

Bar Coding Life

A global project to create a new framework for identifying all biological species gains momentum

Ivan Amato



Photo by Chip Clark

Shelf Life Museum collections of biological specimens like this one maintained by the Smithsonian Institution's National Museum of Natural History could prove valuable for a new project to catalog millions of species by using short genetic signatures.

Every evening during collecting season between March and September for the past six years [Paul Hebert](#) has flicked on an ultraviolet lamp on the grounds of his 2.5-acre property, just outside of Guelph, Ontario. In fluttering droves, the moths have come, alighting onto a white sheet hanging by the light. Each night and just before the following dawn, Hebert has knocked some of the hapless moths into a jar made lethal with a bed of potassium cyanide crystals. Then, in his molecular biology laboratory at the University of Guelph, Hebert and his colleagues remove a leg from each of the lepidoptera specimens, extract DNA from the appendage, and sequence a specific and particularly telling snippet of the genetic molecule. They refer to it as "[the barcode of life](#)."

"From my own backyard, I've collected 20,000 specimens, representing 1,000 species-nearly 10% of all lepidoptera [moth and butterfly] species in North America," says Hebert, lead designer and cheerleader of an audacious project to develop the equivalent of a supermarket bar code for every living, nonmicrobial species on the planet.

Within a few years, some proponents say, you'll be able to go into a consumer electronics store such as Circuit City and buy a handheld widget for reading the biological bar code of just about any living thing you might encounter. One leg of an insect, a puff of fur, or any other sample with some cells in it is all it would take. Using onboard databases or ones accessed via wireless links, you could use this visionary gadget to instantly identify species of the organisms you find. Or you might even find that you happened onto previously unrecognized species whose bar codes are nowhere to be found in the master database.

Proponents have no trouble ticking off benefits: Scientists will be able to draw out the tree of life with more confidence and much finer resolution, down to the smaller branches and twigs. Border inspectors could quickly identify whether insect larvae in a multi-million-dollar shipment of grain are members of an invasive species that could wreak agricultural and economic havoc or are members of a benign species. Making this identification would help the inspectors decide whether to allow the valuable shipment through. Aviation safety specialists could gather flecks of bird remains from the fairing of a jet engine to determine what avian species are posing dangers in the airways and then follow up such identifications with bird-control measures tailored for the species.

Perhaps most important of all, suggests [Mark Stoeckle](#), visiting scientist at New York City's Rockefeller University, where he helps administer the Consortium for the Barcode of Life (CBOL), is the potential for the new technology to transform how the public perceives the living kingdom. If you can go out into a forest or even your backyard and for the first time name the species that surround you, Stoeckle argues, it is more likely that you will care about those species and support organizations and policies aiming to preserve, conserve, and otherwise better manage biodiversity.

Not everyone views bar coding with such optimistic eyes. Critics worry that the bar-coding effort, which so far has attracted about \$5 million in funding, could devalue centuries' worth of observations and traditional classification protocols that they say remain pivotal for making confident species identifications, especially when it comes to organisms that haven't been studied extensively.

Some critics fear that a zeal to bar code as many species as possible will divert already limited human and financial resources from what they say matters more when it comes to

making sound habitat and wildlife management decisions: uncovering the interrelationships among communities of species that comprise ecosystems. One of those critics, [Peter Roopnarine](#), associate curator of invertebrate zoology and geology at the California Academy of Sciences, says, "Knowing the number of species is just one thing; knowing the ecological functions of species is another."

None of these fears have sidetracked Hebert, considered the father of the bar coding of life movement. "I want to put together a compendium of life that completes the Linnaean enterprise," he says. He is referring to the vast enterprise to classify life forms that the Swedish botanist and physician Carl Linnaeus initiated in 1735 with the publication of the first edition of his "Systema Naturae." The most familiar manifestations of this ongoing, 180-year-old project are binomial species names such as *Homo sapiens*, which denotes creatures compelled to name and classify their fellow creatures, and *Drosophila melanogaster*, the fruit fly species that has been a darling of genetics research for a century. So far, taxonomists have named about 1.7 million of the total number of biological species, a voluminous roster estimated to number somewhere between 10 million and 50 million species.

Even in these days of routine genetic screening and DNA sequencing, identifying and describing species and deciding if a newly observed organism is a member of a previously unidentified species still require a degree of skill and judgment that comes only with years of experience, says research associate [Christopher P. Meyer](#) of the Florida Museum of Natural History at the University of Florida, Gainesville. Traditional methods of species identification depend on careful observations, measurements, and comparisons of specimens' shapes, anatomy, physiology, and behaviors. Taxonomy is difficult and time-consuming enough, Hebert notes, that a decade ago he had given up on the idea that it would be possible for the world's small community of taxonomists to ever compile a full listing of life on Earth.



MOLLUSCAN CAUTION A study of the brilliantly shelled group of marine mollusks known as cowries, several species of which are shown, indicates that bar-code data alone cannot always correctly identify the species of an organism.

Photos By Christopher Meyer

That was before he and other molecularly minded taxonomists realized what they might be able to achieve with a relatively tiny stretch of DNA found in almost all organisms. The tantalizing 650-nucleotide segment of DNA, which in a human represents a mere two-millionth

of the genome's more than 3 billion nucleotides, is located in mitochondria, the cell's bean-shaped power plants. This DNA segment is part of the gene *CO1* (also known as *cox1*), which encodes cytochrome c oxidase subunit 1, a protein without which cells would be unable to make adenosine triphosphate (ATP), biology's primary biochemical fuel.

This protein's center-stage position in life explains why its associated gene is found throughout the living kingdom. What's more, say bar-coding proponents, the featured *CO1* segment is short enough to be quickly and cheaply sequenced but long and complex enough that variations from individual to individual and from species to species show up. (In plant species, the gene doesn't vary enough, so CBOL participants are searching for a different stretch of DNA to use for bar coding plants.)

Perhaps most important, notes Lee Weigt of the Smithsonian Institution's Laboratories of Analytical Biology in Suitland, Md., is that the cost of sequencing this little genetic hot spot can be well under a dollar, especially for organisms and specimens already in hand in, say, museum collections. Weigt is just now integrating brand-new, high-throughput robotic equipment that, when up and running, should make possible one of the world's quickest, most cost-effective, and most prolific centers for reading biological bar codes.

As a museum-based scientist, Weigt is particularly intrigued by the opportunity and technical challenge associated with the millions upon millions of biological and medical specimens in envelopes, jars, tubes, drawers, and shelves in thousands of museums. Since the late-19th century, many biological specimens have been fixed in formalin, a preparation that now makes it difficult to extract DNA of sufficient length and chemical integrity for reading bar codes, notes David Schindel, executive secretary of CBOL, whose offices are housed at the Smithsonian Institution in Washington, D.C.

To remedy that situation and open up a potentially enormous supply of on-hand specimens to bar coding, Schindel and other CBOL leaders have charged the National Research Council with the task of organizing a workshop that will bring together chemists, biochemists, biophysicists, geneticists, bioinformaticists, and others to solve this chemistry-related bugaboo. A primary goal of the workshop, Schindel says, is to get the chemistry community interested in the Barcode of Life project and particularly in the DNA extraction problem posed by formalin-fixed specimens.

Meanwhile, some CBOL participants have begun bar coding in earnest in two projects focusing on the birds and fishes of the world. "We can do all the birds for \$2 million," predicts Stoeckle, who is helping to orchestrate the All Birds Barcoding Initiative. Its aim, he says, "is to establish a public archive of DNA bar codes for all birds, approximately 10,000 species, by 2010." Meyer, who recently completed a bar-coding investigation of more than 2,000 cowry (marine snail) species, says costs will be far greater for the enormous swaths of biodiversity that will require feet on the ground all over the planet to carry out the first step for bar coding: getting the organisms. Bar coding every species out there, according to some insider predictions, could cost \$1 billion.

If anyone can see the value of a quick, cheap means of identifying species, it's Meyer, whose job at the Florida Museum of Natural History routinely entails identifying, archiving, and displaying species. "I get random stuff sent to me all the time," he says, adding that the labels on the containers frequently are impossible to read or absent entirely. Bar coding can alleviate

the identification headaches that this lack of information brings, he concedes, but he also warns that bar coding bears scientific pitfalls that could undermine the technique's value.

In particular, he and colleague Gustav Paulay are concerned about the potential for bar coders to establish thresholds of difference in sequence variations as quick and simple metrics for determining if a specimen is a member of a particular species. For example, if the difference in the *CO1* sequence from two specimens is less than 2%, which amounts to differences in about a dozen of the bar codes' nucleotides, the specimens would be deemed members of the same species; if the variation is greater, they would be considered most likely to be members of separate species.

This approach has worked well for some well-studied groups, including birds. But Meyer and Paulay decided to test it on cowries, a large, globally distributed group of mollusk species. Many of them had been well-described and well-characterized over the centuries, and many had not. The advantage of this situation is that it provided a means for testing how well bar coding works for identifying specimens of familiar species as well as how well the procedure works for specimens of less studied or even previously unrecognized species.

"I picked cowries because they had so much history and I could stand on the shoulders of others," Meyer says. The often-beautiful-shelled cowries have been objects of research for centuries, even by Linnaeus, using traditional morphological comparisons as well as other comparative methods, including genetic ones. For their investigation, Meyer and Paulay gathered from around the world more than 2,000 cowry specimens, representing more than 93% of the group's 233 recognized species. To investigate bar-code variation within a species, the researchers collected, when they could, multiple specimens of individual species. They also sought representatives of the same species from geographically separated locations to test for geography-based variation.



MOTH CODE This *Eacles imperialis* specimen is one among thousands of moths and butterflies that Hebert has captured and bar coded (bar code shown below) at his Guelph lab.



Courtesy Of Keith Pickthorn

Their analysis of the *CO1* genes from all of their specimens revealed that bar codes worked quite well for identifying specimens of already well-studied and sampled species. For cowries by setting a threshold for a "same species" judgment at no more than a 3% difference between the bar code of an unknown specimen compared with bar codes of all other

cowries, "we could say with about 98% confidence that the specimen represents an independent evolutionary lineage," the investigators reported on Nov. 29 in the online publication *Public Library of Science (PLoS) Biology* ([dx.doi.org/10.1371/journal.pbio.0030435](https://doi.org/10.1371/journal.pbio.0030435)). In other words, using the bar-code data alone with no additional taxonomic data would only rarely lead to the erroneous splitting of specimens that might actually be of the same species.

However, the data also indicated that the difference in bar-code variation between members of the same cowry species often is comparable with the difference in bar-code variation among members of different species, especially those that have diverged relatively recently. As a consequence, using threshold cutoffs alone for making species judgments carried much higher error rates when it comes to the lumping together of specimens that actually are of different species. "Our data indicate that such use [of thresholds] will overlook at least one-fifth of life's forms that are distinct but less divergent," the researchers warn in their *PLoS Biology* paper.

Hebert has a "glass is half full" take on Meyer and Paulay's analysis of bar-code variation. To him, if it's possible to use bar codes to discern four out of five species in a region, especially ones in understudied places where habitats are disappearing fast, then bar coding still will score a win. The alternative, he says, is to reveal little or nothing about these species. Even with imperfect bar coding, he says, "we could go into taxonomic terra incognita" and still efficiently discover most of the new species there.

Regardless of whether the glass is half empty or half full, the taxonomists, geneticists, systematists, ecologists, biologists, and others whom C&EN contacted for this story agree the biological bar coding is likely to provide a powerful new framework for cataloging, categorizing, and monitoring the planet's biodiversity. And if investigators can open up the vast stores of preserved biological specimens to the procedure, costs of the overall project could go way down, says CBOL's Schindel.

"This is one of the most exciting times to be doing this," Meyer says. His cautionary study of cowries, he stresses, ought merely to "rein in the overzealous," who might want to rely too heavily on thresholds and other quick-and-easy ways of analyzing the data without also doing the hard work of traditional, skilled morphological, behavioral, and other comparisons.

"This is not a proposal to replace all morphological and taxonomic data," Schindel concurs. "We don't believe that species boundaries and names should arise exclusively from molecular data, and especially not from a small gene like *CO1*." For exploring understudied groups of species, he and others say, bar coding could be most fruitfully used as a triage method for identifying specimens that require additional examination by taxonomy experts for determining their species status.

Hebert, whose recent \$3 million grant from the Gordon & Betty Moore Foundation makes his bar-coding effort the best funded one in the world, expects the momentum to accelerate. "Within five years, we will have sequence compendia that serve as the portals to all species that taxonomic science has recognized," he says. Already, he notes, the technology for extracting and amplifying the DNA in tiny biological samples like insect legs is available in portable forms. Already there are powerful ways to store huge amounts of data and wirelessly communicate with databases. The only technology component that has to be

miniaturized for a handheld bar-coding device is a module that can do genetic sequencing. But that, Hebert says with a can-do attitude, is a matter of time.

He envisions taxonomy becoming a new everyday activity for everyone, like bird watching and hiking. Says Hebert: "There will be bar-coding devices for me, the port inspector, and school kids to identify any organism they encounter."



SPECIMEN TROVE New methods for extracting high-quality DNA from preserved museum specimens could reduce the cost of the Barcode of Life project.

Photo By Chip Clark

Chemical & Engineering News

ISSN 0009-2347

Copyright © 2006 American Chemical Society