SMRT® SEQUENCING BRINGS CLARITY TO HIV VACCINE AND TRANSPLANT RESEARCH AT THE WISCONSIN NATIONAL PRIMATE RESEARCH CENTER

THE LEADER IN LONG-READ SEQUENCING

PACBIO®

The Wisconsin National Primate Research Center (WNPRC) is a leading Major Histocompatibility Complex (MHC) typing lab that focuses on monkeys. While many scientists are familiar with the importance of characterizing the histocompatibility region of the human genome for applications like disease research or tissue typing before organ transplantation, fewer are aware of the need to accurately type this region in non-human primates. At the primate research lab, part of the University of Wisconsin-Madison, scientists are analyzing immune regions to help test potential HIV vaccines and AIDS therapies. Their work is essential for understanding the effects of treatment ahead of human trials.

Led by Dave O'Connor, a University of Wisconsin-Madison professor and associate director of research at WNPRC, the scientists perform MHC typing to identify monkeys – primarily macaques – that may be naturally able to control viral replication without medication. "A subset of macaques control replication of simian immune deficiency virus in a way that's very analogous to how a couple percent of humans are able to control HIV infection," says Roger Wiseman, a scientist in the O'Connor lab. These individuals can be infected but control viral replication spontaneously, so excluding them from vaccine or therapy trials is important for preventing confounding results. "It's really critical that you know which animals have host genetics that are consistent with them having a high probability of being able to spontaneously control viral replication," he adds.

But MHC typing for monkeys is no simple task. These genes are extremely polymorphic in mammalian genomes, and monkeys in particular have the challenge of widespread copy number variation and pseudogenes. Therefore, the O'Connor lab adopted Single Molecule, Real-Time (SMRT) Sequencing from PacBio for MHC typing. With its long reads, this technology can generate highly accurate, phased coverage of complete gene transcripts without the need for reference-guided alignments.

Sequencing Evolution

The lab began MHC typing with Sanger sequencing platforms more than a decade ago and eventually shifted to next-generation technologies like 454 and Illumina sequencing. But these approaches have not been sufficient for resolving incredibly challenging immune regions, especially in animals with poorly assembled genomes. For example, in the current rhesus macaque reference, there are as many as 19 tandemly duplicated genes collapsed into a single MHC class I gene.

"Short-read technologies, which are really good for resolving non-complex parts of the genome, don't work very well when confronted with these highly polymorphic immune gene loci," O'Connor says. "Unfortunately, those are often the most important or most interesting parts of the genome for people who are studying infectious disease or transplant rejection or other immunologic phenomena."

In its quest for more accurate results, the team examined PacBio® sequencing data generated by a collaborator and was surprised by the quality and extent of coverage. "You could get the full open reading frame," says Julie Karl, a research technician in the O'Connor lab. "For us, it's been long reads from there on out for allele discovery because we were able to get the super high accuracy that we need."

Rhesus macaques similar to this adult female with her infant at the Wisconsin National Primate Research Center are being used to study the SIV virus and help develop HIV vaccines.
Photo credit: Peter Pierre/WNPRC.

Since then, the scientists have been using SMRT Sequencing to analyze cDNA amplicons about 1.1 kb long. "We get very high-quality, highly accurate reads that allow us to call with single-base precision," Wiseman says. The result is a complete transcript, with no assembly required. With this approach, scientists can reliably distinguish between alleles that may be identical except for a single base.

"Long reads let us peer into these complex immune genes with a level of detail that we've never had before," O'Connor says. "Profiling these highly polymorphic genes and their transcripts by long-read sequencing will let us start to correlate the presence of specific variants with specific phenotypes or biological outcomes."

SMRT Sequencing offers more than just accuracy: for the first time, scientists can phase distant SNPs across the entire open reading frame of full-length MHC class I alleles. This allows the scientists to interrogate and track alleles within an individual to pinpoint their functional impact on disease susceptibility and progression.

"When you need to unambiguously span greater distances to see if SNPs are definitively coinherited, then PacBio sequencing gives a much better answer because anything with shorter reads would just be guessing," Karl says.

O'Connor notes that his own team was initially reluctant about SMRT Sequencing, believing the error rate would be too high for their needs. Seeing the data changed their minds, but "preconceptions about the PacBio platform are an unfortunate legacy of some of the early-generation data," he says. "Today there are applications where, through a combination of increasing sequencing depth or taking advantage of circular

Trent Prall and Hailey Bussan, scientists in Dave O'Connor's group, are excited to be performing MHC genotyping locally on the PacBio Sequel System installed at the University of Wisconsin-Madison.

consensus sequencing, you can get highly accurate long-read sequences that enable types of science that were impossible before."

Reporting SMRT Results

The O'Connor lab has published several papers describing the results of its SMRT Sequencing pipelines for primate research. In the journal Human Immunology, they contributed to a project led by collaborators at the US Army Medical Research Institute of Infectious Diseases and reported validation of the long-read approach for MHC typing of class I transcripts in Mauritian-origin cynomolgus macaques. After showing that SMRT Sequencing results were concordant with previous analyses, they typed 10 new Chinese-origin cynomolgus macaques. That exercise yielded 60 transcripts that were "either newly reported, match previously described sequences that lacked official allele nomenclature, or are extensions of prior named alleles that were not formerly full-length," the scientists wrote. "We believe SMRT-CCS on the PacBio RS II system provides a clear advantage over other next generation sequencing platforms for characterizing full-length MHC class I cDNA transcripts because sequencing is achieved in one continuous read."

In the paper, the team noted that a complete database of full-length MHC class I alleles would be a valuable resource for the scientific community. They also envisioned downstream benefits: "Full-length MHC class I cDNA sequences are required for the creation of reagents such as MHC-peptide tetramers, which will enable scientists to assess the range and functionality of CD8 + T cell responses," they wrote. Their publication was recognized by the journal's editorial board with the Best Paper of 2015 Award.

In another paper, scientists at Wisconsin and other institutes used PacBio long-read sequencing for a large MHC allele discovery project. The effort focused on cynomolgus macaques from Chinese breeding centers, including individuals from Vietnam, Cambodia, and other origins. Scientists generated full-length sequences for 313 class I transcripts, more than 40% of them novel and another 37% that extended shorter sequences already included in the allele database. "This significantly expands the number of … full-length alleles in the official cynomolgus macaque MHC class I database," they reported.

New Directions

With PacBio SMRT Sequencing, WNPRC scientists are generating information that wasn't accessible with previous platforms and are now already branching out to study other challenging genomic regions. "It's allowed us to move to other immune gene families, specifically the Killer cell Immunoglobulin-like Receptors or KIRs," Wiseman says. The team has already published results from a KIR project where they have started utilizing SMRT Sequencing to boost the discovery of KIR gene alleles.

In addition, the scientists are coming up with new ways to apply SMRT Sequencing to study MHC and other regions. Karl is performing target capture with NimbleGen reagents to study genomic DNA, rather than cDNA, of the MHC region in various macaque species. "The end goal of the target capture experiment is ultimately to improve our knowledge of what the MHC coding region looks like in multiple animals," Karl says. While several types of macaques have been sequenced over the years, most assemblies relied on tools built for human data. Those tools work well in regions where the macaque and human genomes are very similar, Karl says, "but they tend to really fall apart in places where there are highly repetitive regions like the MHC." This effort would go well beyond the couple of available reference assemblies for a more comprehensive representation of variation across macaque species. Such a resource would be critical for accurate alignment in resequencing studies conducted with other technologies.

The team is also beginning to make use of the Iso-Seq method for macaque T-cell RNA sequencing. This approach makes it possible to do "a transcriptome project to characterize the most abundant isoforms for a whole cell's worth of RNA instead of going gene by gene," Wiseman says. Much of that work is expected to be done on the university's new Sequel System, which will add significantly to the team's sequencing capacity.

Ultimately, all of these efforts will contribute to a much deeper understanding of primate genetics for improved animal models and better studies. "In the future, every macaque that's used in an experimental research study is likely going to have its genome profiled. We're going to want to know what its MHC looks like, what its KIR looks like, and so on," O'Connor says. "We're not there yet, but putting together these collections of sequences will make it possible for us in the future to get more comprehensive genome profiling out of each individual animal."

REFERENCES

Westbrook CJ, et al (2015) No assembly required: Full-length MHC class I allele discovery by PacBio circular consensus sequencing. *Human Immunology*. 76(12), 891-896.

Karl JA, et al. (2017) Major histocompatibility complex haplotyping and long-amplicon allele discovery in cynomolgus macaques from Chinese breeding facilities. *Immunogenetics*. 69(4), 211-229.

Prall TM, et al. (2017) Improved full-length killer cell immunoglobulin-like receptor transcript discovery in Mauritian cynomolgus macaques. *Immunogenetics*. 69(5), 325-339.

HUMAN BIOMEDICAL RESEARCH

PLANT AND ANIMAL SCIENCES

PN: CS120-062217