Salus 赛陆医疗
innovation in biomed

Sentieon

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

高精度且可负担的全基因组和外显子组测序与个体和群体分析服务

白皮书

Introduction
介绍

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.

全基因组测序 (WGS) 和全外显子组测序 (WES) 已成为多种应用中的关键工具，包括遗传疾病研究、遗传风险评估以及人类群体历史重建等[1]。高通量的双端 (PE) 150bp+ 测序是 WGS 的一个重要里程碑，它能够从较长基因组片段的两端生成序列读取。这种方法能够有效地利用相对较短的测序读长来跨越基因组中的大部分重复序列，提高了测序精度和基因组组装质量。

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]:

Salus 测序平台为下一代测序 (NGS) 提供了具备价格优势且可靠的解决方案。Salus 采用了结合多项自研创新的合成测序 (SBS) 原理[2]：

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

大视野成像：与传统镜头相比，视野范围扩大超过100%，成像时间缩短50%。

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

3D芯片：高密度的 3D 共价键修饰引物显著提高了通量，降低了成本，并且增强了稳定性和适应性。

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

高效测序酶：自主研发的高效测序酶可将读长从 PE150 扩展到 PE300 或 SE400。

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

微量荧光染料：对合成染料的优化使用最大程度地降低了试剂成本。

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

快速化学试剂：超快速测序模式将 SE50+8+8 测序时间缩短至仅 4.8 小时。

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

Salus Pro 是赛陆医疗自研测序仪产品线中的一款中通量测序仪，已获得 CE-IVDR 认证，可在 45 小时内生成高达 300GB 的数据，相当于 3 个全基因组测序 (WGS) 或 20 个全外显子组测序 (WES) 的数据[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.

全基因组测序 (WGS) 项目通常会产生大量的数据（TB级别），这在数据存储基础设施和下游分析方法开发方面都带来了挑战。不过通过现今强大的云计算生态系统和不断发展的生物信息学工具使得数据能够快速且可靠地进行分析。这些工具对于具有挑战性的临床应用至关重要，包括快速且高灵敏度、高精确度地检测小型变异（单核苷酸多态性 (SNPs) 和插入/缺失 (InDels)）。

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].

Sentieon DNAscope 是一款先进的解决方案，专为精准且高效地识别生殖细胞系中的小型变异而设计。它巧妙地将基于单倍型的变异检测技术与机器学习算法融合在一起，从而显著提升了检测精度。作为 GATK 的升级版，DNAscope 在继承其逻辑架构的基础上，优化了活动区域的识别以及局部组装流程，尤其在处理基因组中复杂度较高的区域时，灵敏度大幅提升。而且该工具还借助针对特定测序平台定制的机器学习模型，进一步增强了变异检测的准确性[4]。

Study Overview
研究概览

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.

本研究展示了在 Salus 测序平台上生成的全基因组测序 (WGS) 数据和全外显子组测序 (WES) 数据与 Sentieon DNAscope 流程的整合。为了验证这种组合的性能，对经过充分表征的人类基因组 HG001 至 HG007 共 7 个细胞系 DNA 进行了多次重复测序。选择这些基因组是因为美国国家标准与技术研究院 (NIST) 提供了高质量的变异真值集，这有助于准确测量单核苷酸多态性 (SNP) 和插入缺失 (InDel) 的灵敏度和精确度。使用 Salus Pro 平台和 Illumina NovaSeq 平台对七个样本 (HG001 至 HG007) 分别进行了大约 40X 的 WGS 和 250X 的 WES 测序覆盖，以便进行直接比较。

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.

我们使用 Sentieon 软件包（202308.03版）来训练 DNAscope 模型。训练时所用的参考数据集涵盖了在 Salus Pro 平台生成的 HG001、HG002、HG003、HG005 和 HG007 样本。这些数据经过随机分配，其中20%用作验证集，而第20号染色体的数据则专门留作测试用途。所有数据集均通过 Sentieon BWA-Turbo 映射至 hg38 参考基因组，并对由此产生的 BAM 文件进行了质量控制。

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.

在模型训练阶段，通过对已比对的数据集进行了降采样处理，生成了多个 WGS 训练集，其测序深度从 15X 到 40X 不等，以及 WES 训练集，测序深度从 50X 到 250X 不等，以此来提升模型对不同测序深度的适应性。然后依据 DNAscope 高灵敏度模式下识别出的候选变异，搭建了一个梯度提升决策树（GBDT）。

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.

数据集 HG004 和 HG006 被保留用于验证，并被降采样至常用的 WGS 30X 测序和 WES 120X 测序。变异数据与 GIAB v4.2.1 基准 VCF 文件进行了比较，使用 hap.py v0.3.10 和 RTGtools vcfeval v3.9.2 来计算准确性。

Pipeline Implementation

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

流程实现

我们通过 Sentieon 命令行界面（CLI）执行 DNAscope 流程，具体操作如文档中所述。使用的命令行如下：

```
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).

完成 WGS 数据集的比对后，我们细致地评估了它们的质量指标。Salus Pro 的碱基质量评分表现优异，其中超过 95% 的读取序列得分超过 30 分。在 GC 含量介于 20% 至 80% 的基因组区域，Salus Pro 的 GC 覆盖偏差微乎其微，确保了测序效率的稳定性。特别值得一提的是，Salus Pro 在重复率控制上表现出色，重复率低于 5%，而采用相同文库构建流程的 Illumina 平台，其重复率则约为 30%（详见图1c）。

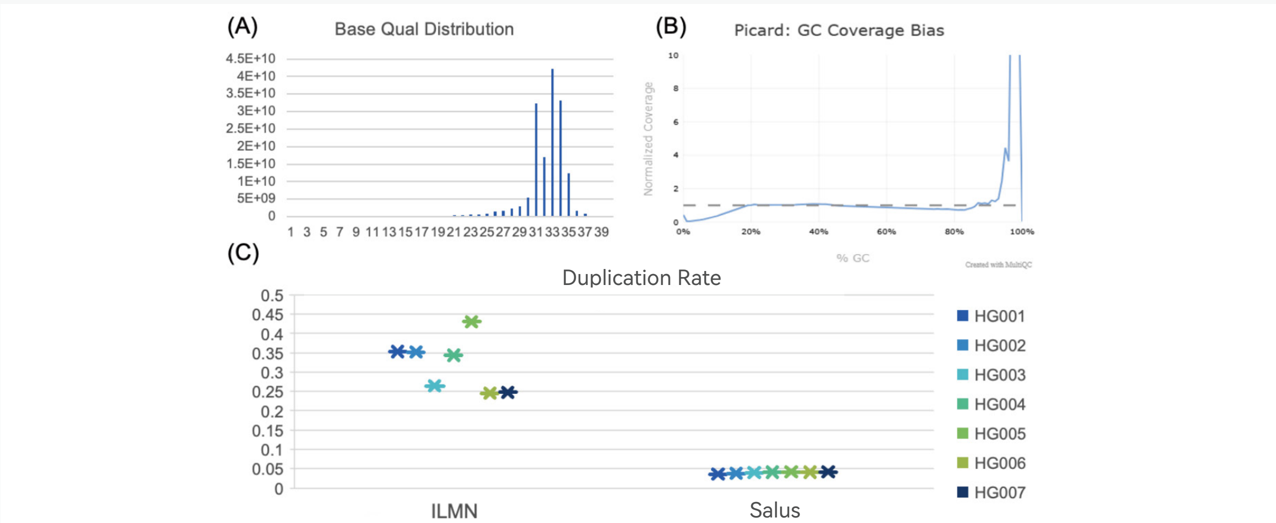

Figure 1. (A) Quality Score Distribution and (B) GC Bias Plot of HG004 WGS dataset. Most reads returned 35+ quality scores Sequencing and coverage is even across most GC windows. (C) Salus showed significantly lower duplication rate, comparing to Illumina (ILMN) dataset whose libraries were made identically.

图1. (A) HG004 全基因组测序 (WGS) 数据集的质量评分分布，以及 (B) GC 偏差图。大多数读取序列的质量评分达到 35 分以上，测序和覆盖在大多数 GC 窗口中分布均匀。(C) 与采用相同文库构建方法的 Illumina (ILMN) 数据集相比，Salus 平台显示出显著更低的重复率。

Single-sample WGS and WES Accuracy
单样本 WGS 和 WES 准确性评测结果

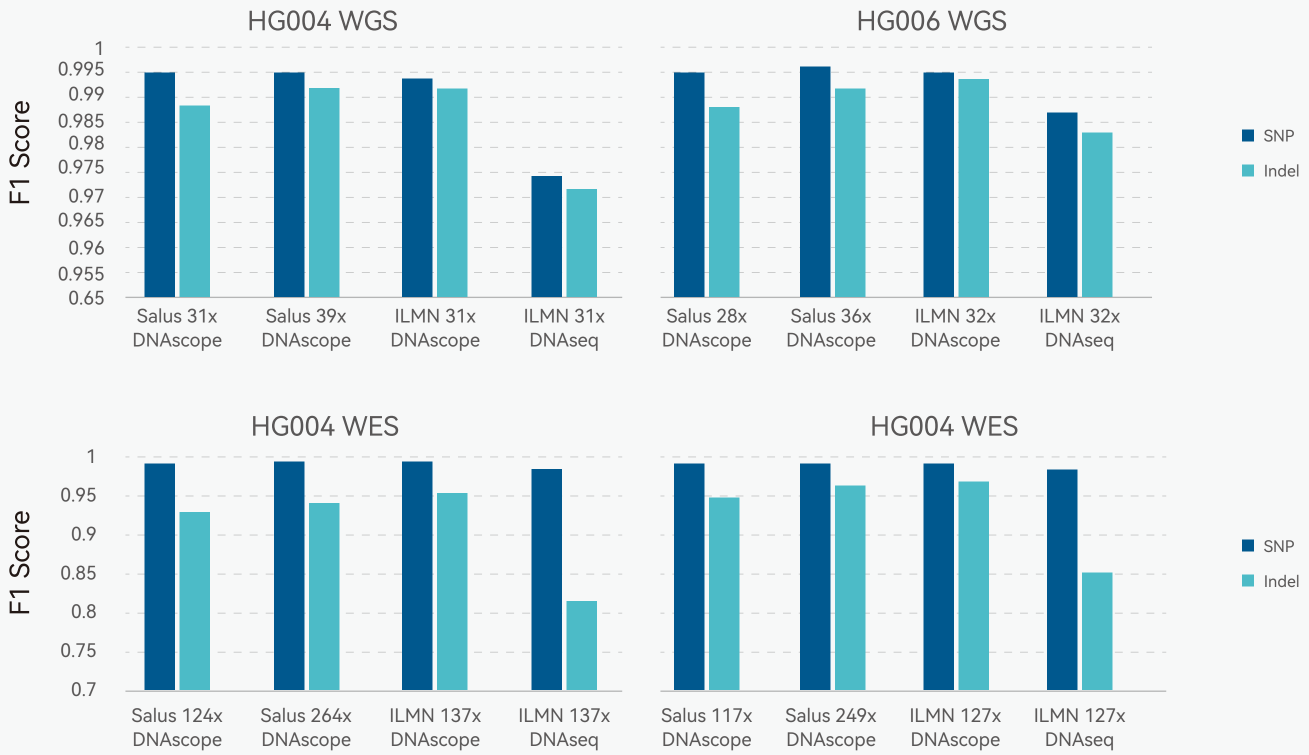

Figure 2. Illustration of the performance of WGS and WES variant calls using models trained on Salus data. Both SNP and InDel F1 scores were consistently high and comparable to ILMN call sets.

图2. 使用在 Salus Pro 产出的数据上训练的模型进行 WGS 和 WES 变异检测的性能示意图。无论是单核苷酸多态性（SNP）还是插入缺失（InDel）的 F1 评分都持续保持在高水平，并且与 ILMN 平台的变异检测结果相当。

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).

尽管两个平台数据集的测序深度并不完全匹配，但 Salus Pro 的 WGS 和 WES 变异检测集在准确性上（见图2、表1）与 Illumina 平台相当。对于 WGS，SNP 的 F1 评分约为 0.995，InDel 的 F1 评分约为 0.990。对于 WES，SNP 的 F1 评分约为 0.992，而 InDel 的评分约为 0.930。增加测序深度主要提升了 InDel 的准确性（见表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	6,898	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	8,127	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.

表1. 经 Sentieon DNAscope 和 DNAseq 处理的选定验证变异检测集的 WGS 准确性基准测试。

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.

我们使用 Sentieon DNAseq[5]（GATK 的另一个版本）处理的 Illumina 数据集作为标准参考。Salus Pro 与 DNAscope 的结合在准确性上显著超过了这一标准。

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.

表2. 经 Sentieon DNAscope 和 DNAseq 处理的选定验证变异检测集的 WES 准确性基准测试。

Joint Genotyping
联合基因分型

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.

为了评估 DNAscope 在减少不同测序平台间差异方面的潜力，我们使用 Salus 平台和 Illumina 平台的 WGS 数据集进行了一次联合变异检测基准测试。该研究的目标是确定 DNAscope 是否能够协调测序过程中的差异，从而使 Salus Pro 成为 Illumina 在队列研究中的一个切实可行的替代选择，即便是在数据收集的进行时也不例外。鉴于 Salus 中通量和高通量测序仪所具有的较低测序成本和更高的可及性，这将使得更多样本的纳入成为可能。

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.

通过将 DNAscope 流程应用于 40 个来自汉族中国人的 WGS 数据集：其中 18 个样本在 Salus Pro 进行测序，其余 22 个样本在 Illumina 平台上测序，这些数据均来源于千人基因组计划[6]。来自两个测序平台的 HG001 至 HG004 的数据被纳入作为非汉族中国人的参照。

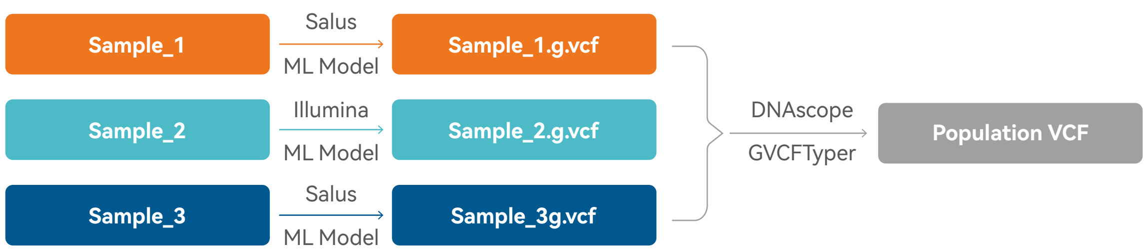

Figure 3. DNAscope joint calling pipeline takes in FASTQ or BAM/CRAM files as input and using Salus or Illumina specific machine learning model to generate individual gvcf files, as well as conducting final joint genotyping.

图3. DNAscope 联合变异检测流程以 FASTQ 或 BAM/CRAM 文件作为输入，然后利用针对 Salus 或 Illumina 平台的特定机器学习模型生成个体的 gVCF 文件，并进行最终的联合基因分型。

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.

DNAscope 联合变异检测流程利用针对每种测序仪定制的预训练机器学习模型来校正该测序平台的错误。该流程生成 gVCF 文件，这些文件随后被处理以进行联合基因分型，最终生成一个群体 VCF 文件，其中包含了在所有样本中识别出的 SNPs 和 InDels。

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.

我们使用主成分分析（PCA）方法来观察那些基于已识别变异的样本聚类情况（见图4）。PCA 图显示，来自 Salus 和 Illumina 平台的数据集相互混合，没有明显的分离，而种族差异主要决定了数据聚类为三个群体。这些结果表明，在队列联合变异检测研究中，使用 DNAscope 处理的 Salus 平台，其测序性能与 Illumina 平台相当。

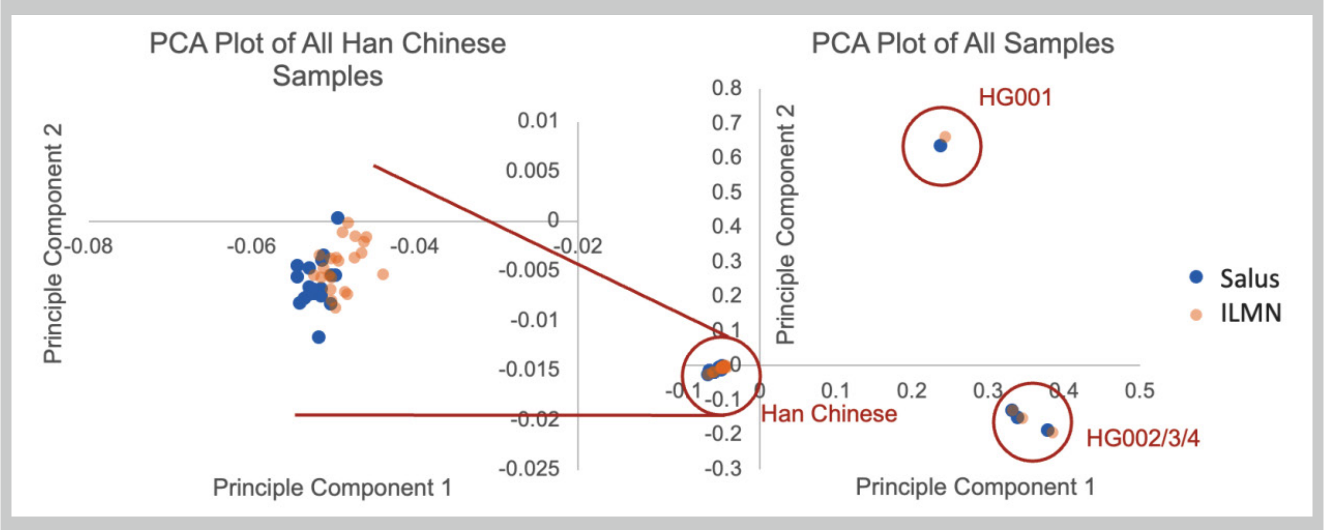

Figure 4. PCA analysis of WGS samples from Han Chinese and other races, sequenced by Salus and ILMN platforms.

图4. 通过 Salus 和 ILMN 平台对中国汉族以及其他种族进行 WGS 样本测序和 PCA 分析的结果。

Summary
总结

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.

本项研究的成果突显了 Salus 平台上训练的 DNAscope 模型在 WGS 和 WES 方面实现了高精度的变异检测。Salus 测序平台与 Sentieon 分析流程的协同整合，确保了 WGS 和 WES 应用在可靠性和数据质量方面的卓越表现。此外，本次联合变异检测基准测试首度证实，经由 DNAscope 处理，不同测序平台产出的数据集得以无缝整合，且在联合基因分型环节未见明显的平台特异性偏差。这些高品质的变异检测成果，为后续一系列应用适配奠定了坚实可靠的基石。

References
参考文献

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

深圳市光明区凤凰街道恒泰裕大厦 1 栋 2001
Floors 7-11, Building 3A & Floors 20, Building 1, Hengtaiyu Research Park, Shenzhen, Guangdong, P.R. China

上海市闵行区闵北路 88 弄 5 号楼 2 楼
2/F, BLK 5, Lane 88, Minbei Road, Minhang District, Shanghai, P.R.China