Adapting Google DeepVariant to Ultima Genomics Reads for Improved Variant Calling

Abstract

DeepVariant is a deep learning-based approach for variant calling which has been successfully applied across many varying sequencing technologies. We optimized DeepVariant for the recently introduced Ultima Genomics sequence data by improving the candidate generation step and extending input data representation to more fully accommodate the rich quality data information encoded in Ultima's data format. Following these improvements, DeepVariant demonstrates high variant calling accuracy on Genome-in-a-Bottle reference samples: F1=99.8% for SNPs and F1=97.8% for Indels in homopolymers up to length 10 across the vast majority (>98%) of the defined high-confidence regions of these samples. The complementarity between the results of DeepVariant and GATK suggest room for additional integration and optimization of both approaches.

Introduction

We recently introduced a novel sequencing platform[1] by Ultima Genomics (UG) with innovative components that enable scalable, high-throughput DNA sequencing and significantly reduce the consumable cost of a sequencing run, bringing the sequencing cost down to $1/Gb in the first implementation, with potential for even lower costs in the not distant future.

DeepVariant[2] is a deep learning-based variant caller that analyzes pileup image tensors containing several layers of information from reads overlapping with candidate variants. DeepVariant has been demonstrated to be a versatile method that can be readily adapted to yield superior variant calling performance to new sequencing technologies.[3] In this report, we describe the development of a modified DeepVariant algorithm optimized for UG data and assess the quality of the variant calls by comparing them to reference Genome-in-a-Bottle (GIAB) samples and their corresponding truth sets.[4]

Ultima Genomics sequencing data

The UG sequencer employs a novel mostly natural sequencing-by-synthesis (mnSBS) chemistry in which each flow cycle consists of a single base from a mostly natural nucleotide (MNN) mix of fluorescently labeled, and unlabeled, non-terminated nucleotides. In each sequencing cycle beads containing identical clonal DNA templates are exposed to the MNN mix and polymerase extension is performed to incorporate 0, 1, or a few bases of a single nucleotide base type (dA, dC, dG or dT) into each growing strand, depending on the length of the respective homopolymer in the corresponding template. mnSBS avoids quenching of fluorescent signals from adjacent labels and instead produces signals proportional to the lengths of homopolymers up to approximately 12 bases.[1]

To generate sequence reads from the optical signal generated per bead by mnSBS, the UG base calling algorithm employs a deep convolutional neural network (CNN) for homopolymer length classification per cycle (i.e 0-12) and outputs the most likely homopolymer sequence as well as probabilities of alternative homopolymer lengths per cycle. This information is used to generate base quality scores calibrated for the specific run. The sequencing base calling error modes produced by mnSBS chemistry differs from standard reversible terminator SBS chemistry. The base substitution error for this type of chemistry is expected to be extremely low since substitution errors could only be generated as a combination of two or more adjacent homopolymer errors, while the dominant errors are homopolymer length misclassifications. Typically, homopolymer calling accuracy is at 99.5% for homopolymer lengths of 1-2 and decreases to 90% at homopolymer lengths of 8.

For accurate calling of germline short variants from Whole Genome Sequencing (WGS) data, a variant calling algorithm should be calibrated to the

specific sequencing errors and biases in the raw data. Hence, standard variant calling methods that were developed specifically for reversible terminator chemistry need to be adapted for accurate analysis of UG data.

Adapting Google's DeepVariant to the Ultima Genomics reads

DeepVariant (DV) is a deep learning-based approach for variant calling. Like other variant callers, the approach is split into two main components: Initially, many potential candidate variants are generated, with emphasis on a high recall, then the candidate calls are filtered to optimize separation between true and false positives. In DV an image classification CNN is used in this latter step to classify each candidate as Homozygous reference (Ref), Heterozygous variant (Het), or Homozygous variant (Hom). The DV calling pipeline consists of three steps: make-examples, call-variants and post-processing. In make-examples, candidate variants are detected (in a similar fashion to GATK's HaplotypeCaller5) using very lenient calling thresholds, and an image containing base-calling and alignment information is generated for the neighborhood of each candidate. Different layers of information (e.g., bases, base qualities etc.) are stored in different image channels (similar to the way color images are represented in RGB channels). In call-variants images are inputted into a trained CNN and classified into one of Ref/Het/Hom classes. Each class is assigned a probability score corresponding to the network's confidence in its classification. In the final post-processing step, the encoded variants and their classifications are analyzed by a script which outputs the VCF file with variant calls and quality.

Two aspects of DV were modified for optimized performance on UG data: First, parameters for detection of candidate variants were tuned. Second, custom UG information channels were encoded to provide the network with additional information unique to UG data.

Candidate generation parameter tuning

The parameters for candidate generation were tuned as follows. Thresholds for minimum-base-quality and for density of realignment windows, which were previously defined based on different data with distinctive characteristics, were both tuned to minimize the number of false-positive variants in the candidate generation stage. This led to an overall increase in recall with a slight decrease in precision. Additionally, an increase in the minimal fraction of reads supporting indels helped to improve precision.

Custom information channels

UG base-qualities are encoded in the standard SAM format, providing probabilities for classification error in homopolymer length in the default base quality (BQ) field, the direction of error (i.e., more likely increase or decrease in length) encoded in a custom SAM tag called TP, and the probability for missed signal in any flow with zero length in a custom T0 tag. The TP and T0 tags are not utilized in the generic DV base quality (BQ) channel implementation and this channel was therefore replaced with three custom channels in the UG-specific implementation (Figure 1):

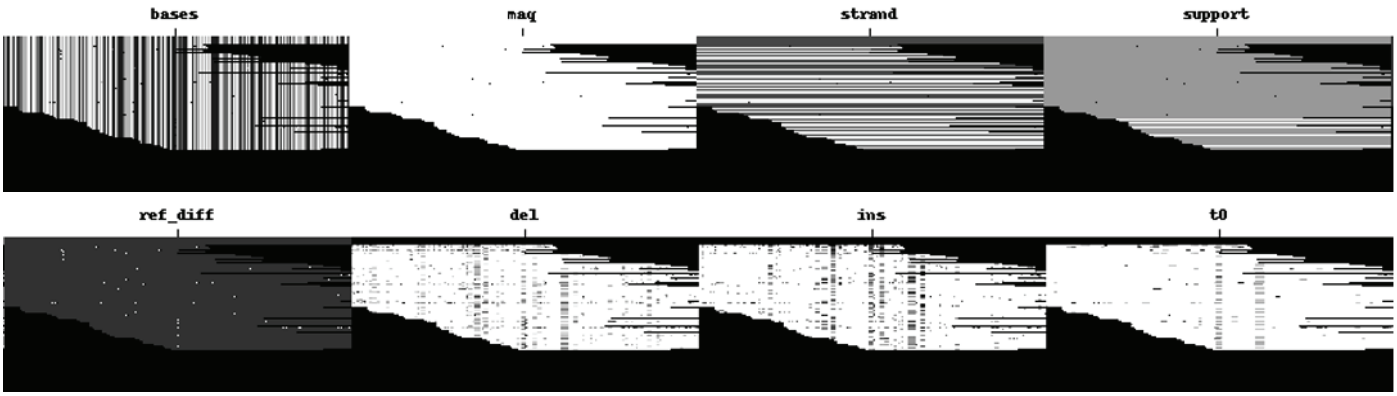

Figure 1. 8 channels in the UG implementation of DV, including the 3 custom channels (del, ins, t0)

- **homopolymer-insertion and homopolymer-deletion channels:** Instead of directly converting the BQ information into a single input channel, BQ and TP data are combined into two separate channels encoding separate probabilities for a homopolymer deletion and homopolymer insertion errors.
- **non-hmer-insertion (t0) channel:** To complete the representation of possible errors, the t0 flag is encoded as is. This channel thus represents the probability of missing one or more additional bases between every two called bases in each read.

Variant calling of Genome-in-a-Bottle reference samples

To assess the quality of the variant calls using DeepVariant, we used the reference dataset of seven standard Genome-in-a-Bottle (GIAB)[4] reference samples HG001-HG007. We trained the UG-optimized DV model using HG001 data and evaluated our performance with HG002-HG007 samples (see Methods). To control for overfitting, the model was trained on chr1-19, and tested on chr20.

We assessed the quality of the variant calls by comparing them to reference GIAB truth sets using the respective high-confidence regions (HCR).[4] To focus only on sequencing accuracy, we excluded homopolymer regions of length≥11, low-complexity regions, and tandem-repeats and low mappability regions. In total, we maintain 98.2% of the original HCR, (referred to as UG-HCR, see Methods).

The overall concordance of SNP variant calls over the UG-HCR was F1=99.8% (recall=99.7%, precision=99.9%) and of indel variant calls it was F1=97.8% (recall=97.1%, precision=98.6%). Variant calling accuracy in the entire GIAB-HCR (excluding homopolymers of length ≥11) was F1=99.2% for SNPs

Table 1. Variant Calling performance of UG-modified GATK and DV algorithms on GIAB reference samples. HG001 is not included since it was used to train the DV model

	GIAB-HCR (excludes HP>=11bp)						UG-HCR					
GATK	INDEL			SNP			INDEL			SNP		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
HG002	90.4%	94.5%	86.6%	99.0%	99.4%	98.6%	96.6%	96.8%	96.4%	99.6%	99.6%	99.6%
HG003	90.8%	95.0%	87.0%	99.0%	99.4%	98.6%	96.8%	97.1%	96.6%	99.6%	99.6%	99.6%
HG004	89.9%	93.9%	86.2%	99.1%	99.4%	98.7%	95.9%	96.4%	95.4%	99.7%	99.6%	99.7%
HG005	91.5%	94.4%	88.7%	99.1%	99.4%	98.8%	96.5%	97.0%	96.0%	99.7%	99.6%	99.7%
HG006	89.9%	93.3%	86.8%	99.0%	99.3%	98.6%	96.1%	96.2%	95.9%	99.6%	99.6%	99.6%
HG007	91.0%	94.3%	87.9%	99.1%	99.4%	98.8%	96.3%	96.7%	96.0%	99.7%	99.6%	99.7%
Average of GATK	**90.6%**	**94.2%**	**87.2%**	**99.0%**	**99.4%**	**98.7%**	**96.4%**	**96.7%**	**96.0%**	**99.6%**	**99.6%**	**99.7%**

GATK	INDEL			SNP			INDEL			SNP		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
HG002	91.6%	97.1%	86.6%	99.1%	99.8%	98.5%	97.7%	98.5%	96.9%	99.8%	99.9%	99.6%
HG003	92.0%	97.3%	87.2%	99.1%	99.8%	98.5%	97.9%	98.6%	97.1%	99.8%	99.9%	99.6%
HG004	91.9%	97.0%	87.3%	99.2%	99.8%	98.6%	97.7%	98.5%	96.9%	99.8%	99.9%	99.7%
HG005	93.3%	97.2%	89.6%	99.2%	99.8%	98.6%	98.0%	98.7%	97.4%	99.8%	99.9%	99.7%
HG006	92.0%	97.3%	87.2%	99.1%	99.8%	98.5%	97.6%	98.5%	96.8%	99.8%	99.9%	99.6%
HG007	92.9%	97.3%	88.9%	99.2%	99.8%	98.6%	98.0%	98.7%	97.4%	99.8%	99.9%	99.7%
Average of DV	**92.2%**	**97.2%**	**87.8%**	**99.2%**	**99.8%**	**98.6%**	**97.8%**	**98.6%**	**97.1%**	**99.8%**	**99.9%**	**99.7%**

Figure 2. Numbers of false positive variants and false negative variants that are unique to each caller (or common to both) were counted in the HG003 genome. In order to focus on variant detection ability, the classification did not consider the assignment of genotypes.

and F1=92.3% for indels, suggesting that a significant fraction of the variant calling errors in this region are indeed related to low-complexity DNA and can likely be improved by optimizing the clonal amplification protocol (Table 1).

Comparing DeepVariant performance with the UG-modified GATK HaplotypeCaller algorithm[1] on the same reference dataset demonstrates slightly better accuracy in SNP calling, but significantly better indel calling performance (both in recall and precision) with F1 increasing on average by 1.4%

A common mode of error of both tools was an error in genotyping (when a homozygous variant was called as heterozygous or vice versa) or in allele calling (when a slightly different allele was called instead of a true one, e.g., insertion of AAT instead of AT). Specifically, in DV call-sets 32-37% of errors were genotyping errors and additional 8-10% of errors were allele calling errors. In the GATK call-sets 30-45% of errors were genotyping errors and 12-14% of errors were allele calling errors. We believe that many of these errors can be corrected by downstream analysis.

Interestingly, many of the FN and FP events are observed by only one of the methods, suggesting that an integrated approach may combine the unique advantages of both methods can improve results even further (Figure 2).

The numbers of false positive variants and false negative variants that are unique to each caller (or common to both) were counted in the HG003 genome. In order to focus on variant detection ability, the classification did not consider the assignment of genotypes.

Summary and Discussion

DeepVariant is a deep learning-based approach for variant calling which has been demonstratively adaptable for use over a range of sequencing technologies with widely varying characteristics. As such, it is not surprising that this framework is readily extensible to accommodate data from the new UG sequencing platform.

Extending the input channels of DeepVariant to accommodate the rich quality information encoded in the UG data format allows the neural network to fully take advantage of the unique attributes of the technology in a way that surpasses the performance of the default network structure.

Using the extended representation, applying DeepVariant variant calling to GIAB standard reference genomes HG002-7 demonstrates high sequencing accuracy for SNPs (99.8%) and Indels in homopolymers up to length 10 (97.8%) across the vast majority (>98%) of the defined high-confidence regions of these samples.

While this performance already compares favorably to other variant calling methods, there is clearly room for additional optimization. Specifically, the difference in erroneous calls (both false positives and false negatives) produced by DV and GATK suggests that the complementarity between the two approaches can be exploited by integrated methods that combine unique data models as well as generic deep learning methodologies.

Methods

DeepVariant training

A customized version of DeepVariant v1.3 which can generate examples with UG custom channels was used to train a DeepVariant inception model on chromosomes 1-19 of HG001 (40X coverage) and validated on chr21: 10000000-20000000. Training and validation examples were generated on GIAB_4.2 high-confidence regions and ground truth vcf file. DeepVariant's make-examples stage was run with the following non-default parameters --min_base_quality 5, --vsc_min_fraction_indels 0.12, --dbg_min_base_quality 0. The model was trained for eight epochs with learning-rate of 0.005, and the best checkpoint was selected according to maximal accuracy on the validation set. The selected model was used to generate vcf files of HG001-HG007 GIAB samples. In inference, the same parameters and channels are used except for an additional parameter: --ws_min_windows_distance 20.

Variant Calling Performance Evaluation

Variant calling performance (recall, precision F1), was calculated using vcfeval[6] to compare single sample variant callset (vcf) with GIAB truth set (v4.2.1)[4] for reference samples HG001-7.

The evaluation region was defined as the corresponding GIAB high-confidence region (HCR v4.2.1), with the following exclusions [% of full HCR]:

- GIAB-HCR (total 99.6% of full HCR):
  - Homopolymer regions of length 11 and higher + 4 flanking bases [0.4%]
- UG-HCR (total 98.2% of full HCR):
  - Homopolymer regions of length 11 and higher + 4 flanking bases [0.4%]
  - AT-rich regions: all 40 bp regions with 95% or higher AT content [0.3%]
  - Short tandem repeats regions, with specific defined thresholds [0.3%]
  - Low mappability and coverage: 50bp regions with mean mappable coverage>15X in 90 independent samples [1.1%]

Bed Files containing specific exclusion areas will be available as supplementary material.

Supplementary Materials

Supplementary data will be made available in the near future at www.ultimagenomics.com/support

Learn more at **ultimagenomics.com/support**

References

1. Almogy, G. et al. Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. http://biorxiv.org/lookup/doi/10.1101/2022.05.29.493900 (2022) doi:10.1101/2022.05.29.493900.
2. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36, 983–987 (2018).
3. https://ai.googleblog.com/2020/09/improving-accuracy-of-genomic-analysis.html.
4. Wagner, J. et al. Benchmarking challenging small variants with linked and long reads. Cell Genomics 2, 100128 (2022).
5. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491–498 (2011).
6. Cleary, J. G. et al. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. http://biorxiv.org/lookup/doi/10.1101/023754 (2015) doi:10.1101/023754.
7. Byrska-Bishop, M. et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. http://biorxiv.org/lookup/doi/10.1101/2021.02.06.430068 (2021) doi:10.1101/2021.02.06.430068.