Genome in a Bottle Consortium

January 2015 Workshop Report

Executive summary

The Genome in a Bottle Consortium held its 5th public workshop January 29-30, 2015 at Stanford University to discuss progress and plan future work developing Reference Materials, Reference Data, and Reference Methods for assessing accuracy of genome sequencing. NIST plans to release its pilot whole genome Reference Material 8398, based on NA12878, in April 2015. For the next set of four NIST Reference Materials, in the next 1-2 months GIAB will have generated a uniquely large and diverse dataset on an Ashkenazim Jewish mother-father-son trio from the Personal Genome Project, and we plan to write a paper for Nature Scientific Data about these data. To lead and coordinate analyses of this valuable dataset, which includes high coverage long and short read sequencing and mapping, GIAB plans to form an analysis group to be led by 2-3 co-chairs. In separate work for somatic mutation calling, GIAB plans to conduct an interlaboratory study to assess the comparability of synthetic spike-in controls to engineered cell lines and real FFPE samples. Also, GIAB is working closely with the Global Alliance for Genomics and Health Benchmarking team to create tools to assess performance against GIAB Reference Materials. This team plans to have a new tool developed for benchmarking in the next couple months. GIAB will also form a group to complete a paper describing examples of and important considerations for using our high-confidence calls. Slides from the workshop are available at http://www.slideshare.net/GenomeInABottle. The next GIAB workshop will be Aug 27-28, 2015 in MD.

Detailed summary

Marc Salit and Justin Zook gave an update on progress since the August workshop, which includes finishing documentation and plans to release the pilot NIST RM 8398 in April 2015 (http://www.slideshare.net/GenomeInABottle/jan2015-giab-intro-and). This RM is ~7500 vials of DNA from a single large batch of Coriell DNA NA12878, and will be available for purchase from NIST at http://www.nist.gov/srm/. They also discussed progress generating sequencing data for the next NIST RMs, which will be based on two mother-father-son trios of Ashkenazim Jewish (AJ) and Asian ancestry from the Personal Genome Project (PGP). Much of this data is already available on the GIAB FTP site at NCBI, and the remaining data should be completed around March 2015. These data will include at least 120x PacBio, 900x Illumina, complete genomics (regular and LFR), long mate-pair Illumina, moleculo, Ion exome, and BioNano optical maps, and are described on slide 20 at http://www.slideshare.net/GenomeInABottle/jan2015-giab-intro-and.

Steve Lincoln gave an example of using the GIAB pilot genome along with 6 other well-characterized samples in clinical validation (http://www.slideshare.net/GenomeInABottle/jan2015-using-the-pilot-genome-rm-for-clinical-validation-steve-lincoln). Less than 0.1% of variants in coding regions of their 29 gene panel were "difficult". However, it is important to include difficult variants (e.g., indels, SVs, complex variants, repetitive regions) in the validation because only 34% of pathogenic variants are simple SNPs and 13% are difficult. Using the well-characterized samples saved almost half of their Sanger validations. However, the 7 well-characterized genomes contained no coding variants in 5 of 29 genes, and only 1 in two other genes. 304 of 310 were SNPs and 6 deletions, all less than 5bp, and no other more difficult variants. Importantly, because there were few difficult variants in the coding regions of the 29 genes, a lab could report 100% sensitivity for GIAB variants and still have 10% false negatives for pathogenic variants.

Bobby Sebra discussed Mt Sinai's PacBio data and analyses for NA12878, as well as the new data that they've been generating for the AJ PGP trio. By combining PacBio reads with bionano, they generated an assembly with N50 of 30Mb. Both they and mark chaisson have also developed methods to call many types of SVs and STRs using PacBio data. The N50 of NA12878 PacBio subreads was about 5kb, compared to 11kb for the AJ PGP trio, so assemblies and variant calling are likely to be even better. Both NA12878 and AJ trio PacBio reads are in the process of being uploaded to the GIAB FTP site at NCBI.

The group then discussed potential ways to organize the analyses of the extensive data GIAB is collecting and making available for the AJ and Asian PGP Trios. The consensus based on previous consortia was that it would be good to have a single analysis group that would be led by 2-3 people. Sub-groups (e.g., for SVs, assembly, etc.) could form as needed after the larger group is organized. We are currently looking for co-leaders of this group, who would provide vision for the analyses and organize and recruit members and writing of potential papers.

Bioinformatics working group (summary slides at **http://www.slideshare.net/GenomeInABottle/jan2015-giab-bioinformatics-summary**)
Chunlin Xiao gave an update on the GIAB NCBI FTP site, which has had over 70,000 files downloaded since August, including a peak of 30,000 in November. The structure of the FTP is being updated so it is easier to find data, and all data submitters are encouraged to submit via SRA when possible using the GIAB BioProject and BioSamples so that data and metadata are easy to find. See slides at http://www.slideshare.net/GenomeInABottle/jan2015-bioinfo-updateonftpsraandusage.

Changhoon Kim talked about the Asian genome project, in which they are working to create an Asian-specific reference assembly. They have generated over 70x PacBio coverage of the genome and performed an initial assembly with good results comparable to previous assemblies. They found interesting results related to a lot of unlocalized sequence, and poor coverage of chrx and chry.

Deanna Church discussed important considerations for GRCh38. The addition of many alt loci to the new assembly can help to reduce false positives and false negatives, but tools are only starting to be developed to take advantage of these alt loci. A small fraction of GIAB high confidence calls are in regions collapsed in GRCh37. Variant calling become difficult with alts

and break tools because they can't distinguish between the allelic duplication added by the alts from the paralogous duplications most are tuned to deal with in some way. NCBI and Ensembl have provided gene annotations on alts, but many other annotations are sparse.
Recent GRCh38 publication: Extending reference assembly models in Genome Biology (http://genomebiology.com/2015/16/1/13)
There was a suggestion that some RNA-seq aligners may be able to perform well around the edges of the alts.

RM Selection and Design working group (summary slides at **http://www.slideshare.net/GenomeInABottle/jan2015-rm-selection-and-design-summary**

In the reference material selection and design group, there were presentations from several groups related to somatic mutations. First Horizon talked about their engineered cell lines, including studies of varying FFPE treatment and new haploid cell lines that will allow faster engineering of mutations. They have new FFPE samples of the GIAB PGP samples, and for GIAB consortium members a special arrangement has been put in place using the promotion code "GIAB2015" at http://giab.horizondx.com/.
Next, AcroMetrix discussed their Hotspot Control in which spike-ins are added to the Ashkenazim son cell line, and an interlaboratory study they conducted with these samples in which one lab discovered they were using a germ line variant caller instead of a somatic caller. AcroMetrix and NCI both presented data demonstrating similar performance between a synthetic DNA spike-in and human tumor samples. Then, the group discussed a proposed GIAB interlaboratory study that would help assess the comparability of spike-ins, engineered cell lines, and real samples for understanding accuracy of somatic mutation detection. More details of this proposed study are in the group summary at http://www.slideshare.net/GenomeInABottle/jan2015-rm-selection-and-design-summary.

Performance Metrics/GA4GH Benchmarking working group

Several member of GIAB are working closely with the Global Alliance for Genomics and health (GA4GH) Benchmarking team to develop standardized metrics and methods to benchmark variant calls against GIAB Reference Materials. Kevin Jacobs presented about the challenges in benchmarking variant calls, particularly due to uncertainty in calls and phasing and due to differences in representing complex variants (i.e., nearby SNPs and/or indels) - http://www.slideshare.net/GenomeInABottle/jan2015-ga4-gh-variant-comparison. He is developing a tool for GA4GH that will take two vcf files and determine whether in each region they match exactly, they are consistent but at least one lacks phasing information, or they do not match. He plans to finish an initial version of this tool in the next couple months. The GA4GH team will then develop accounting tools to report performance metrics like false positive rate based on standardized definitions. The GA4GH team is also working on developing ways to stratify performance into different categories based on variant type, sequence context, function, and characteristics of the data. The team has an evolving document about definitions of performance metrics and stratification methods (https://docs.google.com/document/d/1jjC9TFsiDZxen0KTc2Obx6A3AHjkwAQnPV-BPhxsGn8/edit?usp=sharing). New members of the team are welcome and may contact Justin Zook if interested.

Steering committee meeting

1. How to get the word out about our PGP NIST Reference Material data

NIST will create a Common slide deck and Abstract that others can use in presentations
Monica Basehore will talk @tri-con, Don Baldwin will talk @ ABRF, Horizon will talk at a European conference
(ESHG, MD upcoming)
To coordinate who speaks where, Justin will make a google doc so that everyone can add conferences and see where speakers may be needed.
GIAB might post about the data to Listservs: Champ (informatics sub), ISCB
Who to: (a) end users; b) data which needs analysis
GIAB might effectively use Twitter when the data is available: #bioinformatics, Ruby, Dan
Bioit world and genome web are good places to try to have stories written, perhaps after a press release about the data
Ancillary workshop at ASHG 2015

Link to the abstract we used for the Leiden meeting? Common folder to put such things on GDrive? Justin will set this up

When the reference material is released, we need to get info out about tools to use the Call set data. Ga4gh hopes to have a tool available by the release in April, though probably not yet the tools to make comparisons very easy (e.g., on the web).

2. Coordinating analyses of the data

Needs leadership, and Everything else follows. Nist hopes to find leadership soon and will then post a blog and email to recruit members

3. DREAM challenge

Somatic mutation caller challenge up, and they Propose to do one for germline using GIAB samples
Issue: Xprize didn't work out so well, partly due to lack of good truth and performance metrics, although somatic DREAM project going better
Would need to use one of the PGP samples (all callers are over-trained on 12878) and we don't yet have a truth set for that, but might be developed by the end of the competition.
Two options: 1) single sample, perhaps limited data types; which we in parallel use the family members and all data types to built truth set
2) put all the data into dream challenge and see if they can help with building the truth set (data integration, etc)
Can we get GA4GH to spearhead this? Ask Stephen Keenan to help organize, and possibly David Haussler for support
Issue: Is this a distraction, or a good way to help the community improve callers and improve our callset? The consensus was that ga4gh might be a good place, since many of us are involved there but it wouldn't distract GIAB from the focus on characterizing the genomes.

4. Coordinating GIAB papers

Justin will set up a call to plan paper about using our RMs. Several people expressed interest in participating in the writing group, but we need a leader.
A google doc for collecting details about sequencing being done for the PGP trios is online, and people have started adding to it. We hope to publish this in Nature Scientific Data soon after all the data is collected.

5. ASHG etc
For PGP analyses, we will discuss submitting abstracts in Data analysis group

Good Idea: Do a workshop (a la GRC) – Deanna church contacted Peggi McGovern about this:
The ancillary application will be available on our website beginning in March with a deadline date of May 22, 2015. I have added Justin and Marc name to the list, so they will receive an email when the application is open and may apply for an ancillary event.
Thanks,
Peggi

ASHG Abstract Deadline early June

Also do something at AMP, since important end users are there. We could present at a Session at AMP – needs to go in now to Robyn. Start planning around Valentines day so ASAP - Early bird or workshop or GeT-RM

6. Future of GIAB and new products

Andrew Grupe will continue to coordinate planned interlaboratory Study on spike ins

Justin will set up a call about looking into making Germline spike-ins

Need to identify gaps in medically relevant genes:
Deanna Church will Look at pulling Get-RM data to fill in gaps in important genes in the GIAB data set for a paper she's working on with Lisa
Are there other high quality data sets?
Deanna: Use GTR data – intersect with Medical exome list

7. Other Topics

FDA – new approaches to regulating sequencing mentioned at White House on 1/30
as part of the new personalized medicine initiative

The son in the AJ trio who is himself genetics expert by occupation was present and suggested that more clinical/phenotype data would be good since they are willing to share it publicly: additional cell lines would be good to establish with the purpose of phenotyping and later characterizing influence of drugs and edited in genomic variants (no IPS cell line has been established yet). Father 93, is PGP15 primary colon and metastatic liver cancer survivor (4 yrs in remission) also has tissue samples from fathers tumors as well as whole genome sequence

from tumors and respective control tissue done in 2014 by Cornell Med School. Dad has agreed to donate further samples upon his death. Mother 76. All 4 grandparents died >80. His wife is also a geneticist enrolled in PGP but not AJ (Bulgarian) and they have a 3.5 y.o. son. They are open to other uses and we can contact.