

---

## Genome in a Bottle Consortium: Bioinformatics Working Groups

September 2016 Workshop

# Bioinformatics Working Groups

Thursday 15<sup>th</sup> September 2016 9:00AM - 3:30PM

## OVERVIEW

Current high-confidence calls include SNPs and small indels for ~90% of the genome, but exclude the most difficult variants and regions. The goal of these data jamborees is to make progress characterizing more difficult small variants and structural variants

## SMALL VARIANTS (9:00-11:00 AM)

### Motivation

- Current high confidence small variant calls and regions exclude many difficult variants, so that these cannot be benchmarked.
- Also, only limited local phasing information is provided.

### Schedule outline

- Update on current GIAB high-confidence calls, their improvements, and their limitations (Justin Zook, NIST)
- Use of pedigree calls to improve reference material ground truth calls of NA12878. - now include parents in pedigree analysis and new 300x NA12878 data. (Sean Irvine, RTG)
- Incorporating new calls into Platinum Genomes by kmer analysis; how to incorporate in proper ploidy of small variant calls inside CNVs (Mike Eberle, Illumina)
- Use of linked-read technology data to produce calls in the “dark” regions of the genome and confirmation with long reads (Haynes Heaton, 10X).
  - Singleton vs Trio calling with RTG on 10X (Sean Irvine, RTG)
- Improving automated merging of call sets: variant representation canonization methods (Sean Irvine, RTG)
  - Use vcfeval to build up set of minimal alleles from multiple callsets; then go back and encode individual callsets using these alleles
- Transferring phasing from one call set to another to form high-confidence phased calls. (Sean Irvine, RTG)

---

## Unaddressed needs/questions

- How leverage and make calls on ALT loci, especially on GRCh38?
- Can we incorporate calls in difficult regions supported by only one technology?
- Should indels 20-100bp fall under our small variant integration methods or SVs?

## STRUCTURAL VARIANTS (11:00AM-3:30PM)

### Motivation for discussion

- Structural variants (e.g., indels>20-50bp, inversions, and complex changes) are not represented in current GIAB high confidence calls
- Methods to integrate SV calls and form high-confidence calls and regions need to be developed

### Schedule outline

- 11:00 - 11:30: Overview (Justin Zook)
- 12:00 - 1:00: Callsets/Technologies
  - Speakers (moderated by Ali Bashir):
    - William Salerno - Baylor
    - Jason Chin - Pacbio
    - Alex Hastie - BioNano
    - John Oliver - Nabsys
    - Sofia Kyriazopoulou-Panagiotopoulou - 10X
    - Michael Eberle - Illumina
- 1:30 - 3:30: Visualization/Integration
  - Speakers (moderated by Justin Zook):
    - Andrew Carroll - DNANexus
    - Aaron Wenger - PacBio (Visualization)
    - Nancy Hansen - NIH (Complex Events)
    - Peyton Greenside - Verily (Crowdsourced manual curation)
    - Ali Bashir - MSSM (Hybrid Calling)
  - Integration Discussion (Led by Ali Bashir)

## Unaddressed needs/questions

- What criteria to use to form high-confidence calls?
  - How to deal with uncertainty in predicted breakpoints, type, size, etc.?
- How to form high-confidence regions?
- How best to type SVs found in one technology in other technologies?

- 
- How to deal with uncertainty in breakpoints or exact sequence?

## Notes from Discussion:

- Typing options:
  - Parliament - illumina assembly and PacBio
  - Sviz - map reads to REF and ALT
  - Nabsys - map nanodetector reads to ref with and without variant
  - Find kmers unique to event for typing
  - LUMPY svtyper - new version under development
- Can with add a confidence interval to our high confidence calls? Or bin calls into different precisions
  
- Tier 0: 2 technologies assembled both haplotypes across the breakpoint and they agree
  - If discovery uses 2 technologies, then need a third
- Tier 1: 2 technologies assembled across the breakpoint and they agree
- Tier 2: One technology constructs sequence and other technologies support it
- Make these criteria tags on events. ALso tag with which callers and technologies support the event
- Put merged calls in analysis folder
- Recruit classes for manual curation?
- Find regions where no one makes a call

## Insertions

- Can we call an insertion high confidence without knowing the exact sequence?
  - It's possible these could be complex

## Form 2 SV integration teams:

- Team 1: Develop methods to compare sequence resolved structural variation from different methods
  - Uses outputs of each method (Team 2 will focus on going back to the data)

- 
- Only use breakpoint-resolved methods (global or local assembly-based, or split-read-based for simple deletions)
    - Any method that produces exact sequence (VCF with REF and ALT sequence)
  - Perform multiple sequence alignment between haplotypes from different methods to identify “concordant” assemblies
    - First find exact matches
    - Then move to more complex cases with inexact matches
    - Also later examine support for both haplotypes
    - How should we define “concordant”?
  - Team 2: Apply “SV corroboration” methods to determine if each dataset supports calls found by other datasets, and identify visualization methods for difficult sites
    - Easiest to interpret results from accurate break points (for simple deletions) or assembled sequence (for any type)
    - Possible inputs
      - Just develop strawman set of rules
      - Output of team 1
      - All calls generated by breakpoint resolved methods
      - All calls from every method
      - Output of Justin's deletion integration
      - Develop a method to select non-overlapping inclusive set of calls from all calls.
      - Input a set of false hypotheses to determine sensitivity and specificity for each caller.
        - Random locations
        - Sites where parent is hom ref and child and other parent is variant
    - Possible outputs from each validating caller
      - Output confidence score for each call.
      - Estimate error of of breakpoint or size?
      - Often can infer genotype based on support for REF and ALT haplotypes
    - Existing methods within GIAB for typing include:
      - Parliament (Illumina/PacBio-assembly and PacBio-read-based)
      - sviz - map reads to REF or ALT from Illumina paired end or mate-pair, PacBio, 10X haplotype-separated
      - Nabsys - align mapping data to reference with and without SV
      - Bionano
      - Look for kmers unique to SV (Illumina working on methods to do this in population data)
    - Other potential typing methods to explore

- 
- LUMPY svtyper
  - Spiral “graph genome database”
  - Seven Bridges graph genome aligner/variant caller
  - Other methods to generate features from data
    - Svclassify
    - Personalis - <https://github.com/personalis/cnvthresher>
  - How to handle nearby candidate SVs? Try different phase combinations? Or try to phase using assembly?
  - Use features generated by all of these tools in a machine learning model to classify sites
  - Visualization approaches
    - IGV - new extensions for PacBio and 10X in dev version
    - Dot plots - particularly helpful in repeats and complex SVs
    - PacBio GenomeRibbon.com
    - Verily’s CrowdVariant - could potentially present visualizations from any of the above tools
  - Who to invite to help with the work?