Special News Feature - DNA Sequencing

Since its development, targeted resequencing has made studying large numbers of genes easier for geneticists. Now, evolutionary biologists are adapting these methods to analyze gene families separated by hundreds of millions of years. Patrick Lo takes a look at how their efforts could redefine the “Tree of Life”.

For decades, systematic biologists have been defining evolutionary relationships amongst all organisms and creating phylogenetic “trees” to illustrate their findings. But delineating the exact shape and branching patterns along the so-called “Tree of Life”—the phylogenetic tree describing the evolutionary relationships of all organisms—remains a challenge, even in this era of molecular biology where deep sequencing and genomic analysis are commonplace.

Until recently, classifying an organism along a phylogenetic tree required PCR amplification with degenerate primers followed by amplicon sequencing to study genomic loci in distantly related taxa. This approach, however, limits the number of loci that can be simultaneously examined to significantly less than the hundreds or thousands needed to truly classify very distant taxa.

In order to overcome these limitations and enhance the resolution and size of the Tree of Life, several groups are advancing new methods, combining the power and throughput of next-generation sequencing with novel targeted enrichment strategies to easily capture and sequence thousands of loci in a single experiment.

Ultraconservative

Targeted enrichment uses specially designed oligonucleotide probes to isolate specific sequences from a mixture of DNA molecules through hybridization either on a solid surface or in solution. Isolated sequences can then be used for downstream analyses, including next-generation sequencing. Although developed initially to pull out coding sequences from closely related individuals, evolutionary biologists quickly saw huge potential for targeted enrichment in studies of distantly related species.

“Targeted enrichment techniques allow you to target many hundreds of thousands of loci very easily. What we then needed was a target to go after, and conserved sequences distributed throughout the genome are a good target,” notes Brant Faircloth, an assistant researcher in the Department of Ecology and Evolutionary Biology at UCLA. Faircloth and his colleagues decided to extend targeted enrichment techniques by using ultraconserved elements (UCEs) as probes. Originally discovered in humans, UCEs are somewhat mysterious sequences with unclear functions whose high sequence conservation and other features make them well suited as molecular markers.

Using around 2400 UCE-anchored loci from 9 non-model avian species, Faircloth et al. were able to obtain alignments of nearly 850 loci, recovering the established phylogeny among and within 3 bird lineages that had diverged by 65 million years, providing one of the first demonstrations that next-generation sequencing of UCE-enriched DNA could be applied across the Tree of Life on a fairly deep time-scale.

While his methodology can work across deep time scales, Faircloth actually intends to reverse gears and apply the technique to shallower time-scale questions.

“Phylogeographic studies are of interest to us. We want to start looking at populations of individuals and how those populations may have diverged over shorter timespans. As opposed to 200-400 million years ago, now we’re talking about 10-30 million years ago and less,” explains Faircloth. In addition, the method will be used for population studies. “Among individuals, within populations, we know the same suite of loci will enrich targets that are informative at the individual level. We can actually look at parents and offspring within a population and infer who the parents of offspring are, using these same exact ultraconserved loci that we’ve targeted for phylogenetic questions.”

Please mark the
GMO

It's not always that easy.

Better use BIOTECON Diagnostics'
foodproof® GMO Test Kits.

www.bc-diagnostics.com

BIOTECON Diagnostics

Phylogenetic trees, such as the one above, demonstrate relationships among animals. Developments in sequence capture techniques are now enabling biologists to explore genes and gene families in distantly related species.

Not Quite So Ultra

Annoyance, in part, fueled Alan Lemmon's development of a targeted enrichment sequencing method. "[We] were fed up with having to develop primer pairs for each new non-model species we were working on, or at least getting existing primer pairs working," recalls Lemmon, an assistant professor in the Department of Scientific Computing at The Florida State University.

Lemmon, along with his collaborator (and wife) Emily Moriarity Lemmon, an assistant professor in the Department of Biological Science at The Florida State University, decided six years ago to apply next generation sequencing technology to the study the phylogenetics of non-model organisms. But what they needed was a way to find common sequences in divergent genomes and isolate those for sequencing.

For their anchored enrichment approach, Lemmon used highly conserved anchor regions of vertebrate genomes as probes (2). Through a comparison of the genomes of five model vertebrates, these anchor regions were chosen for being highly conserved single-copy sequences flanked by less conserved regions that were well distributed throughout the genome. The level of conservation of these anchor regions, however, is lower than the UCEs used by Faircloth. This sequence variation is useful for studying shallower time scales.

Using their method, Lemmon's team was able to capture a substantial number of anchor sequences from each of five

non-model organisms when its corresponding model organism in the same vertebrate clade was used to generate the bait probes. The divergence times between model and non-model organisms in each pair ranged from 94 to 254 million years.

Having established the utility of anchored hybrid enrichment, Lemmon now wants to share his lab's expertise and infrastructure. "We really wanted to make sure we built infrastructure and planned for the long term," explains Lemmon. "We could have the sort of system set up so that anyone who wants to work on any species can either work with us or not, but we have really good tool kits for doing the probe identification and probe design, and then the downstream bioinformatics as well." In the end, this approach is ideal for those without expertise in next-generation sequencing or who don't have the time and money to carry it out.

Based on word of mouth, Lemmon's group has undertaken about 12-16 collaborations in the past year. "The next year's going to be a big boom year for the anchored phylogenomics method. We have a lot of people that we're developing kits for which are going to turn into a lot of papers."

Scoring a Touchdown

While the UCE and anchored enrichment work well, each requires identifying and targeting highly conserved sequence elements in species whose genomes are otherwise highly

divergent. This, however, is not ideal for targeting particular genes of interest that are highly divergent from the baits used for enrichment. It was this issue that prompted Gavin Naylor, a professor in the Department of Biology at the College of Charleston, to develop another approach to targeted enrichment sequencing for phylogenetic analysis (3), an approach that can isolate any large group of target genes in evolutionarily divergent organisms.

"We wanted to be able to choose parts of the genome that we thought were interesting and had good properties for phylogenetics. But we wanted to be able to choose our targets; we didn't want to be painted into a corner and only choose those which were ultraconserved," says Naylor.

As it turns out, existing gene capture protocols are not suitable for targeting highly divergent gene sequences. "Gene capture is very specific and very stringent, and while you can pull it out of one species, you try to do it in another one, it's not going to work. So, that's where we try to use relaxing biochemistry protocols to pull them out." Naylor's solution, as devised by Chenhong Li, his post-doc at the time, was to use a touchdown hybridization scheme involving lower stringency hybridization and washing steps for gene capture, thereby allowing the probes to retain more divergent target sequences.

To test their method, Naylor's group first selected target genes in several model vertebrates that were present as unique sequences in all of their genomes. After eliminating possible paralogs, these putatively orthologous sequences were used as baits in two rounds of gene capture with touchdown hybridization to isolate sequences from a distantly related species in the same vertebrate class for each of the model vertebrates. The divergence times for the species in each of these pairs varied from 100 to 300 million years, with similarities of target sequences ranging from 89% to 61%.

Compared to standard conditions, the two rounds of gene capture with the less stringent hybridization and wash conditions greatly increased the number of target sequences captured, and this improvement was especially dramatic when the target species were more highly divergent from the bait species.

With its improved ability to capture sets of pre-specified genes from highly divergent taxa, Naylor envisions that his method will be especially useful for comparative biochemistry and physiology studies. "Imagine you've got 55

genes in a pathway that's associated with cancer, and you can make baits for all of those 55 elements associated with a particular disease condition or a particular morphogenetic pathway... and you've got a candidate gene pathway, say from the human or zebrafish. We can make probes for all of the elements in that candidate gene pathway and interrogate for that set across taxa."

Challenges for the Future

For most researchers, the major advantage presented by these three new targeted enrichment methods over carrying out *de novo* whole genome sequencing (WGS) boils down to costs. While the expense of WGS has dropped considerably in recent years, it is still far less expensive to sequence large numbers of samples with targeted enrichment, especially when it comes to storing and analyzing the much larger, and more diverse, volumes of data generated with WGS.

Gavin Naylor has developed a new hybridization-based approach for targeted gene capture applications.

"There are so many gene families, so many duplications, so many elements of unknown function," notes Naylor. In essence, using targeted enrichment upfront reduces much of the bioinformatic filtering of sequences that needs to be carried out downstream.

Still, data analysis remains a worry as the amount of sequence information being generated with targeted resequencing is nearly beyond the limits of present bioinformatics methods to process efficiently. "Collecting data is not the issue; analyzing the data is the biggest problem that we have. ...I would

imagine in the next two to three years we really see a number of interesting and provocative and hopefully very helpful analytical methods that come on the scene and allow us to analyze the data we have collected," says Faircloth.

Another interesting challenge for the future, according to both Faircloth and Naylor, is how to handle paralogs during sequence analysis. At present, sequences that appear to be paralogs are explicitly eliminated. "We're probably throwing out information that could be massively useful to us, but we're throwing it out because there's really no good mechanism to deal with it," explains Faircloth.

In the end though, these new methods are leading to a new excitement among phylogeneticists around the globe. "Everyone I think has the sense, even those who aren't savvy to bioinformatics or even next-gen sequencing, there's a change happening. Our goal is really just to facilitate science and to help people produce the best possible datasets," says Lemmon.

For Faircloth, targeted resequencing has created a unique and exciting opportunity when it comes to understanding organisms and their evolutionary relationships. "What we can do now is amazing. It really allows us to work with all of these species that for the longest time have constrained our ability to understand the deepest relationships, or the shallowest relationships, or how different taxa spread across a phylogenetic tree differ in population genetic parameters. We can now work with all of these taxa and we're really not limited in terms of data collection and that is a fantastically powerful development."

References

1. Faircloth, B.C., J.E. McCormack, N.G. Crawford, M.G. Harvey, R.T. Brumfield, and T.C. Glenn. 2012. Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Syst. Biol.* 61:717–726.
2. Lemmon, A.R., S.A. Emme, and E.M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.* 61:727–744.
3. Li C., M. Hofreiter, N. Straube, S. Corrigan, and G. J.P. Naylor. 2013. Capturing protein-coding genes across highly divergent species. *BioTechniques* 54: 321–326

Written by Patrick C. H. Lo Ph.D.

BioTechniques 55:175-177 (October 2013)
doi: 10.2144/000114091