Groups closing the gap in reference materials for sequencing assays

William Check, PhD

March 2015—It's a truism in the clinical laboratory that your results are only as good as the reference standards available to QC your assay. For measuring small analytes like glucose that's not a problem.

However, in clinical laboratories the analyte in question increasingly is DNA. In the past five years, next-generation sequencing has been adopted to detect variants in small targeted regions of specific genes, which is useful in oncology and medical genetics. More ambitious applications of NGS—whole genome and whole exome sequencing—have recently begun to enter the clinical realm as well.

At 3 billion base pairs in length, and with any of four bases at every position, human DNA presents an unprecedented challenge to analytic methods and the development of reference standards. And because human DNA is diploid, every base position can be homozygous or heterozygous. Moreover, medical consequences can arise not just from a change in the base occupying a single position—single nucleotide variants—but from larger structural variants, such as the insertion of an extra stretch of DNA where it doesn't belong or the deletion of a small or large segment of DNA, collectively called indels.

Dr. Church

Given this complexity, it's impressive that NGS is as good as it is. But it still falls short for more demanding applications. "Right now whole genome approaches do not really give you whole genome analysis, because the whole genome is not accessible," says Deanna Church, PhD, director of genomics and content at Personalis.

Or, as Arend Sidow, PhD, likes to say, "We kind of pretend to be sequencing whole genomes right now, but we're really not."

"Long-read technologies have the potential to get us much closer to that goal," says Dr. Sidow, who is associate professor of pathology and of genetics at Stanford University. For long-read research, Dr. Sidow's group works with the Oxford Nanopore platform.

"Right now we are really not very good at deciphering structural variants of intermediate size," Dr. Sidow tells CAP TODAY. "We are good at detecting single nucleotide variants in those parts of the genome in which next-generation sequencing works—approximately 75 percent of the genome—and very large variants, hundreds of kilobases. But there is much genetic variation that is between those sizes, and current sequencing technology is not very good at supporting the discovery of those variants. That's where long-read approaches come in."

Dr. Sidow

In 2003 when the "complete sequence" of the human genome was announced, this deficiency was already recognized, although it was downplayed. The answer to a frequently asked question posted on the site of the National Human Genome Research Institute said: "Within the limits of today's technology, the human genome is as complete as it can be. Small gaps that are unrecoverable in any current sequencing method remain…." (www.genome.gov/11006943).

Since then, appreciation for those "small gaps" has grown. As one research group wrote recently, "The human genome is arguably the most complete mammalian reference assembly, yet more than 160 euchromatic gaps remain and aspects of its structural variation remain poorly understood ten years after its completion" (Chaisson MJ, et al. *Nature.* 2015;517:608–611).

Added to this is the great heterogeneity among "normal" genomes. Justin Zook, PhD, of the National Institute of Standards and Technology, tells CAP TODAY: "It's really amazing how much normal variation there is between individuals. We are just starting to understand it. Many of these variations fall into parts of the genome that are difficult to sequence. It can be particularly challenging to detect structural variants."

All of these complexities in the structure of DNA make it difficult to establish a complete human reference genome. And without a complete reference genome, a laboratory can't know whether it has identified all the variations in a patient's DNA. In heritable conditions with no known cause, that complicates the task of assigning pathogenesis.

There are two different references: the "reference genome," a model assembly of the human genome that includes a sampling of variation across the genomes of a number of individuals, and "reference material genomes," which are single genomes well-characterized for variants against the reference assembly. "These reference material genomes are being sufficiently characterized that they will be among the 'best' human genomes, and are intended to act as gold-standard benchmark samples," says Marc Salit, PhD, of NIST.

Dr. Salit is leading a project called Genome in a Bottle, or GIAB, the aim of which is to generate highly accurate reference material genomes as a public-private-academic partnership. As part of this initiative, Dr. Salit, Dr. Zook, and colleagues determined just how much of the genome can be characterized accurately with current methods. NGS in its present incarnations, they estimate, can make high-confidence calls for 78 percent of the genome (Zook JM, et al. Nat Biotechnol. 2014;32:246–251). "[T]here is a need for a highly accurate set of genotypes across a genome that can be used as a benchmark" for human genome sequencing, the authors wrote. NIST will make samples of genomic DNA of these genomes available as reference materials from the NIST Standard Reference Materials Program.

Dr. Ashley

Euan Ashley, MRCP, DPhil, is a collaborator in the GIAB project. "This is a really important area," he says. "My group is focused on genomics for clinical medicine." Dr. Ashley is associate professor of medicine and genetics and, by courtesy, pathology at Stanford University; director of the Stanford Center for Inherited Cardiovascular Disease; and co-director of the Stanford Clinical Genomics Service. He finds "great excitement," he says, in the use of this new technology. "We want to move this technology toward clinical grade. It is transformative for medicine. Yet we need to concentrate more on algorithms to bring the standard of sequencing up to the standard we expect for clinical medicine," he says.

In addition to GIAB, several other projects aim to generate highly accurate human genome sequences. One group that has posted results already is the Genetic Testing Reference Materials Coordination Program, or GeT-RM (wwwn.cdc.gov/clia/Resources/GeTRM/default.aspx). Its coordinator is Lisa Kalman, PhD, health scientist in the Laboratory Research and Evaluation Branch of the Centers for Disease Control and Prevention.

“There are now tests for about 5,000 genetic conditions. But there are very, very few reference materials available for these tests, maybe 50 different materials in 50 different genes,” Dr. Kalman says. “That’s a huge gap. We need reference materials to develop new genetic tests, to validate genetic tests, for QC and also for proficiency testing or alternate assessment activities.”

Fifteen years ago the CDC recognized that gap and started projects to address it, including GeT-RM. “Our program’s purpose,” Dr. Kalman explains, “is to make publicly available highly characterized genomic DNA samples that laboratories can use. Having reference materials will enable labs to validate tests that look at variants in all parts of the genome and to assess the accuracy of the tests.”

Addressing the growth in our understanding of the complexity of the human genome in the past decade, Robert Sebra, PhD, says, “The information content of the genome just gets bigger and bigger as sequencing technology expands. We conduct a variety of R&D using PacBio [Pacific Biosciences] single-molecule sequencing and other long-read technologies toward better detection of structural variants involved in inherited disease.” He is assistant professor of genetics and genomic sciences at the Icahn School of Medicine at Mount Sinai in New York.

“What’s really important,” Dr. Sebra says, “is that no one technology is going to give you the answer to everything. In an ideal scenario, a single technology would offer long reads and accuracy and high throughput. That is not currently the situation. Right now we have to take every case and match it to the current technology that best addresses the types of variants necessary on a given genetic panel. We can do much with available platforms, but for many structural variants and for addressing the unresolved regions of the genome, we will need even longer reads.”

To initiate the GeT-RM effort, Dr. Kalman gathered clinical labs, next-generation sequencing companies, representatives of the National Institutes of Health, and other stakeholders on a conference call. “I asked, ‘How can we address this gap [in reference materials]? Labs are starting to do whole exome and whole genome sequencing but there are no materials to assess their assays.’”

Dr. Kalman

“We selected a genome, NA12878, that many people had chosen to sequence: the HapMap project and the 1,000 Genomes project. So there was already considerable data on it.” They also worked on another frequently sequenced genome, NA19240.

GeT-RM collected preexisting data on those samples—there were many available data sets—and sent samples of NA12878 DNA from the Coriell Institute to many clinical testing laboratories. “We asked them to test it using whatever method they used in their lab and to send us the data,” Dr. Kalman says. “We got lots of data.” To compile the data, they worked with the National Center for Biotechnology Information to regularize the data into a common VCF standard. If labs provided BAM files, those files were converted to a standard format. Dr. Church, who was then at the NCBI, worked with GeT-RM to create a browser containing the aggregated sequence data. By that time, the National Institute of Standards and Technology had produced sequence data and that, too, was included

in the browser.

On the browser, a user can see calls and read alignments for any region of the genome that has been assayed. “You can customize it to your use and look at it in many different ways,” Dr. Kalman says. “You can upload your sequence into the GeT-RM browser and compare it to all existing data sets to see how well you did. Does your data agree with the GeT-RM data sets? If not, you can try to understand how your data differ. You can also download data from the browser into your computer to compare it with your data.”

According to Dr. Salit, GIAB started in about 2009 as an offshoot of NIST’s work in the microarray and gene expression group, making external RNA controls. A joint coordinating center was established at Stanford, called Advances in Biological and Medical Measurement Science (ABMS; https://sites.stanford.edu/abms/).

Dr. Salit, leader of genome scale measurements for the NIST-ABMS program, calls GIAB “an infrastructure for assessment of NGS performance,” and says, “We are developing standards to address the lack of accepted metrics to evaluate the fidelity of variant calls from NGS.” Officially, GIAB’s goal is to “develop reference materials, reference data, and reference methods needed to assess human WGS.”

In a talk about GIAB at last year’s Association for Molecular Pathology meeting, Dr. Salit showed that existing whole genome sequencing platforms disagree about hundreds of thousands of variants; concordance was reached on only 80 percent of calls. “Even where all three platforms make the same call, that doesn’t mean they are right,” Dr. Salit said in an interview. “All three could be wrong.” Bioinformatics programs also have extensive disagreement (O’Rawe JA, et al. *Trends Genet.* 2015;31:61–66).

Dr. Salit notes that when laboratories are sequencing for a particular condition involving only a small portion of the genome and focused on medically actionable variants, they can do highly accurate and reliable work. However, he says, “At the whole human genome scale, there are 3.5 million single nucleotide variants between any individual’s genome and a reference, such as NA12878.” In addition, every person’s genome has thousands of structural variants. When clinical laboratories are doing sequencing in the whole genome or whole exome dimension, substantial disagreement among platforms and bioinformatics programs becomes a factor. Well-characterized reference genomes will help evaluate the performance of several analytical steps, from library preparation to estimation of confidence.

Dr. Salit

“Our goal is to develop large numbers of reference samples that we can distribute to our customers—clinical laboratories and technology companies,” Dr. Salit says. They began with NA12878 as a pilot. However, this genome has only “legacy consent,” which means there is no consent for commercial use. “It may be used in the context of a commercial project, but no products can be derived from it,” Dr. Salit explains.

For commercial development, GIAB plans to work with genome trios—DNA from two parents and a child—obtained by the Personal Genome Project (www.personalgenomes.org). Currently they are working on two ethnic trios, one of Ashkenazi Jewish descent and one Han Chinese. All of these genomes are from persons who have no known pathology or inherited disease. “Because an advanced consent was used in the project, commercial genomes may be developed from these genomes,” Dr. Salit explains. Acrometrix and Horizon Discovery Group are marketing controls derived from GIAB genomes. Release of the genomes as reference materials is planned for the end of this year.

As a member of the Advances in Biomedical Measurement Science program and a collaborator with GIAB, Dr. Zook is developing methods to compare and integrate whole genome DNA sequencing data from multiple platforms and sequencing runs. He led the effort to determine the 78 percent high-confidence calls figure.

Dr. Zook

“Basically, we took data from five different sequencing technologies and developed a process to integrate them,” Dr. Zook says. The integration process was designed to take advantage of the known strengths of the different platforms. Seven read mappers and three variant callers were used. In all, 14 data sets were available. All data were from NA12878, with the GRCh37 genome used for comparison.

Dr. Zook and the NIST collaborators have also done initial work trying to understand different types of structural variants and how to call them from sequencing data. Personalis had generated sequence data from the entire pedigree of the pilot genome (NA12878 comes from a pedigree that has been analyzed) looking for structural variants present in multiple members of the pedigree. “They came up with a list of about 2,300 structural variants for which there was really good evidence,” Dr. Zook says.

Dr. Zook and his colleagues next generated a random set of regions of the genome corresponding to the same sizes as the structural variants. “These will not be structural variants,” he says. They analyzed all of the segments according to several properties, such as depth of coverage, number of discordant paired-end reads, and percent GC (guanine-cytosine) content, and ran them through an annotation program to do unsupervised machine learning and principal components analysis or clustering.

Four distinct groups of DNA segments emerged. Almost all of the random regions and only a few candidate structural variants fell into one group. The Personalis deletions clustered into three separate categories that contained only a few random regions. For instance, one group contained homozygous deletions of alu elements, repeat elements that tend to be highly variable among people.
“In the longer term,” Dr. Zook says, “we hope we can take any candidate structural variant, see where it falls in the clusters, and classify it that way. These structural variants from Personalis are fairly simple. Many other groups are starting to look at more complicated types of structural variants to see if they cluster.”

In a third project, the Global Alliance for Genomics and Health Benchmarking Team, of which Dr. Zook is the chair, a comparison tool is being developed. After a laboratory sequences the NA12878 genome, it can use the tool to put in the data and calculate the sensitivity and specificity of their sequence against the NIST GIAB high-confidence call set. Dr. Kalman is hoping to put a link to that tool on the GeT-RM browser when it comes out in a few months.

Dr. Kalman calls Personalis’ Dr. Church “the main architect of the NCBI GeT-RM browser.” Dr. Church in turn calls participating laboratories “very gracious” for resequencing or validating their variant calls. “They are to be commended for their patience in working with us,” she says. “One challenge we faced was that different laboratories have different processes and procedures, so all laboratories ended up giving us their data in different formats. We had to put all that data into a similar format so we could compare data sets.” Dr. Church notes that bioinformatics formats are flexible, which is good for research. For integrating all data into a common format, however, flexibility was an obstacle. “We had to communicate with laboratories so that we expressed their data faithfully,” she says.

Dr. Church also reports that in some regions there were wide differences in calls from different laboratories. “One fascinating thing is that in some regions of the genome many different methods get the same answer, whereas in

other regions they don't. It may not always be clear what the right answer is.

"It is important to note that the GeT-RM browser focuses on single nucleotide variants and small indels, which make up about 80 percent of all variants," Dr. Church adds. "Our ability to detect structural variants is not as robust."

The ultimate goal of the GeT-RM browser, Dr. Church says, is to help clinical laboratories assess the clinical validity of their NGS assays. "You sequence the DNA, then compare your results to the GeT-RM database." Another important benefit is that it helps in designing a test. "It shows you what your validity is in being able to call a breadth of variants as well as variants throughout the genome," Dr. Church says. "Because in many clinical cases you are not going to know what variant is causing the disorder in that patient or where it is."

People have pointed out that NA12878 is a relatively healthy genome and might not cover all variants that could be seen in a clinical laboratory. "So validating on NA12878 is necessary but may not be sufficient to assess the total validity of your test," Dr. Church says.

She and her colleagues are now proposing to look for data in the GeT-RM data set that could fill in the gaps in the GIAB high-confidence call sequence. "Those gaps make up about 25 percent of the genome," Dr. Church says. "It's an ongoing project. We don't know whether it will work."

Dr. Ashley of Stanford calls the work of GIAB "really important."

"They identified areas of the genome where we can have high confidence in the calls we make. We've been working together, and we have a manuscript in preparation in which we explore the intersection of high-confidence regions and the medical genome. Between 3,000 and 7,000 genes have been clearly associated with human disease. We hoped to find the whole medical genome in high-confidence regions as we define them," he says, "but of course that is not what we found. In a typical group of variants known to be pathogenic, as defined in ClinVar, maybe 20 percent of them are not in high-confidence regions. In the paper we explore why that might be."

In an article published about a year ago, Dr. Ashley and colleagues identified some of the challenges to clinical interpretation of whole genome sequencing and described shortcomings of WGS (Dewey FE, et al. JAMA. 2014;311:1035–1045). "Exome capture kits are very problematic causing some areas of the genome to be not well covered," Dr. Ashley tells CAP TODAY. "Median coverage is commonly 60 or 80 or higher. But some whole exons are effectively missing from exome capture kits. The first exon especially can be challenging, because they are GC rich." Dr. Ashley cites a whole exon effectively missing from KCNH2, one of the genes responsible for long QT syndrome.

According to Dr. Ashley, whole genome sequencing has more even coverage of the genome. "But even there, major medical genes are incompletely covered," he says.

"We need to bring this technology up to clinical grade," he says.

One step in improving NGS could be to incorporate long-read technologies. Dr. Sidow is using the Oxford Nanopore platform to sequence cancer genomes. "If you are missing a large class of variants, you will not be able to identify the cause of all disease," he says. The issue with short-read technologies like the Illumina and Life Technologies platforms is that the genome has repeats, segments that are similar to one another. "You can't unambiguously place those redundant regions unless they are in very large reads," Dr. Sidow says. "Some regions of the genome will always be dark, but most can be accessed with long-read technology.

"Everyone assumes there will be a large amount of disease explainable by structural variants and copy number variation," Dr. Sidow continues. These medium-size variants affect many bases at once, so they have the potential to cause havoc. "They also tend to be rare, and population sequencing is not so good at finding rare variants. If only you and your relatives have that particular variant, it won't ever show up in population studies. For that reason it is quite important to have accurate long-read sequencing."

However, Dr. Sidow identifies an issue with long-read sequencing technologies. “They actually work on single molecules,” he says. “Literally you have a single molecule of DNA for which you are identifying the base sequence. Other next-generation technologies sequence a population of identical molecules, so they have a very good signal-to-noise ratio. Long-read protocols, on the other hand, generate a signal from a single molecule, so they are much noisier. Therefore, they have a much higher per-base error rate.

“When you talk to people who really love these technologies,” Dr. Sidow says, “they downplay this feature. Everyone is very excited about Oxford Nanopore and PacBio right now. They are the first technologies that really make long reads accessible.” Comparing the two instruments, Dr. Sidow notes that PacBio is still very expensive and has a slow turnaround time. “I don’t think that will change,” he says. And Oxford Nanopore is still in what he calls its “technological infancy.” It has a low throughput and can do a yeast genome but not a human genome.

Dr. Sidow identifies a third completely different approach, which gives the benefit of long-read instruments with high base accuracy. Two of these programs are now under development, Moleculo by Illumina and 10X by 10X Genomics. Explaining how they work “gets into the technological weeds a bit,” Dr. Sidow says, “but basically they do a trick. They take long molecules and amplify them as short molecules but retain a great deal of information about where the short fragments originally came from in a long molecule.” Dr. Sidow notes that this approach is in its infancy but he believes it has significant long-term potential.

Dr. Sebra is also doing long-read sequencing, using PacBio instruments to sequence GIAB’s Ashkenazi Jewish trio in collaboration with Eric Schadt, PhD, and Ali Bashir, PhD, of Mount Sinai. They have already generated an NA12878 reference genome using the PacBio data alongside new technology from Bionano Genomics, whose optical mapping platform, Irys, manages long DNA reads.

Long-read sequencing of the Ashkenazi Jewish trio is 75 percent done, Dr. Sebra says. “We have about 60× coverage of the child and 15× coverage of the two parents. Our target is to do a minimum of 60× on the child and 30× on both parents.” When finished, this would be the first Ashkenazi Jewish reference genome at high resolution. Median read length with the PacBio reads is 11,000 nucleotides. In contrast, currently available platforms achieve read lengths in the hundreds.

Dr. Sebra

Aside from his work with GIAB, Dr. Sebra is also using long-read sequencing technology to derive targeted assays for longer, structurally difficult regions of the genome, such as the FMR1 gene or in genes with associated pseudogenes like GBA (glucosidase, beta, acid), responsible for Gaucher’s disease. “We can do that efficiently without having to sequence the entire genome, focusing on long-read sequence panels that target the structural variants of interest,” he says. Such panels can also be made for a variety of inherited disease and oncology applications, negating the need to perform a variety of inefficient analytical methods that remain time-consuming.

Dr. Sebra calls this work “more exploratory.” It would require completing gray areas in the genome in the 1.5 kb to 20 kb range and beyond, “to truly address the unknowns in our current screening capabilities.”

Which brings us back to completing the genome for an accurate reference standard. Dr. Sebra refers to recent work by Evan Eichler, PhD, of the University of Washington, which shows that long-read technology can resolve structural variation in the range of tens of thousands to hundreds of thousands of base pairs (Chaisson MJ, et al. *Nature.* 2015;517: 608–611).

“With longer-read technologies on the horizon, we can offer a more complete human genome,” Dr. Sebra says.

“These new technologies can drive read lengths toward the 100 kbps domain to get to a more complete genome and to begin to address the dark matter in the genome. Right now the significance of those domains is unknown. We won’t know until we discover and study them. But when we start to map those domains, we could very well find more disease-related regions.”

[hr]

William Check is a writer in Ft. Lauderdale, Fla.