Benchmarking variant callers: Towards building a robust exome pipeline

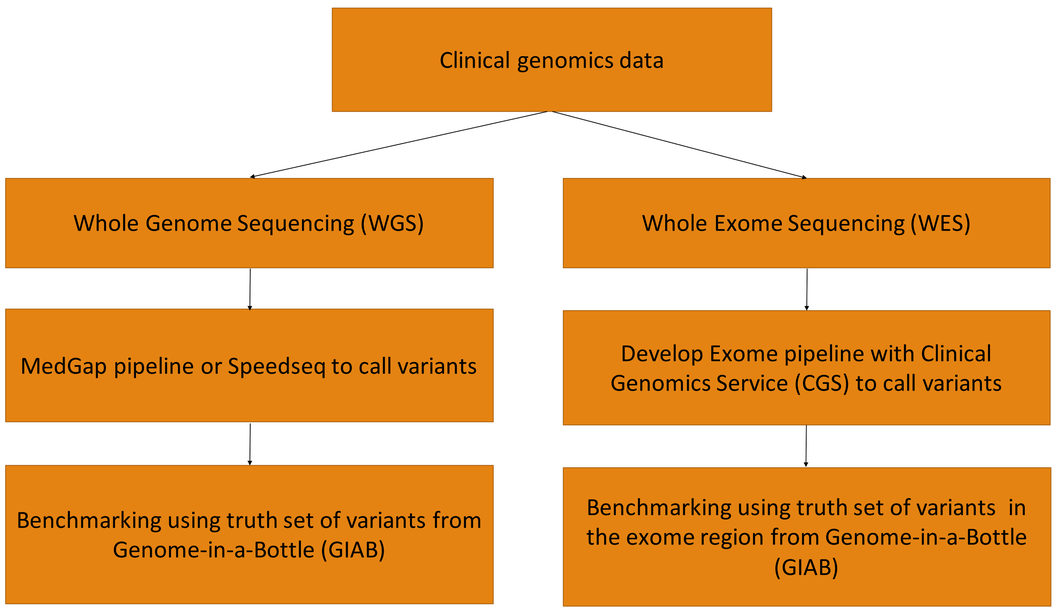
Vandhana Krishnan[1,2], Amin Zia[1,2], Sowmithri Utiramerur[3], Somalee Datta[1,2,3]

[1]Department of Genetics, School of Medicine, Stanford University; [2]Stanford Center for Genomics and Personalized Medicine; [3]Clinical Genomics Service, Stanford Health Care

What is benchmarking?

In the area of method development, it is the method for assessing the performance of software tools by comparing it to gold standards or the best practices in the industry.
One can use benchmarking to assess the performance of software/bioinformatics tools such as variant callers.
Benchmarking will be a continuous process in order to maintain consistent results from the growing versions of software tools meeting the evolving gold standards.
Benchmarking can be achieved by manually comparing the results with gold standards based on specific metrics or using benchmarking tools.

Why do we need benchmarking?



Identify true positives, false positives and false negatives. Decide which tool(s) performs better for calling the variants in the data obtained from clinical cases. Thus, benchmarking is important because it will help identify variants of clinical significance impacting research in personalized medicine.

Data sources: WGS - Illumina's CLIA NA12878, WES - Personalis, Nextera

MedGAP2.4 Pipeline dependencies: samtools 1.19, picard-tools 1.32, GATK 3.4, tabix 0.2.5, bamtools 2.2.3, sjm 1.0, zlib 1.2.7, Python 2.7, Java 8 and Perl libraries (File-Path-2.08, Statistics-Descriptive-2.6, GD-2.45, GDGraph-1.44, GDGraph-histogram-1.1, GDTextUtil-0.86, Math-CDF-0.1). MedGAP2.4.1 has same dependencies but picard-tools 2.4.1 (More details on MedGAP can be obtained from CGS development team)

Speedseq dependencies: g++ and standard C and C++ development libraries, CMake, GNU awk and core utils, Python 2.7 (numpy, pysam 0.8.0+, scipy), ROOT, Variant Effect Predictor, BWA, FreeBayes, LUMPY, Sambamba, SAMBLASTER, Vawk, GNU Parallel, mbuffer

What are the gold standards available to us?

NA12878 Platinum: NISTv2.18, NISTv2.19, NISTv3.2.2 (Jan 2016)
Fasta sequences from the above NIST versions: hg18, hg19, hg38
Bed files from the above NIST versions (high confidence regions)
Variants (SNPs and INDELs) called in the high confidence regions from the above NIST versions

GIAB (NA12878)	No of truth SNPs+indels in GIAB high confidence bed region	No of truth SNPs in GIAB high confidence bed region	No of truth indels in GIAB high confidence bed region
NISTv2.19	3,152,426	2,787,291	365,135
NISTv3.2.2	4,299,935	3,572,851	727,084

Table 1. Number of SNPs and INDELs in NIST NA12878 truth sets obtained using RTG vcfstats

Building Exome gold/truth sets from WGS gold sets

We specifically investigate two exome regions obtained using the UCSC browser for each NIST version:
(1) Coding Exons (2) Exons Plus/Minus 2 (add-on to existing clinical standards)
We intersect each of the bed files above with the GIAB high confidence regions using bedtools to obtain the Exon gold set bed files. Similarly, using vcflib tools the Exon gold set bed files were used to extract the variants in the corresponding regions from the NIST gold set.

Benchmarking

The benchmarking tool used to generate Figure 1A and Figure 1B was Illumina's hap.py.

TP = True Positives
FN = False Negatives
FP = False Positives

Figure 1A

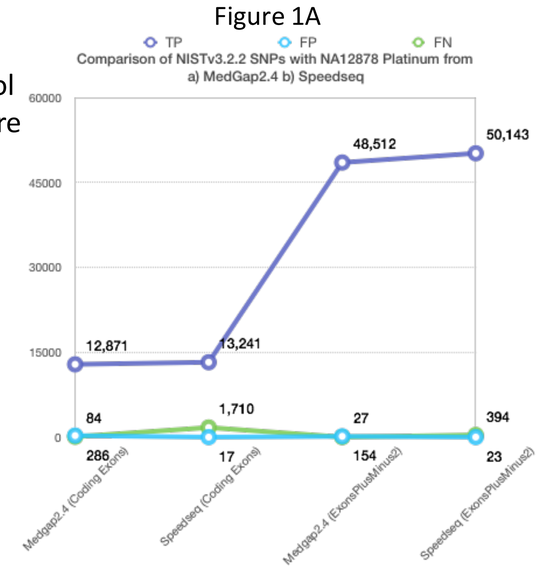


Figure 1B

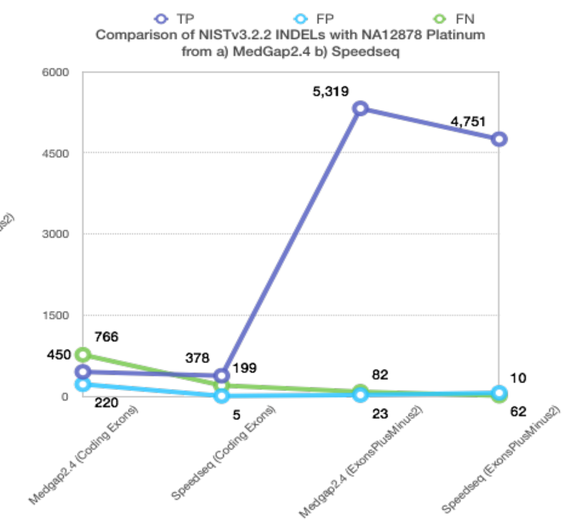


Fig 1A: #Truth Total for SNPs = 13,267
Fig 1A: #Truth Total for INDELs = 459
Fig 1B: #Truth Total for SNPs = 50,226
Fig 1B: #Truth Total for INDELs = 5,516

Future Directions

1. Use RTG vcfeval as an additional benchmarking tool (on-going) and compare against hap.py as they do not generate identical results
2. Use Docker (containerized platform for building apps) for each stage in the Exome pipeline
3. Enhance reproducibility and platform portability: Use GATK best practices. Integrate CGS Exome pipeline with LOOM (workflow engine with job scheduler developed in-house and uses Google Cloud platform)
4. Evaluate CPU and memory requirements, sensitivity and specificity of tools

GIAB Recent Developments (announced Sep 16, 2016)

1. Reference materials (RM) for Ashkenazic Jews (AJ): Son, AJ Trio (father-mother-son), Asians: Chinese son
2. RM for bacterial genomes determined by FDA to have significant impact on public health issues: *Salmonella typhimurium* LT2, *Staphylococcus aureus*, *Pseudomonas aeruginosa* and *Clostridium sporogenes*
3. NISTv3.3 available now (phased calls and phased ID's from GATK - HC, includes 10x Genomics calls, new way to define callable sets that excludes decoy and certain segmental duplications, include some variant and homozygous reference calls found in >10bp homopolymers, data from AJ Trio and Chinese son included
4. GIAB works with Global Alliance for Genomics and Health Benchmarking Team (GA4GH) to incorporate best practices in benchmarking using GIAB genomes

References

1. Zook, J et al. 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology* 32, 246–251.
2. Zook, J et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data* 3, Article number: 160025.
3. Real Time Genomics' RTGcore: https://github.com/RealTimeGenomics/rtg-core
4. Illumina's hap.py: https://github.com/Illumina/hap.py/blob/master/doc/happy.md
5. Chiang, C et al. 2015. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods* 12, 966–968
6. LOOM workflow engine: https://pypi.python.org/pypi/loom-engine/0.1.0
7. Bedtools: http://bedtools.readthedocs.io/en/latest/content/bedtools-suite.html
8. Vcflib: https://github.com/vcflib/vcflib
9. CGS: https://stanfordhealthcare.org/medical-clinics/clinical-genomics/our-team.html
10. GATK Best Practices: http://bit.ly/2cBf825
11. DOCKER: https://www.docker.com

GIAB

Acknowledgements

Funding support from SCGPM and CGS
Nathan Hammond and Isaac Liao: Developers of LOOM
Computational Resources: SCG4 (SCGPM), SUGI (CGS)

Stanford Center for Genomics and Personalized Medicine | NIST National Institute of Standards and Technology U.S. Department of Commerce | JCMB The Joint Initiative for Metrology in Biology | Stanford University | department of GENETICS | Stanford MEDICINE | Stanford HEALTH CARE