Element Biosciences

Sentieon

APPLICATION NOTE

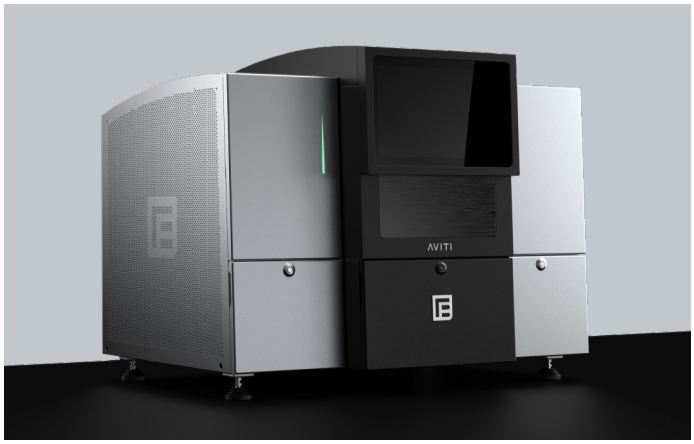
High Quality Whole Human Genome Sequencing with Sentieon DNAScope and AVITI™

Highlights

- **The combination of AVITI sequencing with Sentieon analysis enables highly accurate and efficient whole genome sequencing results**
- **The analysis workflow can be executed in <40 minutes at <$2 in compute costs per 35X Genome**

Introduction

Whole genome sequencing has emerged as an important tool in many applications, such as studying genetic diseases, assessing heritable risk, and reconstructing human population history. Whole genome sequencing projects routinely generate terabytes of data, and the field requires accurate and efficient tools for data analysis. Converting the large amounts of data generated by the sequencer into actionable insights can be a daunting task in both infrastructure and methods development. Fortunately, a rich ecosystem of cloud infrastructure and analysis methods exists and can be leveraged. In this application note, we demonstrate how whole genome sequencing data generated on AVITI™ can be efficiently analyzed using the Sentieon™ secondary analysis software on the AWS™ cloud.



The AVITI system enables two whole genomes to be sequenced per flowcell with 2x150 bp reads to a depth of coverage exceeding 35X per genome. The resulting > 100 GB of sequencing data per genome can be streamed directly to AWS during the sequencing process. An analysis pipeline, specific to the application of interest, can then be used to process the data.

Sentieon provides accurate and efficient pipelines for a variety of Element Biosciences applications including germline and somatic WGS, WES, panels and non-human applications. The software can run on local infrastructure or on any Cloud. The combination of Sentieon pipeline speed, accuracy, and ease-of-use makes it an excellent choice for sequencing analysis.

Methods

To demonstrate the effectiveness of combining AVITI sequencing with Sentieon analysis, we sequenced multiple replicates of the well-characterized human genomes, HG001, HG002, HG003, HG004, and HG005[1]. These genomes were selected because the DNA samples are readily available from Coriell and because high quality variant truth sets are provided by NIST, enabling the measurement of SNP and indel sensitivity and precision.

For the study, we evaluated DNA libraries that were made without the use of PCR, showcasing the AVITI data quality that can be achieved absent upstream, library induced PCR errors. The PCR-free library was constructed using the KAPA HyperPrep™ kit followed by the Adept™ compatibility kit[2]. Additional information about the library preparation methods is available[3]. Multiple 35X data sets from replicate sequencing of HG001, HG002, and HG005 were shared with Sentieon. An AVITI-specific error model was trained using HG001 and HG005 samples, with HG002 held out for evaluation of the trained model. To ensure broad applicability of the model, a wide range of depths, from 15X to 105X, was used in the training. The model was then applied to independent data sets generated at Element to evaluate performance and the ability of the model to generalize to different samples.

Data was down-sampled to 35X coverage to mimic a common customer whole genome sequencing use-case. The number of input read pairs for each analysis was chosen to be 360M, based on the desired 35X coverage, multiplied by the length of the human genome, and divided by the read length. This methodology is supported by the low optical duplicate rate and the high alignment rate of the sequencing data, as >99% of the data is retained and usable. By contrast, other technologies may require a greater number of input reads to attain 35X coverage following the removal of duplicate reads. The data was analyzed with Sentieon DNAScope and the variants were evaluated using the GIAB version 4.2.1, all regions truth set.

Results – Accuracy

The model trained on Element data reduced overall error count by more than 10% and significantly improved performance in low complexity regions of the genome, relative to the default model. Improvements were observed across all coverage levels and genomic contexts. As expected, accuracy improved with coverage depth.

Figure 1: F1 Score across samples

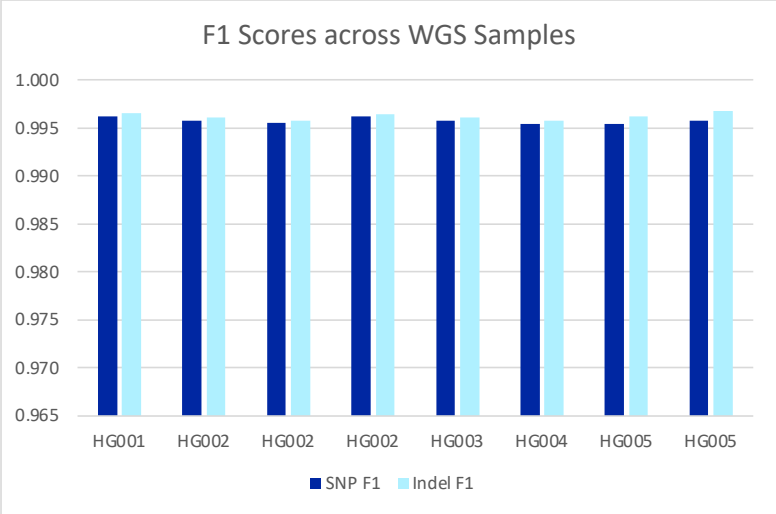


Figure 1 shows the performance of multiple PCR-free data sets using the model trained on Element data. Model training only used data from HG001 and HG005 and none of the data sets shown were used in training, with the exception of HG001, which was only sequenced once. Both SNP and indel F1 score values are uniformly high across sequence replicates and Coriell samples.

Figure 2: F1 score as a function of coverage

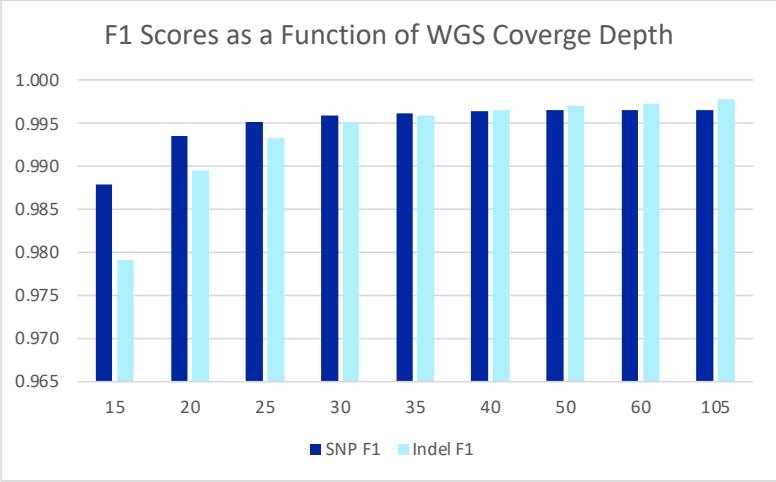


Figure 2 shows the F1 score for HG002 as a function of coverage depth. As previously noted, the model was trained on HG001 and HG005, so HG002 data was not used in the training. Accuracy increased with coverage depth through the entire range, though the increases are small beyond a depth of 30X.

Results – Performance

Table 1: Analysis time and cost
Table 1 shows the speed and cost of analysis. The alignment and variant calling analysis can be completed on AWS in approximately 40 minutes, for less than $2 per 35X genome in AWS compute costs.

AWS Instance Type	hpc6a.48xlarge
Alignment (s)	1641
Dedup (s)	71
DNAscope (s)	610
DNAscope pipeline runtime (min)	38.7
Instance vCPU	96
DNAscope pipeline core-hours	61.9
Instance On-Demand $/hr	$2.88
Sample On-Demand $	$1.86

Figure 4: View of the typical commands used to run the Sentieon pipeline.

```
sentieon bwa mem … | sentieon util sort …
sentieon … --algo LocusCollector …
sentieon … --algo Dedup …
sentieon … --algo DNAscope …
sentieon … --algo DNAModelApply …
```

Conclusions

The combination of AVITI sequencing and Sentieon analysis enables high quality variant calling for whole genome sequencing. An error model based on AVITI data developed by Sentieon, and available in the latest software version, improves accuracy beyond the default model trained on other technologies. The model is broadly applicable across a wide coverage range, with accuracy improving as coverage increases. The entire workflow can be executed in AWS or a comparable compute platform in under 40 minutes at a cost of less than $2 in compute cost per 35X human genome. The high-quality variant calls can then be used as the input for a variety of downstream applications.

References

1. Element AVITI™ System Workflow Guide (MA-00008)
2. Element Adept™ Library Compatibility Workflow Guide (MA-00001)
3. Roche (2022) The KAPA Library Preparation Portfolio allows for high quality sequencing across diverse applications on the Element AVITI™ System [Application Note]
4. For more information about Sentieon, visit: www.sentieon.com

Additional Information

For additional information, visit:

Applications: go.elembio.link/apps

Resources: go.elembio.link/resources

Element Biosciences

Element Biosciences
elementbiosciences.com
Telephone: 619.353.0300
Email: info@elembio.com

Document # LT-00008

For Research Use Only. Not for use in diagnostic procedures.