NIST Consortium Embarks on Developing 'Meter Stick of the Genome' for Clinical Sequencing

September 05, 2012

NIST Consortium Embarks on Developing 'Meter Stick of the Genome' for Clinical Sequencing

By Julia Karow

The National Institute of Standards and Technology has founded a consortium, called "Genome in a Bottle," to develop reference materials and performance metrics for clinical human genome sequencing.

Following an initial workshop in April, consortium members – which include stakeholders from industry, academia, and the government – met at NIST last month to discuss details and timelines for the project.

The current aim is to have the first reference genome — consisting of genomic DNA for a specific human sample and whole-genome sequencing data with variant calls for that sample — available by the end of next year, and another, more complete version by mid-2014.

"At present, there are no widely accepted genomics standards or quantitative performance metrics for confidence in variant calling," the consortium wrote in its work plan, which was discussed at the meeting. Its main motivation 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."

"This is like the meter stick of the genome," said Marc Salit, leader of the Multiplexed Biomolecular Science group in NIST's Materials Measurement Laboratory and one of the consortium's organizers. He and his colleagues were approached by several vendors of next-generation sequencing instrumentation about the possibility of generating standards for assessing the performance of next-gen sequencing in clinical laboratories. The project, he said, will focus on whole-genome sequencing but will also include targeted sequencing applications.

The consortium, which receives funding from NIST and the Food and Drug Administration, is

open for anyone to participate. About 100 people, representing 40 to 50 organizations, attended last month's meeting, among them representatives from Illumina, Life Technologies, Pacific Biosciences, Complete Genomics, the FDA, the Centers for Disease Control and Prevention, commercial and academic clinical laboratories, and a number of large-scale sequencing centers.

Four working groups will be responsible for different aspects of the project: a group led by Andrew Grupe at Celera will select and design the reference materials; a group headed by Elliott Margulies at Illumina will characterize the reference materials experimentally, using multiple sequencing platforms; Steve Sherry at the National Center for Biotechnology Information is heading a bioinformatics, data integration, and data representation group to analyze and represent the experimental data; and Justin Johnson from EdgeBio is in charge of a performance metrics and "figures of merit" group to help laboratories use the reference materials to characterize their own performance.

The reference materials will include both human genomic DNA and synthetic DNA that can be used as spike-in controls. Eventually, NIST plans to release the references as Standard Reference Materials that will be "internationally recognized as certified reference materials of higher order."

According to Salit, there was some discussion at the meeting about what sample to select for a national reference genome. The initial plan was to use a HapMap sample – NA12878, a female from the CEPH pedigree from Utah – but it turned out that HapMap samples are consented for research use only and not for commercial use, for example in an *in vitro* diagnostic or for potential re-identification from sequence data.

The genome of NA12878 has already been extensively characterized, and the CDC is developing it as a reference for clinical laboratories doing targeted sequencing. "We were going to build on that momentum and make our first reference material the same genome," Salit said. But because of the consent issues, NIST's institutional review board and legal experts are currently evaluating whether the sample can be used.

In the meantime, consortium members have been "quite enthusiastic" about using samples from the Harvard University's Personal Genome Project, which are broadly consented, Salit said.

The reference material working group issued a recommendation to develop a set of genomes from eight ethnically diverse parent-child trios as references, he said. For cancer applications, the references may also potentially include a tumor-normal pair.

The consortium will characterize all reference materials by several sequencing platforms. Several instrument vendors, as well as a couple of academic labs, have offered to contribute to data production. According to Justin Zook, a biomedical engineer at NIST and another organizer of the consortium, the current plan is to use sequencing technology from Illumina, Life Technologies, Complete Genomics, and – at least for the first genome – PacBio. Some of the sequencing will be done internally at NIST, which has Life Tech's 5500 and Ion Torrent PGM available. In addition, the consortium might consider fosmid sequencing, which would provide phasing information and lower the error rate, as well as optical mapping to gain structural information, Zook said.

He and his colleagues have developed new methods for calling consensus variants from different data sets already available for the NA12878 sample, which they are planning to submit for publication in the near future. A fraction of the genotype calls will be validated using other methods, such as microarrays and Sanger sequencing. Consensus genotypes with associated confidence levels will eventually be released publicly as NIST Reference Data.

An important part of NIST's work on the data analysis will be to develop probabilistic confidence estimates for the variant calls. It will also be important to distinguish between homozygous reference genotypes and areas in the genome "where you're not sure what the genotype is," Zook said, adding that this will require new data formats.

Coming up with confidence estimates for the different types of variants will be challenging, Zook said, particularly for indels and structural variants. Also, representing complex variants has not been standardized yet.

Several meeting participants called for "reproducible research and transparency in the analysis," Salit said, and there were discussions about how to implement that at the technical level, including data archives so anyone can re-analyze the reference data.

One of the challenges will be to establish the infrastructure for hosting the reference data, which will require help from the NCBI, Salit said. Also, analyzing the data collaboratively is "not a solved problem," and the consortium is looking into cloud computing services for that.

The consortium will also develop methods that describe how to use the reference materials to assess the performance of a particular sequencing method, including both experimental protocols and open source software for comparing genotypes. "We could throw this over the fence and tell someone, 'Here is the genome and here is the variant table,'" Salit said, but, he noted, the consortium would like to help clinical labs use those tools to understand their own performance.

Edge Bio's Johnson, who is chairing the working group in charge of this effort, is also involved in developing bioinformatic tools to judge the quality of genomes for the Archon Genomics X Prize (*CSN 11/2/2011*). Salit said that NIST is "leveraging some excellent work coming out of the X Prize" and is collaborating with a member of the X Prize team on the consensus genotype calling project.

By the end of 2013, the consortium wants to have its first "genome in a bottle" and reference data with SNV and maybe indel calls available, which will not yet include all confidence estimates. Another version, to be released in mid-2014, will include further analysis of error rates and uncertainties, as well as additional types of variants, such as structural variation.

Julia Karow tracks trends in next-generation sequencing for research and clinical applications for GenomeWeb's *In Sequence* and *Clinical Sequencing News*. E-mail her here or follow her GenomeWeb Twitter accounts at @InSequence and @ClinSeqNews.

Related Stories

- U Michigan Partnering with IGC to Create Sequencing-Based PGx Non-Profit
  September 5, 2012 / Pharmacogenomics Reporter
- CAP Publishes Accreditation Checklist for NGS in Clinical Labs
  August 1, 2012 / Clinical Sequencing News
- Life Tech Sees Clinical Promise for NGS Panels, Exomes; Says Multiple Platforms to Drive Dx Strategy
  August 1, 2012 / Clinical Sequencing News
- At AACC, NHGRI's Green Lays out Vision for Genomic Medicine
  July 16, 2012 / GenomeWeb Daily News
- In Educational Symposium, Illumina to Sequence, Interpret Genomes of 50 Participants for $5K Each
  June 27, 2012 / Clinical Sequencing News

footer