Application Note

Accurate Small Variant Calling by DNAscope on MGI DNBSEQ™ G400 Sequencing Platform

Massively Parallel Sequencing using DNBSEQ Technology

The 1990 conception[1] and development of Massively Parallel Sequencing (MPS) technologies in the mid 2000s [2,3], have provided researchers with new insights into genomic and precision medicine[4]. MGI's DNBSEQ technology produces high quality reads via Rolling Circle Amplification (RCA), patterned nanoarrays[5] and cPAS-based sequencing chemistry (cPAS)[6]. These advances have eliminated the need for on-chip PCR clusters resulting in a substantial reduction in sequencing errors and therefore, higher accuracy.

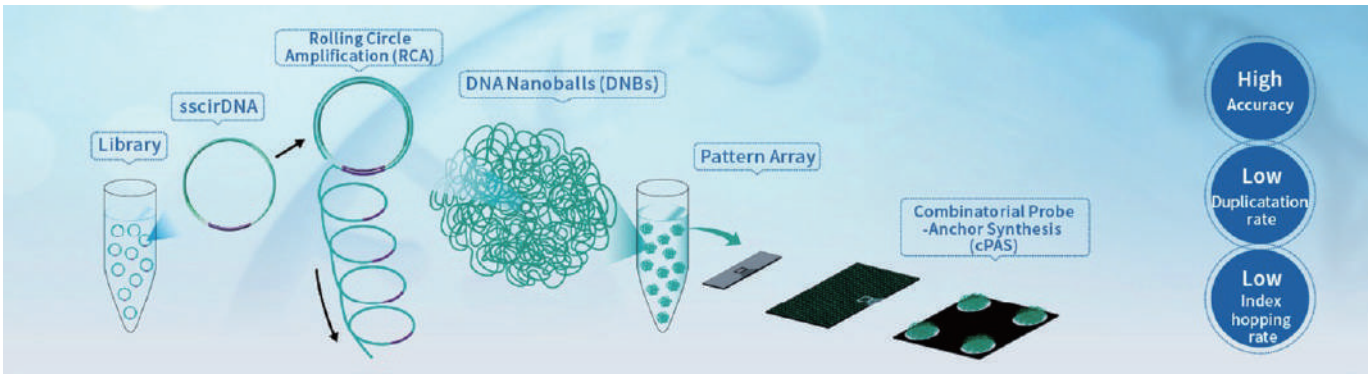

Figure 1. DNA NanoBall Sequencing (DNBSEQ) on patterned array. DNBSEQ technology uses Rolling Circle Amplification (RCA) to generate multiple copies of the fragment insert (eg DNA fragment of interest), and always amplifies from the original copy of template. Therefore, there is no clonal error introduced in the sequencing step as there is with other sequencing technology.

DNBSEQ-G400 sequencing platform

DNBSEQ-G400 sequencing planform is the perfect day-to-day sequencer for various genomic projects. It offers high throughput with data output ranging from 55 to 1440 Gb daily by providing the option to run 1 or 2 flow cells with 2 types of flow cells (550M/reads vs 1800M/reads) and various read length options from SE50 to SE400 or PE300.

DNAscope and Modeling on DNBSEQ-G400 sequencing platform

Bioinformatics tools are also evolving to meet the requirement 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.[9]
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[10], 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. 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.

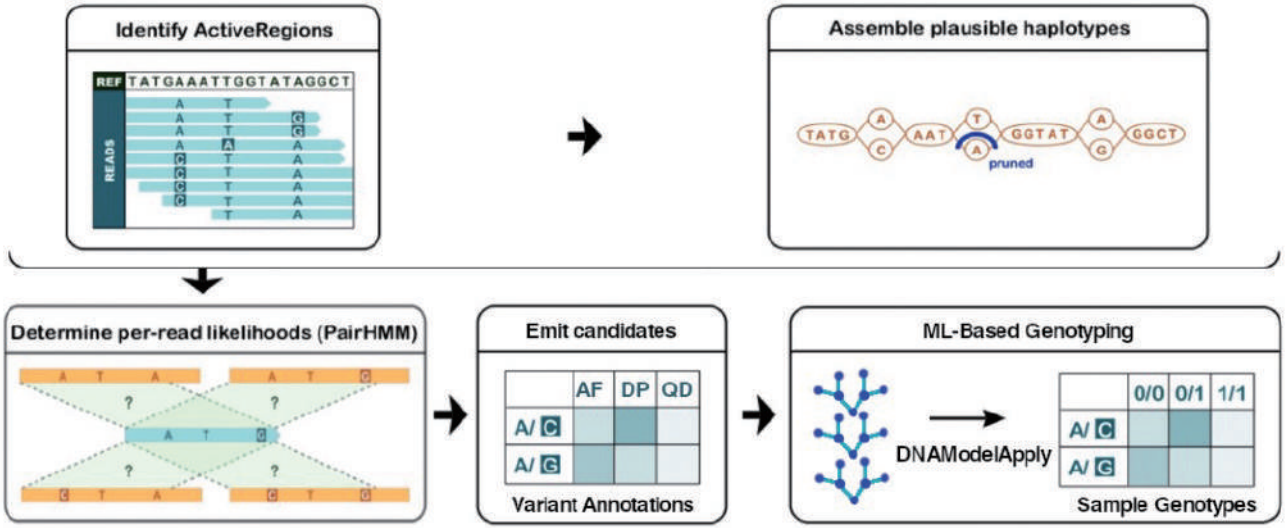

Figure 2. Overview of the DNAscope methodology. DNAscope follows a similar algorithmic flow to GATK HaplotypeCaller. Sites likely to harbor genetic variation are identified as active regions. Sequence reads aligned across active regions undergo local assembly using de Bruijn graphs and read-haplotype 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 uses a robust process for identification of variant candidates. The addition of a platform-specific machine learning model can further improve the overall variant calling accuracy. A MGI model was developed and benchmarked previously on DNBSEQ-G400 and DNBSEQ-T7 using Standard MPS chemistry7. Result showed that the data from DNAscope with MGI Standard MPS chemistry achieved superior SNP and Indel accuracy compared to standard Illumina PCR-free datasets. Here we analyzed accuracy of current PE150 WGS reads generated on MGI's DNBSEQ-G400 sequencers using a new trained model (v0.5) and updated more difficult variant truthset (v4.2.1).

DNAscope MGI Model v0.5 training on DNBSEQ-G400 Sequencing Platform

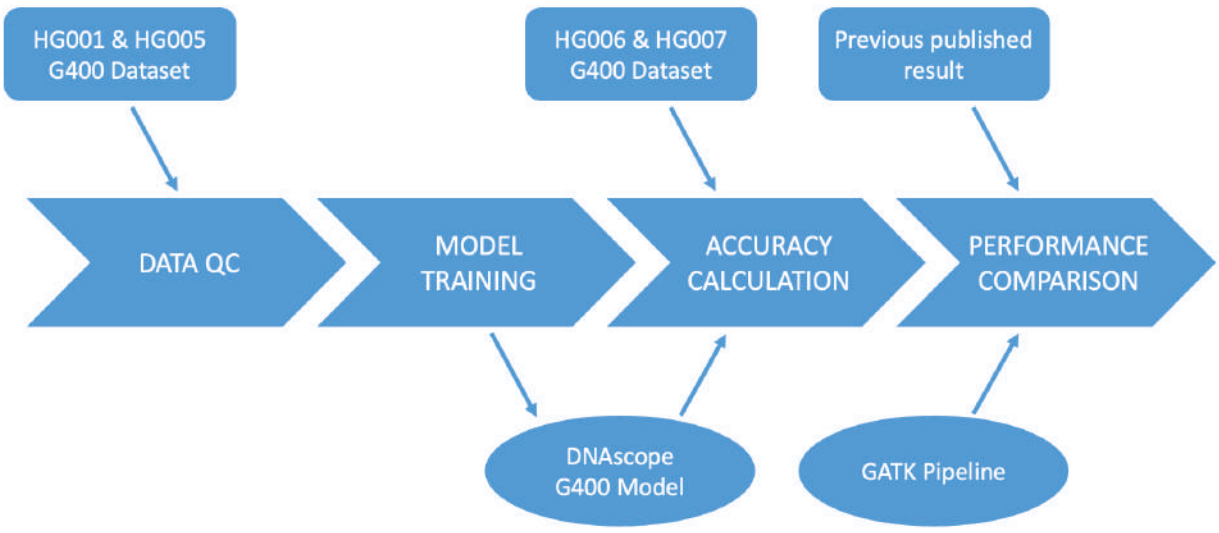

Figure 3. Overview of model training and benchmarking pipeline.

Training of DNAscope model was conducted with an updated model framework introduced in release 202112.01 of the Sentieon software package. Reference Datasets HG001 and HG005 were used as training dataset, with 20% of data randomly split for validation and chr20 held out for testing.

Four HG001 and HG005 PE150 datasets were mapped to the hg38 reference genome using Sentieon BWA. 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.

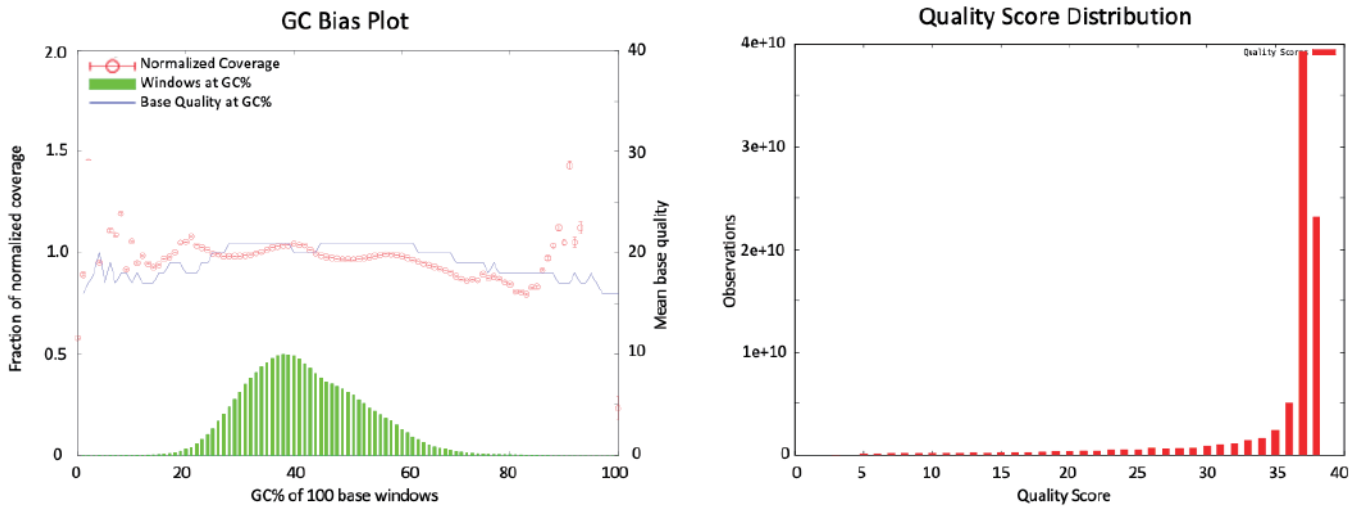

Figure 4. GC Bias Plot and Quality Score Distribution of HG006-PE150 30x dataset. Sequencing coverage is even across most GC windows even for extreme GC%, and most reads returned 35+ quality scores.

The four aligned datasets were down sampled 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.

Accuracy Benchmark

Testing datasets HG006 and HG007 were processed by Sentieon DNAscope pipeline version 202112.05 with the trained MGI 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. In addition, HG006 46x dataset (regularly generated per lane on DNBSEQ-G400 flow cell) returned higher F1-score to its corresponding 30x dataset (~40% less false positive errors), showing that higher depth indeed contributes to variant calling accuracy.

Reference Dataset	Sequencing Platform	Analysis Pipeline	Type	False Negative	False Positive	Recall	Precision	F1-Score
HG006-PE150 30x	DNBSEQ-G400	DNAscope MGI 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 MGI 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 MGI 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 MGI 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-G400	DNAscope MGI model v0.3	SNP	18,041	7,958	0.994	0.998	0.996
			INDEL	2,352	1,162	0.995	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. The DNBSEQG-400 dataset was referred to as "PCR-Free research lib" in Table 3 from the published benchamrk[7]; ILMN dataset accuracy is from recently published DNAscope white paper[9].

To better understand the accuracy improvements offered by the newly trained model and improved sequencing chemistry, we reanalyzed one of the previous published HG001 datasets7 generated from DNBSEQ-G400 with an optimized WGS library protocol and DNAscope MGI model v0.3, using the updated GIAB v4.2.1 benchmark dataset. 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 200MB 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.

The new DNBSEQ-G400 sequencing platform and trained DNAscope model worked well and reached higher accuracy comparing to previously published best WGS dataset and old DNAscope model. Main improvement is 20%-25% reduction of false negative and false positive variant calls. Considering faster processing speed of the new DNAscope model, the improved accuracy level is satisfying. It is also worth pointing out that the GATK pipeline is compatible to DNBSEQ-G400 dataset but accuracy is much lower than DNAscope.

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[9].

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. The ILMN dataset is displayed just for reference and should not be considered as a direct comparison, because 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 "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 in account for further improvements of WGS.

Conclusion

In this study we trained a new DNAscope model for the MGI DNBSEQ-G400 sequencing platform. The combined improvements in the underlying sequencing chemistry and the platform-specific model result in high variant calling accuracy. The elimination of PCR amplification in sequencing array prep greatly reduced errors introduced prior to sequencing. The retrained DNAscope model improves DNAscope's accuracy with DNBSEQ-G400 reads, allows DNAscope to more accurately model systematic error patterns, therefore enabling more accurate discrimination of false positive and false negative variant calls. The overall variant calling accuracy is improved from the previously published accuracy, as well as the current state-of-the-art sequencing and analysis integrated solutions.

References

1. R. Drmanac, R. Crkvenjakov. Prospects for a Miniaturized, Simplified and Frugal Human Genome Project. Scientia Yugoslavica 161(1-2), 97-107, 1990.
2. 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.
3. 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.
4. 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.
5. 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.
6. Fehlmann T, Reinheimer S, Geng C, et.al. cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. Clin Epigenetics. 2016 Nov 21;8:123.
7. Hanjie Shen, Pengjuan Liu, Zhanqing Li, et.al. Advanced Whole Genome Sequencing Using an Entirely PCR-free Massively Parallel Sequencing Workflow. bioRxiv 2019.12.20.885517.
8. Snezana Drmanac, Matthew Callow, Linsu Chen, et.al. CoolMPS™: Advanced massively parallel sequencing using antibodies specific to each natural nucleobase. bioRxiv 2020.02.19.953307.
9. 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.
10. 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.

Complete Genomics, part of MGI Americas | 2904 Orchard PKWY, San Jose CA USA 9513 | +1 (408) 648-2560
en.mgi-tech.com | MGI-service@mgi-tech.com

* FOR RESEARCH USE ONLY. NOT FOR USE IN DIAGNOSTIC PROCEDURES. NOT FOR USE IN USA

** Unless otherwise informed, StandardMPS and CoolMPS sequencing reagents, and sequencers for use with such reagents are not available in Germany, USA, Spain, UK, Hong Kong, Sweden, Belgium, Italy, Finland, Czech Republic, Switzerland and Portugal.