018

precisionFDA Truth Challenge V2: Calling Variants from Short- and Long- Reads in Difficult-to-Map Regions

Ezekiel Maier [1], Nathan D. Olson[2], Justin Wagner[2], Jennifer McDaniel[2], Justin Zook[2], Holly Stephens[1], Samuel Westreich[3], Prasanna Anish[1], Elaine Johanson[4], Boja Emily[4], Omar Serang[3], Sean Watford[1]; [1] Booz Allen Hamilton, [2] National Institute of Standards and Technology, [3] DNAnexus, [4] FDA Office of Digital Transformation/Office of Data, Analytics, and Research

Abstract

The Truth Challenge V2 (May to June 2020) assessed state-of-the-art variant calling in difficult-to-map regions and the Major Histocompatibility Complex. Participants generated variant calls as Variant Call Format (VCF) files for sequencing data HG002, HG003, and HG004 originally given as FASTQ files. Sequencing data were provided from Illumina, Pacific Biosciences, and Oxford Nanopore Technologies, at 35X and 50X coverage, respectively. The variant calls were generated against the GRCh38 version of the human reference genome. Submissions were evaluated based on the harmonic mean of parents' F1 scores for combined single nucleotide variants (SNVs) and insertions or deletions (INDELSs). From the 64 submissions, top performers came from Sentieon, Roche Sequencing Solutions, The Genomics Team in Google Health, DRAGEN, Seven Bridges Genomics, The University of California Santa Cruz Computational Genomics Lab (UCSC CGL) and Google Health, and Wang Genomics Lab. The top performing submissions combined all 3 technologies. The performance of each submission varied across stratifications, in which the best-performing multi-technology call sets had similar performances overall. 90% of submissions for long-read-only used deep-learning (DL)-based methods. The short-read submissions with the best performance used statistical variant-calling algorithms with graph reference. The addition of DL and machine learning (ML) have advanced variant calling by enabling faster adoption of new sequencing technologies.

Introduction

- PrecisionFDA provides access to high-performance computing instances, experts, tools, challenge framework, and virtual shared Spaces where scientists and reviewers can securely collaborate with external partners.
- The first Genome in a Bottle (GIAB) precisionFDA Truth Challenge (2016), asked participants to call small variants from short-reads for two GIAB samples (HG001 & HG002).
  - Benchmarks for HG001 were previously published, but no benchmarks for HG002 were publicly available at the time.
  - This was the first blinded germline variant calling challenge, and results have been used as a point of comparison for new variant calling methods.
  - Performance was only assessed on "easy" genomic regions accessible to the short-reads used to form the v3.2 GIAB benchmark sets.
- Due to advances in genome sequencing, variant calling, and an expanded GIAB benchmark set (mother, father, son), we conducted a follow up truth challenge in 2020.
- The Truth Challenge V2 occurred when the v4.1 benchmark was available for HG002, but only the v3.3.2 benchmark was available for HG003 and HG004.
- The challenge included a short-read dataset (Illumina) and long-read datasets (PacBio and Oxford Nanopore Technologies (ONT)) to assess performance across a variety of data types.
- This challenge used benchmark tools and stratification Browser Extensible Data (BED) files (from Global Alliance for Genomics and Health (GA4GH) Benchmarking Team and GIAB) to assess performance in difficult genomic regions.

Materials and Methods

- Participants were tasked with generating variants calls as VCF files (**Figure 1**) for the GIAB Ashkenazi Jewish Trio (HG002, HG003, HG004).
- Twenty teams submitted 64 unique challenge submissions.
- Challenge participants submitted variant callsets that were generated using one or more sequencing technologies:
  - Illumina
  - PacBio HiFi
  - ONT
- For single technology submissions, Illumina was the most common (55%), followed by PacBio (38%), and ONT (7%).
- Of the multiple technology submissions, Pacbio was used in all twenty, Illumina was used in all but one, and seven submissions used data from all three technologies.
- Submissions used a variety of variant calling methods based on ML (e.g., DeepVariant), graph (e.g., DRAGEN and Seven Bridges), and statistical (e.g., Genomic Analysis Toolkit (GATK)) methods.
- Notably, a majority of submissions used ML based variant calling methods.
  - This was particularly true for long-read and multi-technology submissions, with 37/40 using an ML-based method.
- Submissions were evaluated based on the averaged parents' F1 scores for combined SNVs and INDELs.

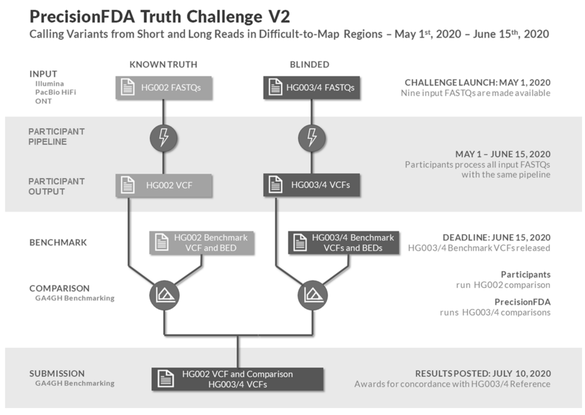

Figure 1. Truth Challenge V2 structure. Participants were provided sequencing reads from Illumina, PacBio HiFi, and ONT for the GIAB Ashkenazi trio (HG002, HG003, and HG004). Participants uploaded VCF files for each individual of the trio before the end of the challenge, and then the new benchmarks for HG003 and HG004 were made public.

Results and Discussion

- In all benchmark regions, the top performing submissions combined all technologies, followed by PacBio HiFi, Illumina, and ONT, with PacBio HiFi submissions having the best single-technology performance in each category (**Figure 2**).
- Variant calls based on ONT performed better than Illumina in difficult-to-map regions despite ONT's higher INDEL error rate (**Figure 2A**).
- ONT-based variant calls had higher F1 scores in difficult-to-map regions than in all benchmark regions (**Figure 2A**).
- Top-performing short-read callsets used graph-based approaches, while top-performing long-read callsets used ML.
- Performance varied substantially across stratifications (**Figure 2B**).
- Top-performing multi-technology call sets had similar overall performance, although with error rates varying by a factor of 10 in the major histocompatibility complex (MHC).
- Comparing performance for blinded and semi-blinded samples revealed possible over-tuning of some methods (**Figure 3**).
- Improved benchmark sets and stratifications revealed innovation in sequencing technologies and variant calling, since the 2016 challenge.
- New stratifications enabled better comparison of method strengths.

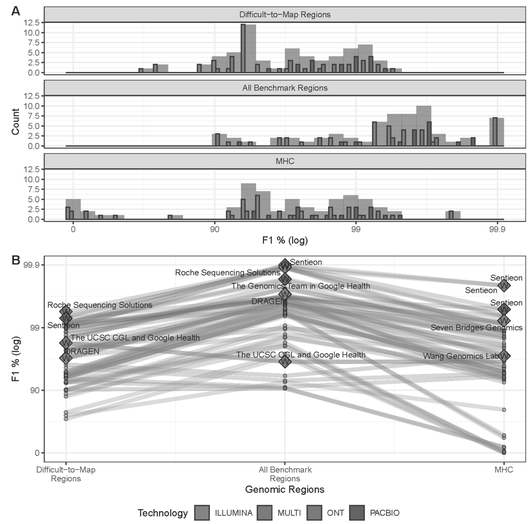

Figure 2. Overall Performance (A) and submission rank (B) varied by technology and stratification (log scale).

Table 1. Summary of Challenge Top Performers. One winner was selected for each Technology/Genomic Region combination, and multiple winners were awarded in the case of ties. Winners were selected based on submission F1 score (SNV plus INDELs) for the blinded samples, HG003 and HG004.

Technology	Genomic Region	Participant	F1
MULTI	All Benchmark Regions*	Sentieon	0.999
MULTI	All Benchmark Regions*	Roche Sequencing Solutions	0.999
MULTI	All Benchmark Regions*	The Genomics Team in Google Health	0.999
MULTI	Difficult-to-Map Regions	Roche Sequencing Solutions	0.994
MULTI	MHC	Sentieon	0.998
ILLUMINA	All Benchmark Regions	DRAGEN	0.997
ILLUMINA	Difficult-to-Map Regions	DRAGEN	0.969
ILLUMINA	MHC	Seven Bridges Genomics	0.992
PACBIO	All Benchmark Regions	The Genomics Team in Google Health	0.998
PACBIO	Difficult-to-Map Regions	Sentieon	0.993
PACBIO	MHC	Sentieon	0.995
ONT	All Benchmark Regions	The UCSC CGL and Google Health	0.965
ONT	Difficult-to-Map Regions	The UCSC CGL and Google Health	0.983
ONT	MHC	Wang Genomics Lab	0.972

* Tied

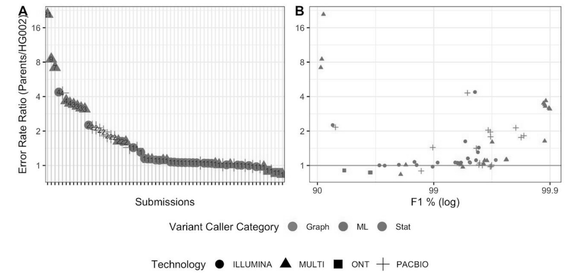

Figure 3. Ratio of error rates using semi-blinded parents' benchmark vs. public son's benchmark. (A) Submissions ranked by error rate ratio. (B) Comparison of error rate ratio to the overall performance for the parents (F1 in all benchmarking regions). Error rate defined as 1– F1.

Conclusion

- Public community challenges, like the precisionFDA Truth Challenges help drive methods development.
- Ground-breaking mapping+variant calling pipelines were developed, optimized, and made available as part of this challenge.
- Innovative ML-based methods were developed for long reads.
- Along with the new benchmark set and sequencing data types, new genomic stratifications were used to evaluate submission performance in different contexts, highlighting methods that performed best in particularly challenging regions.
- This challenge spurred the development and public dissemination of a diverse set of new bioinformatics methods for multiple technologies, thus driving the advancement of research and clinical sequencing.