FOR MARYLAND INVESTIGATORS, LONG READS OFFER NEW PATH TO FINISHED GENOMES

At the University of Maryland's Genomics Resource Center, SMRT® Sequencing is an ideal fit for closing microbial genomes and incorporating methylation data to advance infectious disease studies.

The Genomics Resource Center (GRC) at the Institute for Genome Sciences (IGS) has a scientific pedigree and a sample-to-interpretation service commitment that place it in a league of its own. The team operates under a simple mantra: 'If it can be sequenced, we can do it.'

Both GRC and IGS were founded in 2007 when a high-powered team of investigators formerly at The Institute for Genomic Research (TIGR), led by Claire Fraser, joined the University of Maryland School of Medicine. "The group of faculty and senior staff that came here to start the institute was heavily focused on infectious disease research," says Luke Tallon, scientific director and founding leader of the GRC. "Our primary goal in joining the medical school was to extend our pathogen genomics expertise into host-pathogen studies and direct clinical genomics applications."

In addition to its infectious disease and genomics expertise, TIGR was also renowned for its bioinformatics talent — a trait that continues with the group at IGS. The GRC team of 15 staff members is evenly split between wet lab and bioinformatics, and more than half of the institute's 100-plus employees are bioinformaticians. "One of our strengths

is that we go beyond generating efficient, high-quality sequence data. We have teams of analysts and engineers who can assist investigators with downstream analysis and interpretation," Tallon says. Prior to project initiation, the GRC team consults with each investigator, recommending a custom solution to meet the particular goals of the project.

The GRC was formed both to serve the genomics institute and as a university core facility. "We serve investigators throughout the University of Maryland system as well as across the country and around the world," says Lisa Sadzewicz, administrative director of the facility. The GRC works with hundreds of investigators and has become more visible by presenting and exhibiting at conferences such as the general meeting of the American Society for Microbiology and Advances in Genome Biology and Technology, and by launching a blog last year.

"Our strength is not just our deep history and experience in sequencing and genomics, but our end-to-end service level from the initial project consultation through to publication, including all of the informatics," Tallon says. "That's what sets us apart from other cores."

Choosing PacBio® Technology

The GRC has had a mandate to stay on the cutting edge of sequencing technology since its inception. "We are continuously monitoring and evaluating new technologies," Tallon says. A few years ago, these evaluations led IGS to Pacific Biosciences and its single molecule, real-time (SMRT) sequencer. "We were struck by the potential of the platform," he adds.

As a chargeback facility, GRC must make careful choices about the technologies in which it invests. From Tallon's

UNIVERSITY of MARYLAND SCHOOL OF MEDICINE
INSTITUTE FOR GENOME SCIENCES

Facility name:	Genomics Resource Center
Institution:	Institute for Genome Sciences at the University of Maryland School of Medicine
Staff:	15 scientists, evenly split between wet lab and bioinformatics
Year founded:	2007
Serves:	Hundreds of investigators from the University of Maryland and worldwide
PacBio System installed:	June 2011, upgraded in May 2013 to PacBio RS II
Website:	http://www.igs.umaryland.edu/grc

perspective, the PacBio sequencing platform made sense because it served application niches different than short-read sequencers. The instrument's strength in *de novo* microbial sequencing and other long-read applications made it particularly well-suited to the type of research at IGS. Though there are many factors they consider when choosing a new instrument, an important one is "the relative value of the type of data you're going to get," Tallon says. "Small genomes are such a significant part of what we do, and we're moving more and more into microbial transcriptomes and methylation studies. It's the only platform that allows us to do all of that really well."

Sadzewicz notes that anticipated demand is also an important factor in bringing in a new instrument for a core facility. "We are not driven by only one customer or one small group," she says. "In order to drive down costs, you need to have a wide community from which you can draw projects and samples to fill the capacity of the instrument."

Finishing Genomes

Because of the institute's strong focus on infectious disease, the GRC conducts many sequencing projects for pathogen genomes, human microbiome samples, and other microbial genome applications. As the former lead of the genome finishing team at TIGR, Tallon knows the value of a complete genome sequence.

The PacBio sequencer, which GRC upgraded to the PacBio RS II last spring, is now a workhorse for generating finished or nearly complete microbial genomes as well as genome-wide methylation data. "We're now analyzing base modifications and methylation patterns routinely with most of our small genomes," Tallon says. "We're also doing metagenomic sequencing using the RS II and exploring ways we can use the long reads to get full-length genes and transcripts out of our metagenome and metatranscriptome samples."

The GRC staff has worked hard to optimize its PacBio workflow to get the best results and longest possible reads. The foundation for that is a team of highly trained, talented scientists. Starting with high-quality DNA samples is also critical. The GRC has incorporated automated size selection using the BluePippin from Sage Science to remove smaller fragments from libraries and push average read lengths and SMRT Cell yield even higher. On the *de novo* assembly side, they run data through HGAP and Quiver from PacBio and through Celera® Assembler, comparing results. "For some data sets, HGAP yields the best assembly, and for others, the CA pipeline is the way to go," Tallon notes.

That attention to process has resulted in a high-performance pipeline for finishing genomes. "Prior to PacBio, we couldn't close genomes without manual finishing efforts," Tallon says. With SMRT Sequencing, his team is consistently finishing genomes. "Our biggest challenge is getting sufficient high-quality DNA to start a project. If we get that, the genomes are going to close more often than not."

These achievements have made a real difference to GRC customers. One investigator recently brought in some microbial genomes that had never been sequenced before; given their characteristics, he had little hope that they could be sequenced well. "We were able to quickly generate closed genomes for him using only PacBio data," Tallon says. "He was impressed and told us it would open new doors in his research program."

Future Applications

As read lengths and throughput increase, the PacBio RS II is broadening its application base. At GRC, Tallon and his team are forging ahead with larger genomes, transcriptomes, and even some custom-built analysis tools. "We're starting to do some projects looking at larger *de novo* transcriptomes, using some Illumina® data with PacBio data so that we can look for full-length transcripts in the long reads," he says. They are also exploring the latest long-read chemistry release for human genome haplotype phasing and structural variation detection. Building on the HGAP and Celera pipelines, he adds, "we're developing our own post-processing steps to make even further improvements to genome assembly."

For other teams considering whether SMRT Sequencing is the right choice for them, Tallon says: "If you're going to be working with small genomes, value complete or nearly finished genome sequences, and are looking at base modifications in addition to the genome sequence, there's no better platform out there."

Advice from GRC: Tips for SMRT Sequencing

DNA quality: Start with the highest-quality DNA sample you can get, and run extra MagBead cleanups to ensure optimal results from the sequencing process.

Size selection: The BluePippin platform from Sage Science™ has boosted yield per SMRT Cell and average read length, dramatically increasing the total number of bases generated per run.

Bioinformatics: Your data is only worth what you can get out of the analysis. Invest in experienced bioinformaticians and constantly evaluate and develop new tools.

PN: PF100-110113