# Genome in a Bottle Consortium

*January 2016 Workshop Report*

## Executive Summary

The Genome in a Bottle Consortium held its 7th public workshop January 28-29, 2016 at Stanford University, with approximately 100 in-person and 30 remote attendees. NIST updated the consortium that >150 units of Pilot RM 8398 (aka NA12878) have been sold since its release in May 2015, and there has been great interest in the data characterizing the GIAB genomes. The GIAB ftp at NCBI has had 30k-70k downloads from 600-1800 unique IPs per month over the past several months. The paper describing 12 GIAB datasets for the two PGP Trios, the next GIAB Reference Materials, is at http://biorxiv.org/content/early/2015/12/23/026468, and the Analysis Team breakout discussed analyses of these data. The GA4GH Benchmarking Team is developing standardized benchmarking tools and best practices, as well as online implementations, which it plans to make available in the next year. The RM Selection breakout had useful discussions about related projects and reference samples based on GIAB, as well as potential future GIAB products for somatic mutation calling. The Friday morning session highlighted the use of GIAB products in technology development, optimization, and demonstration. Five different GIAB members presented at this session, all from commercial enterprises. The GIAB platform (genomes, reference materials, reference data, methods) were used in development/optimization/demonstration of increasing read lengths, new technologies to access more difficult variants in more challenging genomic contexts, in developing new bioinformatics methods, and for comparing and optimizing existing bioinformatics. Slides from most presentations are available on the GIAB slideshare site, with specific links below in the detailed summary. The next workshop is currently planned for September 15-16, 2016 on the NIST campus in Gaithersburg, MD.

## GIAB Progress Update

- http://www.slideshare.net/GenomeInABottle/giab-jan2016-intro-and-update-160128
- >150 units of Pilot RM 8398 (aka NA12878) have been sold since its release in May 2015
- Paper describing 12 GIAB datasets for the PGP Trios is at http://biorxiv.org/content/early/2015/12/23/026468 and was submitted in Sep 2015 and is currently in press in the Nature journal Scientific Data - lots of social media attention, and many people already using the data

- From the GIAB ftp at NCBI, there are 30k-70k downloads from 600-1800 unique IPs per month for the past few months
- GA4GH Benchmarking Team (https://github.com/ga4gh/benchmarking-tools/)
    - has 3 sophisticated variant comparison tools under development. It is currently working to integrate the tools so that they give consistent outputs using the performance metrics definitions developed by the team (https://github.com/ga4gh/benchmarking-tools/tree/master/doc/standards).
    - Bed files describing potentially difficult-to-sequence regions are available at (https://github.com/ga4gh/benchmarking-tools/tree/master/resources/stratification-bed-files)
    - Draft work plan for 2016:
        - Q1: Finalize intermediate and final output formats of vcf comparison tools, and use hap.py/quantify to enumerate TP/FP/FNs from hap.py and vcfeval intermediate vcfs, and reconcile differences where possible
        - Q2: Release version 1.0 of GA4GH Benchmarking recommended variant comparison practices, including recommended tool(s) and metrics
        - Q2: Identify and advertise location for benchmarking datasets endorsed by GA4GH (including GIAB's)
        - Q3: Write manuscript describing the importance of the sophisticated variant comparison tools developed by our team (using examples from Brendan O'Fallon/Kevin Jacobs, etc.)
        - Q4: Possibly develop integrated comparison tool with hap.py, vcfeval, and maybe vgraph in one package
        - Q4: Make at least one (public?) online implementation of the GA4GH benchmarking tool, ideally allowing the user to stratify performance interactively

## Analysis Team Breakout

**(http://www.slideshare.net/GenomeInABottle/giab-jan2016-analysis-team-breakout-summary):**

Jiayong Li - Curoverse

- http://www.slideshare.net/GenomeInABottle/jan2016-curoverse-benchmarking-somatic-variantcallingpipelines
- developed methods to benchmark somatic mutations
- determine if called somatic fractions are consistent with benchmark calls
- Currently working on implementation

Yuta Suzuki - University of Tokyo

- Developed methods to analyze allele specific methylation with Pacbio using 10X phasing
- Most CpG islands have both alleles methylated but some are allele specific including known and unknown imprinted genes

Fritz Sedlazeck - JHU/CSHL

- http://www.slideshare.net/GenomeInABottle/jan2016-fritz-sedlazeck-mapping-and-sv-calling-from-pac-bio
- Redesigning sniffles to improve breakpoint precision of alignments
- New NextGenMap-LR mapping method to give more accurate alignments of Pacbio reads around SVs

- Have initial version of bam for AJ trio now and working on refining

Han Cao - BioNano
- http://www.slideshare.net/GenomeInABottle/jan2016-bio-nano-han-cao
- Haplotype aware assembly
- Cross validation of calls with NGS - generally good concordance
- Manually curated every event in NA12878 paper - http://www.genetics.org/content/202/1/351
- New haplotype-aware calls and assembly were just uploaded to GIAB FTP for AJ trio

Small group discussions:
- What criteria should we use to decide when 2 SVs should be considered to be the "same" and merged?
- How should we confirm/validate candidate SVs calls and establish benchmark SVs?
- How can we utilize new sophisticated variant comparison tools to improve our benchmark SNP/indel callsets, and how can we develop high-confidence calls for GRCh38?
- Summary of discussions in: http://www.slideshare.net/GenomeInABottle/giab-jan2016-analysis-team-breakout-summary

## Reference Material Selection Breakout

- NGS RM Panel Project
    - Robust consent practice
        - Existing samples
        - New accruals
        - Priorities are consent for public release of WGS data and consent for commercial redistribution; consent for public release of phenotype information is less important but useful if possible
    - Minimum Information Standard
        - metadata on samples being characterized
    - Sample Accrual Criteria
        - How to pick winners
- Commercial Controls on GIAB Platform
    - Horizon Diagnostics
        - http://www.slideshare.net/GenomeInABottle/jan2016-horizon-giab
        - FFPE of GIAB PGP Samples
        - Engineered Cell Lines
        - "Universal" for RNA/DNA
    - SeraCare
        - http://www.slideshare.net/GenomeInABottle/jan2016-seracare-giab-update-d-yuzuki
        - Surrogates for Oncology Panel in PGP Samples
        - RNA Fusions
        - cfDNA
        - ctDNA, NIPT
- How can GIAB develop RMs for Validation of Somatic Calling?

    - Principles for Sample Selection
        - Needs further discussion
        - Nahid Turan from Coriell developing collection of mostly germline cancer-related mutations
            - What cancer samples should we develop into GIAB products?

## How GIAB is used in technology development, optimization, and demonstration

Mike Schnall-Levin from 10X Genomics

- used GIAB samples as an initial demonstration of their technology
- new improvements in technology (Chromium system)
    - more uniform coverage
    - more barcodes (4 molecules/barcode)
    - compatible with X10
- now 30Mbp phase block N50
- new single cell RNA-sequencing product
- Heterozygous SV calls can become much clearer when sorting reads by haplotype
- Lariat - using long range information to find variants in difficult-to-map regions
    - difficult to validate these methods
    - will be releasing these data publicly in the near future
    - Could be useful for helping to expand GIAB high-confidence regions in the future

Luke Hickey from PacBio

- http://www.slideshare.net/GenomeInABottle/jan2016-pac-bio-giab
- GIAB Reference Materials have been used in multiple parts of the DNA preparation for sequencing, including DNA shearing, size selection, and library preparation.
- Public data from GIAB AJ PGP trio has greatly benefited human bioinformatics development
    - e.g., Ali Bashir et al. at Mt Sinai have assembled each member of the AJ Trio to and 4-5Mbp N50 and call ~24000 SVs in each genome
- Longer read lengths require longer input DNA than the pilot NIST RM 8398
- Mt Sinai sequenced the GIAB pilot NA12878 and were able to combine PacBio, BioNano, and Illumina data to get long N50 assemblies and long TR variation
- Long read methods allow a return to doing human assembly and calling of more difficult variants and difficult regions
    - GRC assembling new human genomes

John Oliver from NabSys

- http://www.slideshare.net/GenomeInABottle/jan2016-nabsys-giab
- Single-molecule electronic mapping technology
- DNA flows through a nanochannel and probes are detected electronically
- Used high-confidence SV calls from svclassify in NA12878 to assess how their technology can be used to confirm the existence, type, size, and zygosity of the events
    - can find evidence for events down to ~300bp currently

- Working on methods for discovery by looking at distributions of distance between probes at every location
- FP and FN rates of detection of the probes is ~4-5%
- Depth of coverage is 77x on AJ son with N50 ~100kb; will talk with GIAB about what genomes would be most useful to do next

Marghoob Mohiyuddin from Bina Technologies

- http://www.slideshare.net/GenomeInABottle/jan2016-bina-giab
- Collaboratively developing new analysis methods and benchmarking methods with GIAB
- Used GIAB high-confidence calls as a basis to simulate variants with VarSim for germline and somatic mutations
    - http://bioinformatics.oxfordjournals.org/content/31/9/1469
- Collaborated with GIAB to use MetaSV and trio analysis to confirm variants found by svclassify
    - http://www.ncbi.nlm.nih.gov/pubmed/25861968
    - http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2366-2

Andrew Carroll from DNAnexus

- http://www.slideshare.net/GenomeInABottle/jan2016-dnanexus-giab-uses-andrew-carroll
- Use GIAB data in a wide variety of ways - a few examples in this talk
- PrecisionFDA
    - framework to enable the community to share bioinformatics methods, datasets, and comparisons to benchmark callsets
    - GIAB callsets are used as benchmark datasets
- Use data from GIAB and GIAB genomes as examples in a variety of pipelines
- Compared many combinations of mappers and variant callers to GIAB high-confidence calls
- Analyzing SVs from AJ trio with Parliament to help find higher confidence SVs with support from multiple technologies

## Technology Panel Discussion

- How can you validate FP and FN rates for new methods that are detecting things for which there are no high-confidence sets?
    - PacBio and Evan Eichler used BAC clones to assess FP rates; FNs are difficult
    - 10X using fosmid sequence, though they are haploid like BACs, which can be challenging; pedigree analyses are also used
    - Nabsys is modifying the reference with variants based on previously found SVs
    - Andrew - analogy to astronomy: there are things you can see directly (like stars); there are things you can infer are there (like black holes); and there are things you have to build gigantic detectors to find (like dark matter)
        - most collaborators prefer to rely on real data rather than simulated data
    - Francisco De La Vega - GIAB should not only use diverse data types to validate each other but also integrate together to have “the whole be greater than the sum of the parts”
        - everyone agrees this is a good idea - the question is how to harness the community to do these integrations
        - There are some starts to this: Bionano+PacBio; Illumina+PacBio

- Krishna Pant from Color Genomics - many of the types of errors found by pedigree-based methods can also be found by sequencing the same sample many times and/or combining different technologies
    - Justin - agree this is often true. One exception is homopolymers, where all technologies have high error rates, and pedigree information can be used to find true variants when there's lots of noise
- Deanna - to access the very difficult regions of the genome, it would likely be useful to work with groups with specific expertise in particular regions
    - John Oliver - still holds out hope that a genome-wide method will work for these difficult regions
- What should our goals be for characterizing more difficult regions of the genome?
    - Mike Schnall-Levin - there are lots of low-hanging fruit we can do in the interim
- How frequently should we release calls and how confident should we be?
    - perhaps use 3 tranches of confidence - high-confidence, possible, and uncertain
    - releasing calls sooner is better
    - Steve Lincoln - better characterizing small variants in clinical regions is also an acute need
    - mike - 3 tranches and every 3 months?
- What should GIAB prioritize in next 6-12 months?
    - Mike - SNPs/indels outside high-confidence regions (both easier and harder); SVs; cancer sample with CNVs/SVs
    - Marghoob - difficult regions; cancer; perhaps better to focus on clinically interesting regions than whole genome for short-term applications, but not necessarily for longer-term applications
    - Luke - combining strengths of all technologies would benefit their own technology development; progress on difficult clinically actionable variants would also be useful
    - John - whole genome characterization is valuable; important to remember that current clinically actionable regions may not be
- Deanna - better to do 1 genome really well or many partial genomes?
    - Andrew - focus analysis on 1 but start sequencing more; ancestral diversity is important
    - Mike - 1 genome is higher priority for them
    - Deanna - trade-off is tuning to 1 genome vs. tuning to limited regions characterized in many genomes
- Brad: current cancer resources we use are ICGC-TCGA DREAM (https://www.synapse.org/#!Synapse:syn312572/wiki/58893) and AML31 (http://aml31.genome.wustl.edu/) and OICR titration (http://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-015-1803-7)

## Steering committee notes

- What would additional funding be used for?
    - Deanna - project manager would be useful to coordinate analyses and communication
    - NIST working on new GIAB website - will get feedback about a draft from steering committee
        - send to CHAMP and other email lists when ready

    - could use google calendar to announce meetings, calls, speakers
- Staffing up GIAB to accomplish our work
    - Steve - bioinformaticians would help
        - Marc and Justin are always looking for NIST-NRC postdoc candidates (US citizens only),
        - In addition, any good candidates for bioinformatics (regardless of citizenship) are welcome to contact Marc and Justin about job opportunities
        - partnering with academic group may be a good option to find a grad student to help, particularly possible at Stanford, but also at other institutions
    - Dedicated program manager would also help manage collaborations, communications, liaisons
    - NIST is exploring options for this

- GIAB could play a role in making a case for why our work is important, which would help
    - Chris Mason could ask 23andMe and Alfred P Sloan Foundation about education to public
    - Steve Lincoln - we also need to educate the clinical community about the importance of using GIAB products properly
    - Nahid Turan organizing a panel about Precision Medicine that we could talk on
    - GenomeWeb is a good forum
        - tie to GIAB data release
    - CAP Today
    - AMP webinar
    - ASHG webinar (Justin is doing in Feb)
    - ACMG?
    - GIAB talk at AMP in addition to GeT-RM workshop?
    - GIAB-GRC workshop at ASHG like 2015
    - For most compelling study, you need to tie stories to biology and/or health, depending on the audience
- Paper about best practices for clinical validation - Steve L will champion with one other person (maybe Birgit)
- Format of future GIAB workshops
    - Deanna - program committee is a good idea
    - some clinically focused sessions may be useful
    - We will form a program committee - we will send out an email to the consortium to ask for volunteers and select 4-6
    - Potential Fri morning session topics: assembly, clinical, solicit ideas
    - Next workshop likely 9/15-16 in MD
    - Get presenters from projects we're liaising with?
- Coriell saw a large growth in orders after pilot GIAB RM was released (4-5x previous levels), including some large batch orders
- GIAB bibliography; including those using GIAB (easy way to collect these papers over time? maybe start with google scholar)
- Cancer genomes

    - work with Dick McCombie and Mike Schatz, and/or Elaine Mardis, and/or TGen?
- Coriell is reprogramming GIAB cell lines into iPSCs
- tumor-normal cell lines consented properly would be great
    - aneuploidy may or may not be a problem depending on use case
    - Marc/Justin will follow up with Sasha