Genome in a Bottle Consortium

September 2016 Workshop Report

Executive Summary:

The Genome in a Bottle Consortium held its 8th public workshop September 15-16, 2016 at the National Institute of Standards and Technology in Gaithersburg, MD, with approximately 100 in-person and 30 remote attendees. Working group meetings (***Small Variant Bioinformatics***, ***Structural Variant Bioinformatics***, and a ***New Sample Thinkshop***) were held on Thursday 15 September, and a plenary session was conducted on Friday 16 September.

NIST announced the release of four new genomic Reference Materials (RMs): RM8391 (Ashkenazim son), RM 8392: (Ashkenazim trio), RM 8393: (Chinese son), and RM 8375: (set of 4 microbial genomes). Standardized benchmarking tools to compare variant calls to these benchmark genomes have been developed by the GA4GH Benchmarking Team

The pilot GIAB genome (NA12878) was released as NIST RM8398 https://www-s.nist.gov/srmors/view_detail.cfm?srm=8398 in May 2015, and >280 units have been sold. The GIAB ftp at NCBI has had ~50k downloads from ~1000 unique IPs per month over the past several months. The paper describing 12 GIAB datasets for the 7 current GIAB genomes was published in June in Scientific Data: http://www.nature.com/articles/sdata201625.

The New Sample Thinkshop https://static1.squarespace.com/static/5739222a27d4bd28d98e3ce9/t/5846df61bebafb624c0595db/1481039714224/GIABSampleThinkshopOutline.pdf recommended that GIAB characterize several new germline samples from African, Hispanic, and mixed ancestries, and that all GIAB products be PGP consented http://personalgenomes.org The thinkshop recommended that GIAB pursue developing cancer reference samples by producing new cell lines from a PGP-consented individual, with one or more tumor segments and a “matched normal” from the same individual.

The Small Variant working group Small Variant working group https://static1.squarespace.com/static/5739222a27d4bd28d98e3ce9/t/5846df85be6594c4abad373b/1481039749568/GIABSept2016BioinformaticsWorkingGroups.pdf focused on refinement of integration methods for complex variants, difficult-to-map regions, and methods to use phased

pedigree information to add more difficult variants and phasing to the GIAB high-confidence calls.

The Structural Variant working group https://static1.squarespace.com/static/5739222a27d4bd28d98e3ce9/t/5846df85be6594c4abad373b/1481039749568/GIABSept2016BioinformaticsWorkingGroups.pdf formed 2 sub-teams to develop high-confidence benchmark SV calls: one team will develop methods to compare breakpoint-resolved SVs from different methods, and one team will coordinate methods to corroborate candidate SVs found by one method in other datasets.

Detailed summaries:

Slides from most workshop presentations are available on the GIAB slideshare site: http://www.slideshare.net/GenomeInABottle

The Steering Committee discussed the frequency of GIAB workshops, and recommended that the next large workshop be conducted in Fall 2017, and a smaller, focused workshop (structural variant science) be held in Spring 2017. Follow-up will be conducted with the GIAB Google groups: https://groups.google.com/forum/#!forum/genome-in-a-bottle

New Sample Thinkshop

- Outline at: http://jimb.stanford.edu/s/GIABSampleThinkshopOutline.pdf
- New Germline samples
    - Existing samples represent limited ancestry
    - What effect does ancestry have on measurements?
        - Samples of African descent have more complexity but will there really be enough variation (in this limited number of samples) to make a difference? What are we measuring? Do different things happen when there are more variants?
    - For “private variants” interesting new samples might be from a recent expansion in Bangladesh, Ancestral Africans.
    - We are not a population genetics project. Look at another project that has sequenced many individuals and pick out an interesting one with a lot of variants. Need an adequate representation of different kinds of variants; some structural variants may impede things.
        - Can we collaborate with measurement science labs in Asia?
        - PGP has best consent, so high priority should be placed on samples from PGP
    - We should get more info on PGP samples that exist now.
        - For example, what percentage African is the African-American sample?

        - General agreement that we need genotypes on all these new samples before going ahead to quantitatively analyze ancestry
    - How valuable is old Sanger sequencing data (e.g., for HuRef)?
        - Probably not very
    - Are additional samples just to keep GIAB busy?
        - No, the point is to be opportunistic and use what we have.
    - Will all companies be onboard with characterizing these new samples?
    - Is the cost of the NIST RMs a problem?
        - It's important to have RMs from a regulatory perspective
    - Cell lines diverge, lose chromosomal arms. 10x genomics would be interested in having the exact same cell pellet. Cell pellets are also good for people to QC their own extraction methods
        - What is the stability of a cell pellet?
        - DNA prep is important for CNVs.
        - NIST will evaluate the possibility of cell pellets as RMs.
- GIAB Cancer samples:
    - Arend Sidow recommended a multi-sample tumor/normal approach
        - Informatics-based drop-outs should affect all samples equally. Multi-sample analysis decreases noise and increases confidence
    - Cancer cell lines have things that are never found in healthy genomes. For example, they may have an allele frequency of 66% because genes are both amplified and mutated
    - A cancer reference material needs:
        - PGP consented sample
        - msi (microsatellite instability)
        - deletions
        - copy variants
        - SNVs
        - fusions (e.g., BCR-abl)
        - translocations
    - The behavior of a cancer cell line depends on the driver. Certain types of cancer have lots of mutations and would be good RM:
        - breast
        - glioblastoma
        - gall bladder
    - A hyper mutated cell line will help out those with targeted clinical gene panels

Small Variant Working Group

- Outline of session: https://static1.squarespace.com/static/5739222a27d4bd28d98e3ce9/t/5846df85be6594c4abad373b/1481039749568/GIABSept2016BioinformaticsWorkingGroups.pdf

- Sean Irvine, RTG

    - Performed full pedigree calling and phasing, including 300x for NA12878
    - How many children are needed? Start to get benefit of pedigree phasing from 5 children
    - Called variants from 10X - in the future could use these in GIAB high-confidence calls
    - Harmonization of calls for integration
        - Adds ~4% of calls to agreement
    - Harmonization of trio calls
        - Reduces number of Mendelian violation
    - Can transfer phase from one vcf to another
        - Phase ~90% of GIABv3.3 with trio or pedigree
        - Keeps annotations and original calls - just adds phase
        - Can add multiple phased callsets
- Mike Eberle, Illumina
    - Platinum Genomes does pedigree-based phasing (similar to RTG)
    - Created k-mer testing method to filter variants if nearby variants within 25bp exist and are missed
        - Using this method to incorporate new calls and expand high confidence calls
- Haynes Heaton, 10X
    - 10X Chromium generates reads that are linked by barcodes
    - Lariat aligner is able to use barcode information to align reads that normally would have low Map Quality
    - Look for reads support ALT in PacBio - require >2 reads
    - Simulate random FP SNPs
- Plans for ongoing work
    - NIST and RTG will work together to transfer phase information from family analyses to high-confidence calls to increase phasing information
    - NIST and Platinum Genomes will work to merge high confidence calls
    - NIST and 10X will work to develop benchmark variant calls in difficult-to-map regions

Structural Variant Working Group

- Outline of session: https://static1.squarespace.com/static/5739222a27d4bd28d98e3ce9/t/5846df85be6594c4abad373b/1481039749568/GIABSept2016BioinformaticsWorkingGroups.pdf

- Will Salerno - Baylor
    - New PB Honey spots results - more sensitive calls
- Jason Chin - PacBio
    - Now can do diploid assembly to get both haplotypes
    - Some SVs are difficult to define except as an assembled sequence
- Alex Hastie - BioNano

    - New, more sensitive calls from optical maps
    - Look for evidence of calls from son in parents, and found it in many cases. Only 2-3 potential de novo changes
- John Oliver - Nabsys
    - Take candidate calls and map reads to ref with and without variant
    - Calls found in more technologies are supported more often
    - Is # of technologies or # of callers a better indication of truth?
    - Sometimes multiple calls fall in the same region between 2 markers
- Sofia - 10X
    - Use HMM, paired end analysis, and barcode overlap to get small deletions in addition to the large SVs found previously
- Mike Eberle - Illumina
    - Discover and type SVs using a population of a 3000 samples
    - Analyzed AJ trio
        - Attempt validation by depth and HWE in population
        - Assembly of reads from samples that are Hom var
- Andrew Carroll - DNAnexus
    - Have reproducible containers for multiple Illumina and PacBio SV callers
    - Use the GIAB deletions supported by 2+ technologies to examine recall at different coverage levels
- Aaron Wenger - PacBio
    - Enhancements to viewing PacBio data in IGV to make it much easier to see SNPs and SVs
    - Genomeribbon.com - new online SV visualization tool
- Nancy Hansen - NHGRI
    - Examining the effect of repeats when aligning SVs in these regions
    - PBRefine - take candidate SVs and extract regions from PacBio assemblies to see if they confirm the SVs
    - Could refine about 20-30% of candidates from Parliament
        - Some missed due to allele dropout in assembly
        - Some might be FPs
        - Some had >2 haplotypes in assembly
    - Visualize assembly vs ref in mummer
        - Easier to see deletions with flanking repeats
- Plans for ongoing work
    - Form 2 teams to work on integration
        - Team 1: SV assembly/breakpoint group https://docs.google.com/document/d/1KmovH5y7zSo-r7w5sxNXaBXipKztQmi4VY0YlbZYE9Y/edit

            - Develop methods to compare predicted sequence change for “sequence-resolved” methods (e.g., that perform assembly and output assembled alleles or use split reads and output REF and ALT sequence in vcf)

                - Works similarly for all types of SVs
                - Use methods to refine breakpoints/SV sequence (e.g., Spiral Genetics, Parliament, PBRefine)
                - Find calls where multiple technologies support the same or similar sequence-resolved calls
            - Team 2: SV corroboration group: https://docs.google.com/document/d/1EiOIxXA231gWeRiXxHU8rIVVPAgg2xy59rUEq0Amn9U/edit

                - Develop methods to corroborate SVs found by any method in other technologies (e.g., using svviz, Nabsys, BioNano)
                - Develop methods to interpret and integrate corroboration results
        - Develop union vcf with all candidate alleles assigned a unique ID
        - Develop union vcf with calls aggregated into a single vcf line if they have similar breakpoints (using SURVIVOR)

Lightning talks

- Hagen Tilgner - Cornell Weill
    - PacBio RNAseq can phase variants at long distances and help to detect gene duplications
- Yves Konigshofer - Seracare
    - Interlab study of spike-ins at different frequencies to test how well different labs can detect difference
    - Developing highly multiplexed reference materials for targeted NGS assays
        - Add spike-ins to GIAB AJ son
        - Also have biomimetics with spike-ins
        - Developing engineered cell lines - looking into using AJ son
- Liang Liu - ThermoFisher
    - Developing ctDNA standards based on GIAB samples
- Tera Bowers - Horizon
    - Developing circulating fetal DNA standards based on AJ mother and son
- Megan Cleveland - NIST
    - Targeted sequencing of GIAB samples with goal of expanding high-confidence regions in regions of clinical interest
- Jeff Rosenfeld - Rutgers
    - Cell lines have many more de novo mutations than blood when examining a trio, so he advocated for sequencing blood of GIAB samples and potentially making GIAB reference samples from blood DNA
- Chris Mason - Weill Cornell
    - Running epigenetic methylation assays on GIAB samples to supplement existing DNA sequencing results
- Yuta Suzuki - Univ of Tokyo

    - Developed methods to use 10X phased variants to phase PacBio reads and find methylation of CpG islands on each haplotype
- Chen Sun - PSU
    - Developed varmatch to compare complex variants with different representations, similar to rtgtools vcfeval but faster and with additional flexibility
- Sharon Liang - FDA
    - created precisionFDA to develop regulatory science for NGS bioinformatics tools, including using GIAB high-confidence calls in two bioinformatics variant calling "Challenges"

Steering Committee Meeting

- Should we continue to have 2 GIAB workshops per year?
    - Consensus was that this is useful
        - Rapidly moving field
        - If tasks to be done for each meeting are clear, it keeps the work moving
        - Try to hold as a satellite meeting to a larger meeting like ACMG, ASHG, or AGBT?
        - Suggestions to have one focused meeting and one large meeting each year
            - Focused meeting in Spring 2017
                - SV integration could be a good topic
                - It would be useful to teach how to use GIAB genomes for benchmarking
- Communications
    - There is a new GIAB website under JIMB (http://jimb.stanford.edu/giab/)
    - Having GIAB representatives talk at meetings is good
    - Maybe write an editorial piece about GIAB?
    - Continue to work more closely with AMP and CAP to disseminate GIAB work
- Scope of GIAB
    - Continue focus on authoritative characterization of whole human genomes
        - Including starting to work on cancer
    - A primary goal of these well-characterized genomes is to benchmark variant calls