Genome sequencing's big fix

Mike Schatz uses the latest technology to assemble and validate large-scale genome sequencing projects.

The age of individual genome sequencing is almost upon us. There are already companies, like 23andMe, whose sole business is to sequence your genome and inform you of potential disease risk factors in your DNA. But to truly make this a globally accessible technology there are still some hurdles to overcome.

CSHL Assistant Professor Michael Schatz has set out to improve genome sequencing by removing some of those hurdles. One important concern is the accuracy of the sequences being produced. Just as your eye might mistake a letter or word when reading, sequencing machines sometimes misread a unit of DNA, known as a base pair, in a genome sequence. Schatz's group made a big splash earlier this year, when it achieved a breakthrough in just this area.

Sequencing technology's rapid advance

Sequencing technology has come extremely far in very little time. It took $3 billion and years of work to sequence the first human genome, which was finished in the year 2000. Now, the cost of sequencing the entire human genome is down to about $3000 and can be completed in a matter of days. This is currently done using a technique known as second-generation sequencing.

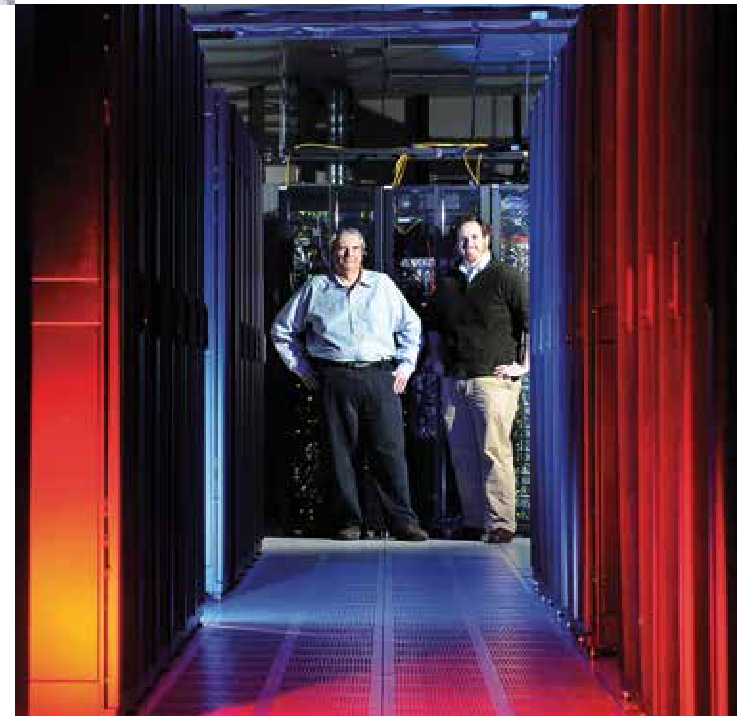
It was the advent of second-generation sequencing technology, or "2nd-gen" as it is called, that rapidly advanced the field. Small DNA molecules — extremely short pieces of the full genome — are copied many times over. These fragments are then analyzed all at once, generating lots of short sequences, also known as "reads" (i.e., genome segments read by a sequencer). Doing it this way ensures high accuracy in the DNA sequence output.

The downside, however, is the length of those short reads, which are typically between 100 and 200 DNA base-pairs long. The human genome is huge, about 3 billion base pairs. "It's like sifting through billions of tiny jigsaw pieces when the overall picture is of a blue sky," says Schatz, "and this becomes a real limitation when you are trying to see the big picture."

Third-generation sequencing, a new technology released for beta testing about two years ago by Pacific Biosciences, offers significantly longer reads. It manages this by sequencing one comparatively much longer DNA molecule at a time. "The very special thing about 3rd-gen sequencing is you can generate these very long reads," says Schatz. Indeed, while they typically range in the thousands of base pairs, the longest read Schatz has seen from a 3rd-gen machine is tens of thousands of base pairs in length.

Comparatively puny, the 100 to 200 base pair 2nd-gen reads are far shorter than the average length of a human gene, which is about 3000 base pairs. The major advantage of long reads is that they often cover much more than the sequence of a single gene, making it easier to resolve the sequence of many genes in a row. This allows researchers to more accurately assemble the full genome sequence.

But even with this improved method there is a crucial trade-off. Third-generation sequencing is considerably less accurate than its predecessor. The error rate can be as high as 15%, meaning roughly 1 in 6 base pairs will be read incorrectly. So while in theory those long reads would be a major leap forward for genome assembly, Schatz notes that "the quality is so low that you can't directly use those reads for a lot of things you want to do."



Schatz and McCombie among the vast computing servers that power their research.

Best of both worlds

Motivating Schatz to find a fix for these errors was the goal of providing highly accurate versions of complete genomes. Such a solution would be extremely relevant across multiple disciplines, e.g., basic research and applied medicine, as well as for biotechnology companies involved in making sequencers.

As a computer scientist, Schatz gravitated toward a software-based solution. He and his collaborators, including CSHL professor W. Richard McCombie, Ph.D., a sequencing pioneer, came up with a hybrid approach that fuses the best aspects of both 2nd- and 3rd-gen sequencing.

"We wrote a software program that takes the short, accurate reads from 2nd-gen sequencing and uses them to polish the long reads generated by 3rd-gen technology," Schatz explains. The software scrubs out the mistakes, reducing them to a paltry rate of about 1 error in 100 base pairs.

Generating sequences is relatively easy; it is assembling them into the full genome sequence that is the hard part. This is made much harder if there are errors in the sequence. So while the software solution Schatz and his colleagues developed means sequencing a sample using both 2nd-gen and 3rd-gen methods, they maintain it is worth the effort. There is a huge benefit, they say, in the amount of time and money saved when assembling the full genome.

After publishing their work in a *Nature Biotechnology* paper earlier this year, Schatz and his team are now eager to apply their software fix to new biological and medical research problems. His group is already engaged in collaborative study with CSHL geneticist Mike Wigler to look for disease markers in the genomes of children with autism.

This research is possible only because of advances in sequencing technology, including Schatz's own software solution. Using these new methods Schatz and Wigler can sequence, assemble, then study the genomes of thousands of children with autism for their project. This gives them a much greater chance of finding something significant.

Edward Brydon