Advancing Benchmarks for Genome Sequencing

Justin M. Zook[1,*] and Marc Salit[2,3,*]
[1]Genome-scale Measurements Group, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[2]Genome-scale Measurements Group, National Institute of Standards and Technology, Stanford, CA 94305, USA
[3]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA
*Correspondence: jzook@nist.gov (J.M.Z.), salit@nist.gov (M.S.)
http://dx.doi.org/10.1016/j.cels.2015.09.004

Several recent benchmarking efforts provide reference datasets and samples to improve genome sequencing and calling of germline and somatic mutations.

Large-scale studies and high-stakes applications, such as clinical decision making, require careful benchmarking of technologies. Here, we highlight recent advances in benchmarking approaches for genome sequencing and related bioinformatics methods. In this issue of *Cell Systems*, Griffith et al. describe the ultra-high-depth sequencing of whole genomes and exomes from normal, primary cancer and relapse tissue from a patient with acute myeloid leukemia (Griffith et al., 2015). The raw data, along with a set of validated somatic variant calls—one of the most comprehensive individual cancer genome-sequence datasets to date—provide a valuable resource that can be used to benchmark somatic mutation calling. In addition, these results complement ongoing somatic mutation-calling ''challenges'' and efforts reported at a recent public workshop from the Genome in a Bottle Consortium on standardizing and benchmarking next-generation sequencing.

One approach for evaluating the performance of a genome-sequencing and variant-calling method is to apply it to a reference DNA sample for which the sequence and variants are known. This is the strategy taken by the National Institute of Standards and Technology (NIST), which in May 2015 released the first well-characterized whole-genome reference material (NIST RM8398) (Nature, 2015). The reference material is based on the NA12878 DNA from the Coriell Cell Line Repository. As immortalized cell lines can be used, it is relatively straightforward to create reference materials for germline genomes.

In contrast, creating cancer reference samples and calling somatic (non-germline) mutations from cancer samples is more challenging. Particular somatic mutations occur in only a small proportion of the measured cells because tumor samples often include a significant proportion of normal cells, and a single tumor itself can contain clones with different variants. Very high coverage is often needed to detect these somatic variants, and even with high coverage, it is difficult to distinguish true somatic variants from systematic errors.

Griffith et al. address these challenges by sequencing several samples from a single patient with acute myeloid leukemia. They use whole-genome sequencing and multiple high-coverage targeted sequencing methods (Table 1 of their paper). In addition, the authors applied several bioinformatics pipelines to explore the variability among methods.

This work yielded a number of valuable resources. The authors manually reviewed a subset of variants to create a ''platinum list'' of variants that can be used for benchmarking somatic mutation callers. Table 2 of their paper summarizes key findings and recommendations. The authors also provide a website to explore the data from this patient sample (http://aml31.genome.wustl.edu/). Although users need permission to download the data from dbGaP, the authors have tried to make this as painless as possible. While these data are limited to a snapshot of current commonly used technologies, bioinformaticians may find these data useful for developing and optimizing somatic mutation-calling methods.

In particular, improved methods are needed to characterize certain types of difficult-to-call mutations in this genome and others. In this respect, the work of Griffith et al. highlights challenges similar to those recently identified by the steering committee of the Genome in a Bottle Consortium.

Genome in a Bottle

On August 27 and 28, 2015, the National Institute of Standards and Technology convened the sixth public workshop of the Genome in a Bottle Consortium (GIAB), with more than 150 public, commercial, and academic stakeholders. The consortium was formed by the National Institute of Standards and Technology in 2012 to develop well-characterized samples that can be used to evaluate DNA-sequencing measurement performance. Recently reported pilot results examined library preparation, sequencing, and bioinformatics (Zook et al., 2014).

The sixth workshop began with the release of extensive data from two mother-father-son trios from the Personal Genome Project that were generated by the consortium since the last workshop. These data are from 11 technologies and included the first public trio sequencing using long-read technology (Table 1). GIAB makes all data public immediately so that anyone can analyze and publish about them. The consortium has formed a team to coordinate analyses, with 15 groups presenting at the workshop and additional groups developing analyses.

A highlight of this workshop was the progress toward developing benchmark variant calls for previously inaccessible regions (e.g., repetitive regions) and variant types (e.g., structural variants) in the genome. The initially released benchmark calls for the pilot sample (Zook et al., 2014) were limited to small variants in about 77% of the genome (with 23% of the genome, including 23% of the clinically relevant genome, inaccessible). To advance beyond this, at the workshop, 15 GIAB members presented algorithms to take advantage of long reads, ''read clouds,'' and other data from Table 1 to

Table 1. Summary of Data Generated by the Genome in a Bottle Consortium for an Ashkenazim Trio and by Griffith et al. for a Normal Sample, Primary Tumor Sample, and Relapse Tumor Sample

Study	Data Type	Technologies	Coverage
GIAB	Paired-end short reads	Illumina WGS, Complete Genomics WGS, Ion Torrent exome, SOLiD WGS	WGS: 50–300×; exome: 1000×
GIAB	Long mate-pair	Illumina WGS	15×
GIAB	"Read clouds"	Moleculo, 10×, LFR	20–100×
GIAB	Long reads	Pacific Biosciences, Oxford Nanopore	<1–70×
GIAB	Optical mapping	BioNano Genomics	50–100×
Griffith et al.	Paired-end short reads	Illumina WGS, exome	WGS: 38–312×; exome: 251–433×
Griffith et al.	Targeted sequencing	Illumina, Ion Torrent	43–13,725×
Griffith et al.	RNA-seq	Illumina	32–542 Gbp

For additional details, see Figure 1 and Table 1 in (Griffith et al., 2015). Certain commercial equipment, instruments, or materials are identified in this paper only to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. WGS, whole-genome sequencing; LFR, long fragment read.

discover structural variants, characterize difficult regions of the genome, and phase variants at long distances. These methods to generate calls, along with methods to integrate calls to form high-confidence structural variants (English et al., 2015), will advance our understanding of difficult variants and difficult regions in the next year.

The Genome in a Bottle Consortium is also considering benchmark samples for somatic mutation calling. There are three relevant commercial products available that are based on candidate GIAB Reference Materials, including formalin-fixed and paraffin-embedded cell lines and cell line DNA with synthetic DNA spike-ins that simulate somatic mutations at different allele fractions. These products are possible because the genomes selected by GIAB are those of individuals from the Personal Genome Project who have consented to permit commercial-derived products from those genomes (Ball et al., 2012).

In a parallel effort, earlier this year, the International Cancer Genome Consortium-The Cancer Genome Atlas DREAM Somatic Mutation Calling Challenge published results from a competition using modified real data to test somatic mutation-calling algorithms (Ewing et al., 2015). They found common sequencing error modes that caused false positives from algorithms. They also found that using an ensemble of multiple algorithms produced the best results. The Challenge provides a public platform to evaluate new submissions as algorithms continue to be refined and new algorithms are developed. In addition, results from a new challenge with real tumor data are expected to be announced in Fall 2015.

Even with good benchmarks, standard definitions for performance metrics and standard sophisticated variant comparison pipelines are essential for users of reference materials and reference data to compare performance. In partnership with GIAB, the Global Alliance for Genomics and Health formed a Benchmarking Team to develop standard definitions and tools for benchmarking variant calls. The team has developed several tools for variant comparisons and is currently reconciling reporting to get the same standardized performance metrics from each tool.

Overall, what is still needed are broadly consented cancer samples that will allow public dissemination of DNA reference material as well as public data from these genomes. In addition to the spike-ins for a small number of mutations discussed above, GIAB has discussed possible ways to address this need, but none are ideal. Cancer samples and genomes are diverse, with a wide spectrum of morphology, cellularity, heterogeneity, number of mutations, types of mutations, and ploidy, so a limited set of samples is unlikely to meet all needs. Gaining confidence in results from tumor sequencing will likely take a variety of types of reference materials and reference data, potentially including samples and data from multiple tumor types, tumor-normal cell line pairs, normal cell line mixtures, and real data modified in silico. The works of the DREAM Challenge and Griffith et al. are important steps toward benchmarking cancer genome sequencing, and they provide a great foundation for future work, including that within GIAB.

REFERENCES

Ball, M.P., Thakuria, J.V., Zaranek, A.W., Clegg, T., Rosenbaum, A.M., Wu, X., Angrist, M., Bhak, J., Bobe, J., Callow, M.J., et al. (2012). Proc. Natl. Acad. Sci. USA *109*, 11920–11927.

English, A.C., Salerno, W.J., Hampton, O.A., Gonzaga-Jauregui, C., Ambreth, S., Ritter, D.I., Beck, C.R., Davis, C.F., Dahdouli, M., Ma, S., et al. (2015). BMC Genomics *16*, 286.

Ewing, A.D., Houlahan, K.E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T.N., Bare, J.C., P'ng, C., Waggott, D., Sabelnykova, V.Y., et al.; ICGC-TCGA DREAM Somatic Mutation Calling Challenge participants (2015). Nat. Methods *12*, 623–630.

Griffith, M., Miller, C.A., Griffith, O.L., Krysiak, K., Skidmore, Z.L., Ramu, A., Walker, J.R., Dang, H.X., Trani, L., Larson, D.E., et al. (2015). Cell Syst. *1*, this issue, 210–223.

The week in science: 15–21 May 2015 (2015). Nature *521*, 264–265.

Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Nat. Biotechnol. *32*, 246–251.