Need for accuracy

Jonas Korlach
Chief scientific officer of Pacific BioSciences

Genomes include regions that are relatively easy to sequence, but they also contain regions that have flummoxed most sequencers. Indeed, to this day, there is not a single complete, perfectly sequenced human genome in existence

As DNA sequencing is increasingly adopted — from clinical applications to species conservation — it may be surprising to learn that accuracy continues to be a challenge for many sequencing technologies.

We often hear the analogy that a genome is like a book, and DNA sequencers are deployed to read each letter and word in that book. However, it's not nearly that straightforward. Genomes include regions that are relatively easy to sequence, but they also contain regions that have flummoxed most sequencers. Indeed, to this day, there is not a single complete, perfectly sequenced human genome in existence.

Still, the sequencing technology development community is making progress. By incorporating the latest computational advances, along with more robust sequencing methods, our march toward perfect accuracy continues. This will be important for a range of clinical and industrial applications.

As new DNA sequencing technologies have emerged during the past couple of decades, the methods underlying them have demonstrated different strengths and weaknesses. For example, some platforms struggle with homopolymers, or long stretches of the same nucleic acid. The ability to distinguish between six or seven guanines in a row is challenging, but it's also essential for generating an accurate genome sequence and making reliable predictions about genome function.

All sequencing tools have some kind of error profile. For those characterised by systematic errors, it is quite difficult to improve accuracy. But for sequencing systems that have random error profiles, a simple approach has been shown to be remarkably effective in boosting accuracy.

This approach relies on the idea that randomly distributed errors can be identified and filtered out with enough data. A single molecule, real-time (*SMRT*) sequencer can be programmed to read the DNA of the same molecule over and over, circling around the same sequence many times. During analysis, each single pass of that molecule can be stacked up against all the others and a consensus sequence can be generated using only the nucleic acid identifications that appear again and again. This effectively washes out any random errors and leads to a highly accurate final genome assembly.

To better understand why accuracy in DNA sequencing makes a difference, consider a few clinical and industrial applications.

Pathogenic variant discovery. Scientists around the world are racing to find the genetic mechanisms underlying previously unexplained diseases. Mounting evidence suggests that some of the diseases that have defied explanation are associated with elements in the genome that have proven hard to sequence, such as repeat expansions or complex structural variants. Filling in the catalogue of pathogenic genetic variants is essential for precision medicine, but it will require the use of highly accurate methods such as *SMRT* sequencing to ensure that each variant is correctly identified. Recently, scientists have found a number of disease-causing variants this way, including tandem repeats associated with a form of X-linked intellectual disability.

Clinical quality control. The rise of two innovative approaches could be transformative in health care: gene therapy and CRISPR gene editing. For both of these, it has become clear that better quality control is sorely needed. For example, scientists have analysed gene therapies with *SMRT* sequencing, finding that they contain far less of the desired DNA sequence than expected, and that genetic errors would have rendered many non-functional. Similarly, analyses of CRISPR gene editing experiments have shown that there are more unwanted edits than expected. In both cases, highly accurate analysis through *SMRT* sequencing could be used for quality control to filter out gene therapies or CRISPR-edited sequences that harbored unwanted DNA changes.

Industrial biotechnology. Many research teams have turned to microbes for the production of energy, chemicals, medication ingredients, and more. Synthetic biology allows scientists to alter microbial genomes, modifying these tiny organisms to optimise the traits needed for a particular kind of production. This approach relies on highly accurate sequencing; even the smallest error in characterising a microbial genome could send scientists down the wrong path. Researchers at the US Department of Energy's Joint Genome Institute recently used *SMRT* sequencing to analyse fungal genomes, looking in particular for bioenergy features and other metabolites of interest. By generating more accurate genome sequences, they were also able to make more accurate predictions about genes and pathways involved in key traits.

So what's next? Highly accurate *SMRT* sequencing has opened the doors to producing better genome assemblies for humans and other organisms, discovering genetic elements that have previously resisted characterisation, and determining the quality of gene-editing experiments, among many other important uses. Now that such accuracy is possible, we can move forward with confidence as DNA sequencing is incorporated into a broader range of clinical and industrial applications.