Genome in a Bottle Consortium

August 2015 Workshop Report

Executive summary

On August 27-28, 2015, the Genome in a Bottle Consortium held its 6th public workshop at NIST with about 170 participants from government, academia, and industry to discuss reference data and reference materials for human genome sequencing, and methods to characterize genomic reference materials. The morning session of the 2nd day was devoted to exploring use of GIAB products for analytical validation of NGS as a clinical test, including discussion of regulatory oversight with a presentation from the FDA. Important outcomes include:

1. Since the January 2015 workshop, GIAB participants have made data from 11 technologies publicly available for two father-mother-son trios from the Personal Genome Project that are candidate NIST Reference Materials. 15 different groups presented methods they are using to analyze these data for structural variants (SVs), de novo assembly, SNPs/indels, phasing, and epigenetic modifications. We set an initial deadline to submit candidate SV calls vs GRCh37 by Sept 4, so that a variety of comparison and integration methods can be applied, including Parliament by DNAnexus/Baylor, MetaSV by Bina, and svclassify by NIST/DNAnexus. We are continuing to accept candidate SV calls, including revised versions, until Nov 1, and then produce an integrated set before the Jan 28-29, 2016 workshop. NIST is also working on producing high-confidence SNP/indel/homozygous reference calls for all genomes this year.
2. Various possible categories of GIAB reference material products were considered by consortium members, including circulating tumor DNA/cell-free DNA, a large set of samples with difficult mutations characterized in clinically important regions, and conducting an interlaboratory study of spike-in controls that simulate somatic mutations. Although these are important projects, the Genome in a Bottle steering committee decided that these projects are out of GIAB scope with present resourcing, and that GIAB priority should be comprehensive characterization of whole human genomes, beginning with the 7 samples presently under study for reference material development.
3. The next GIAB Workshop will be held January 28-29, 2016 at Stanford University.

Updates since the last workshop

(slides at http://www.slideshare.net/GenomeInABottle/giab-aug2015-intro-and-update-150821pptx):

1. NIST released the pilot genome as the first whole genome NIST RM 8398 (http://tinyurl.com/giabpilot)

2. GIAB has generated 11 datasets for the 2 GIAB PGP trios, which are described in a paper submitted to Nature Scientific Data and on biorxiv at http://biorxiv.org/content/early/2015/09/15/026468
3. Preliminary structural variant calls (mostly deletions) were released for NA12878 (http://biorxiv.org/content/early/2015/05/22/019372)
4. The Global Alliance for Genomics and Health Benchmarking team has developed 3 sophisticated benchmarking tools (vcfeval, hap.py, and vgraph) to compare complex variants, and is currently working towards reconciling the outputs of these tools and implementing counting of true and false positives and negatives to output standardized performance metrics. (see https://docs.google.com/document/d/13tK8KJWnnGebljTODIh7zMIlM6iGFbn39d0QLJOobfc/edit?usp=sharing for more information or to get involved)
5. NCBI has restructured the GIAB FTP site to make data easier to find.
6. NCBI is working to add new data for NA12878 and GIAB PGP trios to the GeT-RM browser (http://www.ncbi.nlm.nih.gov/variation/tools/get-rm/)

Analysis Team

1. 15 different groups presented methods they are using to analyze these data for structural variants (SVs), de novo assembly, SNPs/indels, phasing, and epigenetic modifications. These presentations are listed here: https://docs.google.com/spreadsheets/d/1-1Gqt2RcqlAjHvX1mw_0eaG1DltE9nJEGiF-3G1DAxM/edit?usp=sharing and most slides are available on slideshare at http://www.slideshare.net/GenomeInABottle/presentations
2. Most groups have produced calls for GRCh37, so our current focus is on developing high-confidence variant calls of all sizes for GRCh37. Some, but not all, groups are willing to produce calls for GRCh38 as well, and a couple have already done so. We decided to encourage anyone interested to explore using GRCh38 and the benefits it has in practice, which may motivate and enable additional groups to do these analyses. NIST plans to attempt to produce high-confidence SNP/indel calls for GRCh38 using the same methods as GRCh37 after getting the methods to work for GRCh37.
3. Francisco De La Vega discussed the possibility of generating a personalized reference by selecting the ALTs in our samples and adding in novel sequence insertions from assemblies. Anyone interested in pursuing this line of research is encouraged to contact Francisco and Justin.
4. An important area of interest is to characterize more difficult regions of the genome, since GIAB's current calls cover 77% of the genome. Toward this goal, Arend Sidow discussed very promising methods being developed with Alex Bishara and Yuling Liu in Serafim Batzoglou's group at Stanford. They use 10X Genomics and moleculo data to map reads accurately in regions of the genome that are not possible to map with normal short-read sequencing. These new bam files could be used by other groups to call variants as well. We hope to use these methods to call high-confidence variants in more difficult regions, possibly confirming with long read assemblies from PacBio.
5. After developing high-confidence SNP, indel, SV, and homozygous reference calls for the Ashkenazim trio, the Analysis Team will likely write a joint manuscript, possibly having a draft by the January 2016 workshop. Individual groups are permitted (and encouraged) to submit their own manuscripts as well either before or after the GIAB joint manuscript.

6. We discussed methods to form high-confidence variants beyond the current small variants in relatively easy regions of the genome. Instead of using targeted experimental validation to assess accuracy of our high-confidence calls, we have generally compared to other callsets and, for a subset of variants, used manual curation of reads from multiple technologies to determine what causes any differences between methods and technologies. For SVs, tools like svviz can be very helpful for this in addition to normal genome browsers like igv. Baylor and DNAnexus have also developed Parliament to generate candidate SV calls from multiple methods and to look for support by PacBio and Illumina reads for any candidate calls. DNAnexus is currently applying Parliament to the initial set of SV calls submitted to GIAB. NIST and collaborators have developed svclassify to look for evidence in any bam file for candidate SVs, which should be useful for evaluating candidate calls. Bina has also developed MetaSV to find calls with multiple types of evidence.
7. Calls from Illumina Platinum Genomes and RTG pedigree analyses may be a way to expand beyond the current high-confidence regions. NIST attempted this last year, though they found that it became difficult to define high-confidence regions in a robust way when combining these datasets, so more work needs to be done.
8. We discussed ways of documenting methods used by groups contributing calls. Francisco suggested Synapse (https://www.synapse.org/) as one possibility, which has been used successfully by the DREAM challenges. Francisco volunteered to convene a call about this.
9. Chunlin Xiao has reorganized the FTP and is developing a GitHub site to make the FTP easier to search and navigate, which he will send an email about soon.
10. We discussed how often to submit new versions of calls, and Chunlin Xiao said the FTP could handle as many versions as anyone is willing to submit, but the most useful calls will be submitted before 9/4 and 11/1. If people have calls using new methods before 11/1, then they are welcome to submit at any time, and we may be able to go through multiple iterations of evaluating candidate calls.
11. Valerie Schneider said that any novel sequence we find in de novo assemblies would be really useful to submit to the GRC, since they are compiling new sequences and could let us know if others have seen the sequence.
12. The FDA is starting a new SEQC project focused on assembly of human genomes, and they will likely use GIAB samples for this and continue to be a part of GIAB to see how we can best coordinate.
13. PacBio hosted an informatics developers conference the day before GIAB, which included discussions about assembly and forming benchmark SV calls and SV representation. Those interested should join google group at https://groups.google.com/forum/#!forum/smrt_sv making sure to select to receive emails.

Other reference samples breakout

1. Horizon Discovery discussed their product with the GIAB PGP samples embedded in FFPE - http://www.slideshare.net/GenomeInABottle/aug2015-horizon-diagnostics
2. Seracare discussed their new product with synthetic DNA containing tumor-related mutations spiked-in to the GIAB PGP samples
3. The need for controls for circulating tumor DNA was discussed
4. Invitae, Counsyl, Personalis, and others discussed the need for samples with difficult mutations in genes commonly tested by NGS panels, since GIAB samples generally

have easier mutations in these genes. Lisa Kalman of GeT-RM is leading a study to find and/or make these samples.

Analytical validation discussion

Platform for GIAB community to connect with FDA and for guidance on what GIAB products would enable labs to validate their sequencing and bioinformatics
Questions:

1. What RMs would be useful for analytic validation of somatic variants
2. How does targeted sequencing differ from WGS in terms of analytic validation needs?
3. What's the role of benchmarking data sets in validating bioinformatics?
4. Is there a role for "benchmark" or "reference" pipelines?
5. What GIAB products other than RMs should we produce?
    a. Would a product like a whitepaper outlining the common pitfalls in analytic validation for NGS be a good product?
    b. Is there a need for other process controls (e.g., FFPE-embedded, mixtures, etc.)?
    c. What role can spike-ins play in validation? What would they look like? For somatic mutations? For germline mutations?
2. What are the most specific knowledge gaps in how to do analytic validation for NGS?

Session Presentations

1. Marc Salit discussed an overall standards architecture for NGS: http://www.slideshare.net/GenomeInABottle/aug2015-salit-standards-architecture
2. Zivana Tezak from the FDA discussed their work towards analytical validation: http://www.slideshare.net/GenomeInABottle/aug2015-zivana-tezak-analytical-validation
3. Deanna Church from Personalis discussed important considerations for analytical validation: http://www.slideshare.net/GenomeInABottle/aug2015-deanna-church-analytical-validation
4. Steve Lincoln from Invitae discussed Analytical Validation and Performance Monitoring of NGS: http://www.slideshare.net/GenomeInABottle/aug2015-steve-lincoln-analytical-validation
5. Jared Maguire from Counsyl discussed their approaches to analytical validation (slides not available)

Steering committee notes

1. Scope of GIAB was confined to characterizing a small number of genomes as well as possible
2. More funding for GIAB work could be useful. NIST has competitive postdoctoral opportunities if anyone knows someone who may be interested. Chris Mason may try to organize another U41 proposal to get more funding for GIAB-related work
    a. Elizabeth Mansfield (and NIST) would love to have more success stories from GIAB to justify funding
2. Some papers were discussed:

a. Would like to do one when we have SNP/indel/CNV/SV calls, possibly draft in Jan 2016
    b. Still want: Scientific Data Paper (presubmission inquiry was very positive, now submitted and on biorxiv at http://biorxiv.org/content/early/2015/09/15/026468)

    c. Best practices paper (possibly do this as part of GA4GH - will discuss at NYGC meeting Oct 13)
2. Events:

a. GIAB will have 2 talks at AMP GeT-RM meeting
b. GIAB is co-hosting ASHG workshop w/GRC; Justin also has a talk at ASHG
c. Robyn suggested we may do an educational AMP webinar?

2. Abstracts: We should submit more - volunteers are welcome to present about GIAB - contact Justin and Marc and enter rows if interested in this spreadsheet: https://docs.google.com/spreadsheets/d/1-VPmlejDHQSH1ifFGX_rUsRt1G9DZV9JeQ2eQB4M4b8/edit?usp=sharing
3. Similar to talks from clinical labs at this workshop, we may have some tech developers give short presentations at next meeting on how they used GIAB materials