Precise Sizing and SMRT Sequencing Offer Unprecedented Read Length for Clinical Studies

At the Icahn Institute for Genomics and Multiscale Biology, scientists use automated DNA sizing together with long-read sequencing to analyze clinical samples, conduct routine surveillance on microbes, and more.

Robert Sebra, Ph.D.

At the Icahn Institute for Genomics and Multiscale Biology at Mount Sinai in New York City, technology development expert Robert Sebra, Ph.D., sees tremendous need for long-read, high-accuracy clinical sequencing for use in microbial surveillance, detection of repeat expansions, and more. To meet that demand, he relies on Single Molecule, Real-Time (SMRT®) Sequencing from Pacific Biosciences with BluePippin™ automated DNA size selection from Sage Science. Together, these tools offer a powerful solution and industry-leading read lengths that allow Sebra and other researchers to resolve repeat elements and structural variants, rapidly close microbial genomes, and measure epigenetic marks.

Sebra, an assistant professor of genetic and genomic sciences, is no stranger to the SMRT Sequencing platform: he spent five years working at PacBio helping to develop that technology. Ultimately, his belief in the system led him to join the Icahn Institute, where he would get to use the PacBio® sequencer in the field. "There was a lot to be gained by taking the technology and applying it in a clinical setting," says Sebra, who came to Mount Sinai in 2012. "I had experienced firsthand the value of long-read sequencing and wanted to apply it to human and infectious disease research."

"BluePippin is fast and cheap, and it's the only option for size selecting in a high-throughput fashion. We purchased one as soon as it was available."

Since its founding by Eric Schadt in 2011, the Icahn Institute has attracted some 150 leading scientists and clinicians who bring a network-based approach to various biological questions, many of them focused on cancer, Alzheimer's disease, allergy and asthma, and infectious disease. Among the institute's well-stocked core facilities are two PacBio RS II sequencers and a BluePippin instrument, which are used together for projects requiring extra-long reads.

Sebra's idea that this kind of approach would be useful in a hospital environment was prescient. "I can't emphasize enough the tremendous potential that I see for long-read sequencing in tackling hard-to-sequence samples in the clinical arena. The technology has led to novel results creating a rapid growth of interest as data become more accessible," he says. Indeed, the institute has churned through some 1,800 SMRT Cells in the past year and shows no signs of slowing down. Sebra and his colleagues have already demonstrated the extraordinary value of long-read DNA sequencing for microbial and human clinical samples, and they have a slew of other projects in the pipeline.

Technology Focus

The move to a hospital and genomics institute may have offered Sebra many new opportunities to apply long-read sequencing, but it didn't change his passion for technology development. He works with researchers and clinicians throughout the institute, helping them determine which technology solution best fits the biological question they are trying to answer. For Sebra, that means he has to be well-versed in the entire range of applications for next-generation sequencing platforms. "For PacBio, that application

space spans anything from epigenetic profiling to clinical validation and both targeted and whole-genome sequencing of human structural variants to achieve disease associations not achievable without long reads," he says.

In disease research studies, for example, a clinician might be interested in looking at a copy number variant or rearrangement that may be implicated in the disease — Sebra makes that possible by designing the method to capture the genetic element in relevant samples for direct sequencing. He has participated in projects involving hospital surveillance, fast microbial genome finishing, metagenomics, epigenetics, and much more. On the technology front, one of Sebra's particular areas of interest is finding ways to reduce the amount of DNA required across various sequencing platforms, so that sequencing libraries can be generated more easily from low-input samples such as single cells.

As research and clinical projects come his way, Sebra must first ascertain which sequencing platform is the best fit. The PacBio RS II is his go-to system for epigenetic profiling, finishing microbial genomes, and exploring DNA samples likely to have repeats, large structural rearrangements, or ones that require allelic or accessory genome phasing.

Microbial genomes in particular are a sweet spot for the SMRT technology, Sebra says. "Those are easy projects because we can sequence the epigenome and finish the entire genomic assembly in a few days while maintaining a low cost." That genome-plus-epigenome capability explains much of the demand for PacBio sequencing, because no other platform offers the ability to look at genome-wide methylation and other base modifications. Factor in the cost, Sebra says, and it's the obvious choice. "Researchers and clinicians are very aware of cost and turnaround time

Sebra has been pleased with the results of pairing these platforms, noting that the size selection step has exceeded his expectations for overall improvement in read length and throughput of SMRT Sequencing.

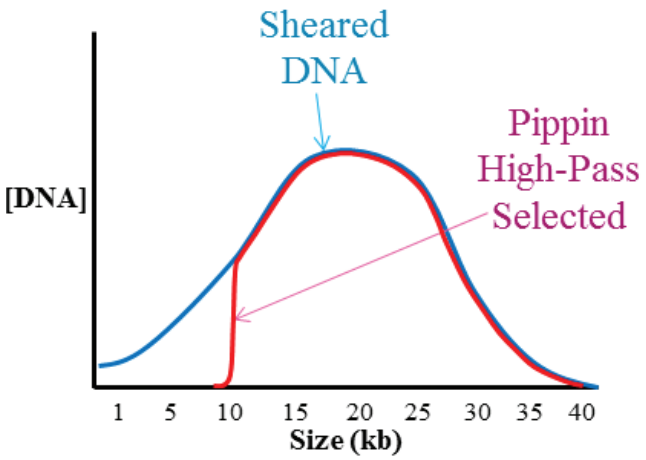

Figure courtesy of Robert Sebra

for diagnostics, and SMRT Sequencing is an obvious win-win for achieving these attributes in the infectious disease arena while also offering potential for novel discovery."

As he applies long-read sequencing to these projects where it will make the biggest impact, Sebra continually looks for ways to generate the longest possible reads. One complementary technology for the PacBio workflow is BluePippin, an automated DNA size selection platform from Sage Science. Removing smaller fragments from the sequencing library ensures that the PacBio platform focuses on the longest fragments, so accurate sizing can improve average read length considerably. "You could do a traditional pulsed field gel every time you're trying to size select, but it takes too much time, doesn't scale well, and the DNA input requirement is really high," Sebra says. "BluePippin is fast and cheap, and it's the only option for size selecting in a high-throughput fashion. We purchased one as soon as it was available."

Since bringing in BluePippin in 2012, Sebra's team has run more than 100 libraries using the BluePippin+PacBio combo — in fact, he says, "For projects requiring near finished genome assembly, I don't think we've prepared a library without BluePippin size select since owning the instrument." He has been pleased with the amount of size-selected library the technology yields, noting that in virtually every experiment it produces more than enough to sequence a genome to completion on the PacBio RS II. He generally excludes all fragments smaller than 10 Kb to target the ultra long fragments, but says that in cases where input DNA is especially low or the genome is quite large and requires more library, he lowers that threshold to 7 Kb.

Pipeline at Work

Sebra has been pleased with the results of pairing these platforms, noting that the size selection step has exceeded his expectations for overall improvement in read length and throughput of SMRT Sequencing. The boost to mean read length from adding BluePippin size selection ranges from about 30 percent to 125 percent, depending on the input quality, he says. Two studies — one microbial, the other human — offer a snapshot of how the pipeline is performing for ongoing efforts at the institute.

In one project, Sebra and his colleagues are working on an ambitious, big-picture study for infectious disease surveillance that could be used internally at hospitals as well as to test external samples. Methicillin-resistant Staphylococcus aureus, or MRSA, is especially important to surveillance programs "because of the potential in characterizing community-acquired isolates," Sebra says. The idea for this type of program is to sequence microbial samples and then conduct a phylogenetic analysis to figure out the source and history of an infection.

In one infectious disease study, the team sequenced multiple MRSA isolates using PacBio with and without BluePippin sizing, finding that prior to sizing, 50 percent of the bases are in reads 5 Kb or longer, while after sizing that number more than doubled to 12.5 Kb. Full sequencing, from sample prep through to genome assembly, took about 48 hours, and cost as little as $300 per isolate, often assembling to a single contig, Sebra notes. "The big take-home message was that we can do low-contig assemblies with just a couple of SMRT Cells," he adds. "We could rapidly assemble isolate genomes, including plasmids, to rapidly source that isolate and improve patient treatment." That's one of the reasons that the PacBio technology is critical for this kind of surveillance program: those long reads allow for phasing clinically relevant plasmids in a separate circular contig. With the success of the MRSA study, Sebra says, it is now easy to "imagine scaling that approach across all infectious disease isolates."

"Size selection reduces the number of SMRT Cells required to achieve a particular sequencing goal, so BluePippin pays for itself pretty quickly."

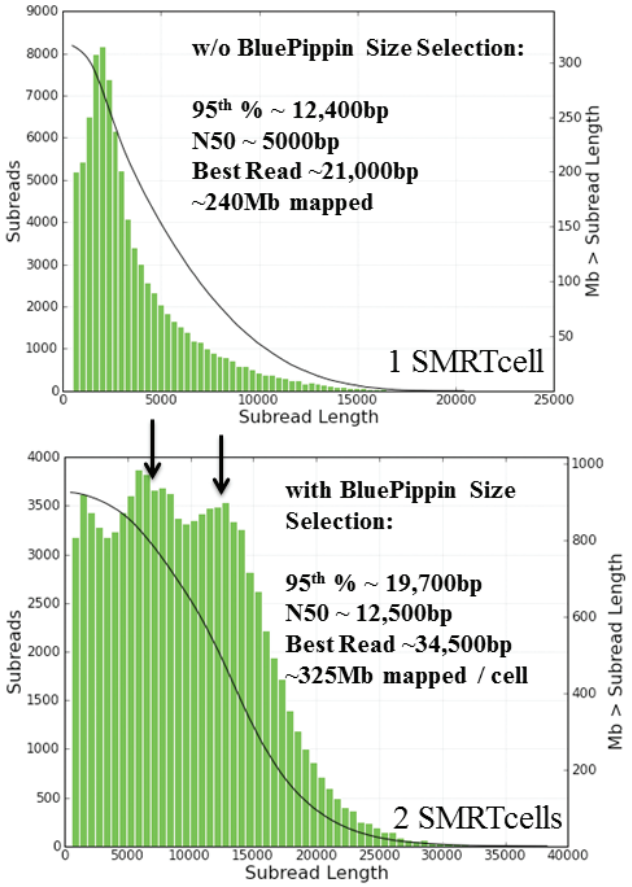

Figure courtesy of Robert Sebra

In a separate ongoing project, Sebra and his collaborators have sequenced a standard human genome sample — known to the scientific community as NA12878 — to above 30x coverage using PacBio with BluePippin. "With informatics strength from the Bashir group, our goal is to better resolve the structural elements larger than 10,000 base pairs that were unachievable with any other technology up to this point," he says. "We want to discover which regions of the genome are missing in the current reference so we can better associate those with disease."

There are many genetic landscapes, from trinucleotide repeats to copy number variants or inserted elements, that are linked to disease severity, Sebra says — but they are impossible to detect in assemblies where the reads are too short to assemble them. By applying long-read sequencing, he and his partners hope to rescue these missing regions. Ultimately, that could make things like genome-wide association studies more fruitful. In the clinic, Sebra envisions working with clinicians to develop targeted panels of genes with known repeats or other structural variants "to better diagnose disease severity" of a patient.

The effect of BluePippin sizing was also significant in the human study, increasing the mean subread length from about 2,800 bp to almost 8,000 bp. Size selection also helps to focus sequencing on pieces of the genome that otherwise may not achieve high coverage due to mapping complexity. "Without size selection, you'll greatly reduce the coverage of redundant regions of the genome," Sebra says. Armed with both platforms, Sebra and collaborators are pushing ahead with their human genome work, hoping to reach even higher coverage with SMRT Sequencing to generate a more complete human reference.

Advice for Others

Many people attribute the success of Sebra's PacBio pipeline to his years working at the sequencing company and assume that these kinds of results are out of reach for new users. That couldn't be further from the truth, says Sebra, noting that the work done on these instruments is reproducible across users with varying levels of expertise. "Other people can absolutely roll out this pipeline," he says. "It's quite scalable and easy to teach these techniques. In particular, user-friendly assembly pipelines such as HGAP2 enable researchers of varying degrees of expertise to conduct complete experiments from isolation to assembly."

He notes that the single most important ingredient for this sequencing workflow is DNA quality. "It really comes down to the DNA prep, and isolating the DNA with care, to avoid physical and chemical damage before going into the BluePippin size-selection cassette and then onto the PacBio system for sequencing," he says. That helps to optimize both technologies to ensure the longest reads possible for the highest-quality assemblies.

As for whether the BluePippin addition is right for other scientists, there's a simple way to determine that, according to Sebra. "If your throughput of runs is high enough, a BluePippin is really pretty affordable. Size selection reduces the number of SMRT Cells required to achieve a particular sequencing goal, so it pays for itself pretty quickly."

Suite 3150
500 Cummings Center
Beverly, MA 01915
978.922.1832
www.sagescience.com

sage science