Genome in a Bottle Consortium: New Sample *Thinkshop*

September 2016 Workshop

New Samples

Thursday 15th September 2016 1:30PM - 3:30PM

OVERVIEW

GIAB is planning to develop new whole-genome reference samples for both germline and somatic contexts. This "thinkshop" will decide the principles for selecting, prioritizing, and developing these new samples. A whitepaper, possibly to be developed into a 'marker paper' for publication in a widely read journal, will be used to guide the next phase of GIAB work.

- These samples are intended to serve as enduring scientific assets for development of genome-sequencing measurement science.
    - They are intended to act as benchmarks and references, or "***etalons***."

Germline samples: an initial proposal is to develop ~8 new germline samples from diverse ancestry groups from individuals in the PGP.

Cancer samples: an initial proposal is to develop tumor/normal pairing genomes from a fully consented individual, using multiple, different cell lines derived from the same tumor.

Sample format: existing GIAB RM samples are gDNA from large batches of cells in ~50-100 kB fragments; should new samples accommodate longer-read sequencing methods?

GERMLINE

Motivation for new samples

- The existing GIAB genomes represent limited ancestry groups.
    - Human genomes of different ancestries will pose different measurement challenges.
    - Overrepresentation of limited ancestry groups can affect the development of measurement science, reference genome resources, and bioinformatics tools.

Characteristics of new samples

- Similar in all ways to current GIAB samples

    - Authoritatively characterized
    - Stable and homogeneous
    - Publically available
    - Renewable
- PGP-consented (Explicit consent for public release of whole genome sequencing data and commercial redistribution)
- Additional ancestry groups - particularly African, Hispanic, and mixed ancestries
- Available as cell lines (new samples would require fresh blood collection)

Plan for sample development and characterization

- Work with PGP and Coriell to establish appropriate cell lines and materials
    - Proposal based on PGP cell lines currently available from Coriell (single individuals):
        - African American male
        - Puerto Rican male
        - Indian/White female
        - European/Brazilian/Portuguese female
        - Colombian male
        - PGP1 White male - has a BAC library, as well as fibroblast and iPSC lines at coriell
        - Chinese/Filipino/Hispanic male
        - European, deeply phenotyped and characterized at Stanford
        - HuRef (not currently PGP)?, European, sequenced with Sanger and other technologies
- GIAB community characterizes with multiple technologies
- NIST uses integration process to form high-confidence variant calls

Unaddressed needs/questions

- How does this effort relate to other efforts or other reference samples?
- Space to identify opportunity for GIAB portfolio development

SAMPLE FORMAT

Motivation for discussion

- As technologies evolve, what should GIAB samples look like?

RMs from NIST

- Authoritative samples of GIAB genomes

- Existing RMs (pilot, AJ Son, AJ Trio, Asian Son) are gDNA extracted from large batch cell cultures
    - Batch of gDNA established as homogeneous and stable
    - Characterization is done largely from this gDNA batch
        - Some characterization (10X, BioNano, Complete Genomics LFR) done from cell pellets
- Are these RMs needed?
- What's the best format for dissemination?

Materials from biorepository

- Source cell lines for NIST RMs are available as cell lines, cell pellets, and extracted DNA from Coriell in the NIGMS Repository

Commercial reference sample products

- Horizon GIAB FFPE samples
- Acrometrix Oncology Hotspot Control - AJ son with spike-ins
- Seracare - AJ son with somatic spike-ins, circulating tumor DNA, and cardiomyopathy mutation spike-ins

SOMATIC

Motivation for new samples (Arend Sidow, Stanford/JIMB)

- Facilitate technology development for calling somatic variants of all types in cancer genomes
- There is limited availability of reference-grade characterized cancer genome samples
    - *There are de facto references, but no formal references*
- No currently available tumor-normal cell lines are explicitly consented for public release of whole genome sequence data or commercial redistribution
- Starting to characterize cancer genome reference samples will enrich community experience with difficult somatic variants

Experience with Tumor/Normal Pair Genomes as "*A somatic reference standard for cancer genome sequencing*" (David Craig, TGen, USC)

- *Scientific Reports* **6,** Article number: 24607 (2016) - doi:10.1038/srep24607

Experience with somatic variation across a large tumor (Noah Spies, JIMB)

Characteristics of new samples

- Explicit consent for public release of whole genome sequencing data
    - Is it possible to use a PGP consent?
- Homogeneous and stable DNA
    - Choose tumor types/cell lines that do not have high mutation rates (or alternatively grow large batches of cells and extract DNA?)
- Tumor-normal pairs - ideally multiple tumor cell lines from one individual?

Plan for sample development and characterization

- Develop multiple cell lines from a single individual: tumor and normal
    - Develop 16 tumor cell lines?
        - Different cell lines from spatially distinct/distinct regions of a single tumor
        - Alternatively, create single cell line and culture it into different batches that have different somatic mutations that occur during culture
- GIAB community characterizes with multiple technologies
- Cancer genome reference sample development is highly likely to enrich community experience with difficult variants in difficult regions

Unaddressed needs

- How does this effort relate to other efforts or other reference samples?
- Space to identify opportunity for GIAB portfolio development