

precisionFDA Truth Challenge V2: Calling variants from short- and long- reads in difficult-to-map regions

Nathan D. Olson¹, Justin Wagner¹, Jennifer McDaniel¹, Justin Zook¹, Holly Stephens², Samuel Westreich³, Prasanna Anish², Elaine Johanson⁴, Boja Emily⁴, Omar Serang³, Sean Watford², Ezekiel Maier²

¹ National Institute of Standards and Technology, ² Booz Allen Hamilton, ³DNAnexus, ⁴ FDA's Office of Data, Analytics, and Research/Office of Health Informatics



FDA

Abstract

The precisionFDA Truth Challenge V2 aimed to assess variant calling in difficult-to-map regions and the Major Histocompatibility Complex (MHC). Twenty challenge participants applied their variant calling pipelines and submitted 64 variant callsets for one or more sequencing technologies (Illumina, PacBio HiFi, and Oxford Nanopore Technologies). 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 v3.3.2 benchmark was available for HG003 and HG004.
- The challenge included short-read and long-read datasets from two technologies 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

- Figure 1** illustrates the challenge participant workflow
- 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 (ONT)
- For single technology submissions, Illumina was the most common (24 out of 44), followed by PacBio (17), and ONT (3).
- 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

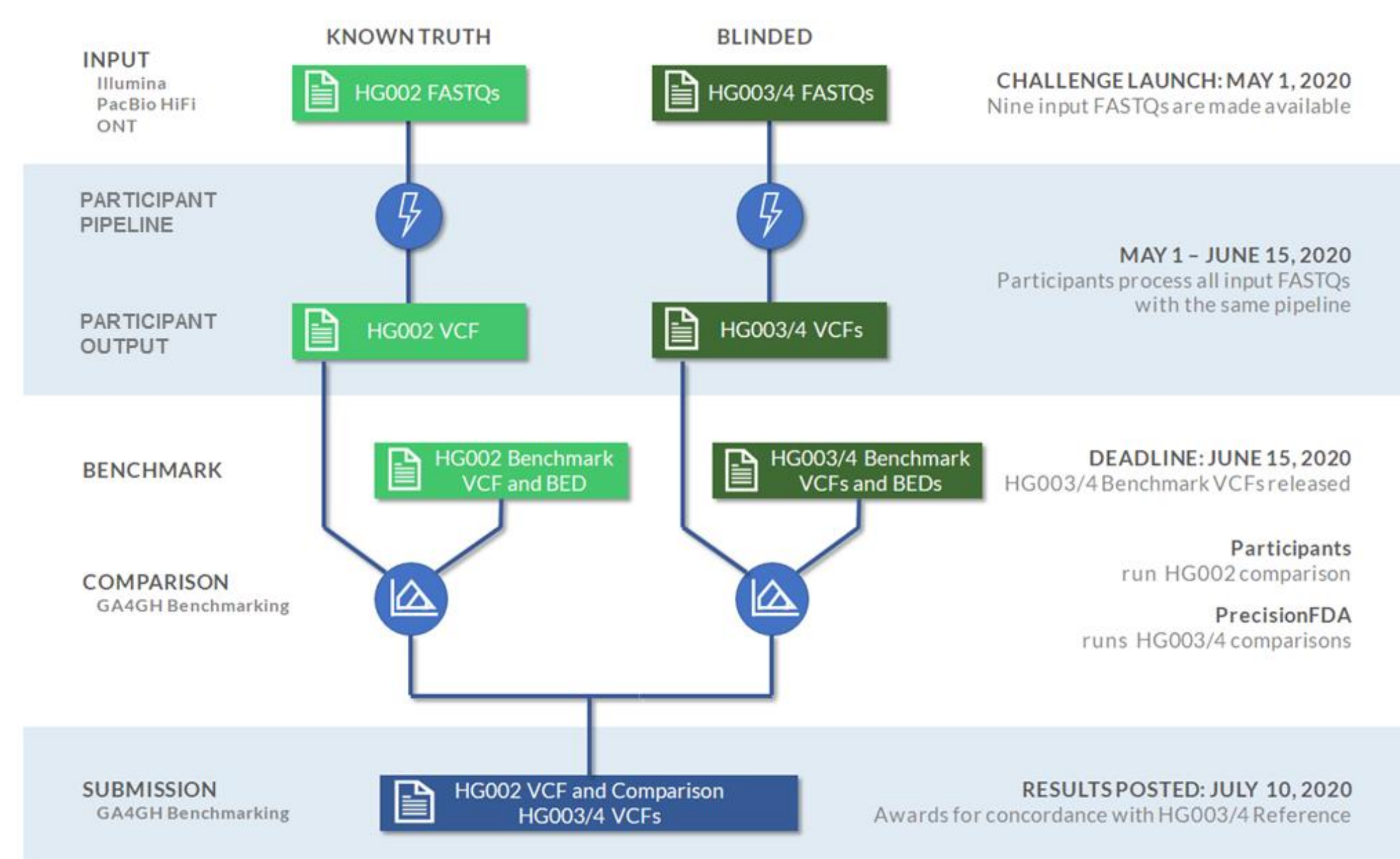


Figure 1. Truth Challenge V2 structure. Participants were provided sequencing reads from Illumina, PacBio HiFi, and ONT for the GIAB Ashkenazi trio (HG002, HG003, and HG004). Participants uploaded VCF files for each individual before the end of the challenge, and then the new benchmarks for HG003 and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

Results and Discussion

- Figure 2** shows overall performance and rank (F1) by technology and stratification
- Table 1** shows a summary of top performers
- Figure 3** shows F1 performance across genomic stratifications relative to the MHC
- The MHC region is more easily resolved with long-read-based methods as these are more likely to map in this region of high variability
- 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.
- Variant calls based on ONT performed better than Illumina in difficult-to-map regions despite ONT's higher INDEL error rate.
- ONT-based variant calls had higher F1 scores (accuracy) in difficult-to-map regions than in all benchmark regions.
- The best-performing short-read callsets used graph-based approaches, while best-performing long-read callsets used ML
- Performance varied substantially across stratifications, with the best-performing multi-technology callsets having similar overall performance
- 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.

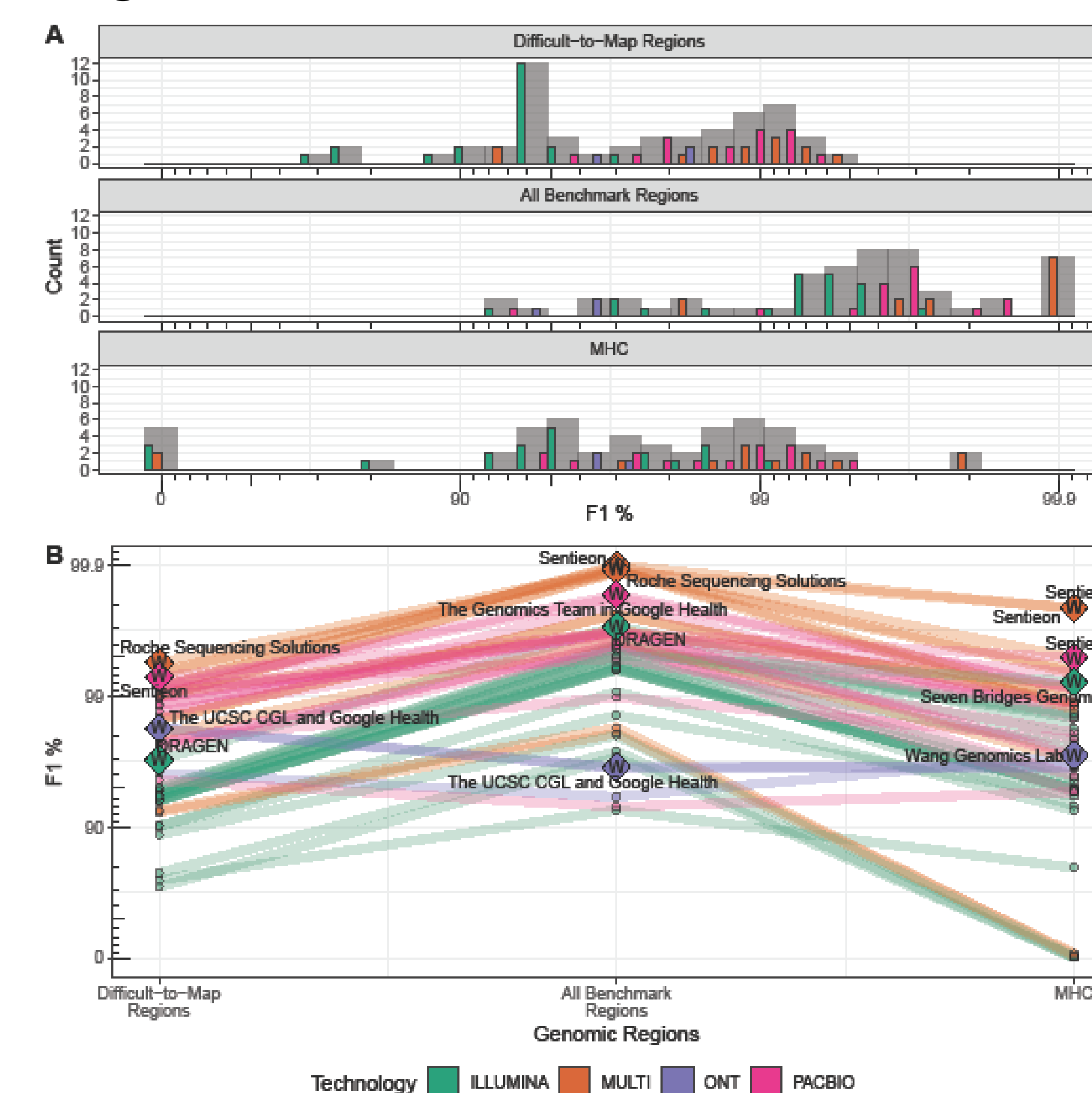


Figure 2. Overall Performance (A) and submission rank (B) varied by technology and stratification (log scale).

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	0.989
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	0.98
PACBIO	MHC	Sentieon	0.98
ONT	All Benchmark Regions	The UCSC CGL and Google Health	0.98
ONT	Difficult-to-Map Regions	The UCSC CGL and Google Health	0.98
ONT	MHC	Wang Genomics Lab	0.97

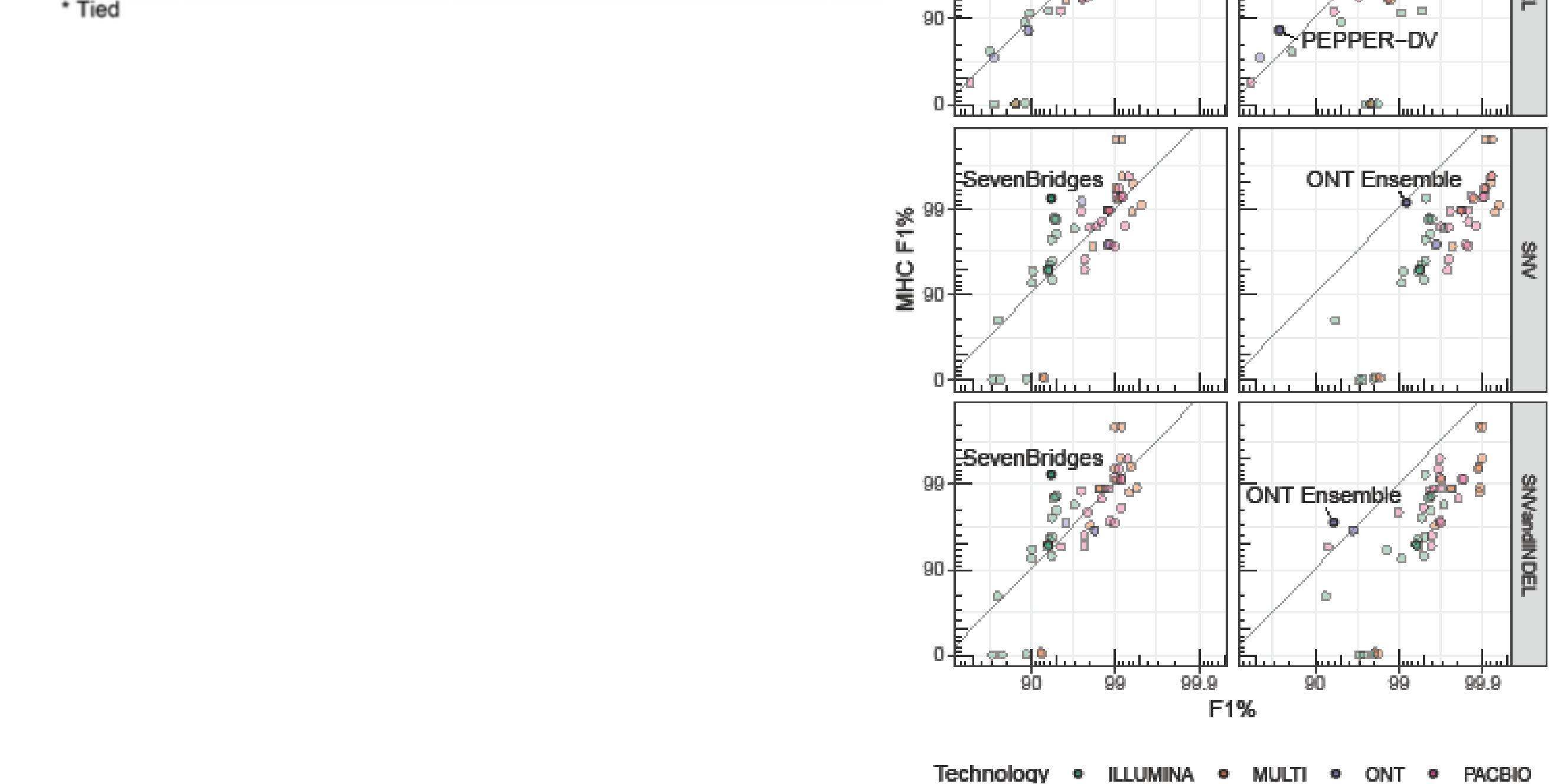


Figure 3. Submission performance comparison for the F1 metric between all benchmark regions and difficult-to-map regions, relative to MHC. F1 is plotted on a scale with axis labels and ticks indicating F1 percentage values. Points above the diagonal black line perform better in MHC relative to all benchmark regions or the difficult-to-map regions.

Conclusion

- Public community challenges, like the precisionFDA Truth Challenges help drive methods development.
- A number of 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.