FOR ANTARCTIC GENOME, SCIENTISTS FIND UNIQUE CAPABILITY IN SMRT® DNA SEQUENCING

Scientists at the Korea Polar Research Institute used the PacBio® *RS* for the successful *de novo* assembly of a GC-rich bacterial genome that couldn't be pulled together with short-read technology

Hyun Park, Ph.D., has just completed a genome assembly that he says could not be generated with any sequencing technology other than the PacBio *RS*.

Dr. Park, a senior scientist and project leader at the Korea Polar Research Institute who studies organisms from Antarctica, wanted to sequence *Streptomyces*, bacteria whose genome is made up primarily of GC content. But even after applying Illumina® sequencing and Sanger reads, he was still left with far too many gaps to assemble the 7.6 Mb genome.

"The gaps generated from both Illumina reads and Sanger reads in these high-GC content regions could not be filled with the same technology," Dr. Park says, calling it "impossible" to close the gaps with those sequencing technologies.

Polar Exploration

Dr. Park, who has been a senior research scientist with the Korea Polar Research Institute (KOPRI) since 2005, and is also an associate professor at the University of Science and Technology in Korea, has published numerous papers describing genomic studies of organisms found in Antarctica.

Dr. Hyun Park, Senior Scientist, Korea Polar Research Institute

KOPRI operates scientific research stations in both the Antarctic and the Arctic. As just one part of its many polar research programs, the institute aims to play a key role in enhancing the understanding of global climate change.

> Dr. Park and his colleagues write that PacBio's SMRT technology "shows better performance with high GC content organisms and is expected to be the new tool to improve de novo sequencing and assembly."

In Dr. Park's view, the benefits of polar research could span a variety of fields, including medicine. The genetic modifications of *Streptomyces* in Antarctica, for example, offer a host of lessons — one of the most notable in antibiotics research, he says. "The lower optimal temperature of enzymes of the Antarctic *Streptomyces* could also be widely used in biological engineering," he adds. "The lower temperature of enzyme reactions would multiply the specificity of enzyme reactions. By doing so, the side effects of reactions could occur at a much slower rate or not at all."

Recently, Dr. Park has been focusing on *Cladonia borealis*, the dominant species of lichens found in Antarctica. Like all lichens, the moss-like *Cladonia* is a composite organism that can consist of algae, bacteria, and fungi living in a symbiotic relationship. In this organism, algae are responsible for gathering nutrients through photosynthesis, which also feeds the fungus. The bacteria help the organism maintain the mechanisms through which the symbiosis functions. They also confer some of Cladonia's heightened resistance to the harsh polar elements: frigid temperatures, UV light, and a very dry climate.

Streptomyces, the target of Dr. Park's genome project, is one of the bacterial strains found in *Cladonia borealis*. The ultimate goal is to perform full-scale functional genomics to better understand the bacterium and its role in *Cladonia*, so

isolating the bacteria from a sample on King George Island and determining its genome sequence was "a top priority" for Dr. Park.

The Sequencing Challenge

But being a high research priority didn't mean that *Streptomyces* was willing to give up its secrets so easily. The bacterium is known for its very high GC content — 71 percent — so Dr. Park and his colleagues were well aware that they would have to generate significant coverage to produce an assembly.

And it still wasn't enough. The team generated 200x coverage of the *Streptomyces* genome using the Illumina platform and wound up with an assembly-resistant 185 contigs. Even the far more expensive Sanger sequencing resulted in too many gaps for a proper assembly. According to Dr. Park, it was "impossible" to fill the gaps with any of the sequencers they tried.

So, with the help of the Korean collaborator and service provider DNA Link, Dr. Park and his team turned to a completely different type of sequencer. Known for its very long reads and ability to sequence through GC-rich regions, the single-molecule, real-time (SMRT) technology from Pacific Biosciences offered an intriguing alternative to the difficult genome.

They approached *Streptomyces* with two flavors of PacBio sequencing: the high-accuracy circular consensus short reads, and the ultra-long continuous reads averaging 1.5 Kb. Dr. Park and his team generated 15x coverage using just four SMRT Cells, fed the information through the Celera® Assembler, and emerged with only 26 contigs. For the first time, they were able to produce a useful assembly of the organism's genome. "PacBio *RS* reads fill the gaps, where Illumina reads can't fill them with 200x genome coverage." Dr. Park says. "The PacBio long reads are very useful in improving the results of the assembly."

Dr. Park and his team generated 15x coverage using just four SMRT Cells, fed the information through the Celera Assembler, and emerged with only 26 contigs. For the first time, they were able to produce a useful assembly of the organism's genome.

In a poster describing their work, Dr. Park and his colleagues write that PacBio's SMRT technology "shows better performance with high GC content organisms and is expected to be the new tool to improve *de novo* sequencing and assembly."

Bigger Genomes Ahead

Dr. Park recommends that other scientists working on GC-rich genomes consider the PacBio *RS*. "The long reads make it possible to fill the gaps from high GC content," he says, adding that these reads are also very helpful in closing gaps generated by short-read sequencing technologies.

Even as he continues to analyze the *Streptomyces* data, Dr. Park is already looking forward to other organisms to which he can apply the PacBio technology. He's got his sights set on a much larger one this time: the 600 Mb codfish found in Antarctica. "The results of assembly represent the tendency of increasing error rate according to genome size," Dr. Park says. "Therefore, PacBio's long reads are much better for the assembly of larger genomes than high-throughput, short-read sequencing."

***Cladonia borealis* is the dominant species of lichens found in Antarctica**

KOPRI Korea Polar Research Institute

DNA Link

www.pacb.com/denovo