April 20, 2012

NIST “Genomes in a Bottle” Workshop Summary

On April 13, 2012, NIST convened the workshop “Genomes in a Bottle” to initiate a consortium to develop the reference materials, reference methods, and reference data needed to assess confidence in human whole genome variant calls. The principal motivation for this consortium is to enable science-based regulatory oversight of clinical sequencing.

In the morning, NIST introduced potential tasks of the consortium – to establish use scenarios for standards and performance metrics, select appropriate materials, and characterize the reference materials. NIST also presented a “strawman” work plan to solicit input regarding how the consortium should proceed. Representatives from FDA, CDC, and CAP discussed their perspectives regarding uses of reference materials in regulated and accredited clinical applications. Elaine Mardis presented several applications of whole genome sequencing to oncology at Washington University, and Mickey Williams discussed spike-in controls his lab has been developing for targeted sequencing of cancer-related mutations. Finally, Justin Johnson discussed the Archon Genomics XPrize work characterizing genomes and validating pipelines, Ryan Poplin discussed Broad’s pipeline to analyze genomic data with particular focus on NA12878 as a reference sample, and Justin Zook discussed NIST’s work towards integrating genotype calls from multiple sequencing runs and platforms. The slides for the presentations are attached.

In the afternoon, four breakout groups (breakout group leaders in parentheses) discussed the topics outlined below:

1. Use scenarios (Elaine Mardis, Washington University)
    a. In the next two years, targeted and exome applications likely will predominate the diagnostic test offerings
        i. We need validated SNVs and focused insertion/deletion mutations <u>today</u> for clinical applications
    b. A few important use scenarios:
        i. Actionable targets in cancer Tumor:Normal pairs
        ii. *De novo* mutations in pediatric diagnoses
        iii. Apparent genetic etiology (Mendelian)
        iv. Genetic disorder with multiple gene possibilities causative for the phenotype
        v. Genetic disorder but failure of previous testing (fetal also)
    c. Use of genomic reference materials for research applications and instrument/pipeline validation will likely speed clinical translation
2. Selections of Genomes (and other DNA) for reference materials (Mickey Williams, NIH/NCI)
    a. Cell lines and synthetic DNA are most practical due to renewability

   b. Advantages of synthetic DNA include easier/faster characterization, ability to blind proficiency testing, can be spiked into any clinical sample
   c. A family trio or quartet might be useful, with one genome characterized to a higher degree than the others
   d. 10 ug of DNA/vial and 1 vial is sufficient for most applications, and more units of a single vial is more desirable than fewer multi-vial units.
      i. Sufficient units must consider global use, extending beyond anticipated US-only use
   e. Possible genomes for reference materials include the four 1000 Genomes canonical trios (CEU from Utah, YRI from Nigeria, PUR from Puerto Rico, and KHV from Vietnam), samples from the Personal Genome Project, and tumor:normal cell line pairs

3. Characterization of Materials (Lisa Brooks, NIH/NHGRI)
   a. Existing data should contribute to characterization and to understand the value added by each method (e.g., fosmids and optical mapping)
   b. New sequencing should be collected from existing and upcoming technologies – use a variety of methods, including mate-pair and fosmid libraries
   c. Other methods: SNP genotyping, aCGH, HLA chip, and STR typing
   d. For validation, higher throughput methods (possibly at at one or more large-scale sequencing centers), such as RainDance/PacBio could be followed by PCR/Sanger on the most difficult sites
      i. Choose some sites randomly and some based on discrepancies
      ii. Check several hundred to a few thousand sites of each variant type
   e. Further discussion is needed regarding the format in which characterization should be presented (e.g., assembly vs. VCF vs. a new file format)

4. Performance Metrics and Figures of Merit (Justin Johnson, EdgeBio)
   a. This consortium might best be focused on metrics that can be obtained specifically from reference materials (whole genomes and spike-ins) – documentary standards that describe methods for using the RMs might be a useful output
   b. Possible metrics might include size distribution of pairs, pairing error rates, assembly completeness/phasing, truth tables, qq plots, base quality score accuracy, genotype quality score accuracy, etc.
   c. New data formats might be needed both to represent characterization of the reference material and to assess performance
   d. Is it possible to evaluate parts of the sequencing/analysis pipeline independently so that some parts of the pipeline could be changed without revalidating the entire pipeline?
   e. Mapping NA12878 reads to an NA12878 reference assembly might be an interesting way to evaluate performance of mapping/alignment software