BARCODE OF LIFE

Inspired by commercial barcodes, DNA tags could provide a quick, inexpensive way to identify species

BY MARK Y. STOECKLE AND PAUL D. N. HEBERT

Wandering the aisles of a supermarket several years ago, one of us (Hebert) marveled at how the store could keep track of the array of merchandise simply by examining the varying order of thick and thin lines that make up a product's barcode. Why, he mused, couldn't the unique ordering of the four nucleic acids in a short strand of DNA be mined in a similar way to identify the legions of species on earth?

Ever since Carl Linnaeus began systematically classifying all living things 250 years ago, biologists have looked at various features—color, shape, even behavior—to identify animals and plants. In the past few decades, researchers have begun to apply the genetic information in DNA

AMADEO BACHAR

to the task. But both classical and modern genetic methods demand great expertise and eat up huge amounts of time. Using just a small section of the DNA—something more akin to the 12-digit barcode on products—would require far less time and skill.

So we set a challenge for ourselves: to find a segment of DNA—the same part of the same gene for every species—that would reliably distinguish one animal species from another. Looking ahead, we expect that soon a handheld barcode reader, similar to a GPS device, will "read" such a segment from any tiny piece of tissue. An inspector at a busy seaport, a hiker on a mountain trail, or a scientist in a lab could insert a sample containing DNA—a snippet of whisker, say, or the leg of an insect—into the device, which would detect the sequence of nucleic acids in the barcode segment. This information would be relayed instantly to a reference database, a public library of DNA barcodes, which would respond with the specimen's name, photograph and description. Anyone, anywhere, could identify species and could also learn whether some living thing belongs to a species no one has ever recognized before.

Why We Need Barcoding

Morphology—the shape and structure of plants and animals—has enabled scientists to designate some 1.7 million species, a remarkable feat, and morphology remains the foundation of Linnaean-type taxonomic diagnosis. Relying on morphology alone to describe life's diversity has limits, however. The nuances that distinguish closely allied species are so complex that most taxonomists specialize in one group of closely related organisms. As a result, a multitude of

KEY CONCEPTS

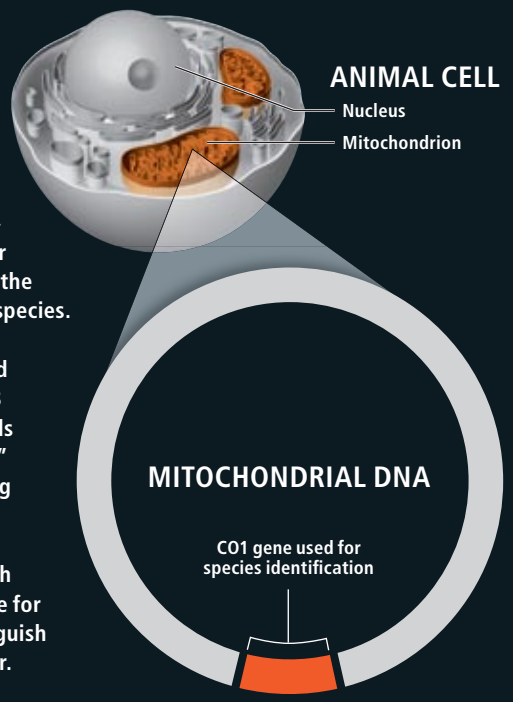
- Traditional methods for classifying plants and animals demand great skill and time. Examining a small portion of the DNA is faster and easier.
- This new method is called barcoding, because it was inspired by the barcode on products.
- The authors propose that a segment of mitochondrial DNA can distinguish animal species. They imagine a day when a handheld scanner (similar to a GPS device) will link to a database of the barcodes of all species. Then, by inserting a snippet of tissue into the scanner, anyone can get an instant identification of a creature or plant.

—The Editors

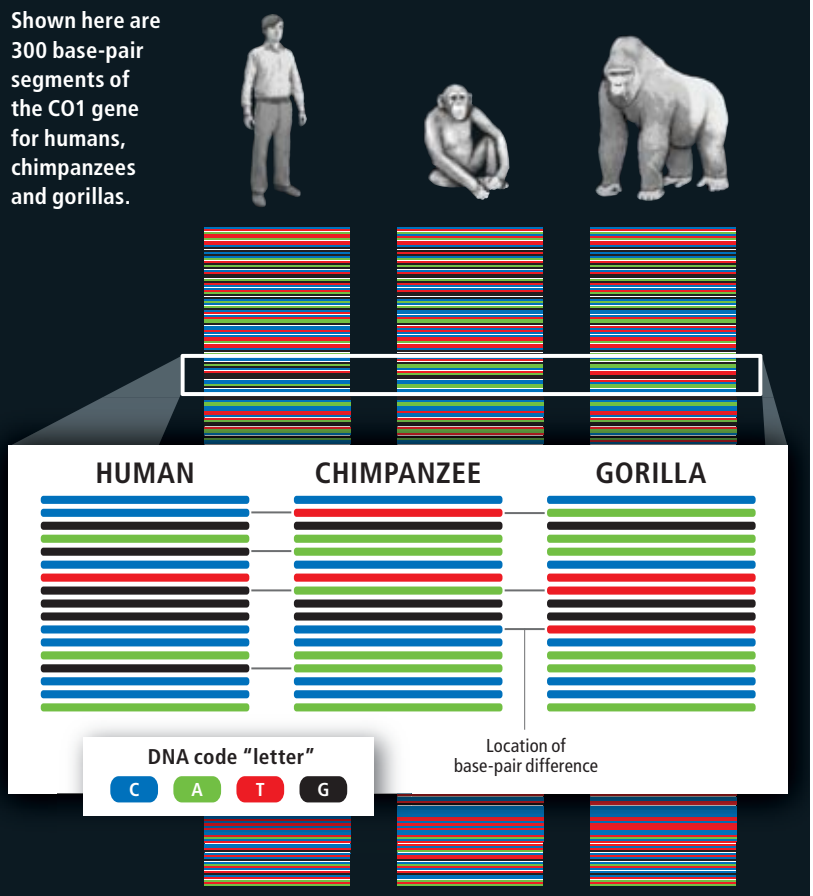
[HOW BARCODING WORKS]

STREAMLINED GENETICS

Each cell from an animal contains DNA in both the nucleus and the mitochondria. The authors and their colleagues selected a small segment of DNA from the mitochondria—the same short strand for each species—to use for the identification of animal species. The segment they chose comes from a gene called CO1. It contains only 648 base pairs of nucleic acids (essentially, the "letters" of the DNA code), making for quick reading of its DNA sequence. But the small piece varies enough from creature to creature for the differences to distinguish one species from another.



Shown here are 300 base-pair segments of the CO1 gene for humans, chimpanzees and gorillas.



taxonomic experts are needed to identify specimens from a single biodiversity survey. Finding appropriate experts and distributing specimens can be time-consuming and expensive. Web-based databases with high-resolution images help with the logistics to some extent, but other problems persist.

For example, biologists estimate that some eight million species have not yet been described, and as the encyclopedia of morphological characterizations expands, simply determining whether a specimen matches a known species will become increasingly difficult. Furthermore, eggs and juvenile forms, which are often more abundant than adults, may have no distinguishing characteristics and must be reared to maturity (if that is possible) to be identified. In some species, only one sex can be identified. For plants, a specimen may be readily classified from flowers, whereas roots and other vegetative parts are indistinguishable. A quick and easy standardized method of using genetic information could bridge these problems.

Making It Work

The first step toward discovering whether a pared-down method of using genetic information made sense was finding a short piece of the DNA that could actually deliver identifications—one that was long enough to contain information that would distinguish species but short enough to be fast and efficient to use. After some trial and error, we were able to settle on a particular gene segment as the standard reference for animal species. (Plants are another story [*see sidebar at top of opposite page*].) This segment is part of a gene housed in mitochondria—energy-producing subunits of cells, which are inherited from the mother. The gene we selected gives rise to an enzyme called cytochrome *c* oxidase subunit 1, or CO1 for short. The CO1 barcode region is small enough that the sequence of its nucleic acid base pairs (the "rungs" of the famous double helix) can be deciphered in one read with current technology. And although it is a tiny fraction of the DNA inside each cell, it captures enough variation to tell most species apart.

In primates, for example, each cell has about 3.5 billion base pairs. The CO1 barcode is only 648 base pairs long, yet examples taken from humans, chimpanzees and the other great apes harbor enough differences to distinguish the groups. Humans vary from one another at one or two base pairs in the barcode region, but we

TOMMY MOORMAN (*illustrations*); SOURCES: CORIELL INSTITUTE FOR MEDICAL RESEARCH (*chimpanzee and gorilla sequences*); BIODIVERSITY INSTITUTE OF ONTARIO (*colored barcodes*)

diverge from our closest relative, chimpanzees, at approximately 60 sites and from gorillas at about 70 sites.

Mitochondrial DNA proved especially suitable, because sequence differences among species are much more numerous than in the DNA of a cell's nucleus. Thus, short segments of mitochondrial DNA are more likely to parse separate species. In addition, mitochondrial DNA is more abundant than nuclear DNA and therefore easier to recover, especially from small or partially degraded samples.

To prove that this small DNA tag could actually identify a species, we, along with our colleagues, have tested the effectiveness of the CO1 barcode in diverse animal groups from land and sea, from the poles to the tropics. We have found that CO1 barcodes by themselves distinguish about 98 percent of species recognized through previous taxonomic study. In the remainder, they narrow identification to pairs or small sets of closely allied species, generally lineages that only recently diverged or species that hybridize regularly.

Now that we have found a barcode, the next step is to compile a reference library of this segment from specimens whose identity is already firmly established. By comparing barcode DNA from some creature against these "voucher specimens," researchers can determine whether the organism is a member of a known species or is a new find. The mechanics of creating the library are simple: someone obtains DNA from a tissue sample, determines the base-pair sequence of the barcode segment, and enters the information into a barcode database. The acquisition of specimens is more complex. The extent of variation within each species, though low, nonetheless suggests that at least 10 individuals per species should be analyzed to register this diversity. Even though the world's museums hold more than 1.5 billion specimens, most were not prepared with DNA recovery in mind, and many are simply too old to yield full barcode sequences. For older museum specimens that serve as original references for taxonomic names, amplifying a mini barcode of 100 to 200 base pairs, a size that can often be recovered from old or damaged DNA, will usually provide enough information to demonstrate membership in the same species as younger specimens with full barcodes. To aid construction of the barcode library, researchers at many institutions have begun assembling large tissue banks stored under conditions that preserve DNA.

Keeping track of so many specimens and their sequences is an engineering challenge in itself. But the process has already begun with the

THE UNIQUE CHALLENGE OF PLANTS

The gene used for barcoding animals is not practical for plants, because the plant genome has evolved quite differently. Also, an inability for two groups to mate productively with each other commonly defines animals as separate species, but many plant species can hybridize, which blurs their genetic boundaries. Scientists from museums, universities and botanical gardens around the world are now testing several highly promising gene segments that might serve as a barcode for all plant life.

[PRACTICAL USES]

BARCODING IN THE REAL WORLD

Once a handheld barcode reader is available for examining a tissue sample and is connected to a database, scientists foresee many practical uses:

- Biologists could identify organisms in the field to quickly assess biodiversity.
- Public health authorities could identify mosquitoes carrying infectious agents, such as West Nile virus, and other disease vectors, enabling timely application of targeted control methods.
- Restaurant owners and consumers could check fish to be sure what they are buying is what is advertised.
- Taxonomists could spot genetically distinct specimens, speeding up cataloguing of new species before they become extinct.

- Farmers could identify pest species invading their fields, and port inspectors could intercept shipments harboring harmful species at borders.
- Doctors could rapidly diagnose fungal pathogens and parasites, such as the one that causes malaria.
- Museums could analyze the large backlogs of collected specimens, helping them find undescribed species lurking in museum drawers.

- Regulatory agencies could test animal feed for forbidden items likely to spread illnesses such as mad cow disease.

TOMMY MOORMAN (*leaf*); NOAH PORITZ *Photo Researchers, Inc.* (*mosquito*); SCOTT CAMAZINE *Photo Researchers, Inc.* (*bugs*); TOM HOPKINS *Aurora Photos* (*fish fillets*)

[CASE STUDY]

PARSING BUTTERFLIES

Caterpillars (*photographs, below left*) of the skipper butterfly (*Astraptes fulgerator*) in Costa Rica differ in appearance, habitat and favored foods, but the adults all look very similar (*below right*), and scientists had long thought they belonged to a single species. Barcoding tells a different story, however. Because variation in the CO1 gene correlates with appearance, lifestyle and chosen foods of the caterpillars, researchers determined that, despite the outward appearance of the adults, the butterflies actually divide into 10 separate species.

CATHERINE HART (*Stoeckle*); B. ST. JACQUES (*Hebert*); DANIEL JANZEN *University of Pennsylvania* (*caterpillars and butterfly*)

[THE AUTHORS]

Mark Y. Stoeckle (*left*) **is an adjunct faculty member in the Program for the Human Environment at the Rockefeller University. A graduate of Harvard Medical School, he is also clinical associate professor of medicine at Weill Medical College of Cornell University. He is an accomplished nature photographer. Paul D. N. Hebert** (*right*)**, best known for founding the concept of DNA barcoding, completed a Ph.D. in genetics at the University of Cambridge and currently holds a Canada Research Chair at the University of Guelph, where he also directs the Biodiversity Institute of Ontario. In his free time, he enjoys chasing small life in exotic places—gathering moths in Australia is his current favorite.**

establishment of a public database called the Barcode of Life Data systems, or BOLD (online at www.barcodinglife.org). BOLD now has over 460,000 records from more than 46,000 species spanning the animal kingdom, with particularly dense records for birds, fishes, butterflies and moths. Each of these records contains the species name, barcode sequence, collection location, links to the voucher specimen, photographs and other biological data. To help coordinate the enormous effort involved in the assembly of such a comprehensive library, the Consortium for the Barcode of Life (CBOL) was established in 2005; it includes 150 institutions from 45 countries that support the development of DNA barcoding as a global standard for the identification of species. The actual assembly of records will be driven by the International Barcode of Life Project: a 25-nation alliance that plans to process five million specimens from 500,000 species by 2014.

What We Have Learned So Far

As E. O. Wilson points out, despite 250 years of effort we do not know, even to the nearest order of magnitude, how many species live on earth. DNA barcoding is already helping to speed cataloguing of biodiversity. One of the major findings so far is that there are more species—each more narrowly specialized—than scientists had realized. This revelation has come about through new information that barcoding has provided on so-called cryptic species, organisms that look alike but show genetic differences indicating they are separate species.

DNA barcode surveys have revealed cryptic species lurking in museum drawers in every group studied so far. For example, Hebert, together with Daniel Janzen, a biodiversity ecologist at the University of Pennsylvania, and John Burns, a taxonomist at the Smithsonian Institution, and their colleagues in Costa Rica, found that what was thought to be one species of skipper butterfly, *Astraptes fulgerator,* was actually at least 10 different species [*see box above*]. Because the adults are extremely similar, scientists did not realize they were so different genetically. Similarly, Alex Smith of the Biodiversity Institute of Ontario and his colleagues discovered that three morphologically recognizable species of flies that parasitize diverse insects are in fact an assemblage of 15 species, with each lineage specializing on a few hosts. Work by one of us (Stoeckle) showed that even in a very intensively studied group, North American birds, about 4 percent of named species contain genet-

ically distinct lineages that are likely to be separate species.

One of the most striking early findings is the surprisingly low level of mitochondrial genetic diversity within most animal species. This discovery confounds a prediction from population genetics theory that older or larger populations should show more diversity. Low levels of variation are often thought to indicate recent population bottlenecks. For example, scientists thought the relative absence of mitochondrial variation in human populations indicated a near-extinction of early humans in eastern Africa 150,000 years ago. According to this hypothesis, all modern humans trace their origin to a single female from this time, the so-called mitochondrial Eve. The discovery that similarly impoverished levels of genetic diversity are the rule across the animal kingdom raises doubts about the Eve hypothesis and presents a larger unsolved scientific question: What forces limit mitochondrial diversity within species? We and others believe that the consistently low levels of sequence divergence reflect frequent "selective sweeps," in which new, advantageous mutations displace ancestral variation, pruning diversity within species.

Our research so far has demonstrated that barcoding can speed up the survey of biodiversity. The fact remains, however, that formal descriptions of new species can take years to complete. The generation of sequence data is thus running far ahead of official species descriptions. We view barcoding as creating a map of DNA diversity that will serve as a framework for subsequent detailed study. Just as the speed and economy of aerial photography caused it to supplant ground surveys as the first line of land analysis, DNA barcoding can be a rapid, relatively inexpensive first step in species discovery. The "ground truthing" will take more time. But linking these approaches will produce an integrated view of the history and present-day existence of life on earth and help to shepherd life's full magnificence into the coming centuries. ■

[CONTROVERSY]

The Authors Answer Some Common Concerns

A big science initiative such as barcoding will make funding even more scarce for already underfunded disciplines such as taxonomy.

There is no evidence that barcoding is draining away support. On the contrary, the sources of funding for barcoding, including private foundations (Alfred P. Sloan Foundation, Gordon and Betty Moore Foundation) and government agencies (Genome Canada), are new sources of funding for taxonomy.

DNA barcodes frequently cannot distinguish between closely related or recently diverged species.

Very young species that cannot be distinguished by DNA barcoding make up a small fraction of species and are often difficult to separate by traditional methods as well. Some of these cases represent single species incorrectly split or species in the process of formation, and DNA barcoding can help flag such cases for taxonomic review.

The mitochondrial gene used as a barcode does not differentiate accurately between all animal species and does not work at all for some taxonomic groups.

Taxonomic groups demonstrating effectiveness of barcodes include bats, bees, chitons, clams, copepods, fish, frogs, fruit flies, mayflies, nematodes, spiders, sponges and springtails. Results so far show there are very few animal groups not well distinguished by DNA barcodes. In fact, many animal species cannot be distinguished by traditional methods or require expensive equipment or advanced training, and yet they are readily identified by DNA barcoding.

Barcoding is not really new; it is just a marketing device.

The underlying concept goes back 30 years to Carl Woese of the University of Illinois, who first showed that DNA sequences could be used to reconstruct the Tree of Life. But the idea of establishing an identification system for all plant and animal life using genetic sequences from a uniform locus was first proposed in 2003, and the DNA barcode reference libraries have begun to accumulate only in the past three to four years. What is also new, and what makes the system work, is attaching a uniform set of data to each barcode record.

MORE TO EXPLORE

The paper that launched a thousand barcoders: **Biological Identifications through DNA Barcodes.** Paul D. N. Hebert, Alina Cywinska, Shelley L. Ball and Jeremy R. deWaard in *Proceedings of the Royal Society B*, Vol. 270, No. 1512, pages 313–321; February 7, 2003. Available at **http://journals.royalsociety.org**

The Barcode of Life Data systems is a workbench for researchers, with public links to published projects, an "identification engine," a taxonomy browser, Google maps, and more: **www.barcodinglife.org**

Consortium for the Barcode of Life (CBOL), an international initiative devoted to developing DNA barcodes as a global standard for the identification of biological species, is based at the National Museum of Natural History: **www.barcoding.si.edu**

Mark Y. Stoeckle's DNA Barcode Blog is a weekly illustrated scientific blog about short DNA sequences for species identification and discovery: **http://phe.rockefeller.edu/barcode/blog**