IN BROAD INSTITUTE STUDY, PACBIO® *RS* DEMONSTRATES "OUTSTANDING" ACCURACY FOR SNP VALIDATION

Scientists at the Broad Institute evaluate the PacBio *RS* for SNP validation and discovery, adding the instrument to their standard validation pipeline to make use of its high sensitivity and specificity

When Mauricio Carneiro, Ph.D., joined the Broad Institute in December 2010, he was given responsibility for evaluating the new Pacific Biosciences® sequencing platform — and he expected to spend his time cataloging a list of the particular sequence errors generated by the instrument. But instead of finding a slew of systematic errors, he says, "the big surprise was the fact that the errors were random. This is really a unique property that no other sequencer out there has."

Carneiro knew this meant one thing: "This is huge," he says. "Random error mode gives you a tool to determine what is right and what is wrong."

Realizing that the PacBio *RS* sequencer's random error mode provided a means for making highly accurate calls given enough depth and the right mathematics, Carneiro decided to put the machine through its paces with two SNP calling studies: one for validation and another for discovery.

Dr. Mauricio Carneiro, Computational Biologist, Broad Institute

In a paper entitled "Pacific Biosciences sequencing technology for genotyping and variation discovery in human data," published in *BMC Genomics* in August 2012, lead author Carneiro and his colleagues at the Broad detail their findings from those validation and discovery studies, comparing the PacBio results to data from the Illumina® MiSeq® platform. They conclude, "These data show excellent utility for follow-up validation and extension studies in human data and medical genetics projects, but can be extended to other organisms with a reference genome."

> "These data show excellent utility for follow-up validation and extension studies in human data and medical genetics projects."

Since the time the paper was written, Carneiro says, the PacBio *RS* has become a standard validation tool in the Broad Institute's pipeline.

First Meeting

In some ways, Mauricio Carneiro's path was bound to meet the PacBio *RS*. He joined the Broad's Program in Medical and Population Genetics just a few months after the PacBio sequencer was installed, and his background in computer engineering and computational biology made him a perfect candidate to assess the data generated by a novel sequencing platform. "We got this brand new machine, and we wanted to know if the data was useful and if we could actually use it for human studies," he says.

Carneiro was no stranger to next-gen sequencing technologies, and he knew that evaluating a new technology meant dealing with data that had not been fully optimized. "Every sequencer that comes to market starts out with a lot of problems," he says, remembering how hard it was to work with Illumina data when that platform first emerged. "But we developed this whole suite of tools to deal with that and now it's our standard workhorse."

He anticipated a similar situation for the PacBio platform, and was pleasantly surprised to find that "it really didn't seem like we had miles and miles to go" to produce useful data. "We didn't even change any code. With a few adjustments here and there, we were able to make really good SNP calls right off the bat," he says.

"We didn't even change any code. With a few adjustments here and there, we were able to make really good SNP calls right off the bat."

What really struck Carneiro from the first few PacBio data sets was the random, rather than systematic, error mode of the instrument. From a mathematical perspective, not all errors are equal. The DNA sequencers that Carneiro had previously used had systematic errors, or mistakes that were made again and again in particular sequence contexts, such as GC-rich regions and homopolymer runs. "Error modes that are systematic are really hard to work with," Carneiro says. "You can't mathematically decouple the error from the truth."

Random error mode, on the other hand, "is the perfect error model you want to have in any instrument," he says. "If the error is random, the more you sequence that site, the less likely you will be to see the error again. With adequate coverage you will always dilute the error into the truth."

That's why the raw sub-read error rate doesn't faze Carneiro. "People hear there's a high error rate and think that they can't deal with this data," he says. "But the error rate really doesn't make that much difference as long as you have enough depth and the error is random. The error rate doesn't bother me at all."

Because of its yield and its novel approach to sequencing DNA, the PacBio *RS* struck Carneiro right away as a good orthogonal validation tool. "It's a very different technology; it could very well pick up artifacts from other technologies," he remembers thinking. "It seemed attractive and we wanted to know how well it performed."

Validation & Discovery

Carneiro designed two targeted-sequencing studies to determine how well the new instrument performed for SNP calling, a major need for the medical and population genetics group. To test the machine's function as a validation tool, he fed it regions of DNA known to include hard-to-call SNPs. This list of SNPs, which had been used in the past to test other sequencers, is a continuously updated set of variants that have been notoriously difficult to call correctly across any number of platforms. Nobody knows quite why these SNPs prove so challenging, Carneiro says, "but they are consistently wrongly called by sequencers."

In the validation study, the PacBio *RS* performance was "outstanding," Carneiro says. The paper reports the library of 98 hard-to-call variants fed to both the PacBio *RS* and the Illumina MiSeq as a comparison. Of these, the PacBio platform correctly called 96 sites, resulting in data with 97 percent sensitivity and 98 percent specificity, the team noted in the *BMC Genomics* paper. The MiSeq, by contrast, accurately genotyped 93 of the 98 sites, providing 100 percent sensitivity and 91 percent specificity. "Pacific Biosciences *RS* data performed well by all metrics, and at a similar quality to Illumina data, demonstrating that the *RS* is a powerful tool for follow up validation or extension," the authors wrote.

The second study focused on the instrument's performance for SNP discovery, again running the same DNA on both PacBio and the MiSeq. Carneiro and his team used 61 amplicons covering 177 kb of DNA sequence that covered high-variance regions that are not considered particularly difficult to call. For this test, the sequencers performed comparably, Carneiro says.

"The only problem we found with the PacBio data is a reference bias," he says. "We show in the paper that if it wasn't for the reference bias, the numbers would be exactly the same or better than MiSeq."

That's because while the PacBio instrument did indeed find most of the SNPs, the alignment software hid some of those SNPs in insertions; the ultimate list of SNPs called, therefore, does not fully represent the sequencer's accuracy. Of 225 SNPs in the amplicons, the PacBio *RS* correctly called 197, while MiSeq correctly called 222 of the variants. Closer inspection of the SNPs that the PacBio *RS* did not call indicated that the majority of them were in the sequence data but lost in the alignment process due to this software bias.

The success of the PacBio data for SNP calling in targeted resequencing means that the instrument has rapidly "became a standard here at the Broad for validation," Carneiro says. "This was really a success. It's a very attractive tool for validation studies."

"That's why in the paper the tone on reference bias is fairly mild," Carneiro adds. "We think we have a good idea of how to solve it."

The reference bias, Carneiro explains, "is an artifact that's created by the way we align reads that come out of the instrument to the reference genome." Because the main error found in PacBio *RS* data is insertions, aligner software has a tendency to hide accurately called, true SNPs as insertions during the alignment process. Going through to inspect these reads manually, Carneiro says, "on some reads you'll see the SNP clearly right there, but on others, the aligner has hidden the SNP as an insertion."

The solution is to adapt the aligner software to make it better

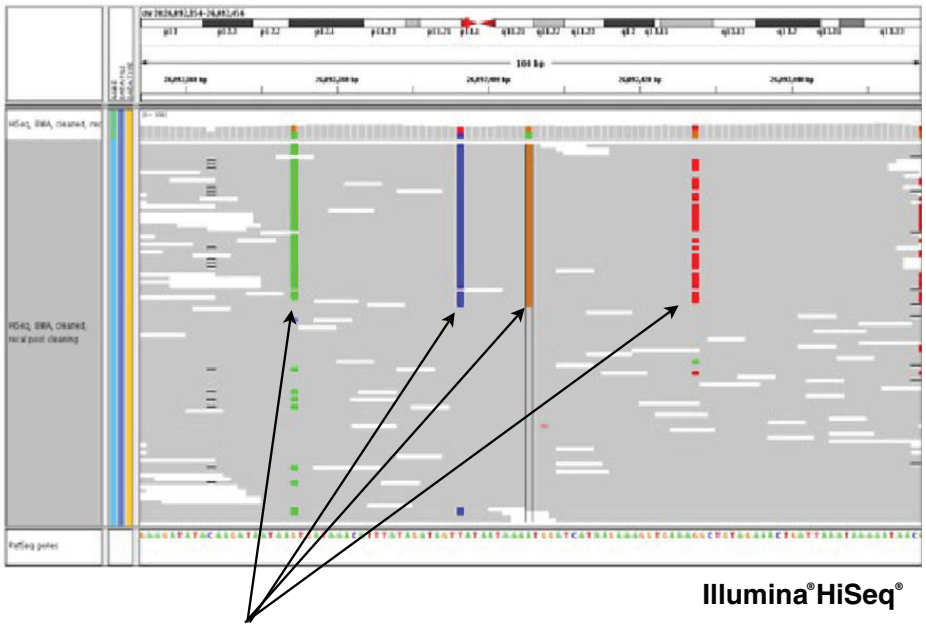
at analyzing PacBio data — and therefore less likely to hide real SNPs recorded in the sequence generated by the instrument. Toward that end, the Broad team recently released the beta version of HaplotypeCaller, a new tool in the Genome Analysis Toolkit (GATK) software package that may serve as a first step in optimizing aligners to more accurately handle PacBio data. "It reassembles all of the reads that cover a region that's suspected to have a SNP before making the actual call, so if there's anything hidden inside insertions it'll be expanded," Carneiro says.

Part of the Pipeline

The success of the PacBio data for SNP calling in targeted resequencing means that the instrument has rapidly "became a standard here at the Broad for validation," Carneiro says. "This was really a success. It's a very attractive tool for validation studies."

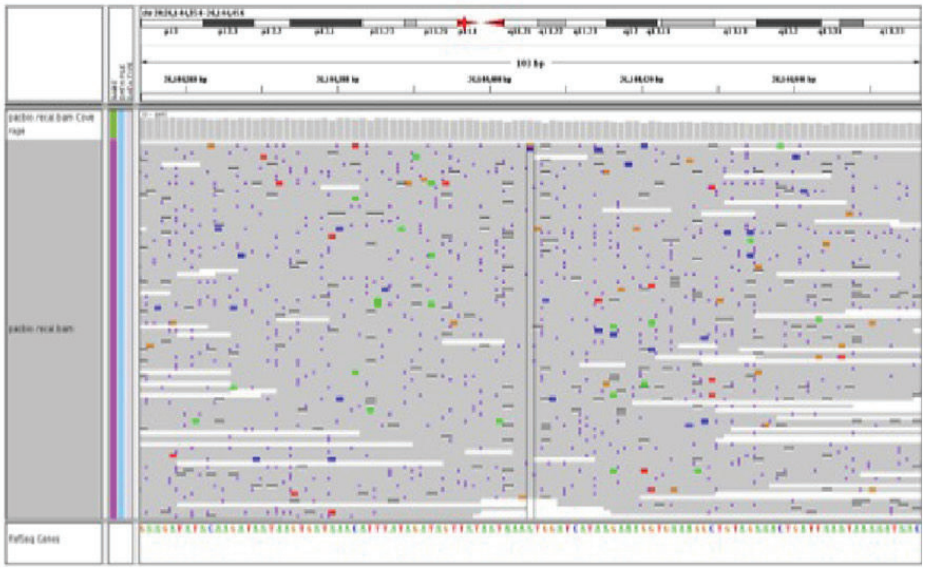
In fact, even before Carneiro's paper came out in *BMC Genomics*, other papers had been published using the Broad's PacBio *RS* validation pipeline — including a high-profile *Nature* publication entitled "Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations." In that effort, lead author Trevor Pugh and a number of collaborators at the Broad and other institutions used exome sequencing to analyze 92 primary medulloblastoma/normal pairs for mutations, following up on those with PacBio resequencing.

Carneiro and his co-authors see plenty of room to expand the use of the PacBio platform. As the instrument's yield improves and their own work advances in optimizing HaplotypeCaller's ability to recognize the SNPs called by the PacBio *RS*, the authors write, this "will further increase its utility in the field of human DNA sequencing."



Illumina® HiSeq®

SYSTEMATIC ERROR



PacBio RS

RANDOM ERROR

This figure from the *BMC Genomics* paper shows the systematic error found in Illumina® HiSeq® data compared to the random error found in PacBio® *RS* data.

www.pacb.com/target