Sequencing the Tree of Life to Understand Evolutionary History

Using the MiSeq® system, Dr. Alan Lemmon, Ph.D. sequences hundreds of genes across many non-model species using a new anchored enrichment approach that circumvents marker development.

Working in the field of phylogenetics, Dr. Alan Lemmon, Ph.D. uses next-generation sequencing (NGS) to sequence hundreds of genes from many species in an effort to define the relationships that make up the tree of life. In some cases, these species might be 250 million years divergent from the nearest model species. In order to access homologous genomic regions from diverse organisms, his group uses an anchored enrichment approach that targets highly conserved regions with non-conserved flanks.

Q: What is unique about your projects' sequencing requirements?

Alan Lemmon (AL): In phylogenetics we need to sequence thousands, even millions, of species, but require just a few hundred genes (e.g., 0.05% of a mammalian genome). Recently, I've been working on different methods of enrichment, adapting sequence capture technology to phylogenetic applications, and to non-model species for which a reference genome is not available.

Q: How does the MiSeq system improve upon previous methods?

AL: We had outsourced to a core lab that used the HiSeq® system and the turnaround time was several months, so we'd have to wait for a lane to fill. Most of their customers were studying humans and doing resequencing, or short, single-end reads. We typically do paired-end 100 bp or longer reads so the waiting game was getting frustrating.

Q: What samples did you select to test on the MiSeq system?

AL: I submitted samples that were going to be used as proof of principal to test our new anchored enrichment application, including ten indexed pooled libraries, five from model species for which we know the genomes (those are the ones from which we designed the probes), and then five additional non-model species with different degrees of divergence from the nearest model. All samples were pooled and then captured. For the MiSeq project, we pooled 10 samples in one lane, but we anticipate being able to sequence up to 48 samples in a single lane. This really speeds things up.

"We found that we can perform a phylogenetic project for ~1% of the cost of traditional PCR + Sanger approaches and ~5% of the cost of modern Amplicon + NGS approaches."

Alan Lemmon, Ph.D. is an Assistant Professor in the Department of Scientific Computing at Florida State University where he uses next-generation sequence data to develop molecular tools for phylogenetics, phylogeography, and population genetics.

Q: Did you find anything new and interesting?

AL: We found that we can perform a phylogenetic project for ~1% of the cost of traditional PCR + Sanger approaches and ~5% of the cost of modern amplicon + NGS approaches. The paper describing this new anchored enrichment approach will be published in a symposium issue of Systematic Biology on applications of next-generation sequencing for phylogenetics and phylogeography.

Q: Does the MiSeq data quality meet your expectations?

AL: The MiSeq data was great. We expected the quality to be good in general, but with the typical decrease in quality from the 5′ to 3′ end of the read. MiSeq produced 1/10th or so of the number of reads of HiSeq, but we had a longer read length. We obtained paired 100 bp reads from HiSeq and paired 150 bp from MiSeq. The quality scores, in terms of relative position, were pretty much identical for HiSeq and MiSeq. Overall, we were quite pleased with the quality of the data.

Q: How was quantity of data from the MiSeq system?

AL: I think we got a more from MiSeq than we expected in terms of total Gb. It definitely produced the quantity of data that we needed. MiSeq yielded somewhere around 2 Gb, if you add across all the libraries we pooled. That's above what was promised, which is great.

Q: Did you see any improvements in downstream analysis?

AL: When we need read lengths longer than 100 bp, we'll typically use an insert size of about 150 bp. This way we'll have overlapping reads when we do paired-end sequencing and get really high-quality reads due to the 3′ end overlapping of the two reads. This is useful because it allows us to assume that the large majority of the reads are perfect, and that there's very little indel error relative to homopolymer problems, so that we can turn to a different kind of bioinformatics. This method of analysis is a lot faster and allows us to do estimates of the transcriptomes quickly because we don't have to worry about the nuances of having a lot of errors in the reads. This was really useful for our snake venom evolution project.

Q: What method do you use to prepare libraries?

AL: Currently we're using a Meyer and Kircher protocol published in Cold Spring Harbor Protocols in 2010 for our library preparation. It allows us to pool hundreds of individuals and do indexed paired-end sequencing. We're using this protocol because we need to have longer insert sizes to stretch out and capture fragments farther away from our coding region. I'm excited about the Nextera® kit because of the 15-minute hands-on time for library preparation. We look forward to being able to use the more streamlined protocol in applications requiring shorter insert sizes. For example, in one upcoming project we plan to estimate the relationships of hummingbirds from start to finish in one week by combining anchored enrichment, Nextera, and MiSeq technologies. The previous phylogenetic estimate required years of work.

Q: What do you see as the biggest benefit of the MiSeq system?

AL: Just having a fast turnaround really made all the difference. Up until now we've been hobbling along developing a protocol for this project, with a few months between revisions. With MiSeq, we're looking forward to sequencing a library in one day rather than in ten days with the core lab, and not having to wait for the lanes to fill. When we get the data back we can do the bioinformatics in a day and quickly turnaround improvements to the protocol. This will be really useful in allowing us to get through a lot of diverse types of projects.

Q: With such a fast turnaround time, how do you see the research program changing?

AL: My research program is going to change a lot because now we can optimize the capture efficiency for different sized genomes. We're interested in amphibians and we usually test frogs. But salamanders have huge genomes and many researchers are interested in applying the anchored enrichment approach to them. This spring we'll do some library preparations in different ways with salamander genomic DNA sequenced on MiSeq. Because we'll have a quick turnaround time, we can prepare samples in different ways and see which approach improves enrichment efficiency.

"It's amazing how much nicer it is to look at the Illumina assemblies. They are just really clean."

"...in one upcoming project we plan to estimate the relationships of hummingbirds from start to finish in one week by combining anchored enrichment, Nextera, and MiSeq technologies. The current estimate of phylogeny took years of work."

Q: How does the MiSeq system affect your pooling strategy?

AL: We calculated the number of individuals we could comfortably pool on HiSeq versus MiSeq. It depends on the genome size of the taxonomic group. Bird genomes are fairly simple and, as long as the capture is efficient, we can pool dozens of individuals and run them on MiSeq, which will reduce costs and save us time. Taxonomic groups with larger genomes, like frogs, are still going to require running on HiSeq. But we can do preliminary tests on MiSeq with our samples to make sure that they're ready to go for HiSeq. We're looking forward to that use of MiSeq as well.

Q: What features impacted your decision to select the MiSeq system over alternatives?

AL: I spent a lot of time thinking carefully about which technology to choose. We stumbled across a little catch in the Ion Torrent system. From my understanding, the read length you get off the machine is a function of the bead size, and the bead size determines how tightly you can pack the wells. The larger the bead size, the fewer wells you can have. To increase the read length, they're going to have to decrease the density of the beads, which means a smaller number of reads in the area. It really shows that there's a trade-off there.

In addition, we're comfortable with the Illumina library prep and didn't want to go to emulsion PCR. We hadn't done it before, and it would have been a huge headache to switch over. Also, we had collected some data for the snake venom project with the 454 and had issues with the homopolymer errors causing problems, especially when we were reconstructing transcriptomes and trying to get the whole reading frame. We would have cases where the incorrect number of bases were called for a particular homopolymer, which would lead to a premature stop codon, and that would lead to an incorrect estimation of the open reading frame, even with really high coverage. Because of those issues I wanted to steer clear of the homopolymer issue.

I teach a course in next-gen sequencing every spring and when the students look at the two types of data, for example 454 or Ion Torrent data versus Illumina data, it's amazing how much nicer it is to look at the Illumina assemblies without the homopolymer errors in there. They are just really clean. It's easy to see and you avoid a lot of issues. I was just naturally averse to the type of data generated by Ion Torrent and 454, so I've been pushing for Illumina all along.

Q: How will having a MiSeq system impact processes in your lab?

AL: Having MiSeq in our lab is huge. The waiting time will go way down. It will accelerate the pace at which we can do projects, and troubleshooting. Also, because using MiSeq is so streamlined, our lab technicians can be trained to use it versus HiSeq, which might require more technical skills. That is a huge deal. Having more of the machines spread across universities and campuses will reduce the burden imposed on genome centers and data transfer. With HiSeq, we'll typically submit our order and then have to send the core lab a hard drive to get the data back. Because MiSeq is on site, we can deal with the data transfer a lot more easily.

Q: What would you like to do on the MiSeq system that you don't expect to be able to do now?

AL: The number of reads isn't the issue, but being able to get longer reads is important. I'm really excited about The Broad Institute report on their 300 bp read run. Some of our projects will benefit from having longer reads, using the overlapping approach to get high-quality 150 bp or longer reads. If we can basically throw out any read that has any sequencing error, it makes the bioinformatics a lot less messy. With longer reads we could switch to this overlapping read approach and the downstream bioinformatics would be a lot easier.

Q: What projects would benefit most from the MiSeq system?

AL: There's a large group of people who haven't really had access to next-gen sequencing. They've maybe done a little transcriptome sequencing and a little amplicon sequencing, but I think a large group of people are waiting to use the technology. The MiSeq system would be ideal for universities that are not processing a lot of samples.

Learn more about the MiSeq system at www.illumina.com/miseq

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 770-2012-007 Current as of 16 April 2012

illumina®