APPLICATION NOTE

Achieving High-Quality Whole Human Exome Sequencing with Sentieon DNAScope and Genesense StarSeq100 AI Sequencer

Highlights

- The integration of StarSeq100 AI Sequencer sequencing and DNAscope analysis delivers exceptionally accurate and streamlined whole exome sequencing outcomes.
- This workflow can be efficiently performed on the StarSeq100 AI Sequencer in approximately 30 minutes per human exome.

Introduction

Whole exome sequencing (WES) and whole genome sequencing (WGS) have become indispensable tools across various domains, ranging from investigating genetic diseases to assessing heritable risk and delving into human ancestry information. With whole exome sequencing projects routinely yielding terabytes of data, the field necessitates accurate and efficient analytical tools. However, translating the massive data generated by sequencers into actionable insights presents formidable challenges in both infrastructure and methodological development. Achieving high accuracy in variant calling is highly desirable, given its crucial role in downstream annotation and reporting. Hence, focusing on variant calling accuracy is just as crucial as optimizing pipeline efficiency.

In this application note, we showcase the precise analysis of whole exome sequencing data generated on the StarSeq100 AI Sequencer using the Sentieon™ secondary analysis software DNAscope.

Utilizing cutting-edge technology, the StarSeq100 AI Sequencer revolutionizes sequencing through its dual technological pillars. Firstly, it boasts exceptional adaptability for various clinical applications, thanks to its interchangeable high-quality optical and chip systems. Secondly, it employs a sophisticated hierarchical base calling pipeline driven by deep learning on a heterogeneous computing platform (CPU + GPU). This pipeline encompasses image preprocessing, cluster detection, crosstalk correction, and hierarchical base

calling modules, surpassing traditional methods (such as traditional base calling) across diverse data densities. Notably, previous disclosures highlight that this innovative approach delivers exceptional performance metrics, including the highest effective throughput (12.18% more clusters), the lowest error rate (0.0175% on average), and the highest average%Q30 score(99.27%)[1, 2]. Moreover, recent optimizations have significantly enhanced its efficiency, resulting in a notable uplift of 30% to 40% in effective throughput.

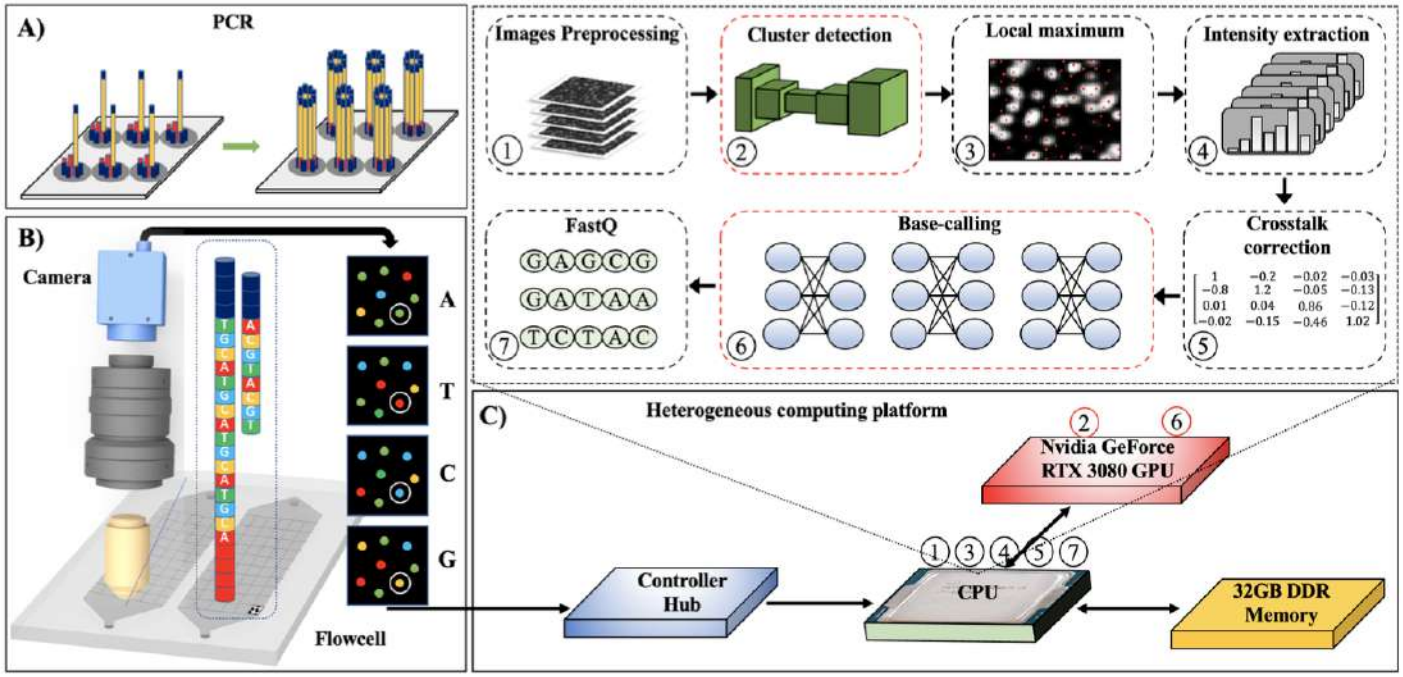


Figure. 1. The architecture of the deep learning-based processing pipeline[2].

As a result, StarSeq100 AI Sequencer outputs three whole exomes per flowcell, generating up to 2x150 bp reads with a coverage depth exceeding 100x per exome. The resultant FASTQ files typically exhibit high accuracy, with Q30 scores surpassing 85%. Moreover, the over 10 GB of sequencing data per exome can be seamlessly streamed directly onboard or transferred to detached analysis hardware for secondary analysis.

Parameters	StarSeq 100
Read length	SE75, PE100, PE150
Flow cell	1 or 2
Single FC reads number	40M、80M、125M
Double FC reads number	80M、160M、250M
Q Score (SE75)	Q30>80%
Sequencing time (SE75)	~ 10.5h

Figure. 2. Specification of Genesense StarSeq100 AI Sequencer capability and capacity[3].

Sentieon offers accurate and efficient pipelines catering to diverse Genesense applications, encompassing germline and somatic WES, panels, and non-human sequencing. The software seamlessly operates on local infrastructure or any cloud environment. Sentieon DNAscope pipeline enhances accuracy through improved active region detection, robust local assembly of reads, and the integration of pre-trained machine learning models tailored to specific sequencers[4]. The combination of Sentieon's pipeline speed, accuracy, and user-friendly modularized design makes it an exceptional choice for sequencing analysis.

Methods

To validate the efficacy of coupling Genesense sequencing with Sentieon analysis, we conducted sequencing on multiple replicates of whole exome sequencing (WES) datasets from two well-characterized human reference samples, HG001 and HG007. These reference samples were chosen due to the provision of high-quality variant truth sets by NIST, facilitating the assessment of SNP and indel sensitivity and precision[5].

In our study, DNA libraries were enriched using the Agilent SureSelect Human All Exon V6 kit. Four datasets from replicate sequencing of HG001 and HG007 (two for each genome), utilizing 75bp single-end reads, were shared with Sentieon as a first batch. A Genesense-specific error model was trained using this initial batch. Subsequently, this model was applied to a testing dataset consisting of one HG001 and one HG007 sample from a second batch sequencing, featuring slightly different chemistry and base caller, to evaluate performance and the model's ability to generalize to diverse datasets.

Specifically, all datasets underwent mapping to the hg38 reference genome using Sentieon BWA-turbo (machine learning guided alignment, version 202308.01). Sorting and deduplication procedures were performed using Sentieon utility modules, followed by quality checks on the resulting BAM files. Analysis of base quality score distribution and mapping rates indicated high read quality, with consistent base quality observed across regions with varying GC content.

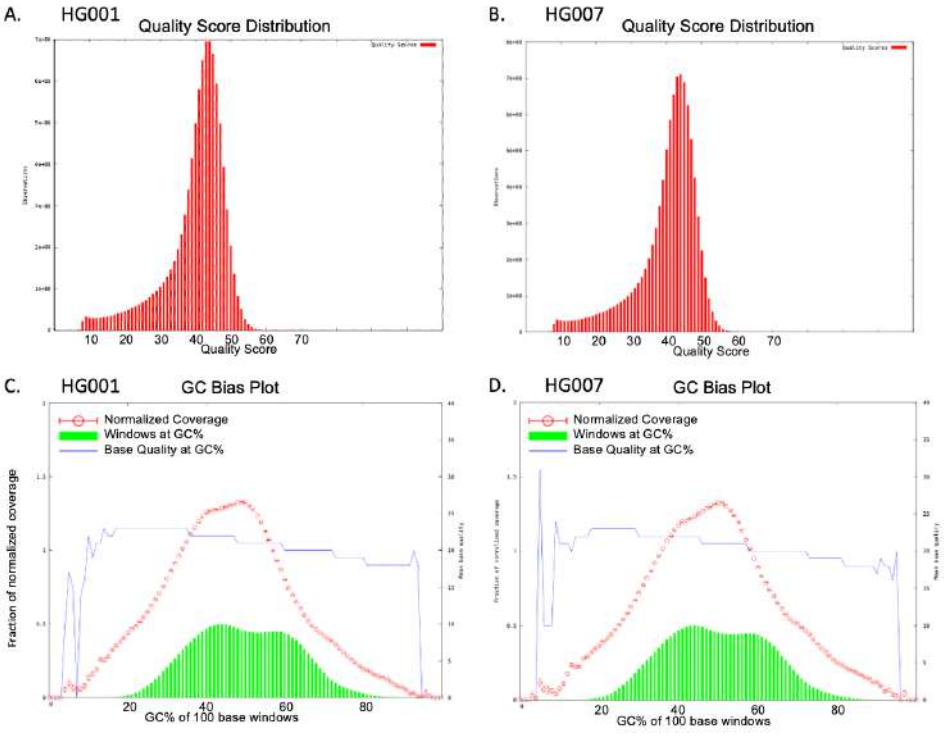


Figure 3. The GC Bias Plot and Quality Score Distribution of the testing datasets HG001 (A) and (C) and HG007 (B) and (D). Most reads exhibited quality scores of 40 or higher, with consistent quality scores observed across most GC windows.

The four aligned training datasets were utilized for model training, wherein a gradient boosting decision tree (GBDT) was constructed on candidate variants generated by DNAscope's highly sensitive mode. This utilized the Genome in a Bottle (GIAB) v4.2.1 benchmark VCF files to produce the DNAscope Genesense v0.1 model. Subsequently, the two testing datasets were analyzed with DNAscope using this model, and the variants were evaluated against the GIAB version 4.2.1 all regions truth set. The calculated accuracy metrics were compared with datasets sequenced by Illumina platform and processed by GATK[6].

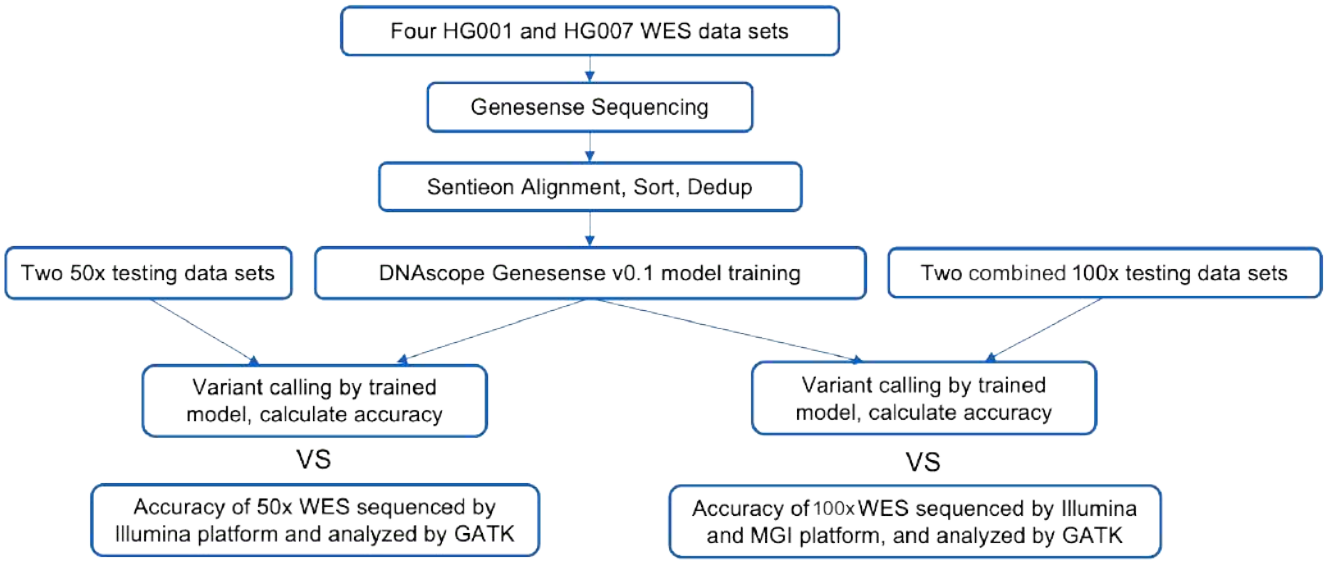


Figure 4. Overview of model training and benchmarking pipeline.

It's noteworthy that due to the nature of single-end 75bp reads, post-deduplication coverage is limited to less than 75x, with an average coverage of approximately 50x (Table 1).

Table 1. Sequencing coverage of the deduped testing datasets

Dataset ID	Total Aligned Bases	Coverage Mean	Coverage Granular Q1	Coverage Granular Median	Coverage Granular Q3	Bases Above 15x
HG001	3,167,055,403	52x	36x	53x	70x	94.1%
HG007	3,201,434,038	53x	37x	53x	70x	95.3%

Further, to obtain a 100x WES dataset, we combined half of the training datasets with the corresponding testing dataset from the same reference sample. These two combined 100x (102.6x for HG001, 100.3x for HG007) datasets were then processed to calculate accuracy.

Results – Accuracy

The model trained on Genesense data markedly enhances variant calling accuracy compared to both the standard GATK pipeline (results matched by the Sentieon DNAseq pipeline) and DNAscope with models trained for other platforms (data not shown).

The overall F1-scores exhibit high similarity between the HG001 and HG007 datasets, suggesting that the trained model is not biased towards specific samples. The majority of errors stem from false negatives, likely attributable to insufficient coverage of some regions post-WES panel enrichment.

Additionally, the accuracy surpasses that of the Illumina platform analyzed by GATK (DNAseq), yielding over two times fewer SNP errors and three times fewer INDEL errors. It's worth noting that the Illumina dataset serves solely as a reference and should not be directly compared due to different reference genomes in analysis and the origin of the Illumina dataset from the Google Brain Genomics project, which may not reflect current standards.

Table 2. Accuracy benchmark and comparison using 50x validation datasets processed by DNAscope. Illumina dataset is from Google Brain project[7].

Sequencing Platform	Analysis Pipeline	Reference Sample	Coverage	Variant Type	Ture Positive	False Negative	False Positive	Recall	Precision	F1 Score
Genesense StarSeq100 AI Sequencer	DNAscope with Genesense model v0.1	HG001	53x	SNP	44,431	1,798	273	0.961	0.994	0.977
				INDEL	3,078	418	107	0.880	0.966	0.921
		HG007	53x	SNP	44,857	2,028	271	0.957	0.994	0.975
				INDEL	2,901	352	90	0.892	0.970	0.929
Illumina Novaseq	GATK (DNAseq)	HG003	50x	SNP		4,900 (combined)				0.966
				INDEL		136 (combined)				0.855

Analysis of 100x testing datasets, derived from combining 50x testing datasets with half of training datasets, reveals slightly higher F1 scores for SNPs and significantly higher scores for INDELs, compared with previous 50x testing datasets. This improvement aligns with expectations of enhanced accuracy with increased coverage depth. Moreover, compared to 100x Illumina and MGI datasets, the two Genesense datasets (only show HG001 in Fig 5) exhibit notably superior accuracy, particularly in the INDEL category although this benchmark serves as a reference and should not be construed as a direct comparison.

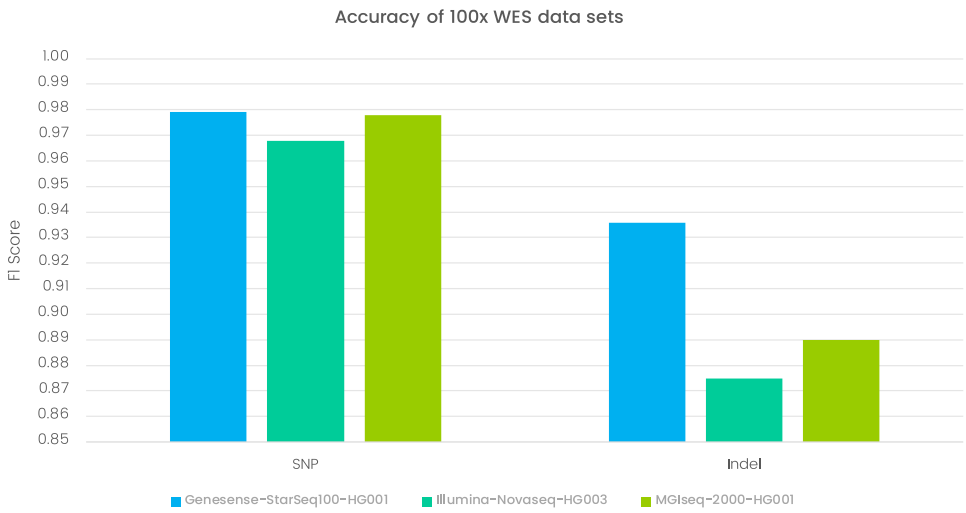


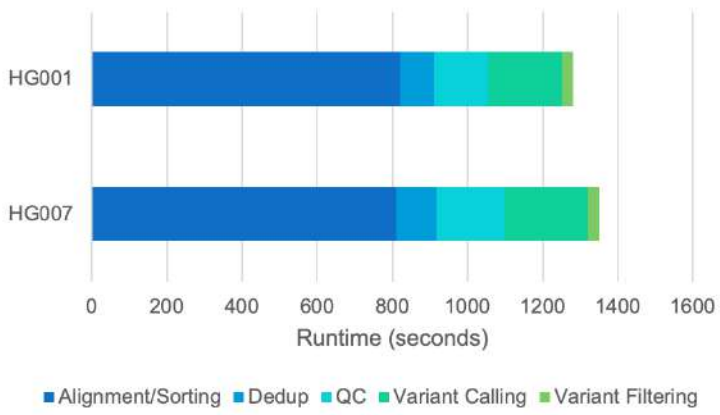
Figure 5. Accuracy benchmark and comparison using 100x validation datasets processed by DNAscope. Illumina dataset is from Google Brain project[7]; MGI dataset is downloaded from China National GeneBank Public dataset G400-HotMPS-PE100-NA12878-WES[8].

Results – Speed

Leveraging redesigned algorithms and more efficient programming languages in its implementation, the Sentieon pipeline has achieved notably accelerated processing speeds compared to standard open-source tools, on generic CPU hardware environments. In this study, we monitored the runtime of the two testing datasets running on a local 32-thread server. The overall runtime averaged approximately 22 minutes (equivalent to approximately 1300 seconds), with the majority of computational time allocated to the alignment step.

Considering that the Genesense StarSeq100 AI Sequencer is equipped with an Intel-i9 CPU, potentially featuring up to 24 threads, its inherent analysis capability enables the processing of a WES dataset in around 30 minutes. This translates to a daily capacity of 48 WES analyses.

A.



B.

Control Computer

Configuration	
CPU	Intel-i9
RAM	64G
Storge	2T SSD
Operating system	Win10

Figure 6. A) Runtime of two testing datasets, benchmark was conducted on a 32 threads 64Gb local cluster; B) Specification of StarSeq100 AI Sequencer Computer[3].

Conclusions

The integration of Genesense sequencing with Sentieon analysis facilitates the generation of high-quality variant calls for whole exome sequencing. Sentieon's development of an error model based on Genesense data enhances accuracy beyond what is achievable with data generated by other sequencing platforms analyzed through standard pipelines. This streamlined workflow can be completed on the StarSeq100 AI Sequencer in approximately 30 minutes per human exome. The resultant high-quality variant calls serve as valuable inputs for a multitude of downstream applications.

References

1. *S. Li et al., "BaseFormer: Transformer based Base-Caller for Fast and Accurate Next Generation Sequencing," 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, 2022, pp. 463-466, doi: 10.1109/EMBC48229.2022.9871730.*

2. *S. Luo et al., "Hierarchical DNN with Heterogeneous Computing Enabled High-Performance DNA Sequencing," 2022 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Shenzhen, China, 2022, pp. 35-40, doi: 10.1109/APCCAS55924.2022.10090281.*

3. *https://gene-sense.com/en/pro_detail.html*

4. *Donald Freed, Renke Pan, Haodong Chen, Zhipan Li, Jinnan Hu, Rafael Aldana. DNAscope: High accuracy small variant calling using machine learning. bioRxiv 2022.05.20.492556.*

5. *Zook, J., Catoe, D., McDaniel, J. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 3, 160025 (2016). https://doi.org/10.1038/sdata.2016.25.*

6. *McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303.*

7. *Google Brain Genomics Sequencing Dataset for Benchmarking and Development was accessed from https://registry.opendata.aws/google-brain-genomics-public.*

8. *https://db.cngb.org/search/experiment/CNX0431708/*

9. *https://www.nvidia.com/en-us/gpu-accelerated-applications/?search=genesense*

10. *https://mp.weixin.qq.com/s/QxsEP4dpLFe9Z730K-0FGQ*

11. *G. Chen et al.,"Deep Learning based Methods and Systems for Nucleic Acid Sequencing, US PCT,17/681,672*

12. *S. Luo et al., "Methods and systems for enhancing nucleic acid sequencing quality in high-throughput sequencing processes with machine learning",PCT/CN2022/104105.*

GENESENSE

www.gene-sense.com

Email: Sales@gene-sense.com

Hongkong:

Address: Building 1E, Hong Kong Science Park

Tel: +852 3460 5372

Shenzhen:

Address: Loop Shenzhen-Hong Kong Science and Technology Innovation Cooperation Park

Tel: 4008 206 806

Shanghai:

Address: Zhangjiang Hi-Tech, Shanghai

Tel: 021-50931201