Retraining and Benchmarking DNAscope with Complete Genomics DNBSEQ Whole Genome Sequencing for Accurate Small Variant Calling

Introduction

The 1990 conception and development of Massively Parallel Sequencing (MPS) technologies in the mid 2000s [1,2] have provided researchers with new insights into genomic and precision medicine.[3] Complete Genomics' DNBSEQ™ sequencing platforms produce high-quality reads through DNA Nanoball (DNB) technology on patterned DNA nanoarrays [4], utilizing cPAS-based sequencing chemistry. These advances have eliminated the need for on-chip PCR clusters, resulting in a substantial reduction in sequencing errors (Figure 1).

Paired-end (PE) sequencing is another important invention in MPS and is very useful for Whole Genome Sequencing (WGS). The process generates sequence reads from both ends of longer genomic fragments captured in the library and bridges over most of the genomic repeats with relatively short reads. With DNBs, high-quality PE sequencing of single-stranded DNA is enabled on DNBSEQ and CoolMPS technology.

Bioinformatics tools are also evolving to meet the requirements of challenging clinical applications, including fast and accurate data processing and small variant calling for whole genome sequencing data. A preferred data processing pipeline is Sentieon DNAscope, which provides accurate and efficient germline small-variant calling.[5]

Highlights

- Sentieon DNAscope pipeline provides accurate and efficient germline small variant-calling for WGS
- DNAscope and Complete Genomics DNBSEQ workflow achieved superior SNP and Indel accuracy compared to standard Illumina PCR-free datasets
- DNBSEQ technology leverages rolling circle replication (RCR) and DNA Nanoballs (DNBs) to prevent clonal errors and index hopping during sequencing for accurate, reproducible results

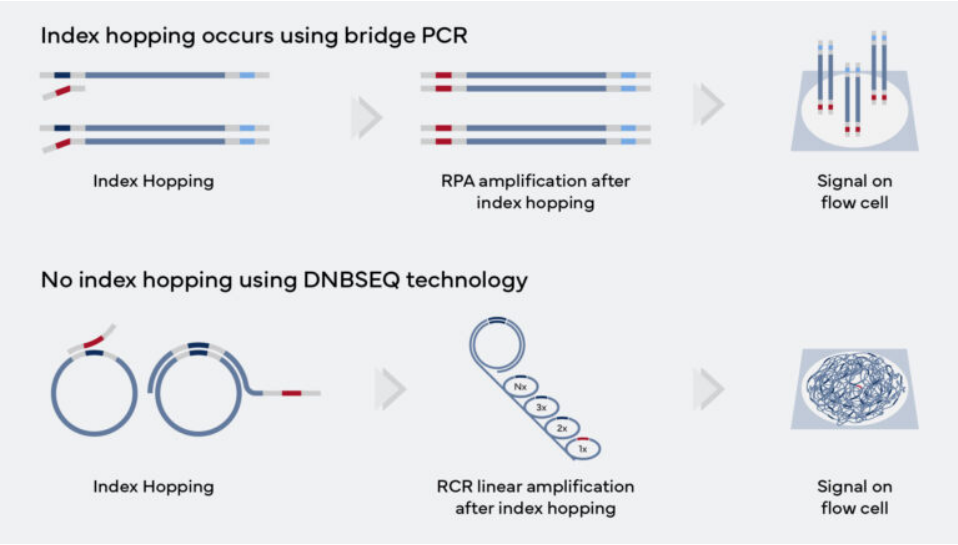

Figure 1. DNBSEQ™ rolling circle replication (RCR) is used to increase signal intensity during array formation, which is followed by sequencing DNBs with DNBSEQ technology. Individual copies from the same DNB are replicated independently using the same ssCirDNA template. In this way, amplification errors cannot accumulate.

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

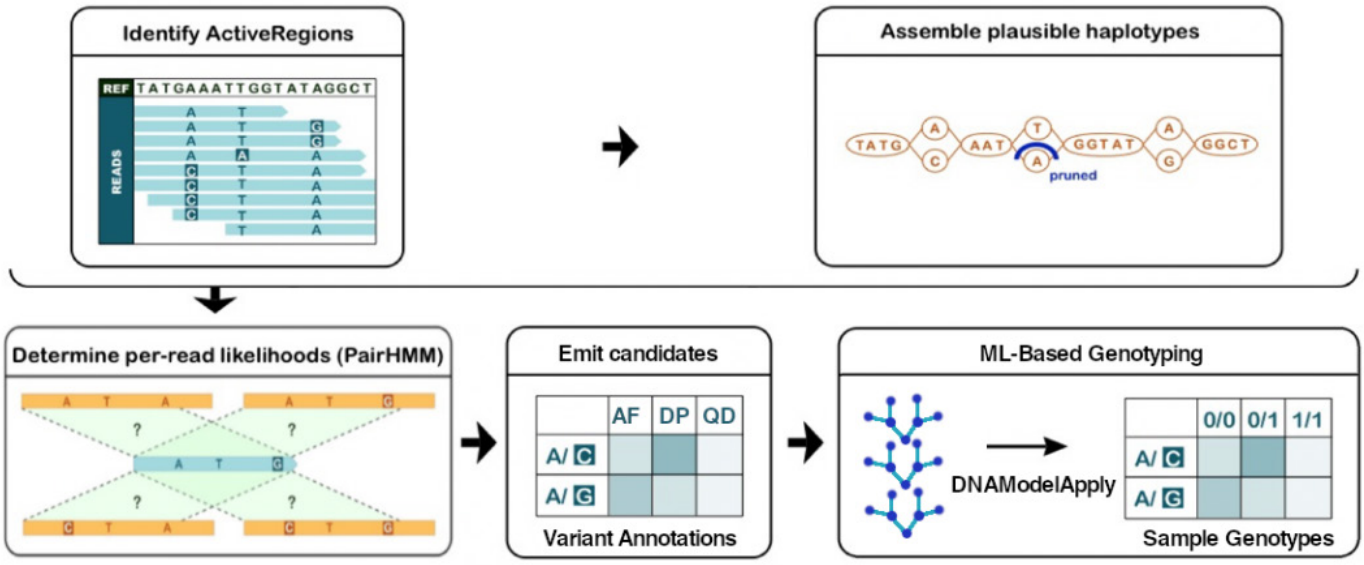

Figure 2. Overview of the DNAscope methodology. DNAscope follows a similar algorithmic flow to GATK HaplotypeCaller. Sites likely to harbor genetic variation are identified as active regions. Sequence reads aligned across active regions undergo local assembly using de Bruijn graphs and read-haplotype likelihoods are calculated through PairHMM. Variant candidates are then annotated and emitted. Machine learning-based genotyping processes variant candidates to determine the correct variant genotype.

DNAscope uniquely combines the well-established methods from haplotype-based variant callers with machine learning to achieve improved accuracy. As a successor to GATK HaplotypeCaller[6], DNAscope uses a similar logical architecture, but introduces improvements to active region detection and local assembly for improved sensitivity and robustness, especially across high-complexity regions (Figure 2). When a machine learning model is applied, DNAscope outputs candidate variants with additional informative annotations. Annotated variant candidates are then passed to a machine learning model for variant genotyping, resulting in improvements in both calling and genotyping accuracy. Advances in the underlying algorithms for genomic data processing and a robust implementation help make DNAscope five to ten times faster than the GATK best practices pipeline.

DNAscope uses a robust process for the identification of variant candidates. The addition of a platform-specific machine learning model can further improve the overall variant calling accuracy. A Complete Genomics model was developed and benchmarked previously on DNBSEQ-G400 and DNBSEQ-T7. The results showed that the DNAscope + CG model achieved superior SNP and Indel accuracy compared to standard Illumina PCR-free datasets. Here, we analyzed the accuracy of current PE150 WGS reads generated on Complete Genomics sequencers using a new, trained model (v0.5) and updated a more difficult variant truthset (v4.2.1). The v4.2.1 dataset is a substantial improvement over the GIAB v3.2.2 benchmark dataset used in the earlier publication and includes approximately 200 MB of challenging genomic regions that were excluded from the previous v3.2.2 datasets. Accordingly, the F1-scores reported in this study are lower than the corresponding scores reported in the earlier manuscript.

Methods

DNAscope Complete Genomics Model v0.5

Training of the DNAscope model was conducted with an updated model framework introduced in release 202112.01 of the Sentieon software package. Reference Datasets HG001 and HG005 sequenced from DNBSEQ-G400 platform were used as a training dataset, with 20% of data randomly split for validation and chr20 held out for testing (Figure 3).

Four HG001 and HG005 PE150 datasets were mapped to the hg38 reference genome using Sentieon BWA and a quality check was conducted on generated BAM files. The base quality score distribution and mapping rate indicated high read quality, and the datasets had even coverage across regions of different GC content (Figure 4).

The four aligned datasets were downsampled to generate multiple training datasets with sequencing depths from 15x to 45x. The HG001-150PE 30x dataset was saved for validation. During training, a gradient boosting decision tree (GBDT) was built on candidate variants generated by DNAscope's highly sensitive mode, using the Genome in a Bottle (GIAB) v4.2.1 benchmark VCF.

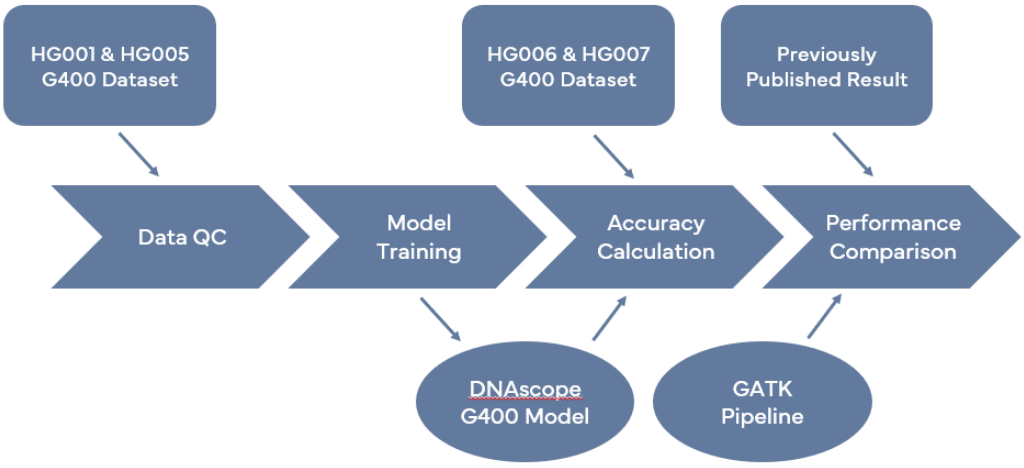

Figure 3. Overview of model training and benchmarking pipeline.

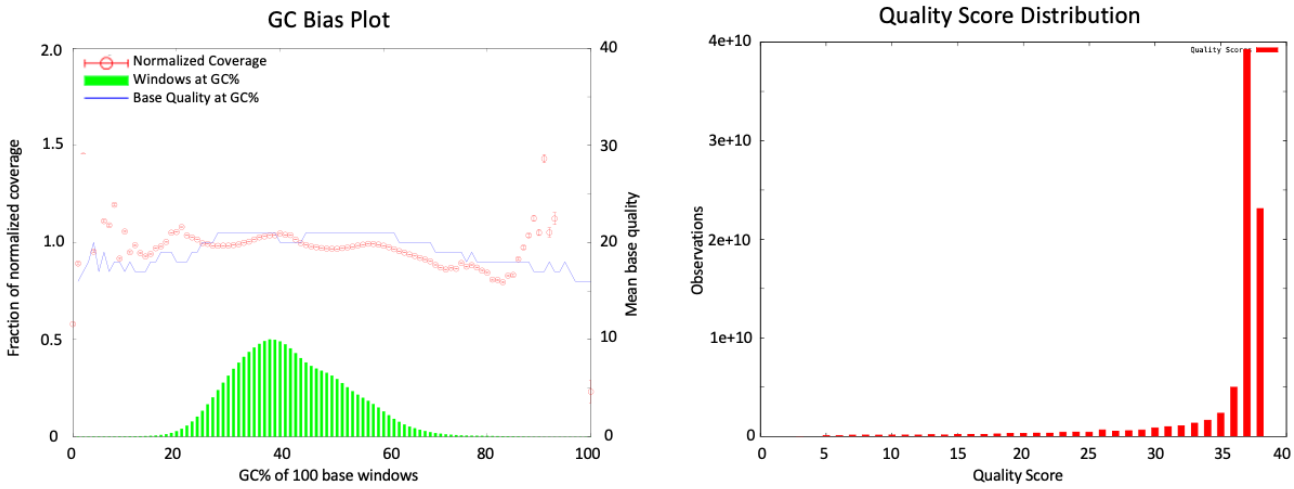

Figure 4. GC Bias Plot and Quality Score Distribution of HG006-PE150 30x dataset. Sequencing coverage is even across most GC windows, even for extreme GC%, and most reads returned 35+ quality scores.

Results

Testing datasets HG006 and HG007 sequenced from DNBSEQ-G400 were processed by Sentieon DNAscope pipeline version 202112.05 with the trained CG v0.5 model. The resulting variants were compared to the GIAB v4.2.1 benchmark VCF using hap.py version 0.3.10 with RTGtools vcfeval version 3.9.2 as the variant comparison engine. The trained model was also validated using the held-out HG001-PE150-30x dataset. The overall F1-score of the held-out HG001 dataset was very similar to the HG006 and HG007 dataset, indicating that the trained model is not overfit to the HG001 and HG005 reference samples which were used in model training (Table 1). In addition, HG006 46x dataset (regularly generated per one lane on DNBSEQ-G400 flow cell) returned higher F1-score than its corresponding 30x dataset (~40% less false positive errors), showing that higher depth indeed contributes to variant calling accuracy. An additional dataset from HG001 was also sequenced on the DNBSEQ-T7 sequencer and analyzed by a later-released DNAscope CG model v1.0 to show consistency across sequencing platforms (Table 1). The accuracy is also comparable or better to Illumina platform (e.g., over two times less, especially false positive, indel errors in DNBSEQ-G400) working with either DNAscope or DNAseq9 (Table 2). The ILMN dataset is displayed for reference and should not be considered as a direct comparison since reference genomes are different and the ILMN dataset is from NIST GIAB project which is not updated to current. In addition to sequencing accuracy, there are many other factors influencing variant call accuracy, especially in so-called "difficult" genomic regions. WGS library quality, including insert length, clonal PCR errors and sequence read length, are recognized as critical factors that need to be taken into account for further improvements of WGS.

Reference Dataset	Sequencing Platform	Analysis Pipeline	Type	False Negative	False Positive	Recall	Precision	F1-Score
HG006-PE150 30x	DNBSEQ-G400	DNAscope CG model v0.5	SNP	18,049	6,084	0.994	0.998	0.996
			INDEL	1,783	836	0.996	0.998	0.997
HG006-PE150 46x	DNBSEQ-G400	DNAscope CG model v0.5	SNP	18,215	3,560	0.994	0.999	0.997
			INDEL	1,407	460	0.997	0.999	0.998
HG007-PE150 30x	DNBSEQ-G400	DNAscope CG model v0.5	SNP	19,101	5,904	0.994	0.998	0.996
			INDEL	1,856	786	0.996	0.998	0.997
HG001-PE150 30x	DNBSEQ-G400	DNAscope CG model v0.5	SNP	14,054	6,065	0.996	0.998	0.997
			INDEL	1,802	919	0.996	0.998	0.997
HG001-PE150 30x	DNBSEQ-T7	DNAscope CG model v1.0	SNP	14,467	6,732	0.996	0.998	0.997
			INDEL	2,619	1,146	0.994	0.998	0.996
HG002-PE150 30x	ILMN NovaSeq	DNAscope ILMN model v1.0	SNP	20,368	8,571	0.994	0.997	0.996
			INDEL	3,633	2,062	0.993	0.996	0.995

Table 1. Accuracy Benchmark using selected validation datasets processed by DNAscope. DNBSEQ-T7 dataset was downloaded from CNGB database as "CNX0200271"; ILMN dataset accuracy is from recently published DNAscope white paper.[5]

Reference Dataset	Sequencing Platform	Analysis Pipeline	Type	False Negative	False Positive	Recall	Precision	F1-Score
HG001-PE150 30x	DNBSEQ-G400	GATK (DNAseq)	SNP	21,390	25,077	0.993	0.992	0.993
			INDEL	3,380	2,897	0.993	0.994	0.993
HG002-PE150 30x	ILMN NovaSeq	GATK (DNAseq)	SNP	33,446	28,933	0.990	0.991	0.991
			INDEL	15,032	10,196	0.971	0.980	0.976

Table 2. Accuracy Benchmark using selected validation datasets processed by GATK (represented by DNAseq). ILMN dataset accuracy is from recently published DNAscope white paper.[5]

Summary

The results of this study demonstrated that the newly trained DNAscope model for the Complete Genomics DNBSEQ platform, along with the combined improvements in the underlying sequencing chemistry resulted in high variant calling accuracy. The elimination of PCR amplification during both library prep and sequencing array prep greatly reduced errors introduced prior to sequencing. The retrained DNAscope model improves DNAscope's accuracy with DNBSEQ reads, allowing DNAscope to more accurately model systematic error patterns andenabling more accurate discrimination of false positive and false negative variant calls. The overall variant calling accuracy is improved from the previously published accuracy, as are the integrated sequencing and analysis solutions.

References

1. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005 Sep 9;309(5741):1728-32.
2. Bentley DR, Balasubramanian S, Swerdlow HP, et.al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008 Nov 6;456(7218):53-9.
3. Tucker T, Marra M, Friedman JM. Massively parallel sequencing: the next big thing in genetic medicine. Am J Hum Genet. 2009 Aug;85(2):142-54.
4. Drmanac R, Sparks AB, Callow MJ, et.al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010 Jan 1;327(5961):78-81.
5. 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.
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.

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

Ordering Information

Category	Product	Cat. No.	Vendor
Library Prep	DNBSEQ Universal Library Conversion Kit	940-000963-00	Complete Genomics
Instruments	DNBSEQ-G400 Genetic Sequencer	900-000641-00	Complete Genomics
	DNBSEQ-T7 Genetic Sequencer	900-000698-00	Complete Genomics
Sequencing Reagents	DNBSEQ-G400 High-Throughput Sequencing Set (FCL PE150)	940-000810-00	Complete Genomics
	DNBSEQ-T7 High-Throughput Sequencing Set (FCL PE150)	940-000836-00	Complete Genomics
Software	DNAscope Complete Genomics Pipeline	-	Sentieon

Contact Sentieon for more information

Sentieon

Website: https://www.sentieon.com
Tel: 650-254-6360
Email: info@sentieon.com

To learn more, visit completegenomics.com

Contact us:
info@completegenomics.com

Technical support:
United States: US-TechSupport@completegenomics.com
Canada: CA-TechSupport@completegenomics.com

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

