Genome in a Bottle Consortium Work Plan

Executive Summary

Clinical application of ultra high throughput sequencing (UHTS) or "Next Generation Sequencing" for hereditary genetic diseases and oncology is rapidly growing. At present, there are no widely accepted genomic standards or quantitative performance metrics for confidence in variant calling. These are needed to achieve the confidence in measurement results expected for sound, reproducible research and regulated applications in the clinic. On April 13, 2012, NIST convened the workshop "Genome 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. A principal motivation for this consortium is to develop widely accepted reference materials and accompanying performance metrics to provide a strong scientific foundation for the development of regulations and professional standards for clinical sequencing.

At present, we expect the consortium to have four working groups with the listed responsibilities:

(1) Reference Material (RM) Selection and Design
    select appropriate cell lines for whole genome RMs and design synthetic DNA constructs that could be spiked-in to samples
(2) Measurements for Reference Material Characterization,
    design and carry out experiments to characterize the RMs using multiple sequencing methods, other methods, and validation of selected variants using orthogonal technologies
(3) Bioinformatics, Data Integration, and Data Representation
    develop methods to analyze and integrate the data for each RM, as well as select appropriate formats to represent the data.
(4) Performance Metrics and Figures of Merit
    develop useful performance metrics and figures of merit that can be obtained through measurement of the RMs.

The products of these working groups will be a set of well-characterized whole genome and synthetic DNA RMs along with the methods (documentary standards) and reference data necessary for use of the RMs. These products will be designed to help enable translation of whole genome sequencing to clinical applications by providing widely accepted materials, methods, and data for performance assessment.

An open meeting will be held August 16-17, 2012 at the National Institute of Standards and Technology in Gaithersburg, MD to receive public comment on this proposed work plan for the "Genome in a Bottle" Consortium.

Genome in a Bottle Consortium

Vision of the future

Clinical application of ultra high throughput sequencing (UHTS) for genetic, genomic, and oncology tests is rapidly growing. At present, there are no widely accepted genomic standards or quantitative performance metrics for confidence in variant calling. These are needed to achieve the confidence in measurement results expected for sound, reproducible research and regulated applications in the clinic. 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. A principal motivation for this consortium is to develop widely accepted reference materials and accompanying performance metrics to provide a strong scientific foundation for the development of regulations and professional standards for clinical sequencing.

Use Scenarios

In the next two years, targeted and exome applications will likely dominate diagnostic test offerings. Therefore, we need validated SNVs and focused insertion/deletion mutations today for clinical applications. A few important near-term use scenarios are:

- Actionable targets in cancer Tumor:Normal pairs
- *De novo* mutations in pediatric diagnoses
- Apparent genetic etiology (Mendelian)
- Genetic disorder with multiple gene possibilities causative for the phenotype
- Genetic disorder but failure of previous testing (including fetal)

In addition, development of reference materials characterized on a whole genome scale for research applications and instrument/pipeline validation will likely facilitate translation of new methods including whole genome sequencing to clinical applications.

Products

Reference Materials

Two classes of Reference Materials will be developed by this consortium:

- Whole genomic DNA from human cell lines or tissues – the consortium will select and prioritize human genomes from multiple population groups and tumor/normal cell line pairs. Sufficient DNA will be obtained to provide useful batches of homogeneous material for RMs.
- Synthetic DNA constructs that could be spiked in to any sample will also be designed and synthesized. These synthetic structures might represent particular mutations of clinical interest, or structural motifs of interest and potential utility in methods development, optimization, or proficiency testing.

The whole genome RMs will initially be released as NIST RMs, which will be characterized for genomic sequence (variants against a reference, and possibly de novo sequence), homogeneity between vials and stability. After further characterization, verification of a subset of variants, and estimation of confidence in genotype calls, any RMs can be established as NIST Standard Reference Materials (SRMs),

internationally recognized as certified reference materials of higher order. This working group will also be responsible for planning renewal of RMs to ensure an enduring supply.

Reference Data

All Reference Materials will be characterized extensively by a variety of sequencing and other methods. The consortium will develop methods to form consensus genotype calls from the datasets, with selected sites validated by orthogonal methods. These data, the consensus genotypes and associated confidence levels will also be publically released as NIST Reference Data.

Reference Methods/Documentary Standards

The consortium will determine performance metrics to be determined from the Reference Materials, and will develop well-characterized reference methods describing how to use the Reference Materials to assess performance of any particular sequencing method. These methods might include both experimental protocols and open source software for comparing measured genotypes to the consensus genotypes determined by the consortium. Some of these methods may be developed as reference methods into documentary standards through organizations such as CLSI.

Product	**Estimated Completion Date**
Initial data integration methods and set of SNP genotype calls (without well-characterized uncertainties) released for NA12878 based on existing data	Nov. 2012
DNA from large batch of NA12878 available for characterization	Dec. 2012
Complete Report of Analysis for NA12878	Jun. 2013
Release of NA12878 as first RM	Dec. 2013
Initial data integration methods for indels and set of indel genotype calls (without well-characterized uncertainties) released for NA12878	Jul. 2013
Complete whole genome sequencing of NA12878	Oct. 2013
Perform verification of selected variants in NA12878	Feb. 2014
Release NA12878 genotypes with uncertainties to convert it to a NIST Standard Reference Material	Jul. 2014
Procure DNA for subsequent RMs	??
Open-source software to obtain performance metrics from RM	May 2013

Ways of Working

An open meeting will be held in June or July 2012 to receive public comment on this proposed work plan for the “Genome in a Bottle” Consortium. The consortium will then form the four working groups outlined below to work on specific tasks. The working groups will have periodic conference calls/webinars to discuss progress and any issues that arise. The full consortium will also meet periodically.

Working groups

Reference Material Selection and Design

Tasks to be completed at August 2012 meeting:

1. Layout work plan (task list) for the working group
2. Establish a timeline for the tasks
3. Determine who will be responsible for each of the tasks
4. Whole genome RMs
    a. Establish a strategy for developing a whole genome RM portfolio
        i. Sources of RMs to consider: 1000 Genomes Project, PGP, HuRef, non-cell line?, others?
        ii. Prioritize ethnicities, genders, families, tumor-normal pairs, genomes with medically actionable mutations, etc.
        iii. Should some genomes have family member genomes also developed as RMs but characterized to a lesser degree?
        iv. Should multiple genomes be characterized in parallel, or an initial genome RM developed as a prototype?
        v. Make a recommendation regarding commercial development and dissemination of whole genome RMs from this consortium.

            Entities other than NIST could produce "secondary" Reference Materials based on NIST Reference Materials, ensuring a renewable supply for clinical applications. These secondary materials might in fact be better suited to the end-use for a variety of reasons: packaging, technical support for their use, purity, chemical or physical form or matrix, price, commercial availability in "kit" form, quantity, or other market-driven considerations.
5. Synthetic DNA RMs
    a. Discuss what types of synthetic DNA RMs might be most useful
        i. What measurement questions should they answer?
            1. variant types, sequence contexts, phasing
        ii. How long should they be?
        iii. How many should there be?
    b. Consider NCI actionable cancer mutation spike-ins from Mickey Williams' group for targeted sequencing

Future Tasks:

1. Whole genome RMs
    a. Determine priority of future RMs
2. Synthetic DNA RMs
    a. Ongoing recommendations on synthetic constructs design and application

Measurements for Reference Material Characterization

Tasks to be completed at August 2012 meeting:

1. Layout work plan (task list) for the working group
2. Establish a timeline for the tasks
3. Determine who will be responsible for each of the tasks
4. Develop a sequencing strategy for whole genome characterization
    a. Determine sequencing platforms and library preparation protocols (fragment, mate pair, others?) to be used for each RM. For now, Illumina, Complete Genomics, and ABI 5500, but consider upcoming technologies like Ion Proton and Oxford Nanopore, others??
    b. How much sequencing should be dedicated to difficult regions of the genome to result in a "finished" diploid genome?
    c. Should targeted sequencing be performed alongside whole genome sequencing?
    d. Should fosmid sequencing be performed?
    e. In addition to NIST measurements, who will contribute sequencing data?
5. Determine other characterization methods to be used
    a. SNP arrays, aCGH, optical mapping

Future Tasks:

1. Determine strategy to verify genotype calls (collaborative with Bioinformatics group)
    a. Which sites and how many should be verified with orthogonal methods?
    b. What methods should be used (e.g., Sanger, fosmids, Ion Torrent, 454, PacBio, Sequenom)?
    c. Which methods can be performed in a high-throughput manner?
    d. Who will perform the confirmation tests?

Bioinformatics, Data Integration, and Data Representation

Tasks to be completed at August 2012 meeting:

1. Layout work plan (task list) for the working group
2. Establish a timeline for the tasks
3. Determine who will be responsible for each of the tasks
4. Discuss possible methods to map and integrate data
    a. Gather existing data for NA12878 to develop methods and form preliminary consensus genotype calls?
    b. Should multiple mapping algorithms be used for each dataset?
    c. Focus first on SNPs, then indels, complex variants, and structural variants?
    d. Discuss methods to arbitrate between datasets that disagree
    e. How should difficult regions of the genome be addressed?

Future Tasks:

1. Determine methods to map and integrate data
    a. Determine how to arbitrate between datasets that disagree
    b. Develop methods to determine confidence in genotype calls and/or assembly
    c. Create a strategy to easily incorporate new data
    d. Should de novo assembly or hybrid methods be used?
2. Determine strategy to verify genotype calls with Measurements working group
    a. Which sites and how many should be verified with orthogonal methods?
    b. What methods should be used (e.g., Sanger, fosmids, Ion Torrent, 454, PacBio, Sequenom)? Which will have the most orthogonal errors and biases?
3. Determine how the characterization of the genomes should be represented
    a. Should it be represented as genotype calls, an assembly, or both?
    b. Determine how confidence in homozygous reference calls should be indicated
    c. Select and/or create file formats for representation

Performance Metrics and Figures of Merit

Tasks to be completed at August 2012 meeting:

1. Layout work plan (task list) for the working group
2. Establish a timeline for the tasks
3. Determine who will be responsible for each of the tasks
4. Identify important performance metrics that can be obtained from whole genome and synthetic RMs
    a. Should metrics for individual components of the process (e.g., library preparation, sequencing, and bioinformatics) be generated in addition to metrics for the process as a whole?
    b. are there different metrics needed for targeted sequencing?
5. Identify how these performance metrics might be obtained
6. What types of reference data for the RMs (or "Electronic RMs") might be useful and how should it be presented?
    a. Perhaps an NCBI browser?
    b. How could the reference data best be used to assess bioinformatics pipelines?

Future Tasks:

1. Develop methods to obtain performance metrics from sequencing of whole genome RMs and Reference Data
2. Develop methods to obtain metrics from synthetic DNA constructs
3. Release Reference Data for the RMs
4. Release open-source software for calculating metrics