Study highlights need for better characterized genomes for clinical sequencing

March 1 2016

This is the end result of a DNA sequencing process. Each color represents one of the four base chemicals that make up DNA (adenine, guanine, cytosine and thymine). NIST's genome reference material is a benchmarking standard that can help labs determine how well their DNA sequencing processes are working. Credit: Gerald Barber, Virginia Tech University

A new study that assesses the accuracy of modern human-genome-sequencing technologies found that some medically significant portions of an individual's DNA blueprint are situated in complex, hard-to-analyze regions that are currently prone to systematic errors.

These genes and gene segments lie in yet-to-be-benchmarked regions that presently make up almost a fourth of the human genome's 3.2 billion pairs of chemical building blocks.

Stanford University and National Institute of Standards and Technology (NIST) researchers write that their findings should be a "call to arms for those interested in clinical grade technical accuracy for genome sequencing." As genome sequencing transitions from research to clinic, they say, it is essential to have methods to benchmark performance in all regions that are sequenced for diagnostic or other medical purposes.

Challenges in benchmarking difficult, but clinically important regions of the genome are reported in today's issue of *Genome Medicine*. The results underscore the need to extend benchmarking references against which sequencing data and analyses can be compared and validated.

In effect, these types of standards are quality-control and quality-assurance tools. They are necessary for checking the accuracy of sequencing data and analyses—and preventing false positives and false negatives. However, genome-sequencing technologies aimed at the large health care market are advancing so quickly that efforts to develop the field's underpinning benchmarking tools must race to keep up.

Central to the Stanford-NIST study, one such tool is the genomic reference material created by NIST and its partners in the Genome in a Bottle consortium. The NIST reference material—NIST RM 8398,

Human DNA for Whole-Genome Variant Assessment—currently has about 77 percent of the genome characterized with high levels of confidence.

"The harder-to-characterize regions that we can't yet sequence with confidence include regions known to be clinically important," explains NIST biomedical engineer Justin Zook. "This means that our benchmark genome cannot currently be used to assess performance for more challenging genes and other difficult regions of the genome that already are being tested or for which new sequencing methods are being developed."

"The good news is that, in this case, 77 percent of the donor's genome was reliably sequenced using current methods," says lead author Rachel Goldfeder from Stanford University. "The challenge now is to focus our efforts on the other 23 percent—namely, on regions of the genome that remain elusive. Only then can we realize the full potential of precision medicine."

In their study, Stanford and NIST researchers used data from whole genome sequencing and whole exome sequencing methods. Exome sequencing focuses only on the protein-encoding portions of genes, comprising less than 2 percent of the entire genome.

Both types of these so-called next-generation sequencers follow a similar process. Paired strands of DNA are uncoupled and randomly chopped into short segments. Numerous copies of the segments are made and then are sequenced by recreating the missing paired strand for each copy. The matches are analyzed to determine their sequence of letters from the from the four-letter genetic alphabet: A (adenine), C (cytosine), G (guanine) and T (thymine).

Then, bioinformaticians apply complex mathematical algorithms to

determine where the decoded pieces originated. The pieces can then be compared to a defined "reference sequence" to identify variations in stretches of letters and where letters have been deleted or inserted in specific genes. When differences are found, a "variant call" is logged.

For RM 8398, the Genome in a Bottle consortium had catalogued high-confidence variant calls in the well-characterized regions of the benchmarking genome. The Stanford-NIST team compared these calls with variant calls made with two sequencing systems. Of particular interest were differences in 56 "medically actionable" genes that the American College of Medical Genetics and Genomics (ACMG) recommends for reporting.

Accuracy of variant calls within high-confidence regions depended on the genome region; type of difference—say, an inserted or substituted letter; extent of coverage (number of times a specific DNA segment has been read); and analytical methods.

In whole genome sequencing, for example, false negative calls—unidentified variations or mutations—resulted largely from software tools used to filter out errors in sequencing data, the researchers found. Most false negatives in whole exome sequencing stemmed from poor coverage—not enough reads to generate data of sufficient quality.

In some ways, significant parts of the genome are largely uncharted territory. Only about 5 percent of the 19,000 to 21,000 protein-encoding genes are situated entirely within portions of the human genome currently characterized with high confidence.

Unlike many research studies of groups of people, a "false call on a clinical report" can result in harmful consequences for patients, their families and even groups of people at risk for specific diseases, the researchers explain. Therefore, they say, it is critical to understand how

accurately all regions of interest can be tested.

The team also points out that because current sequencing technologies are prone to systematic errors at certain genome locations, some variants reported in publically available genome-sequencing databases may actually be false positives, or it may be difficult to distinguish between real variants and systematic sequencing errors.

The Stanford-NIST team found that, on average, about a fifth of each of the 56 disease-related genes flagged by ACMG is situated outside well-characterized, high-confidence regions of the NIST reference genome. Addressing this "sobering" state of affairs, the researchers write, requires working toward consensus across technologies or "at the very least," transparency in communicating the confidence level for every variant call.

More information: Rachel L. Goldfeder et al. Medical implications of technical accuracy in genome sequencing, *Genome Medicine* (2016). DOI: 10.1186/s13073-016-0269-0

Provided by National Institute of Standards and Technology