:

```
sentieon-cli dnascope [-h] \
  -r REFERENCE \
  --r1-fastq R1_FASTQ ... \
  --r2-fastq R2_FASTQ ... \
  --readgroups READGROUPS ... \
  sample.vcf.gz
```

Results

BAM Quality Control

After aligning the WGS datasets, quality metrics were evaluated. The Salus platform demonstrated high base quality scores, with over 95% of reads having a score >30. GC coverage bias was minimal, showing consistent sequencing efficiency across genome regions with GC content between 20–80%. Notably, the Salus platform exhibited a significant advantage in duplication rates, with <5% duplication compared to ~30% observed with Illumina, despite identical library preparation protocols (Figure 1c).

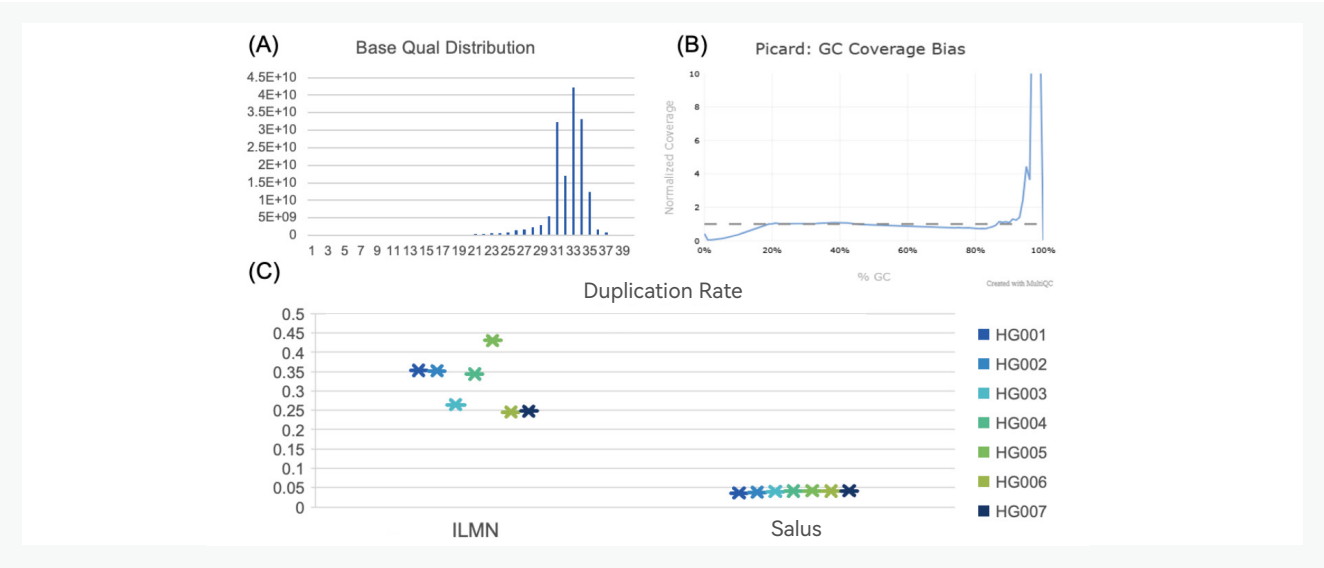

Figure 1. (A) Quality Score Distribution and (B) GC Bias Plot of HG004 WGS dataset. Most reads returned 35+ quality scores Sequencing and coverage is even across most GC windows. (C) Salus showed significantly lower duplication rate, comparing to Illumina (ILMN) dataset whose libraries were made identically.

Single-sample WGS and WES Accuracy

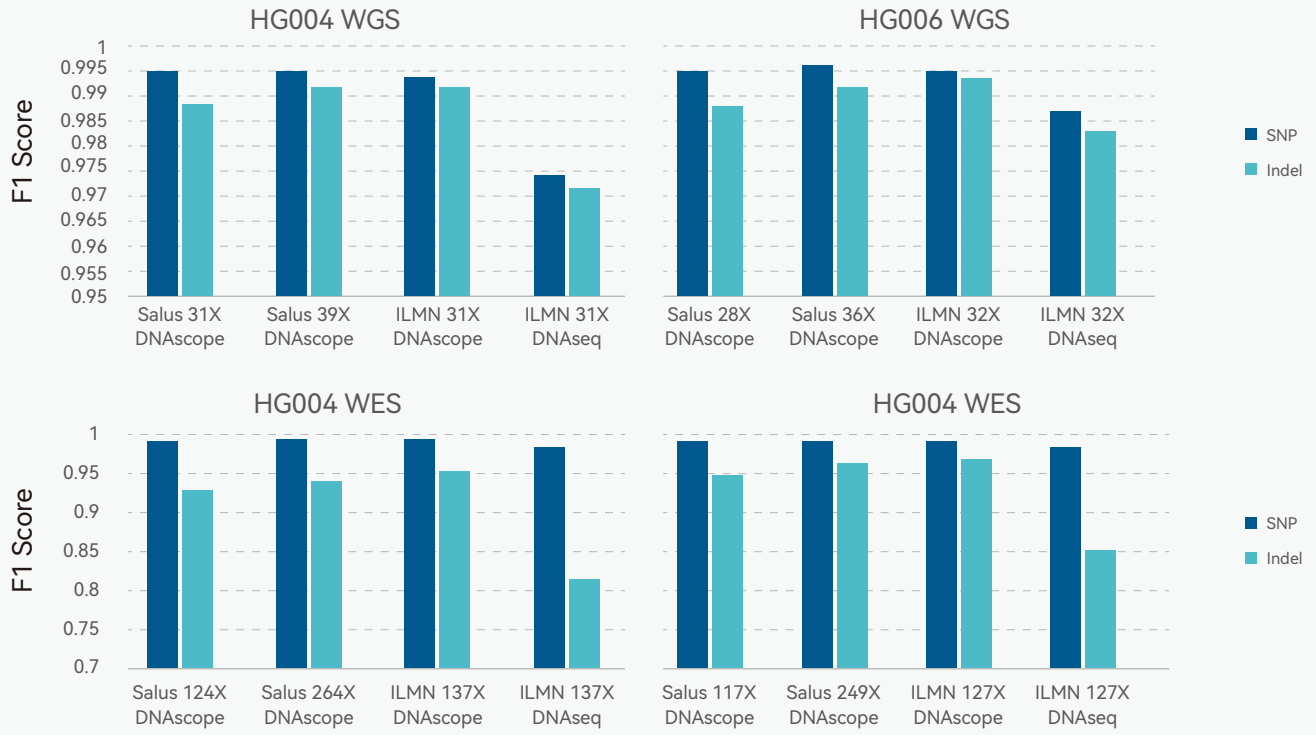

Figure 2. Illustration of the performance of WGS and WES variant calls using models trained on Salus data. Both SNP and InDel F1 scores were consistently high and comparable to ILMN call sets.

Although sequencing depths for Salus and Illumina datasets were not perfectly matched, the Salus WGS and WES call sets achieved similar accuracy (Figure 2, Table 1) to Illumina, with SNP F1 scores of ~0.995 and InDel F1 scores of ~0.990 for WGS. For WES, the SNP F1 score was ~0.992, and the InDel score was ~0.930. Increased sequencing depth primarily improved InDel accuracy (Table 2).

Sample	Sequencer	Depth	Analysis Pipeline	Indel			SNP		
				False Negative	False Positive	F1 Score	False Negative	False Positive	F1 Score
HG004	Salus	31X	DNAscope Salus WGS Model v0.9	8,271	3,904	0.988	26,469	6,893	0.995
HG004	Salus	39X	DNAscope Salus WGS Model v0.9	6,091	2,573	0.992	26,229	4,513	0.995
HG004	ILMN	31X	DNAscope Illumina WGS Model v2.2	4,958	2,817	0.992	25,668	11,857	0.994
HG004	ILMN	31X	DNAseq	8,124	20,380	0.972	32,984	141,775	0.974
HG006	Salus	28X	DNAscope Salus WGS Model v0.9	6,614	3,310	0.988	24,321	8,702	0.995
HG006	Salus	36X	DNAscope Salus WGS Model v0.9	4,699	2,065	0.992	23,233	5,659	0.996
HG006	ILMN	32X	DNAscope Illumina WGS Model v2.2	3,465	1,749	0.994	22,901	7,991	0.995
HG006	ILMN	32X	DNAseq	5,533	9,113	0.983	30,044	55,527	0.987

Table 1. Accuracy WGS benchmark using selected validation call sets processed by Sentieon DNAscope and DNAseq.

For reference, Illumina datasets processed using Sentieon DNAseq[5] (a GATK reimplementation) served as the gold standard. The Salus platform combined with DNAscope achieved accuracy significantly exceeding this baseline.

Sample	Sequencer	Depth	Analysis Pipeline	Indel			SNP		
				False Negative	False Positive	F1 Score	False Negative	False Positive	F1 Score
HG004	Salus	124X	DNAscope Salus WES Model v0.9	161	51	0.930	555	144	0.992
HG004	Salus	264X	DNAscope Salus WES Model v0.9	136	43	0.941	529	128	0.993
HG004	ILMN	137X	DNAscope Illumina WES Model v2.2	113	30	0.953	530	103	0.993
HG004	ILMN	137X	DNAseq	181	456	0.814	564	801	0.985
HG006	Salus	117X	DNAscope Salus WES Model v0.9	103	30	0.950	583	212	0.992
HG006	Salus	249X	DNAscope Salus WES Model v0.9	82	15	0.964	512	151	0.993
HG006	ILMN	127X	DNAscope Illumina WES Model v2.2	59	19	0.971	563	134	0.993
HG006	ILMN	127X	DNAseq	99	347	0.852	641	860	0.984

Table 2. Accuracy WES benchmark using selected validation call sets processed by Sentieon DNAscope and DNAseq.

Joint Genotyping

To evaluate the potential of DNAscope in mitigating differences between sequencing platforms, we conducted a joint calling benchmark using Salus and Illumina WGS datasets. This study aimed to determine whether DNAscope could harmonize sequencing discrepancies and enable Salus to serve as a viable alternative to Illumina for cohort studies, even during ongoing data collection. The lower sequencing costs and higher accessibility of Salus mid- and high-throughput sequencers would allow for the inclusion of more samples.

We applied the DNAscope pipeline to 40 WGS datasets from Han Chinese samples: 18 sequenced on the Salus platform and 22 sequenced on the Illumina platform, obtained from the 1000 Genomes Project[6]. Data from HG001-4 from both sequencing platforms were included as “not Han Chinese” reference points.

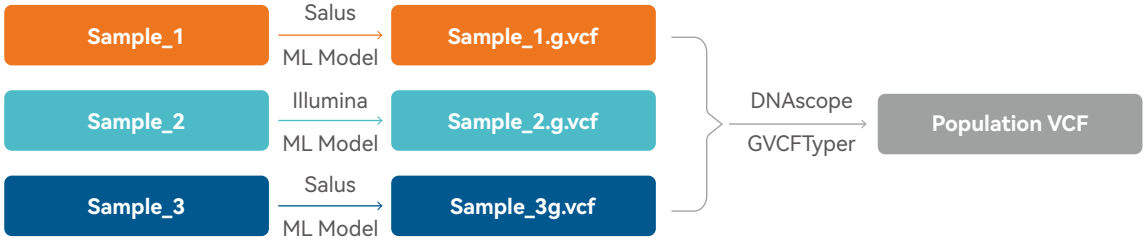

Figure 3. DNAscope joint calling pipeline takes in FASTQ or BAM/CRAM files as input and using Salus or Illumina specific machine learning model to generate individual gvcf files, as well as conducting final joint genotyping.

The DNAscope joint calling pipeline corrects sequencing platform-specific errors using pre-trained machine learning models tailored to each sequencer. It produces gVCF files, which are subsequently processed for joint genotyping to generate a population VCF file containing SNPs and InDels identified across all samples.

We performed principal component analysis (PCA) to visualize sample clustering based on identified variants (Figure 4). The PCA plot revealed that datasets from the Salus and Illumina platforms were intermixed without distinct separation, whereas ethnographic differences predominantly shaped the clustering into three groups. These results demonstrate that the Salus platform, when processed using DNAscope, delivers sequencing performance comparable to that of the Illumina platform in a cohort joint calling study.

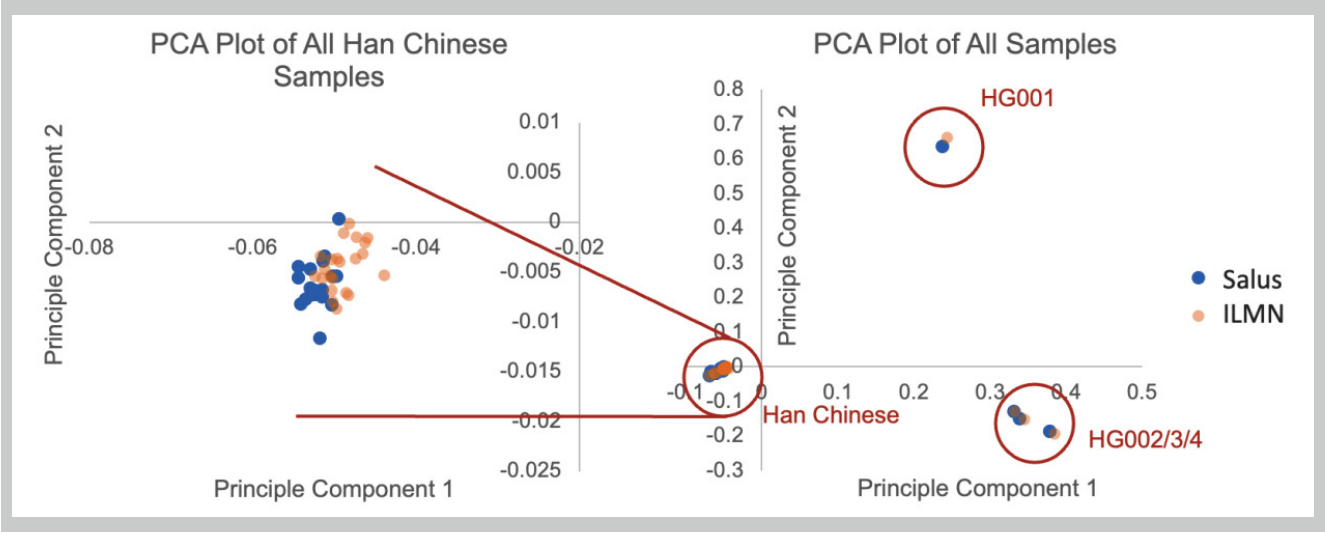

Figure 4. PCA analysis of WGS samples from Han Chinese and other races, sequenced by Salus and ILMN platforms.

Summary

The results of this study demonstrated that the pre-trained DNAscope models for WGS and WES on the Salus platform achieved high variant-calling accuracy. The integration of the Salus sequencing platform with the Sentieon analysis pipeline enabled reliable, high-quality variant calling for both WGS and WES applications. For the first time, the joint calling benchmark showed that datasets from different sequencing platforms could be combined without introducing significant platform-specific biases in the joint genotyping results when processed using DNAscope. These high-quality variant calls provide a robust foundation for various downstream applications.

References

1.Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17(5)* 333–351. https://doi.org/10.1038/nrg.2016.49

2.Self-Developed Technology. Retrieved from https://nj.gzwhir.com/szslyl202311221360/SelfdevelopedTechnology/index.aspx

3.Sequencing Platform. Retrieved from https://nj.gzwhir.com/szslyl202311221360/SequencingPlatform/list.aspx?lcid=110

4.Freed, D., Pan, R., Chen, H., Li, Z., Hu, J., & Aldana, R. (2022). DNAscope: High accuracy small variant calling using machine learning. *bioRxiv* . https://doi.org/10.1101/2022.05.20.492556

5.Freed, D., Aldana, R., Weber, J. A., Edwards, J. S. (2017). The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv* , 115717. https://doi.org/10.1101/115717

6.Fairley, S., Lowy-Gallego, E., Perry, E., & Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research* , 48(D1), D941–D947. https://doi.org/10.1093/nar/gkz836

Salus 赛陆医疗
innovation in biomed

0755-2374 5832

info@salus-bio.com

http://salus-bio.com

Floors 7-11, Building 3A & Floors 20, Building 1, Hengtaiyu Research Park, Shenzhen, Guangdong, P.R. China

2/F, BLK 5, Lane 88, Minbei Road, Minhang District, Shanghai, P.R.China

Sentieon

(650) 254-6360

info@sentieon.com

https://www.sentieon.com

160 E Tasman Dr STE 208 San Jose, CA 95134-1619 United States