NEW VIEW OF CHICKEN TRANSCRIPTOME OFFERS CLUES TO HEART DEVELOPMENT

Scientists at the Gladstone Institutes were early adopters of SMRT® Sequencing for transcriptome studies. In a recent study, they used full-length isoform sequence data to overhaul the annotation of the chicken genome, thus providing heart biology researchers with a valuable new reference tool for future studies.

Neither Alisha Holloway nor Sean Thomas is an expert in chicken biology, but together they have developed an invaluable genomic resource for the chicken biology research community. Their discovery of a large number of previously inaccessible transcripts will serve as an important foundation for further studies of the model organism and of heart biology.

Holloway and Thomas are both based in the Bioinformatics Core at the Gladstone Institutes in San Francisco. They work with the Bench to Bassinet Program, a collaborative effort started by the National Heart, Lung, and Blood Institute to study heart biology and congenital heart problems. "The idea of the consortium is to incorporate clinical and basic science to understand heart development," says Holloway, Director of the Bioinformatics Core. "We are part of the basic science side of that consortium."

In a project that began in 2012, Holloway and Thomas, a Staff Research Scientist, worked with collaborators using chicken as a model for heart development. They planned to use short-read sequence data to add to existing resources of chicken transcripts, but quickly realized that the data could not conclusively link the exons in any given gene.

Meanwhile, Holloway and Thomas had been keeping an eye on Pacific Biosciences due to the unique long-read sequencing capabilities of the platform. The PacBio® sequencer had already been shown to contribute significantly for genome assembly improvement, but at that time its potential for improving genome annotation was not yet established. Holloway and Thomas saw the potential of the technology and were willing to try full-length isoform sequencing on their samples. They emerged with enough chicken transcript information to revolutionize what's possible for scientists studying gene expression in the organism.

A Model Chicken

Chicken is an ideal research organism for scientists in the Bench to Bassinet Program because it has long been an excellent model for embryogenesis and is useful in elucidating heart biology as well. Genes found to be important in chicken heart development may help scientists find orthologs in human and other species.

Dr. Alisha Holloway, Director of the Bioinformatics Core at the Gladstone Institutes in San Francisco, helped develop a new reference tool that will assist heart biology researchers in future studies of the model chicken organism.

Holloway and Thomas do not usually conduct studies on chicken, but their bioinformatics expertise allowed them to dive into existing sequence resources and figure out how they could help their consortium partners. Immediately, they saw a major problem. "Even at a gene level, the annotation was really bad for chicken," says Holloway, who estimates that half the genes were missing. "Our collaborators were having a lot of trouble with the analyses they were doing because they couldn't rely on a good gene annotation like you have for mouse and human."

Thomas says the problems were quite obvious: RNA lining up to unannotated parts of the genome indicated missing genes, while some genes that had been annotated automatically "were much longer than any gene would ever be."

So Holloway and Thomas signed on for a major undertaking: overhauling the annotation so that researchers everywhere — not just those in the Bench to Bassinet Program — would have a reliable starting point for their studies. But they hit a snag of their own. "We originally thought that we would try using the short-read data to see if we could reconstruct transcripts," Holloway

Sean Thomas, a Staff Research Scientist at the Bioinformatics Core at the Gladstone Institutes, used SMRT Sequencing data to span entire transcript isoforms.

says. "The problem with short-read data is that the reads don't span more than two or three exons, even if you're using paired-end data, so we weren't able to reliably determine which exons were connected to each other for transcripts that have a complex exon structure."

They turned to the only long-read sequencer on the market for what would become one of the first major transcriptome studies using SMRT Sequencing. "We started exploring the idea of using PacBio data to cover a full transcript," Holloway says. With long reads, she thought, "we would have the full structure of the gene." PacBio's Iso-Seq™ method generates data that span entire transcript isoforms from the 5' end to the 3' poly(A) tail, providing critical information about gene structure. The results are also far more accurate than traditional RNA-seq since full-length transcripts eliminate the need to infer isoforms from short-read data.

Going Long

With a few initial pilot runs, the Gladstone team became familiar with the PacBio data and applied it to the study of heart tissue from five adult chickens. "It's a completely different sequencing platform and has different strengths and weaknesses compared to other platforms," Thomas says.

The work paid off: Holloway and Thomas published their results in a *PLoS One* paper in 2014. They wrote, "By generating new long-read sequences and incorporating existing short-read and EST sequences, we identified thousands of transcript isoforms as well as hundreds of genes not currently included in the Ensembl annotations."

Because SMRT Sequencing data were still relatively new in the transcriptome field, the scientists added short-read data to call exon junctions observed in both data sets for additional validation. "We saw really good agreement" between short-read and long-read data, Thomas says, noting that the long reads allowed the team to link multiple exons and clearly define full genes.

The researchers were very pleased with the amount of new gene information generated in this experiment — specifically for heart, the main tissue of interest for this investigation. The publication cites 9,221 new transcript isoforms, including 5,930 genes with new annotations and 539 entirely new genes. They also found 2,337 exons in new genes, and 5,299 novel exons in known genes. Compared to known transcripts — there were only some 16,000 in Ensembl and 5,000 in RefSeq when they started the project — this marks a major step forward for the chicken annotation. "It's a huge improvement over what was there," Thomas says. "It's really important that we were able to identify hundreds of new genes — complete genes that had not been annotated previously."

They also followed up on the newly identified genes, checking whether they were expressed in other tissues in chicken and whether similar genes were known to exist in other species. "While searches against other databases yielded homologs for three of the new gene regions ... the remaining genes remain uncharacterized, including the 121 gene regions that exhibited tissue-specific expression and might play key roles in chicken biology," they noted in the paper.

"We were focused on new transcripts and genes in the heart, so we knew that the sequencing we were doing was not going to identify every transcript in the chicken," Holloway says. "But we feel that we have a very good estimate now of what's expressed in the developing heart in chicken."

Next Steps

Now that Holloway and Thomas's findings have been made publicly available, scientists everywhere can benefit from the vastly improved chicken annotation. As reference material, it will serve as a critical starting point for researchers trying to understand gene expression in the organism. The information will also be useful in mapping transcripts to other species to look for orthologs. "That will help us learn how they are related evolutionarily to other genes that we know are important to heart development in mammals," Holloway says. "We can make comparisons between species more easily."

That will come in handy for the Gladstone team, which continues its work with the Bench to Bassinet Program. In the meantime, Holloway and Thomas's publication adds to the growing body of work demonstrating that long reads provide a uniquely comprehensive view in transcriptome studies.

www.pacb.com/isoseq