Genome in a Bottle Consortium

January 2018 Workshop Report

Executive Summary:

The Genome in a Bottle Consortium held its 9th public workshop January 25-26, 2018 at Stanford University in Palo Alto, CA, with approximately 90 in-person and 20 remote attendees.

- Day 1 featured an update on GIAB progress and a roadmap of future work, and 16 presentations about evaluations of draft large variant calls, data visualization, and new methods for difficult genomic variation.

- Day 2 featured a panel discussion about "Principles for Dissemination of GIAB Samples" and a discussion of work towards future somatic and germline samples.

This report describes highlights of progress since the September 2016 workshop, highlights of the future roadmap for GIAB work, detailed summaries and links to slides from presentations at the workshop, and a summary of the steering committee meeting discussion.

Progress:

- Best practices to use GIAB genomes to benchmark variants now published with GA4GH
- New manuscript about GIAB high-confidence small variants
    - Extensively for technology development, optimization, and demonstration
    - ~15,000 unique users of data in 2017 (~30 % increase per year since 2014)
- GIAB has enabled >30 innovative reference samples from 3 companies for clinical assay validation
- New data available and in progress from linked, long, and ultralong read technologies for GIAB samples
- Open science project iterating on draft benchmark large variants (latest draft)
    - 7 presentations giving feedback about quality + 4 presentations about data visualization

Road Ahead:

- Improve small variant calls - ongoing collaborations with several groups using new methods for:
    - Challenging regions (difficult-to-map regions, complex variants, tandem repeats, phasing)
- Develop and publish benchmark large variant callset
    - Evaluate its utility as a benchmark with GIAB Analysis Team
- Sample development of broadly consented tumor reference materials
    - Developing experimental protocols using cell lines derived from organoids

Detailed summaries:

Slides from most workshop presentations are available on the GIAB slideshare site.

GIAB Update and Roadmap

- Slides available
- What GIAB is: authoritative characterization of human genomes to provide benchmarks
- What GIAB isn't: population genetics; exhaustive disease-specific samples; non-human
- Prior workshop takeaways (report is available)
    - Work towards sequence-resolved benchmark large indels and SVs
    - Continue to improve small variants: phasing, difficult to map, etc.
    - Select germline samples of additional ancestries and develop somatic samples with broad consents
- Draft GA4GH Benchmarking manuscript almost read to submit, with >15 active co-authors
    - Robust, sophisticated benchmarking tools on GitHub and PrecisionFDA
- GIAB Benchmark calls are being used: 286 citations of the 2014 Paper
- There are currently 31 products from 3 different companies based on the GIAB PGP cell lines
    - GIAB has NIST leadership with a commitment to authoritative characterization and an industry participation that is agile
- Unique open, public datasets are being used in methods development: 82 citations of the scientific data paper
- New data collected and made public since 2016:
    - 10x genomics → Chinese trio
    - BioNano 2-enzyme for AJ trio
- New data underway and/or planned for 2018:
    - PacBio Sequel → ~60x on Chinese son and 30x on each Chinese parent; 30x additional on AJ son and mother; insert read length N50 of 16 to 19 kb
    - BioNano (new DLS labeling method)
    - Oxford Nanopore: Collaboration with NIST, Nick Loman, and Matt Loose labs for ultralong reads (>50kb N50) on AJ trio
    - Strand-seq - in collaboration with Jan Korbel lab for phasing and inversion detection
    - Hi-C
- GIAB Roadmap: Sample Development and Maintenance
    - Explore broadly consented somatic sample development with tumor-derived cell lines
        - Experimenting with deriving cell lines from cancer organoids
        - Collaboration with Arend Sidow lab at Stanford
        - Longer term: develop methods to characterize benchmark somatic variants in cell lines
            - Small variants
            - Large variants and SVs
            - Allele fraction
    - Develop germline samples from additional ancestries into GIAB reference samples
- GIAB Roadmap: Genome Measurement Science

    - Ongoing collaborations with several groups using new methods to improve small variants
        - Integrating phasing from linked reads, long reads, and inheritance
            - Whatshap (Marschall lab) and HapCut2 (Bansal lab)
        - Integrating variants in difficult-to-map regions from linked and long reads
            - 10x, Marschall lab, Bansal lab, Edico
        - Using assemblies and graphs to improve calling of clusters of variants
            - Seven Bridges, Sentieon, Octopus, Texas A&M
        - Integrating specialized tandem repeat calling methods
            - Eberle team at Illumina, Gymrek and Bafna labs
        - Improve identification of somatic variants in cell line DNA
        - New small variant callset incorporating these improvements planned for 2018
    - Will develop collaborations and methods for using ALT loci
    - Continue work toward publication of benchmark SV callset
        - Define confident regions to enable false positive assessment
        - Evaluate its utility as a benchmark to identify FP and FN calls
        - Develop examples of benchmarking SV calls
- GIAB Roadmap: Performance Metrics for Genome Characterization
    - Publish best practices for benchmarking germline small variants
    - Develop benchmarking methods for large indels and structural variants
    - Develop benchmarking methods for somatic variants
    - Develop benchmarking methods for diploid assemblies

Manual Curation of SVs

- svviz2 presentation → Noah Spies, NIST
    - svviz maps and visualizes reads on the SV allele and the reference allele
    - Now autogenerates dotplots for ref vs alt, ref vs ref, several sampled long reads vs each allele, and ref/alt vs any regions of homology elsewhere in the genome
- PacBio visualization → Aaron Wenger, PacBio
    - Manual Curation: IGV, dotplots
    - IGV 2.4.x ← everyone should upgrade to the latest version which includes improved support for long noisy (eg PacBio) reads and 10x reads
    - How to select samples for manual curation?
        - Random sample
        - Compare to another set and curate differences
        - Plot a distribution on some reasonable dimension and curate tails
        - Subcategorize by some characteristics (e.g., repeats, size, type)
- Manual curation of CNVs and SVs using population level data → Andrew Gross, Illumina
    - Trying to deliver SV results in the clinic
    - Side note: Cell line data is much noisier than whole blood. Someone suggested this might be because of the replication timing/S-phase.
    - Replication-based SV curation:

        - many small events are too noisy to curate in a single sample
        - grouping a population of samples by rough genotype can be a powerful means of curation
- SVCurator web-based app ➔ Lesley Chapman, NIST
    - Demo of a web app that displays multiple visualizations of the evidence for an SV and questions to collect feedback about the accuracy of the SV call
    - Should be useful for training machine learning models, assessing v0.5.0 and later callsets, etc.
    - An online "curatathon" is being planned for Spring/Summer 2018 using the SVcurator app

Feedback on GiaB SV calls version 0.5.0

NIST released v0.5.0 of the draft “straw man” large indel and SV callset in early January, and asked GIAB volunteers to evaluate its accuracy, comprehensiveness, and utility for benchmarking. Eight individuals presented feedback about these calls at the workshop:

Fritz Sedlazeck, Baylor College of Medicine

- Comparison to other published call sets:
    - 1000G, older GiaB
    - Given fudge factor of 1kb or 100bp
    - 60-70% re-found
- Comparison of sequence-resolved and non-sequence resolved
    - When all six technologies are required to support an event, all 96 events from non-sequence-resolved were also in sequence-resolved; as we require fewer datasets to support, we start to see events that didn’t make it into 0.5.0 set
- Look at overlap with genes; a few outlier genes with huge numbers of SV calls overlapping (greatest outlier was PTPRN2)
- Overlap with repeats: 30% of SVs are overlap by 70% a repeat
- Sniffles - now can genotype from pacbio
    - Very few without any evidence from pacbio
    - Could be similar to svviz genotyping
- CrossStitch: merges pacbio with 10X to provide diploid analysis of long reads
    - Work in progress...

Aaron Wenger, PacBio

- At the bulk scale, things look pretty good!
- Question to answer: Has v0.5.0 met the GIAB goal of: Take any call set, look at differences with GiaB, GiaB version should be true
- Took PBSV calls, used survivor to compare with 0.5.0
- PBSV making some extra smaller calls not in GIAB
- GIAB doing better with LINE insertions
- Took 20 calls in PBSV but not in GIAB:
    - Putative false negative: a tandem repeat expansion that was missed (though near another event)
    - Putative false negative: biallelic locus, tandem repeat expansion+contraction
    - Potential erroneous call: non-tandem but nearby duplication

- Conclusion: specificity is good, sensitivity is still not quite there
    - Regions of disagreement tend to be in large tandem repeats, often with other nearby calls
    - Biallelic (ie compound heterozygous) are also a challenge
- Still to-do: figure out which criteria could be improved to increase sensitivity in final callset

Nancy Hansen, NIH/NHGRI

- Made “SV calls” by comparing GRCh38 vs 37 using svrefine
- 312 total *simple* events - these should be “true” events found in 37 (though false against 38)
- Size distribution: mostly small
- Missing about half of these changes, though about 25-40% of changes are exactly the same in our calls
- We have some assemblies that support our version, so maybe HG002 may better match 37 than 38
- These are some of the hardest places, though!
    - May need some “high-confidence regions” for SVs as well, to punt on really difficult regions

Chunlin Xiao, NIH/NCBI

- redundant/overlapping SVs and cases
- 50bp+ calls: ~8k deletions, ~11.5k insertions, some complex
- Many more insertions than deletions; should (in theory) be balanced
- Several hundred deletions are overlapping with another event; mostly deletion with deletion, but occasionally deletion with complex event
    - Two nearby deletions: these can be tricky and different callers will call as 2 deletions, 1 large deletion, or 1 large complex event
- Many insertions also cluster with other events; though does not explain difference in counts between insertions and deletions
    - Some of these have same coordinate, just different length

John Oliver, Nabsys

- Comparison with Nabsys
- *Verification of deletions* - “SV-Verify”
    - Construct reference and variant alleles, map reads against those (similar to svviz)
    - Use SVM to make final decision
    - Analyzed 2783 deletions ≥300 bp in size
    - Caveat: looking at total length differences, can’t parse out complex events accurately
- Verification of pacbio only or illumina only:
    - 300-1000bp - PacBio: 58%, Illumina: 30% validated
    - >1000: PacBio: 70%, Illumina 12%
- ~500 parent-only calls:
    - 151 of which they validate in the son (perhaps breakpoint resolved in parents, not in son? We should genotype everything from parents in the son, no?)
- Very good agreement on deletion size
- Aggregate size measurements provided by nabsys highlight regions with multiple events (complex variants)
- *de novo assembly*
    - Enables variant calling, comparisons to GIAB

    - Gives an idea of how to assess multiple neighboring events

Peter Krusche, Illumina

- Structural variant validation using population data
- Polaris - WGS data for a larger cohort, 220 total samples
- Paragraph - graph realigner for SV breakpoints, joint call/genotyping across individuals
- Start with >1000 unrelated individuals
    - Discard events with too many heterozygous calls
- Compare GIAB AJ calls to polaris calls (manta candidates)
    - The non-validating ones seem to be STRs (which raises the question of whether they're genotyping them accurately or we're calling them accurately...)

Andrew Carroll, DNAnexus

- Discovery and force-calling of SVs with GIAB 0.5.0
- Discovery is hard, confirmation/genotyping is much easier (but not easy!)
- Commonly heard: "all I want is a list of disrupted genes, deletions and duplications. I don't care about a list of breakpoints!"
- Parliament2: fast, accurately, population scale
    - 60 cpu hours per sample
    - Using for 100k WGS samples
- Precision and recall: breakdancer, manta, delly, cnvnator are all under 0.5
- Orthogonal genotyping using svtyper, improve precision for callers to ~.75
- Force apply svtyper to 0.5.0 calls
    - Subset Illumina datasets to 30x depth
    - Correct call/genotype: good for deletions (better than their own calls, as expected)
    - Recall is okay, smaller variants are more difficult
    - Force calling on pacbio results in precision of ~.99, though genotyping is slightly less concordant
    - Recall is also pretty good, no difficulty with smaller events
- Comparing 0.5.0 with 0.4.0
    - Force call with same methods
    - Precision generally goes up, recall goes down
- Take-away: 0.5 is slightly better, includes a few additional more difficult events

Ian Fiddes, 10x Genomics

- Longranger 2.2 SV calling and analysis of GIAB 0.5.0 calls
- New longranger: uses HMM over barcode coverage to detect deletions
- Using 0.5.0 set to benchmark new version of longranger...

New Approaches to Characterizing Difficult Variants and Regions

Charles Lee, Jackson Laboratory

- Human Genome Structural Variant Consortium
- Analyzing a smaller set of trios at great depth across technologies

- Partitioning PacBio reads into haplotypes; about 67% of reads could be assigned to a haplotype
- Integrative phasing gets best results: dense/local + sparse/global
    - Combination of 10x and strand-seq give good coverage at different resolutions
- Short-read methods alone missed about 15% of short indels (700k vs 800k)
- Larger SVs: 10k with short reads alone, increasing to 32k with long reads/optical mapping
    - But found Illumina was still finding some high-confidence events not found by long-read technologies
    - For example: PacBio was missing a lot of the CNVs and known segmental duplications
- Conservatively only found 306 inversions per individual
- Discovered full-length L1s, many of which were present in all children

Tobias Marschall, Max Planck Institute

- Maximum likelihood haplotyping
- Integrative haplotyping across technologies, using strand-seq and more dense data
    - Can get 95% of all SNVs into the largest phase block on a given chromosome
- Horizontal genotyping: genotyping across haplotypes
    - Idea is to split long reads into haplotypes, then within each haplotype identify new variants (or at least genotype candidate variants)
- Using this approach to identify SNVs from pacbio in hard-to-detect regions
    - Found 50k novel SNVs not found by short-reads (GATK), still need to validate
- Working with GIAB to determine how to expand high-confidence regions

Karen Miga, UCSC

- Centromeric regions: a source of new, unexplored human sequence variation
- Predicting higher order repeats from pacbio reads
    - Collapse higher order repeats into a graph, look at frequency of reads supporting each edge between repeat units
    - Chromosome-specific assignment using flow-sorted sequencing data, 10X Genomics data anchoring to p- and q-arm sites, and low throughput experimental assays
- Read depth estimates of average satellite array size; see similar distributions in population level data
- Can detect reads supporting variants within the higher order repeats, look at frequency
- Looking for variants in HORs present in AJ trio

Brock Peters, BGI/Complete Genomics

- Single tube long fragment read technology
- Combination of cpt-seq with a beads: transpose into long fragments, capture on individual beads, then transfer barcode from bead onto
- Majority of reads are found in barcodes with 100+ reads
- 50kb median fragment length, all the way up to 300kb
- 85% of barcodes label a single 20kb+ fragment
- 50 million compartments, 1ng DNA, though high duplication rate
- Transposition results in 9bp overlap, which may help in de novo assembly
- happy to provide data, help people getting your own lab going with protocol
    - Large start-up cost is the beads, though produces reagents sufficient for 400k libraries; willing to share small aliquots of the beads

Discussion of Ongoing and Future Work

Andrew Gross, Illumina

- Adjudicating between repeats of different lengths
    - Can we split out the high-complexity regions, for benchmarking purposes?
    - Could use high-confidence bed files (originally intended for small variants) on SVs
    - Nancy Hansen's svanalyze/refine annotates repeat expansions & contractions
    - Stratification will be really important for understanding results from benchmarking SVs, because different methods will perform differently in different genome contexts
    - Stratifying performance inside long stretches of homozygosity could also be interesting

Steve Lincoln, Invitae

- how do the SV results inform our small variant call sets (v3.3.2)?
    - What about the confident regions?
        - Eg, we exclude all regions around variants ≥50bp (old methods) - we should update these! Previously, we used the union of all SV methods, which was both too broad because it was the union and too narrow because new methods were not included. We may want to iterate a bit more on the SV callset, but this should improve the next version of small variants
    - We exclude many 20-50bp variants from the high-confidence regions in the small variant call set; some of the large indel & SV callers are doing well in this size range, so we likely could improve these calls
    - Long-range: how do we deal with diploid assembly? Creating good assemblies, representing them?

Andrew Carroll, DNAnexus

- what's the plan for getting to SV callset v1.0?
    - When we're ready for publication?
    - We aren't yet confident that it passes the GIAB goal (most putative false positives and false negatives should be errors in their calls, not in the benchmark), especially without delineating high-confidence regions
    - Need more comparisons being done with other callsets on our individuals, especially other long-read (in addition to PBSV) or linked-read
    - Know that there are some complex regions, need a short-term fix to be helpful

Fritz Sedlazeck, Baylor College of Medicine

- collect confident false positive variants, would be a useful catalog
    - Most of these events are close to true variants, though some are out on their own

Aaron Wenger, PacBio

- our numbers seem to be a bit short of those in HGSVC?
    - Aaron's work suggests we're missing some stuff...
    - Charles Lee: HGSVC callset should be conservative; think they're getting 90% of SVs, excluding the complex tandem repeat regions which are called as variable but the exact call isn't quite understood
    - GIAB goal is to provide a benchmark (correct!) rather than a comprehensive collection

Charles Lee, Jackson Laboratory/HGSVC

- let's get end-to-end genome references!
    - We're starting out conservatively in GIAB SV efforts, but working towards this is good

Francisco de la Vega, Stanford University

- we want to enable labs that are doing clinical testing now to test their tests
    - Need shorter term deliverables that are useful right away to demonstrate quality of their process
    - Longer term goal is of course to expand into more regions, be more comprehensive
    - Arend Sidow: technology developers also need benchmarks!

Principles for dissemination of GIAB samples

Panelists:

- Kara Norman - Acrometrix
- Russell Garlick - SeraCare
- Linda Kahl - BioBricks
- Brittany Wright Schuck - FDA
- Sasha Zaranek-Curoverse (PGP co-founder)

Linda Kahl - BioBricks

- Presentation on OpenMTA
- Uniform Biological MTA (UBMTA, 1995)
    - Academic or non-profits
    - No commercial use
    - No redistribution
    - Openmta.org
        - New MTA open for comment that allows commercial use and redistribution

Panel Discussion highlights

- What measures should GIAB establish to assure quality of genomes, cells, propogated cells, derived products, edited cells, etc.?
    - Should there be documentary standards to help assure "traceability" of redistributed products?
- Does every GIAB sample need to be open for unrestricted distribution and use?
    - How essential is enduring, open, unrestricted availability?
    - Does NIST need to establish an independent repository with terms supporting this?
    - Are we at risk depending on the NIGMS repository at Coriell for long-term unrestricted availability?
- Kara Norman - 5 fold (quant) difference in suppliers for EBV compared to WHO, so quality of redistributed products can be important
- Brittany (FDA) reference samples are necessary but not sufficient. Clinical samples are still needed
    - Commutability is important for some samples - means that the reference sample behaves in the same way as a clinical sample would
- SeraCare/Acrometrix care more about the further characterization of the current genomes than about adding new ones, based on what they've heard from their customers
    - But... Francisco De La Vega said, as a customer, that more germline samples are needed from different ancestries, because pipelines may be over-trained on the current samples
    - Steve Lincoln agreed - There is definitely a role for more samples

        - The new AMP/CAP guidelines say labs must have:
            - At least 59 samples in a validation (mix of clinical and reference samples, but no specific number for percentage of each)
            - 59 examples of each variant type that the assay is supposed to detect (e.g., SNVs, indels, complex)
            - Getting 59 is regarded as a challenge, and more GIAB samples could help
    - Arend Sidow - Population genetic diversity is needed. Unrestricted distribution should be a principle of GIAB

Future GIAB samples

Justin Zook, NIST

- Germline samples
    - NIST worked with Stanford to perform an ancestry analysis (using rfmix) of PGP individuals with Coriell cell lines and self-reported non-Caucasian ancestry
        - Found 5 with some African, East Asian, South Asian, and/or Native American ancestry
    - PGP1 also now has family members in PGP, as well as BAC libraries and extensive characterization
    - PGP now has the ability to recruit, which may enable families from additional ancestries

Arend Sidow, Stanford Medicine

- Somatic samples
    - Starting a research project to see if they can develop cancer cell lines for GIAB
        - Establishing new cell lines from blood and tumor tissue of broadly consented patients is a huge logistical challenge, so exploring possibility of making cell lines from organoids originating from tumor tissue
        - Idea is to seed organoids into 96 well plates and see which ones grow as cell lines
        - Collaborating with Calvin Kuo (PI whose lab has experience with organoids)
        - Akshay Sanghi is a grad student working in Arend's lab on this

Steering Committee Meeting

Roadmap

- See GIAB Update and Roadmap, and slide 33 for roadmap summary
- Continuing to improve small variants is important
    - STRs, Clusters, Small Variants, Hard to map
- Add? How to benchmark assembly
- Characterizing somatic samples was discussed, not high on “near-term things to do” because we want well-consented samples first, but on roadmap. (JZ)

Next 2-3 workshops

- Last workshop September 2016
- Agreement that annually is a good frequency
- Having calls has helped a lot, annual feels good, back to east coast?

    - More access by federal staff, FDA, but also many stakeholders on west coast, and January is better on the west coast
- Decided next workshop likely in Stanford, January 2019
- Should there be a Keynote? Not necessarily - it is nice to hear from many individuals involved in GIAB work

Liaisons

- Might be time for a Webinar with AMP about benchmarking?
- AMP presentation (probably more when we have somatic samples)
- NSGC presentation?
- CAP
- An Interlaboratory paper of clinical labs using GIAB RMs could be a way to reach the clinical community
- AGT (technicians)
- AMP Europe
- ESHG might be a good meeting when we have capacity
- HGSVC (co-PI Charles Lee attended this Steering Committee meeting)
    - They could put their people to work on the GIAB samples, but need funding.
    - Many coordinating issues. How to make the HGSVC methods more standardized and exportable? Who would do the calling and integration? JZ Question about how this would impact data release & publication?
    - Took 2.5 Years to do their 3 trios. Less now but integration still an ongoing research project.
    - Use experimental validations to understand accuracy.

Samples/Consent/Repository

- Mission: Research vs. Standards making. Arend – the goal of GIAB is benchmark call sets and standard samples. HGSVC goals somewhat different.
- Coriell, ability to distribute PGP with re-distribution issue.

Synthetic diploid discussion

- New manuscript about synthetic diploid constructed from 2 haploid cell lines
- Andrew Carroll warned of community potentially being confused about uses of GIAB vs synthetic diploid
- Decided to discuss different use cases in the GA4GH Benchmarking paper
- Current synthetic diploid is limited substantially because the cell lines are not in a public repository and there appeared to be no clear path to doing this when Justin inquired. However, there is significant value in characterizing the 2 haploid cell lines separately, so GIAB should look into the possibility of getting broadly consented and available hyditaform mole haploid cell lines.