EVOLUTION

# Large-Scale Gene Comparisons Boost Tree of Life Studies

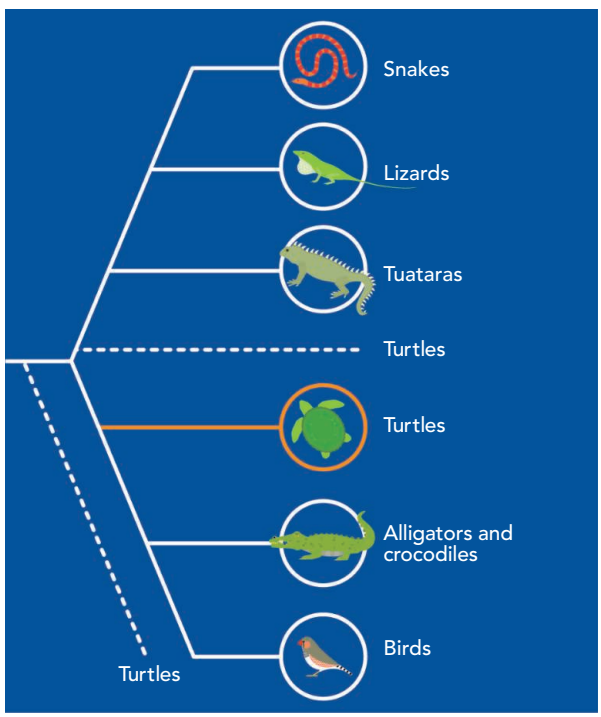
Emily Moriarty Lemmon will never forget how she struggled as a grad student to build a family tree for chorus frogs. To figure out how these sonorous amphibians are related to one another, she needed to assess as many DNA differences between species as she could, but time and money greatly limited the number of comparable sites she could identify and test. "It was really frustrating and kind of pointless," Lemmon recalls. In the end, she could not assemble enough data to build a satisfactory chorus frog tree, or phylogeny.

So in 2009, when she and her computer science–savvy husband, Alan Lemmon, began their faculty positions at Florida State University in Tallahassee, they decided to come up with a technique that would spare their students and others similar pain. They developed kits of genetic markers, identifiable in the DNA of many species, that provide a set of standard reference points for measuring genetic differences between species—and hence how closely they are related. Over the past 6 months, the couple embarked on a sort of promotional tour for the method at several major evolution meetings, sometimes wearing black T-shirts sporting their method's name: anchored phylogeny.

Their technique and similar approaches relying on predetermined sets of markers found in the genomes of multiple species are streamlining the study of the tree of life. "It's amazing," says Chris Simon, a molecular systematist at the University of Connecticut, Storrs. "You can spend your time working on the data analysis rather than spending huge amounts of time working on data collection." By making data collection for tree-building more comprehensive and faster—a matter of weeks rather than years—the new techniques promise to clarify our picture of the ancestry of familiar creatures. Already, one team pioneering the approach has redrawn the family tree of reptiles, showing that turtles are more closely related to crocodiles than to lizards.

Scientists originally drew family trees based on visible features—the number and placement of limbs or fins, or the shape of jawbones in fossilized remains, for example. Then genetics took over phylogenies, with biologists determining species relatedness from how DNA sequences differed among species. But pulling out even a few genes or DNA regions that are appropriate for such comparisons used to take years. As a result, there's often been too little data to resolve tricky branches of a tree, such as where the species in question are very distantly related or evolved very quickly. In theory, tree-builders can now simply sequence the whole genomes of multiple species and compare them, but that brute-force approach is still too expensive and complicated for many organisms other than microbes, with their simple genomes.

The strategy embraced by the Lemmons and others offers a middle ground: hundreds, even thousands of DNA regions to compare for a fraction of the cost of a whole genome. The key is identifying sequences common to many species that researchers can easily capture using specific probes—short pieces of DNA that home in on matching DNA. These shared sequences are sometimes too similar across species to help much with building a family tree, but they are typically flanked by variable DNA that can be much more revealing.




**Turtle triumph.** A method using ultraconserved DNA shifted the putative turtle branches (dashed lines), placing turtles (orange) closer to crocodiles and birds than to lizards and snakes.

Travis Glenn, an evolutionary biologist at the University of Georgia, and Brant Faircloth, now an evolutionary biologist at the University of California, Los Angeles, have focused on ultraconserved elements (UCEs), 100- to 200-base sections of DNA that are virtually the same across many vertebrates. Identical probes work in species separated by hundreds of millions of years of evolution. These probes fish out not just the UCEs but also flanking variable DNA. With sequence capture, "what you are sequencing is what you want … and [you] can look at the same markers in a wren, a snake, and a turtle," Faircloth says.

In early 2012, he and his colleagues demonstrated the technique on the placental mammal tree, showing that it placed multiple species on their correct branches. Later in the year, in *Biology Letters*, they published a UCE analysis that recast the reptile family tree, potentially resolving a longtime controversy by showing that turtles are more closely related to crocodiles and birds than to lizards and snakes. They have also refined the avian family tree, and in a just-published study of five rainforest birds, they showed that the technique can reveal the beginnings of new branchings, when species start to split.

The Lemmons use a different set of conserved markers. They didn't think such highly uniform markers as UCEs were needed and so looked for sequences that were similar in specific groups of animals but not necessarily exactly the same. To come up with their first set, they compared the human genome with those of the chicken, green anole, Western clawed frog, and zebrafish, identifying 512 DNA sequences in common to all and existing as single copies in the genome. They then paid a company to make thousands of probes for a "vertebrate kit," which targeted the DNA regions shared by the five species.

The probes work in many other vertebrates as well—including the Lemmons' chorus frogs. By using the probes to fish out and compare several hundred regions of DNA, Emily Lemmon was finally able to build a credible family tree, she reported at an evolution meeting in Ottawa in 2012.

Since then, the Lemmons have built a kit that includes more probes from reptiles and amphibians and, together with collaborators

CREDIT: G. GRULLÓN/*SCIENCE*





Published by AAAS

**Phylogenetics family.** Emily and Alan Lemmon want biologists to use their approach to building family trees.

in Australia, they are pulling out DNA from all that continent's frogs, as well as many of its snakes and lizards. "They've got a very clean operating system that they've developed so that anybody can get plugged into a great assembly line process to get their data," says David Weisrock, an evolutionary biologist at the University of Kentucky in Lexington.

They have worked with Simon, of the University of Connecticut, to develop a kit for a group of insects, the Paraneoptera, which includes true bugs, lice, thrips, and cicadas. Simon hopes the effort will resolve the relationships among the several hundred cicada species she studies. There are also kits for beetles; Hymenoptera, the group that includes wasps, bees, and ants; annelids; and one in the works for plants.

Alan Lemmon, a theoretical evolutionary biologist, estimates that they now have 50 collaborators, with more inquiries coming in every week. Collaborators provide the DNA of organisms for which they want to create a tree, and the Lemmons design the probes, sequence the DNA, and do bioinformatics quality checks and preliminary analyses at cost (about $175 per sample)—in exchange for co-authorship. In 2012, they processed 300 samples, and they expect to process 5000 a year by 2014. Some have criticized this as a "pay to play" approach, but Alan Lemmon emphasizes there's no profit made and says that the collaborators get a useful service. "We can get their feet wet without them having to spend a lot of resources," he says.

Faircloth and Glenn's UCEs are also gaining traction, as they and others use them to identify variable DNA in organisms ranging from fish to primates and eventually, plants and invertebrates. They take a different approach to pushing out their technology, having created a website where the probe sequences, protocols, and software are free for all to use. So many other researchers have contracted one company, called MYcroarray, in Ann Arbor, Michigan, to put together probes that the company just started offering several probe kits as catalog items.

As to which is better: "They're the Coke and Pepsi" of molecular phylogenetics, says Bryan Carstens, an evolutionary biologist at Ohio State University, Columbus.

Other researchers are coming up with their own flavors of sequence capture. The diversification is just what you'd expect from a successful innovation. Says R. Alexander Pyron, an evolutionary biologist at George Washington University in Washington, D.C.: "It's really light years ahead of what we were doing in the past."

**–ELIZABETH PENNISI**

BIOMEDICINE

# NIH Seeks Better Database for Genetic Diagnosis

In 2011, Heidi Rehm, a molecular geneticist at the Harvard-affiliated Brigham and Women's Hospital in Boston, was asked to help physicians follow up on a prenatal ultrasound scan that showed a "nuchal translucency," a low-density area near the spinal cord of a fetus. It's viewed as a sign that the fetus might have a disfiguring condition called Noonan syndrome. So Rehm's lab did a DNA test, which came back positive for a gene variant listed in the lab's internal database as Noonan-related. The parents ended the pregnancy.

Many months later, however, the lab learned that the researcher who linked the variant to Noonan syndrome had concluded it was benign after all. But there was "no easy way to put that information in the public domain," Rehm says. So it had been set aside.

To make that less likely to happen in the future, the National Institutes of Health (NIH) last week awarded $25 million for a new project called the Clinical Genome Resource, or ClinGen. The plan is to create a single reference point for data on medically important gene variants. Rehm, one of 10 researchers funded as part of the award, hopes the project will reduce uncertainties in genetic diagnosis.

She recently took part in an unpublished experiment in which three clinical labs tested their prowess at making diagnoses from the same DNA sample. When they compared how they rated genes for medical significance, Rehm says, the labs disagreed on 20% of them.

According to Lisa Brooks, overseer of the project at NIH's National Human Genome Research Institute, there are now about 2000 databases on genes and diseases worldwide. Each lab concentrates mainly on its own work. ClinGen will scoop up clinically relevant information on gene variants from as many databases as will cooperate, review them, and share interpretations. The aim, Brooks says, is to create a "curated and annotated" collection of all medically relevant human gene variants. The ambitious effort builds on a preexisting public database called ClinVar at NIH's National Center for Biotechnology Information. ClinVar already contains more than 51,000 reported gene variants from 60 signed-up contributors. Johns Hopkins University in Baltimore, Maryland, heads the donor list with 23,642 entries. Rehm's lab is second, with 6996.

ClinGen is supposed to expand this data collection and improve its quality. (ClinVar is not curated now.) One funded effort—including Rehm's lab; Robert Nussbaum at the University of California, San Francisco; and others—will develop standards and solicit genetic data and associated medical records from patients and doctors. Keeping personal information private while sharing data online will be a big challenge, Rehm says.

A second group will classify gene variants according to medical significance and call on specialist panels to make calls about specific cases—for example, to decide whether a variant is pathogenic or not. The third group will work on computerized data classification and release. One ambitious goal is to devise algorithms that predict the medical relevance of variants.

Sherri Bale, managing director of GeneDx in Gaithersburg, Maryland, who took part in the DNA interpretation experiment with Rehm, says that DNA testing companies like hers recognize the value of the project. "Without a curated database of variants, we are not going to be able to move into the whole-genome world," she says. "It has just got to happen."

**–ELIOT MARSHALL**

CREDIT: KRISTIN ENGEBRETSEN