Genome in a Bottle Consortium

28th - 29th March 2019

EXECUTIVE SUMMARY

The Genome in a Bottle Consortium held its 10th public workshop on the 28th - 29th March 2019 at Stanford University in Palo Alto, CA with approximately 80 in-person attendees and 25 remote attendees.

Day 1 set the context for GIAB by featuring a progress update, roadmap, and reports from related genomic reference sample characterization efforts. There was also a thinkshop on distributing GIAB samples as cells, talks on new data and analysis methods, and new benchmark sets. Day 2 featured a discussion of the GIAB product development strategy, a thinkshop on genome assembly benchmarking strategies, and the GIAB steering committee meeting.

GIAB Repository

There was enthusiastic discussion of establishing a new dedicated repository to assure that the GIAB-Personal Genome Project (PGP) cell lines will be broadly available in an enduring manner for research, commercial use, and redistribution of products created with them. This topic was addressed in both the cell-based samples thinkshop and steering committee meeting. Coriell and GIAB will together investigate governance and policy guidelines, and funding models for a such a GIAB repository. GIAB needs to establish protocols that assure accurate use of cell-based materials that maintain the integrity of GIAB characterizations. Developing a publication on proper use was recommended. An interlab study to evaluate protocol reproducibility and growth-to-growth variability was discussed, with the potential to both establish methods and performance expectations.

New small variant calls

Benchmarking artifacts were apparent in the evaluations of the version 4α draft small variant benchmark set that was distributed in advance of the workshop. Performance and anomalies described in segmental duplication regions show that characterization and benchmarking are hard in the regions GIAB is advancing into! To release a new v4.0 benchmark set, GIAB is working on methods to address the challenge of benchmark calls in segmental duplications. Ultralong read assemblies could help resolve some of the segmental duplications that aren't in the reference. GIAB will explore using knowledge from GRCh38 to fix calls in GRCh37.

Community needs

Clinical lab practitioners in attendance strongly called out that GRCh38 is still sparsely adopted in their market. For the time being, GIAB needs to assure resources are disseminated for both GRCh37 and GRCh38. Relevant databases such as gnomAD may be moving to GRCh38 soon, so it is also important for GIAB to keep up. While benchmark small variants are available for both references, the draft SV calls, v4alpha small variants, stratifications for benchmarking, and phased bam files for long reads are only available for GRCh37.

There was a call to “productionize” the small variant benchmarking process, perhaps using an enhanced precisionFDA platform (easier access and use). There was also a call for easier data access, searching, and more datasets, e.g. different types of exome data. Clinical applications would also benefit from gene-focused performance metrics and statistics about coverage of genes by GIAB benchmark sets.

Data can be found through the Genome in a Bottle Github repository here, and in the GIAB Bioproject at NCBI here.

Genomics Standards Landscape

There is an active landscape of projects related to GIAB. There were presentations and discussions of other projects and the unique GIAB role in both the plenary session and Steering Committee meetings. The consensus was to continue GIAB’s unique work in authoritatively characterizing a small number of genomes (existing portfolio and incremental additions, as identified in past workshops). Critical partnerships and collaborations that were identified include HGSVC, GRC, NHGRI/HGRP, MDIC, SEQC2, and Telomere2Telomere.

GIAB Charge

The Steering Committee and community at large recognize GIAB’s priority to provide reliable benchmark sets for as much of the genome as possible.

ROADMAP

2019

- Structural Variants
    - Manuscript Describing V0.6
    - Restructuring the SV pipeline to make it easier to run on the 7 current GIAB genomes and multiple references (GRCh37 and GRCh38)
    - AJ Trio SV callsets for GRCh37 and GRCh38 using same methods as V0.6 with new input callsets
- Small Variants
    - HG002 V4.0 callset, a mature version of V4α presented at the workshop, for GRCh37 and GRCh38
    - Reimplementation of integration pipeline in python
    - Benchmarking
        - GRCh38 Stratifications
        - Draft report template for use in interpreting benchmarking results.
- New Datasets
    - Release of the Ultra-long read ONT for HG002 and PromethION for all 7 genomes
    - Release of CCS datasets for all 7 genomes
    - Getting new datasets not necessarily for material characterization but for public use in benchmarking - e.g. exome data.
- Data Availability
    - Work with NCBI, NIST, and GIAB consortium to develop a data management plan for storing, documenting, and public release of GIAB raw data and analysis.
- Materials
    - Establishing a new repository for GIAB cell lines.
- Other
    - Should we have a GIAB-focused hackathon?

2020

- Structural Variants
    - Manuscript on best practices for benchmarking SVs
    - NA12878 and Chinese Trio SV callsets for GRCh37 and GRCh38 using same methods as V0.6 with new input callsets
- Small Variants
    - NA12878 and AJ and Chinese Trio V4.0 callsets for GRCh37 and GRCh38
- New Datasets
    - Publications on the Ultra-long ONT and CCS datasets

- Materials
    - Addition of new GIAB cell lines for increased diversity
    - Broadly-consented Tumor-Normal cell lines (hopefully)

2021+

- GIAB diploid assemblies - Telomere to telomere
- GIAB reference graph representation
- Benchmarking assemblies
- Tumor-Normal cell line characterization
- Characterization of additional GIAB genomes for increased diversity

DETAILED SUMMARY: Workshop Day 1

How GIAB Fits in the Rest of the World

Presentations about other efforts at characterizing genomic reference samples

- Carolyn Hiller presented from the Medical Device Innovation Consortium (MDIC) about plans and progress to generate a somatic reference sample resource for particular somatic variants.
- Li Tai Fang presented about the Somatic Mutation Working Group of the SEQC2 Consortium work defining the somatic mutation truth set for a tumor-normal cell line pair.
- Karen Miga presented the new Telomere to Telomere Consortium's efforts to produce a complete human genome assembly from long reads, with a current focus on haploid CHM13 cell line.
- Charles Lee from the Human Genome Structural Variation Consortium presented on work identifying structural variants in 3 trios from the 1000 Genomes project using multiple sequencing methods.

Thinkshop - Cells as GIAB Samples

The first of two thinkshops was a group discussion on the distribution of GIAB materials as cells.

Discussion points:

- Some genome sequencing and mapping methods have needed to start with cells rather than NIST reference materials, since the DNA was too short. Should the next version of GIAB reference materials be cells?
- Coriell is exploring new methods to ship long DNA, so packaging in cell might be unnecessary, though it's unclear if this will be sufficiently high throughput
- Consent of PGP samples enables using GIAB cells for genome editing - could GIAB try to limit risk of unethical uses?
- Methylation of DNA was affected by temperature changes as small as 5 degrees during cell culture. Cell culture protocol is extremely important, at least if samples are used for epigenetics. Is this also important for DNA sequencing since methylation affects single molecule sequencing signals?
- While we haven't seen batch effects for small variants or structural variants yet, as we move to characterize increasingly challenging variants in repetitive regions like homopolymers, tandem repeats, segmental duplications, and centromere and telomere, we may find these regions change more rapidly during cell line propagation.
- Can nuclei be distributed?

New Data from GIAB Genomes

Presentations on the new data which has arisen through the characterisation of GIAB Genomes.

- Nate Olson presented on current work with using Oxford Nanopore Technology to sequence ultra-long reads from the GIAB samples, and data for HG002 will be public soon after QC.
- Karen Miga presented sequencing 11 reference genomes in 9 days using PromethION nanopore sequencing, and are sequencing all GIAB samples.
- Aaron Wenger presented on the PacBio Circular Consensus Sequencing that has been done for HG002.
- Peter Lansdorp presented on Strand-seq, which could help GIAB characterize phasing and large inversions.

New Methods to Characterize GIAB Genomes

Presentations on evaluating v4alpha small variants and early results from NCBI's Pangenome Hackathon.

Justin Wagner presented a draft small variant benchmark set (v4alpha) that covers segmental duplications and other regions that are difficult to map with short reads.

Lightning Evaluations of v4alpha Small Variant Benchmark

Billy Rowell, Andrew Carroll, Ian Fiddes, and Yih-Chii Hwang presented evaluations of the draft benchmark set, finding that it may include some questionable calls in segmental duplications.

Lightning Talks from the NCBI Human Pangenome Hackathon

Ben Busby presented on the NCBI Human Pangenome Hackathon to use graph methods to represent the human genome.

Jason Chin presented on a group project in the Hackathon to generate a fully phased diploid assembly of the MHC region for the GIAB HG002 sample.

Shilpa Garg presented new methods for diploid human genome assembly.

GIAB Product and Tool Roadmap

Presentations from NIST on GIAB's product and tool development roadmaps.

Justin Wagner presented on the roadmap for development of the small variant benchmark integration pipeline.

Nate Olson presented on the roadmap for develop of the structural variant benchmark set pipeline.

DETAILED SUMMARY: Workshop Day 2

Product Development Strategy

A prompted discussion of GIAB's development strategy moving forward.

GIAB members discussed product development strategy. Topics discussed included what MTA should be used, cell line access and potential for a new GIAB repository, ongoing leaderboard for precisionFDA challenges, need for clinical sequencing-focused metrics, data, and interest in different sequencing measurements for the same sample.

There is also interest in making similar benchmarks for non-human genomes. While this is outside the GIAB scope, perhaps it is possible to find a way to collaborate with other relevant bodies.

Thinkshop: Benchmarking Genome Assemblies

The second of two thinkshops consisted of presentations and discussion on different approaches and challenges associated with benchmarking genome assemblies.

Arend Sidow presented on the need for new performance metrics for human diploid assembly distinct from non-human assemblies.

Jason Chin described methods and tools for assessing assembly accuracy as well as debugging assembly issues.

Benedict Paten discussed the difficulty of genotyping all of the genome and need for assembly comparison software.

DETAILED SUMMARY: Steering Committee Meeting

Policy

- Should NIST stand up a repository?
- Remarks from Coriell
    - Coriell has a lot of experience with this, and distributing samples is not a small job
    - Concern from NIH about making cell lines available in an unrestricted manner since they have had prior experience with irreproducible samples
- New repository considerations
    - NIST has been talking with Coriell about starting a GIAB repository
    - A new option is Coriell offered to sponsor a sustainable collection with OpenMTA
        - Team would have the capacity to bring in new samples
        - Could create a pipeline to bring in samples from different populations or types
    - Need to identify purpose and governance as a group
    - An MTA makes it so do not have to perform policing by explicitly stating what can be done from beginning
    - GIAB PGP samples have had extensive investment in data and analysis, and we want to make these broadly available to academic and commercial entities to promote innovation
    - Ethical framework is important to consider
    - What are the governance expectations?
    - Plan to establish governance guidelines for cell repository at Coriell
    - At Coriell, anybody can establish a repository with whatever guidelines needed so long as it's funded
    - Steering Committee motion decided to establish Marc, Justin, and Coriell for governance development and develop strategy for sustainable repository funding
- PGP LCLs have been approved by Corriell to redistribute 6 months ago
    - iPSCs not covered by this policy
- Further concerns for completely open GIAB cell line
    - potential unethical uses
    - Cell lines are concerning because CRISPR/CAS9 and other modification mechanisms
    - people may start redistributing a commercial product
    - Cell line authentication is not widely done
    - Redistribution and not following best practices has many potential downsides
    - Suggestions to address concerns for open GIAB cell line
    - Advisory to rename cell line that is a declared a derivative upon redistribution
    - Publish paper on best practices for use and redistribution
    - When redistribution of cell lines is permitted, Coriell recommends not allowing derived products to use the same ID as the parent cell line
- Publications
    - Nature Biotechnology has 6 months non-open access period, and there was a concern about GIAB publishing in a non-open access journal, but the consensus was that this is

ok if GIAB continues to use twitter, emails, and website to advertise the authors' links to for free public access to the papers.

Communications

Let Justin know if there are any concerns or suggestions related to use of @GenomeInABottle twitter or for www.genomeinabottle.org website.

Strategy

Discussion Points:

- How many more samples should GIAB characterize?
    - Most important priority would be small set of best characterization possible
    - Would be more valuable having more data (exome, etc) for a limited set of samples
    - Focus on process so the reproducible data generation is possible once new samples are available
- Potentially convene GIAB Hackathon to tackle:
    - Difficult regions
    - Benchmarking
- Have few samples characterized as best as possible
    - Having truth set is a value add for a technology developer
    - Fewer but well characterized samples
- Comments on workshop
    - Overall did well
    - Thinkshop strategy planned more ahead of time
    - Planning - January/February less busy than Spring
- Next year's workshop will be important because GIAB will need to determine how it differs from and works with NHGRI's new HGRC grantees
- Somatic complex mixture materials are very important