THE NEXT FRONTIER IN ASSEMBLY: LONG READS OFFER FINISHED GENOMES

PacBio® long reads are changing the game in genome finishing. A new error correction pipeline developed by veterans in genome assembly allows scientists to add long-read data to their short reads and finally finish all those incomplete genomes.

A decade in genome assembly would have some scientists eager to move on to a different challenge. Not so for Michael Schatz, who says that recent innovations with long-read technology are reinvigorating the community. "The frontier for assembly is changing rapidly," he says. "It's a really exciting time to be sequencing genomes."

Schatz is an assistant professor at Cold Spring Harbor Laboratory who contributed to a recent assembly project led by Adam Phillippy and Sergey Koren at the National Biodefense Analysis and Countermeasures Center. Their goal: apply the long reads generated by the PacBio® *RS* instrument to dramatically improve the quality of the assembled genomes and even assemble high-quality finished genomes directly from the reads. The results of their work were published in *Nature Biotechnology* on July 1, 2012.

Phillippy and Schatz started out in genome assembly more than 10 years ago, and Schatz recalls from their time at The Institute for Genomic Research that in the microbial assembly projects they worked on, "it was extraordinarily expensive to do that final step of finishing — to try to close every gap."

Dr. Michael Schatz, Assistant Professor
Cold Spring Harbor Laboratory

That was in the days of Sanger sequencing, considered the gold standard foundation for high-quality genome assemblies. When short-read technologies became popular a few years later, Schatz and his colleagues saw assembled genomes take a turn for the worse as the number of contigs increased significantly and repeats, segmental duplications, and gene families became more difficult to assemble correctly. Since then, Phillippy, Koren, and Schatz have been on a mission to get back to high-quality genomes — but without the substantial expense of Sanger finishing.

"The team has proven the effectiveness of the pipeline with a range of genomes, from microbes to higher eukaryotes, and 'it works extremely well'."

When they heard that a startup company called Pacific Biosciences would be launching a long-read sequencing platform, "we were really excited about the idea that these reads would have the capability to solve a lot of challenges that we'd been facing for years," Schatz says.

As with any new technology, the novelty of PacBio's SMRT® sequencing approach meant that the scientists first had to learn how to assess and apply the data. Raw read error rates were higher because of the single molecule nature of the platform.

"We were really excited about the idea that these long reads would have the capability to solve a lot of challenges that we'd been facing for years."

Just as short-read sequencers improve individual read accuracy by layering many reads together and reporting only the consensus base calls, Phillippy, Koren, and Schatz were convinced that it would also be possible to optimize the accuracy of the PacBio reads in the same way. They decided to update the Celera® Assembler program to address this new type of data — and, in the process, realized that the long-read data would indeed be their opportunity to build higher quality, cleaner assemblies.

The team's major breakthrough is an error correction pipeline that takes advantage of the long-read data offered by the PacBio *RS*, mixes in high-accuracy short reads, and runs all of it through the updated Celera Assembler to generate a high-quality assembly. "The pipeline that we converged on draws the best from all possible ideas and does an excellent job with the data," Schatz says. "Despite the apparently high error rate initially, we can almost totally compensate for it." The Nature Biotech paper reports that through this pipeline, read accuracy improves to better than 99.9 percent and median contig sizes double compared to short-read assemblies.

The Long Read Bet

In some ways, Phillippy, Koren, and Schatz's firm belief that long reads would be critical for high-quality genome assemblies ran counter to the trend in the scientific community. Most scientists with short-read sequencers were simply generating higher and higher coverage with their platforms in an attempt to improve the assemblies for their organisms of interest.

So why didn't Phillippy, Koren, and Schatz take the same approach? Their in-depth background in genome assembly told them it simply wouldn't work. "We just knew that there wasn't enough information in the short reads," Schatz says. "If we could extract the information from the long reads, we were certain that we would be able to put together a good assembly."

The scientists understood that long reads would be critical to proper assemblies, and that short-read sequencers would never be able to improve results to multi-kilobase reads. "I'd be interested in having a sequencing-by-synthesis technology where the reactions can go out 10,000 bases long, but there's just no way that the chemistry is going to sustain that many cycles," Schatz says. "To get long reads, it has to be single molecule sequencing."

The trade-off, of course, is that single molecule sequencing inherently has a higher raw error rate, Schatz adds. "Because we're imaging a single molecule at a time, we're going to see any sort of error whatsoever." By contrast, individual errors are masked in short-read sequencing systems by taking a consensus of the sequence generated across the cluster of many different molecules; single-molecule error rates for those systems are never reported.

One major factor worked in their favor. Unlike some of the short-read sequencing platforms that generated data with systematic errors, the errors in PacBio data were randomly distributed. For people skilled in informatics, that's a world of difference — random errors can be identified and corrected by algorithms, while systematic errors cannot.

Schatz notes that single molecule sequencing has advantages beyond genome assembly. In their Nature Biotech paper, Phillippy and Koren present some preliminary analysis on the corn transcriptome generated by their collaborator Zhong Wang at the Joint Genome Institute. For that work, Schatz says, "instead of trying to infer what the alternative splicing is, we can just directly read it off. There's no way to do that without single molecule sequencing." Having the error correction pipeline, therefore, makes possible several applications that would not be feasible any other way.

Building the Pipeline

The team behind this project formed long ago: Phillippy, Koren, and Schatz were all students of Steven Salzberg and Mihai Pop at the University of Maryland as well as veterans of TIGR and later the J. Craig Venter Institute (JCVI). Those connections also included co-authors Erich Jarvis, now a Duke scientist using the parrot as a model for language development, and Brian Walenz, currently at the JCVI.

In developing the pipeline for PacBio data, the scientists evaluated several methods for correcting the long reads. One variable was timing: at what point does the error correction take place? "A really popular strategy is to do an Illumina-only assembly and then align the PacBio reads to the assembly — that's what we call the hybrid scaffolding approach," Schatz says. "The process of aligning the long PacBio reads to Illumina contigs effectively error-corrects the long reads."

But that method didn't work as well as Phillippy and Schatz wanted. "What we found was that if there are any sort of problems in the short read assembly at all — that the repeats are being collapsed, there's any sort of chimeric contigs, or if your assembly was scattered into a lot of pieces — it was really hard to effectively use those long reads," Schatz says. "That prompted us to instead focus on doing error correction up front."

Indeed, the final pipeline calls for mapping the short reads onto the PacBio long reads first, and then assembling the corrected reads. Mapping the short reads onto the long reads in an efficient way proved to be a challenge, and "we ended up using somewhat of a brute force approach using very short, exact matches," Schatz says. "We were able to retrofit the Celera Assembler to do that."

Another complex problem was aligning short reads when the long read consisted primarily of repetitive sequence. "Especially if it's a repeat with more than 99 percent identity, it gets to be really tricky to correctly identify which short reads should be aligned to those long reads." The team designed some techniques to deal with this by evaluating the top alignment candidates for every short read, and then carefully assessing the alignment coverage to determine the best match. "We really had to spend a lot of time to try to get that algorithm just right to be able to separate out the repeats," Schatz says.

All of the code developed for this project is open source and available with documentation through the Celera Assembler on SourceForge at http://wgs-assembler.sourceforge.net.

Calling All Short Reads

One of the other variables the team evaluated was which short reads worked best at correcting the PacBio data — but they ended up without a strong preference, Schatz says. "It works very well with PacBio CCS reads and with Illumina or 454 reads." Whatever the platform, he recommends that users of the pipeline have 25x to 50x short read coverage, and then add in "even moderate coverage" of PacBio long reads.

The error correction pipeline is a boon not just for scientists planning future genome sequences, but also for every researcher who has generated data through the years on Illumina® or 454® systems but hasn't managed to get a high-quality assembly. All of that data can be dusted off and could pay dividends when combined with PacBio long reads. "There is enormous pent-up demand to get time and access on these PacBio instruments," Schatz says.

For scientists who have short-read data and are sequencing a single organism, Schatz says, "the error correction pipeline runs really well out of the box." The team has proven the effectiveness of the pipeline with a range of genomes, from microbes to higher eukaryotes, and "it works extremely well," Schatz adds.

"It's really the case where you can take these reads with 15 percent error, run one command, and then suddenly these reads are basically perfect," he says. "It's pretty spectacular to see those before-and-afters."

For projects that are more complex, such as tracking alternative splicing or studying metagenomics, Schatz suggests that researchers get in touch with one of the paper's authors for advice on tuning the pipeline effectively. The tool is certainly useful for transcriptomics or metagenomics, he says, but the pipeline available on SourceForge was "really designed and tuned for individual genomes."

For more information, read the team's *Nature Biotechnology* paper, which includes the *de novo* assembly of the 1.2 gigabase parrot genome. Schatz notes that the data analyzed in the paper is about a year old, and advances in the PacBio technology since then have led to assembly improvements "that are even more dramatic." He says that Sergey Koren in particular "has had some really exciting new results where microbial chromosomes get assembled into individual contigs. That's the absolute best you can ever hope to get."