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

Nathan D. Olson[1], Justin Wagner[1], Jennifer McDaniel[1], Justin Zook[1], Holly Stephens[2], Samuel Westreich[3], Prasanna Anish[2], Elaine Johanson[4], Boja Emily[4], Omar Serang[3], Sean Watford[2], Ezekiel Maier [2]

[1] National Institute of Standards and Technology, [2] Booz Allen Hamilton, [3]DNAnexus, [4] FDA's Office of the Chief Scientist/Office of Health Informatics

Abstract

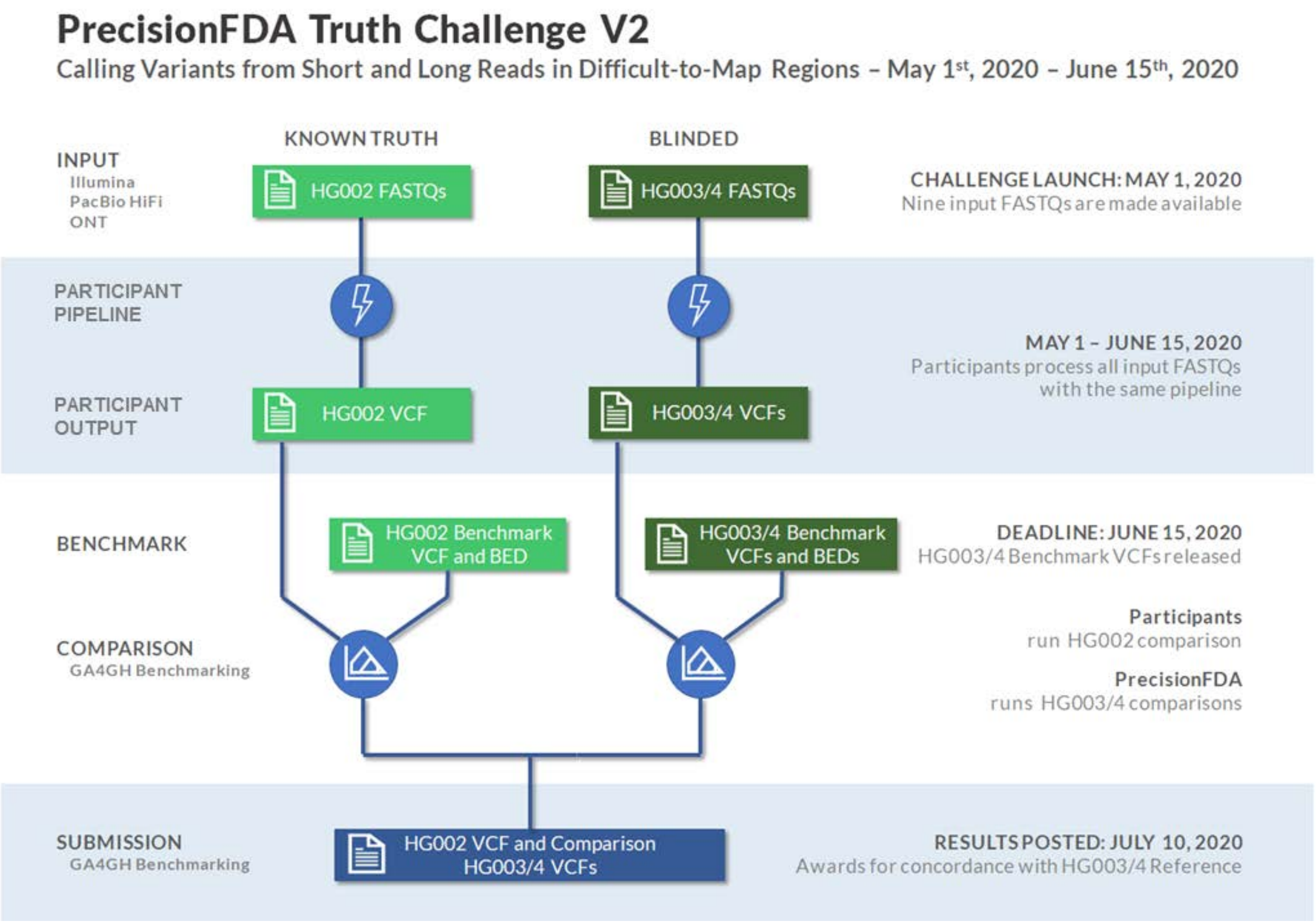
The precisionFDA Truth Challenge V2 aimed to assess variant calling in difficult-to-map regions and the Major Histocompatibility Complex (MHC). Sixty-four submissions were received from twenty participants. A submission included a variant callset for one or more sequencing technologies (Illumina, PacBio HiFi, and Oxford Nanopore Technologies) as well as a methodology description. Submissions were evaluated following best practices for benchmarking small variants with the new Genome In A Bottle (GIAB) benchmark sets and genome stratifications. Challenge submissions included innovative methods for all three technologies, with graph-based and machine-learning methods scoring best for short-read and long-read datasets, respectively. New methods out-performed the Truth Challenge V1 winners. Recent developments in sequencing and variant calling enabled participants to benchmark variants in challenging genomic regions, paving the way for the identification of previously unknown clinically relevant variants.

Introduction

- The precisionFDA platform provides access to high-performance computing instances, a community of experts, a library of publicly available tools, a challenge framework, and virtual shared Spaces where FDA scientists and reviewers can securely collaborate with external partners
- The first 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 from two technologies (PacBio and ONT) to assess performance across a variety of data types.
- This challenge used benchmark tools and stratification BED files developed by the GA4GH Benchmarking Team and GIAB to assess performance in difficult genomic regions

Materials and Methods

- Participants were tasked with generating variants calls as variant call (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
    - Oxford Nanopore Technologies (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 machine learning (ML; e.g., DeepVariant), graph (e.g., DRAGEN and Seven Bridges), and statistical (e.g., GATK) methods
- Notably, a majority of submissions used machine learning (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


PrecisionFDA Truth Challenge V2

Calling Variants from Short and Long Reads in Difficult-to-Map Regions – May 1st, 2020 – June 15th, 2020

KNOWN TRUTH

BLINDED

INPUT
Illumina
PacBio HiFi
ONT

HG002 FASTQs

HG003/4 FASTQs

CHALLENGE LAUNCH: MAY 1, 2020
Nine input FASTQs are made available

PARTICIPANT PIPELINE

MAY 1 – JUNE 15, 2020
Participants process all input FASTQs with the same pipeline

PARTICIPANT OUTPUT

HG002 VCF

HG003/4 VCFs

BENCHMARK

HG002 Benchmark VCF and BED

HG003/4 Benchmark VCFs and BEDs

DEADLINE: JUNE 15, 2020
HG003/4 Benchmark VCFs released

COMPARISON
GA4GH Benchmarking

Participants
run HG002 comparison

PrecisionFDA
runs HG003/4 comparisons

SUBMISSION
GA4GH Benchmarking

HG002 VCF and Comparison HG003/4 VCFs

RESULTS POSTED: JULY 10, 2020
Awards for concordance with HG003/4 Reference

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 (Fig 2)
- Variant calls based on ONT performed better than Illumina in difficult-to-map regions despite ONT's higher INDEL error rate (Fig 2A)
- ONT-based variant calls had higher F1 scores in difficult-to-map regions than in all benchmark regions (Fig 2A)
- Top-performing short-read callsets used graph-based approaches, while top-performing long-read callsets used ML
- Performance varied substantially across stratifications (Fig 2B)
- Top-performing multi-technology callsets had similar overall performance, although with error rates varying by a factor of 10 in the (MHC)
- Comparing performance for blinded and semi-blinded samples revealed possible over-tuning of some methods (Fig 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

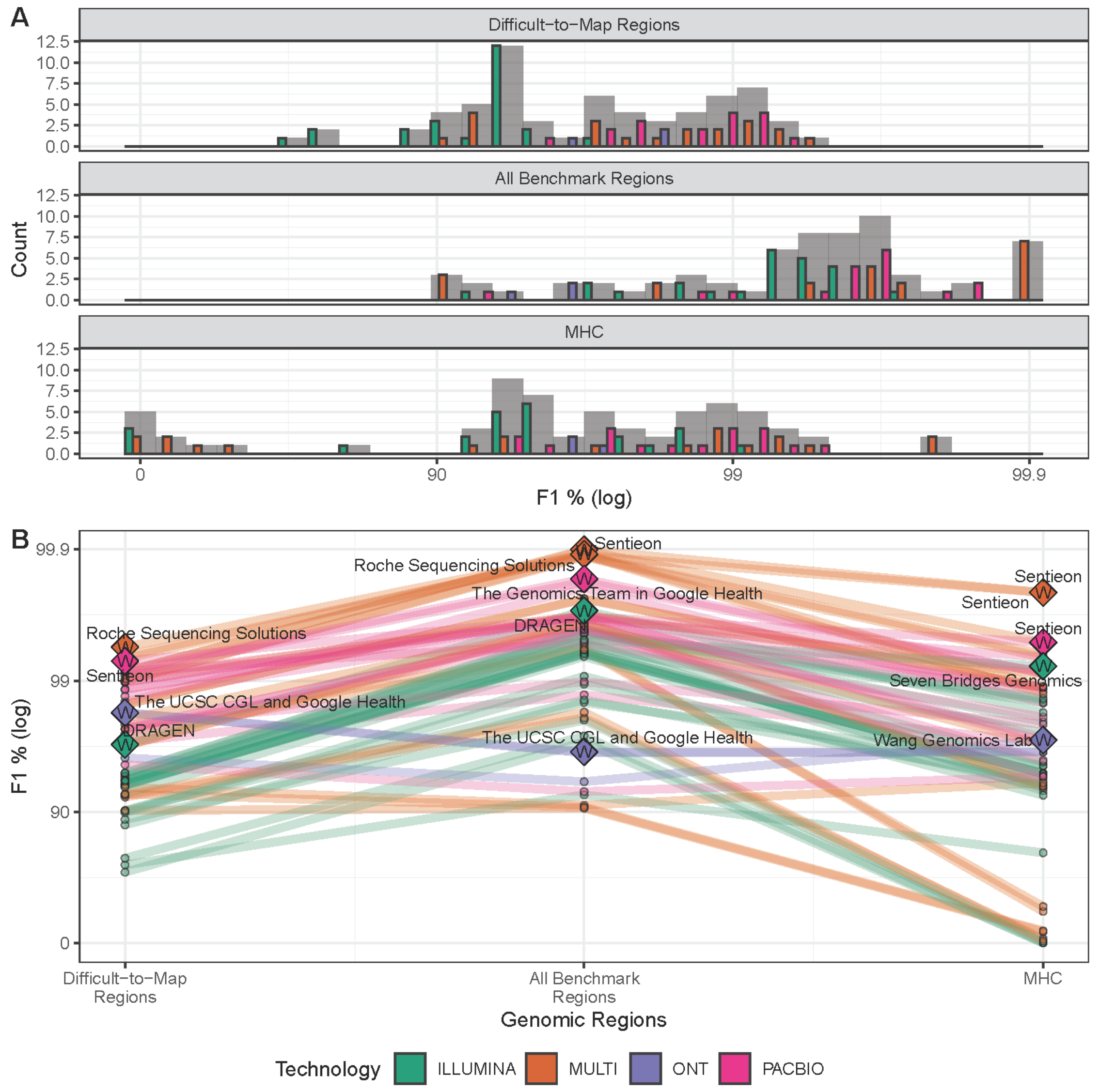


Figure 2. Overall Performance (A) and submission rank (B) varied by technology and stratification (log scale).

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

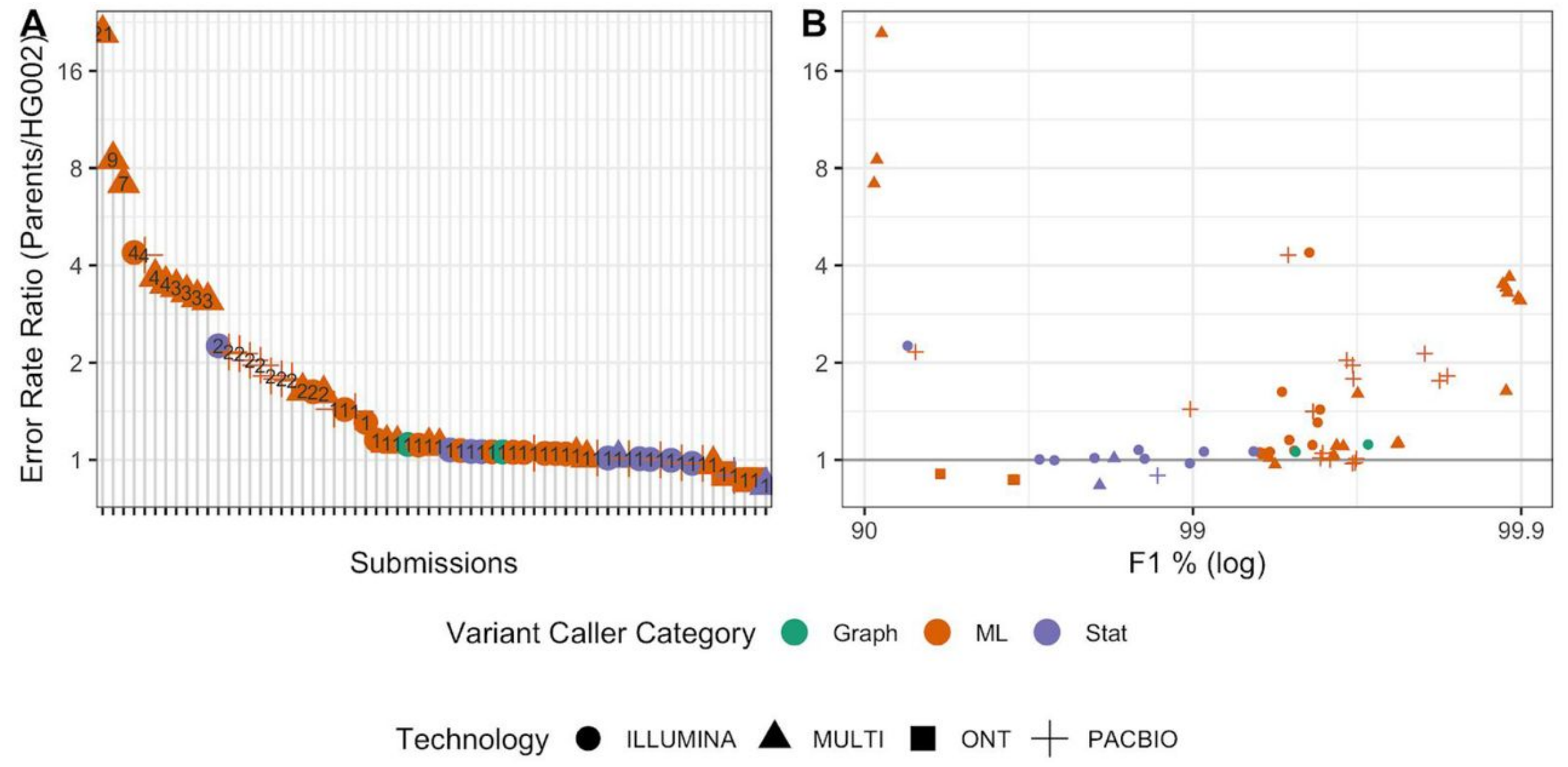


Figure 3. Ratio of error rates using semi-blinded parents' benchmark vs. public son's benchmark. (A) Submissions ranked by error rate ratio. (B) Comparison of error rate ratio to the overall performance for the parents (F1 in all benchmarking regions). Error rate defined as 1– F1.

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 machine learning-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