DNA SEQUENCING

No Genome Left Behind

A project to sequence 10,000 vertebrates has just been launched, but sequencing technologies are not yet up to the task

Times have changed. A decade ago, it took several rooms full of sequencing machines and millions of dollars to decipher the genome of a small nematode. This month, two sequencing machines and $500,000 yielded a draft of the far more complex DNA of the cod. Now, a group of genome and museum experts are calling for the unraveling of 10,000 vertebrate genomes, about one per genus across the backboned animal world. They have already set in motion a global effort to gather the DNA needed for such an undertaking. "This is the most comprehensive experiment that anyone has ever proposed to analyze the evolution of genomes," says Oliver Ryder, a conservation biologist at the San Diego Zoo in California.

The Genome 10K plan, formally announced this week, is short on details: where funding will come from; what sequencing strategy to use; and how to process and make use of the data generated. But supporters are gung ho. "It's a grand plan that will happen," predicts Joel Cracraft, an ornithologist at the American Museum of Natural History in New York City.

New sequencing technologies aided the rapid completion of the cod genome project, one of about a half-dozen efforts using advanced, low-cost methods to analyze large vertebrate genomes, originally considered too complex to be unraveled with the newer technologies. And better tools are coming online. Besides, says Byrappa Venkatesh of the Institute of Molecular and Cell Biology in Singapore, there's a precedent for being impetuous: "After all, when the Human Genome Project was launched, there was neither an appropriate technology for sequencing a complex genome nor a suitable algorithm for assembling and annotating" the results.

Beyond the human genome

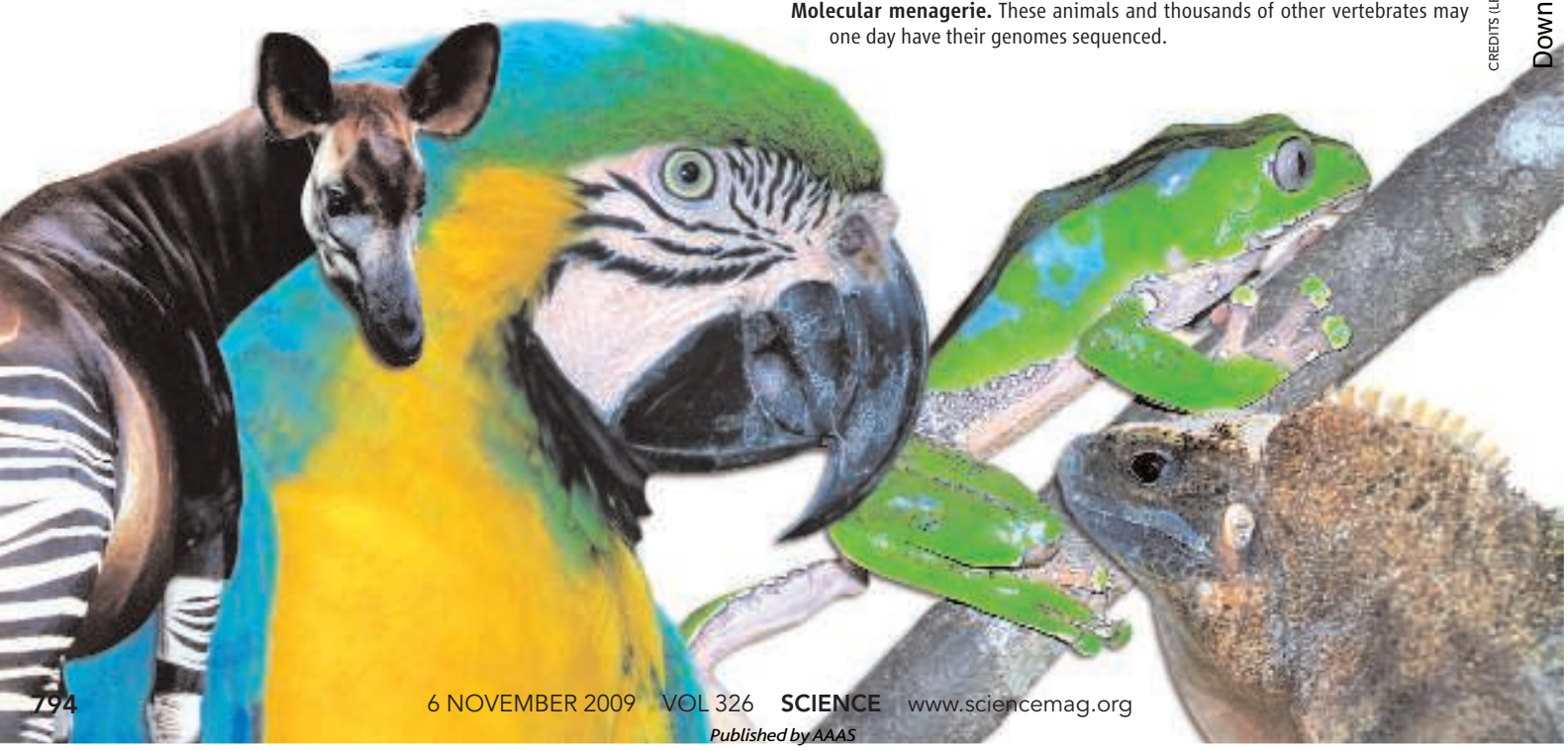
Ever since they finished the human genome sequence, researchers have been champing at the bit to sequence other genomes to compare with human DNA. Such comparisons help identify conserved regions that likely serve key roles in survival and also regions that differ between species and likely represent adaptations to a particular way of life.

With that in mind, the National Human Genome Research Institute (NHGRI) 5 years ago began to assemble a list of 32 mammals and 24 other vertebrates that would make good candidates for analysis. David Haussler of the University of California, Santa Cruz, and Stephen O'Brien of the National Cancer Institute in Frederick, Maryland, were part of the selection committee. "One of the most difficult things was to get specimens that had good DNA," O'Brien recalls. Anticipating that genome sequencing will get much cheaper, Haussler, O'Brien, and Ryder decided to prepare for a full-out assault on vertebrates.

The three organized a meeting in April at which 50 participants from the United States, Canada, South and Central America, Europe, and Asia came up with a list of 10,000 candidates, one-sixth of the known vertebrates. After an evening of hashing out reservations and concerns, they divided into animal-specific groups to see what suitable DNA existed, and where. To their surprise, they concluded that freezers around the world held DNA from more than 16,000 species.

As described online 5 November in the *Journal of Heredity*, the Genome 10K project has compiled a list of tissue samples from more than 43 institutions. For 600 species, cell lines already exist; the plan calls for establishing cell lines for 2000 more species. In addition, the researchers want to sequence several individuals for at least one species per order.

Advocates say the project will reveal new information about the human genome and basic biology. With dense sampling, "the insights into genome evolution and speciation from an evolutionary perspective would be extremely powerful," says William Murphy of Texas A&M University in College Station. Cracraft, for example, is curious about grebes and flamingos, which look nothing alike. "The question is how to account for that disparity in biology and form," says Cracraft.



Molecular menagerie. These animals and thousands of other vertebrates may one day have their genomes sequenced.

CREDITS (LEFT TO RIGHT): SAN DIEGO ZOO/WWW.SANDIEGOZOO.ORG; AMNH; ANDREW J. CRAWFORD; SAN DIEGO ZOO/WWW.SANDIEGOZOO.ORG

Published by AAAS

"Whole genomes are really going to spur things on."

Others believe the endeavor is essential to aid conservation efforts. "Wait another 20 years and it will be too late for several species," says Olivier Hanotte, a conservation biologist at the University of Nottingham in the United Kingdom, who has pushed to have endangered species included on the list.

Sequencing challenges

But waiting might be a good idea. So-called next-generation technologies have revolutionized sequencing during the past few years, providing cost and time savings (*Science*, 17 March 2006, p. 1544). But there have been tradeoffs: Efficiency brought a shorter "read length"—the number of consecutive bases that can be sequenced—and lower accuracy for each base determined. The accuracy can be compensated for by sequencing the genome multiple times to detect errors. Short reads aren't much of a problem for human DNA, because the existing reference human genome can help sort out where these bits of sequence belong. But assembling large genomes from scratch is quite a challenge.

Undaunted, about a dozen groups are pushing the limits of the new technologies, taking a variety of approaches, with varying degrees of success. "We're trying to figure out how to best do this," says Richard Wilson, director of the Genome Sequencing Center at Washington University School of Medicine in St. Louis, Missouri. Right now, there's a penalty for using cheaper methods: The least expensive sequencer generates the shortest reads, which are the hardest to assemble. Illumina technology produces a base of sequence for about one-tenth the cost of Roche 454 technology, but 454 has the advantage of generating reads about 350 bases long compared with Illumina's 75 bases.

When they tackled the cod genome in late 2008, Kjetill Jakobsen, a genomicist at the University of Oslo, and his colleagues used Roche 454. It took several months using two machines to generate the bulk of the 750-million-base cod sequence, covering the genome 30 times over. When they were done, "we used brute force," relying on extensive computing power to put the pieces together, says Jakobsen. The price to sequence: $500,000.

At least two sequencing centers have bet that they can sequence large genomes even with Illumina's very short reads. In January, the Beijing Genomics Institute in Shenzhen announced that in a month it generated 150 billion bases and stitched them together into a 3-billion-base genome of the Olympic mascot, a panda. They came up with their own computer program for assembling the genome, and the draft consisted of thousands of pieces averaging 300,000 bases long.

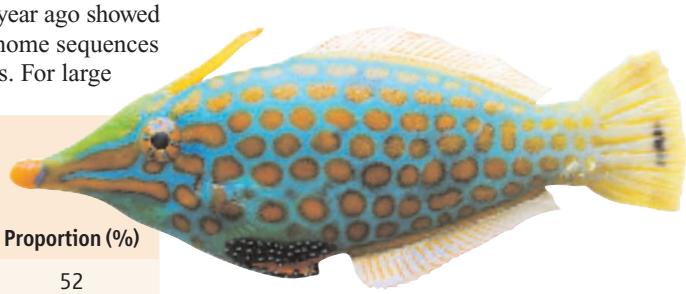
The Broad Institute in Cambridge, Massachusetts, has assembled the stickleback genome sequence this way and is now working on the bush baby. "We've believed in this concept for a while, even when people thought it was crazy," says the Broad Institute's Chad Nusbaum, who more than a year ago showed that bacterial and fungal genome sequences were doable with short reads. For large genomes, "it's not solved yet, but there are enough indications that it is going to work that we're convinced this is the way to go."

Several teams are using a mix of sequencing technologies, hoping this hybrid approach will boost accuracy and efficiency. As Washington University School of Medicine starts the vervet monkey, it expects to use both Illumina and 454 machines and to rely on the traditional capillary-sequencing technology used on the human genome for sequencing the ends of 100,000-base bacterial artificial chromosomes representing the monkey's genome. "For mammalian-sized genomes, that's going to be very helpful," says Wilson. Richard Gibbs of Baylor College of Medicine in Houston, Texas, has adopted a similar strategy for the baboon. That hybrid approach can speed up finishing, "but it's not a knock-it-out-of-the-park solution," Gibbs points out. "Difficult sequence is always difficult sequence."

Other potentially revolutionary tools could be on the market in the coming year. For example, the Menlo Park, California–based Pacific Biosciences is working on a technology that tracks the enzyme that puts DNA together, noting each new base added. It can sequence two to five bases per second, says the company's Stephen Turner. Last February, he reported read lengths of 3000 bases. "We're absolutely going after the de novo assembly market," says Turner.

Wishful thinking?

Even with all the advances, "the technology isn't there yet" for the kind of sequencing envisioned by Genome 10K, says Adam Felsenfeld of NHGRI. O'Brien, Haussler, and Ryder agree. They need the price tag to drop to $2500 for sequencing a large genome. Their goal is to spend $50 million for the whole project.



Coming up with the cash could be a challenge. Murphy, Haussler, and O'Brien, among others, hope private philanthropists will foot the bill. Hanotte is calling for public funding to ensure the unrestricted access to the data. Venkatesh thinks funding agencies that support biological, conservation, or biodiversity research will chip in. But NHGRI, which has supported much of the sequencing of large genomes in the United States, is noncommittal. "I don't know how much we will see it fitting in with our priorities," says Felsenfeld.

Furthermore, although Felsenfeld and others applaud Genome 10K for identifying tissue sources early, bioinformaticists are worried that insufficient attention has been paid to data management for the assembly and analysis of sequences. To assemble 10,000 genomes in 5 years as proposed will require processing a genome a day, warns Webb Miller, a computer scientist at Pennsylvania State University, State College. "There's a real problem here," he notes.

Despite large gaps in their plan, the Genome 10K leaders are supported by an enthusiastic cadre of peers. Participants at the April meeting and their colleagues are busy checking out the samples and deciding which ones have top priority. "We've got real momentum now," says Haussler. He hopes to do a pilot study demonstrating that these samples are suitable DNA sources for sequencing. Sooner or later, those genomes will be sequenced, Turner predicts. "I don't think it's a question of 'if' but 'when.'"

–ELIZABETH PENNISI

SPECIES ON THE SEQUENCING DOCKET

Group	Known Species	10K Species	Proportion (%)
Birds	9723	5074	52
Reptiles	9002	3297	37
Mammals	5416	1826	34
Amphibians	6570	1760	27
Fish	31564	4246	13
Total	62275	16203	26

CREDIT: ANDREW C. BENTLEY/BIODIVERSITY RESEARCH CENTER/UNIVERSITY OF KANSAS; SOURCE: ADAPTED FROM GENOME 10K COMMUNITY OF SCIENTISTS

Published by AAAS