New diploid assembly-based Genome In A Bottle

HG002 T2T “Q100” small and structural variant benchmark sets.

N. D. Olson[1], N. Dwarshuis[1], J. McDaniel[1], J. Wagner[1], J. M. Zook[1], T2T Q100 Consortium, and the GIAB Consortium

[1]National Institute of Standards and Technology (NIST)

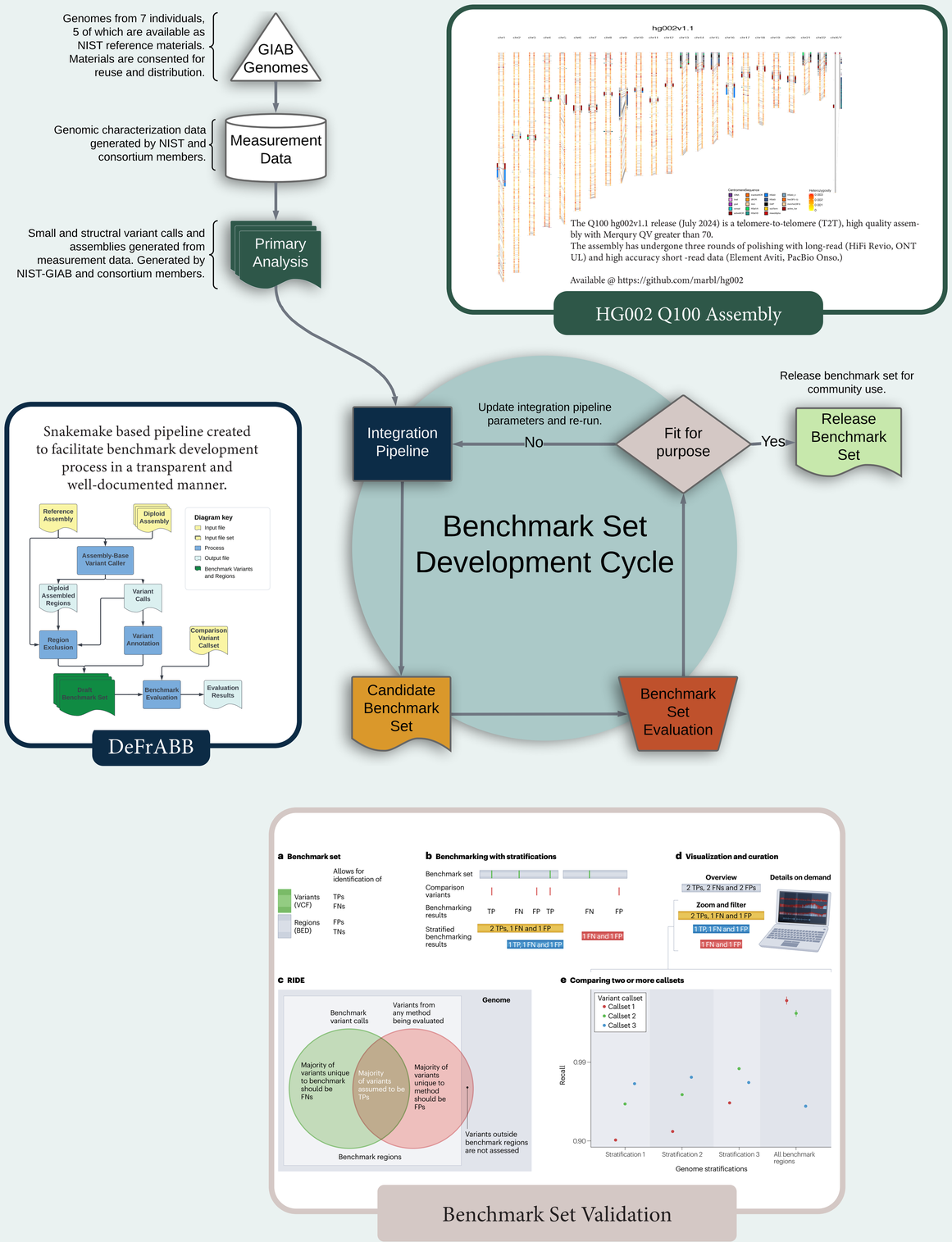
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

TELOMERE-TO-TELOMERE CONSORTIUM

ABSTRACT

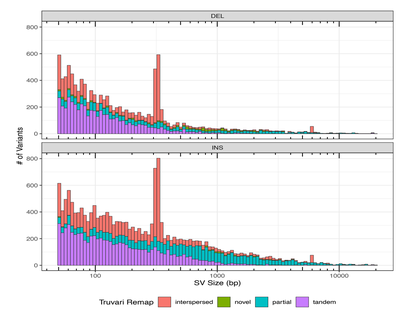
Genome in a Bottle was started over ten years ago to develop well-characterized human genome reference materials for validating genome sequencing and variant calling methods. GIAB has produced genomic DNA reference materials and benchmark sets composed of high-confidence variant calls and regions confidently identified as homozygous reference or as high-confidence variants. These benchmark sets are widely used to evaluate and train variant calling methods. Traditionally, these benchmark sets are developed by integrating variant callsets generated using different sequencing technologies and variant calling algorithms. However, this process involves mapping reads to a reference, which is limited in its ability to characterize more complex genomic regions. With recent advances in sequencing methods and genome assembly algorithms, creating high-quality diploid de novo genome assemblies is possible. A subgroup of the T2T consortium that generated the first complete human genome assembly, CHM13, is working on a highly polished and curated T2T diploid genome assembly of GIAB HG002 (son of the Ashkenazi Jewish trio.) While previous GIAB benchmark sets have used diploid assemblies for particular genomic regions, we use the HG002 Q100 T2T diploid assembly to generate the first whole genome assembly-based small and structural variant benchmark sets. These benchmark sets were generated using a new Snakemake-based pipeline and include regions of the genome that were excluded when using read-mapping-based benchmark generation methods, e.g., the highly polymorphic MHC, gene conversions, and more complex SVs. The small variant benchmark set includes ~ 89.8% (2.74 Gbp) of the HG002 assembly with >3.6 million SNVs and ~950,00 indels (>50bp), >700,000more variants than the previous v4.2.1 benchmark. The structural variant benchmark set covers nearly 90.3% (2.75 Gbp) of the HG002 assembly and nearly 30,000 variants. This represents a significant improvement over the previous v0.6 HG002 SV benchmark, which covered only 2.51Gbp and included ~10k variants. Draft versions of the benchmark sets are publicly available and undergoing external validation before official release.

BENCHMARK SET GENERATION



VARIANT BENCHMARK SETS

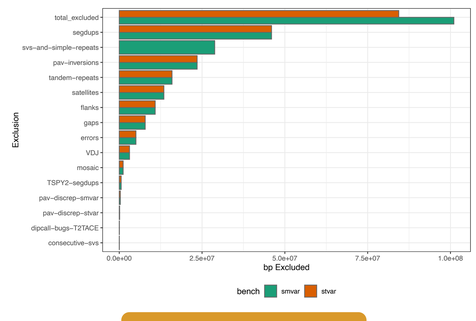
Type	# Vars	SNPs	INDELs
Small	4,504,410	3,644,968	859,442
SV	28,232		28,232



Benchmark Variants

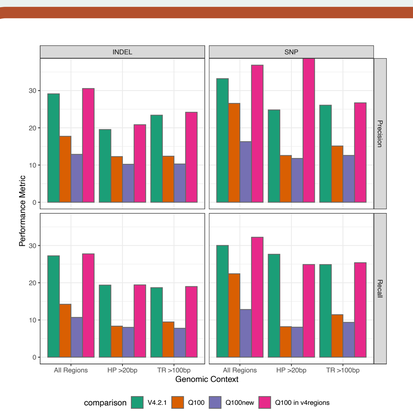
Summary of variants in draft Q100 variant benchmark set for GRCh38. The draft benchmark set includes 16% more variants compared to v4.2.1 and ~3X more variants than the v0.6 SV benchmark.

Type	Regions (bp)	Cov. (%)
Small	2,739,688,363	92.9
SV	2,756,335,521	93.5



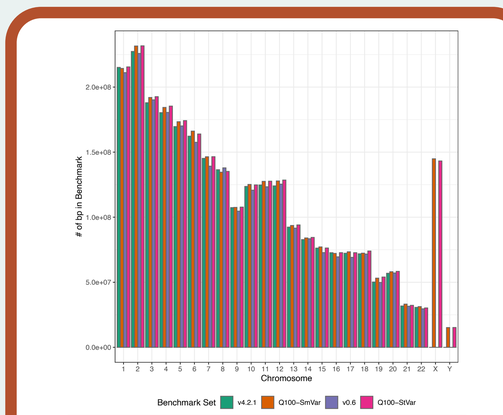
Benchmark Regions

Number of bases and percent of Q100 assembly covered by the draft variant benchmark set for GRCh38.



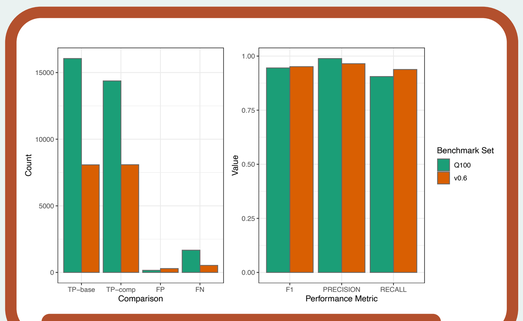
Sm. Variant Benchmark Comparison

Stratified benchmarking results for DRAGEN v4.3.4 small variant callset compared to v4.2.1 and Q100 GRCh38 benchmark set. v4.2.1 and Q100 benchmarking results were similar in v4.2.1 benchmarking regions. Lower performance in Q100 benchmark regions and regions unique to Q100 benchmark set due to the Q100 benchmark set covering more challenging genomic regions.



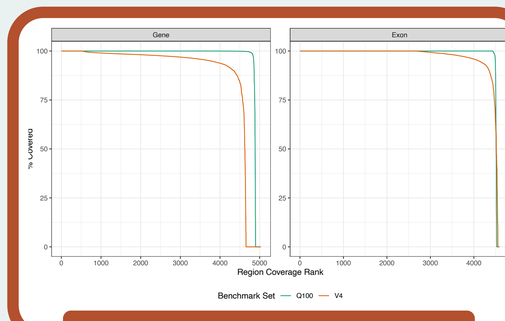
Benchmark Coverage Comparison

Number of bases covered by Q100 draft small and structural variant benchmark compared to previous small (v4.2.1) and structural (v0.6) benchmark sets.



SV Benchmark Comparison

Assessment of SV caller performance on the HG002 GIAB draft Q100 SV benchmark set for PacBio HiFi PBSV variant calls.

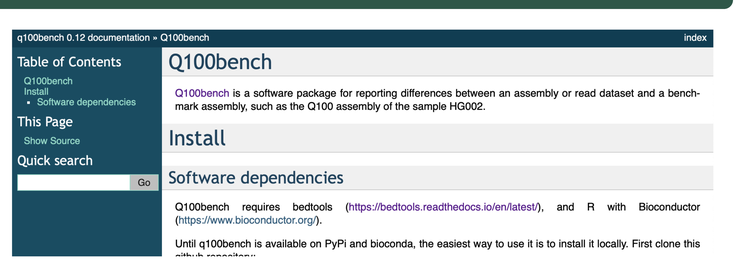


Medically Relevant Gene Coverage

The Q100 benchmark regions includes more medically relevant genes than the v4.2.1 small variant benchmark regions.

GENOME BENCHMARKING

Benchmarking software (q100bench - git://github.com/nhansen/q100bench) available to compare test assemblies to the hg002v1.1 benchmark and report completeness, continuity, phasing, and base-level accuracy. q100bench can also be used to benchmark read sequence and quality value accuracy.



ACKNOWLEDGMENTS

We would like to thank Adam Phillippy and Nancy Hansen as well as our collaborators from the T2T and GIAB Consortium. Grilled Cheese Nate thanks Phil Lesh for the music and memories, fare thee well.

Scan here for access to a github repo with additional details and references.

Scan here for more information on the Genome In A Bottle Project.
https://www.nist.gov/programs-projects/genome-bottle