A botanical macroscope

Jesse H. Ausubel[1]

The Rockefeller University and the Alfred P. Sloan Foundation, 1230 York Avenue, New York, NY 10065

In 2003 zoologists proposed a short, standardized DNA sequence from a uniform locality on the genome to serve as a DNA barcode to identify the species of tissue from almost all animals—reliably, quickly, and economically. In 2008 two New York City high school students, Kate Stoeckle and Louisa Strauss (1), used the accumulating barcode reference library to prove that about ¼ of fish items they purchased in Manhattan were inaccurately labeled. Their "Sushigate" investigation earned front page coverage in the *New York Times* and attention from top managers at major food retailers and caterers.

Equal curiosity would surely attach to revelations of the contents of jars of plant powders that herbalists sell. Or to seedlings in a garden, a regrowing forest, or an exotic jungle. "A DNA barcode for land plants," by the Plant Working Group of the Consortium for the Barcode of Life (CBOL) in this issue of PNAS (2) hastens the prospect that many more people can speedily obtain good identifications.

To extend barcoding to plants, 52 authors from 24 institutions in 9 nations, led by Peter Hollingsworth of the Royal Botanical Garden in Edinburgh, propose a pair of short sequences totaling about 1,450 base pairs from *rbcL* and *matK* as the foundation for a DNA barcode library for plants. Using *rbcL* + *matK* in the sample of 907 specimens examined, the Plant Working Group discriminated species in 72% of cases, with the remaining species being matched to groups of congeneric species with complete success. The U.S. Food and Drug Administration and many other organizations around the world concerned with commerce in plants should take notice. Barcoding will ease their job to protect consumers, and justify a large societal investment in building a global reference library for plants as well as animals and fungi. Meanwhile, entrepreneurs are likely to make handy tools available not only for sushi eaters but for the millions of people who worry about food allergies and the tight group of investigators who worry about trade in endangered woods. DNA forensics for nature are a practical boon.

The ability of a consortium to field a team promptly for an orderly examination of a new technique speaks volumes. Bravo. Instead of haphazard scouting, the Plant Working Group, building on a prior search cultivated by Robyn Cowan of the Royal Botanical Garden (Kew), marshaled knowledge of a wide geographic range without delay. They applied the crucial criteria of

- Universality: which loci can be routinely sequenced across the land plants?
- Sequence quality and coverage: which loci are most amenable to the production of bidirectional sequences with few or no ambiguous base calls?
- Discrimination: which loci enable most species to be distinguished?

The team pooled data across laboratories, including sequence data from 907 samples representing 445 angiosperm, 38 gymnosperm, and 67 cryptogam species, no small accomplishment.

The barcode of life provides another master key to knowledge about a species.

In much biological discovery, monitoring, and research, identifying species remains the front line. Since Linnaeus, biologists have gathered distinguishing features in taxonomic keys to assign binomial species names, such as *Homo sapiens*, to mystery specimens. Then, as a master key opens all of the rooms in a building, the binomial name opens all knowledge about a species. From mollusks to wasps to fishes, evidence now shows that short DNA sequences from a uniform locality on genomes can be another distinguishing feature. As a Linnaean binomial is an abbreviated label for the morphology and other characteristics, the short DNA sequence is an abbreviated label for the genome of the species. The barcode of life provides another master key to knowledge about a species. Compiling a public library of sequences linked to named specimens, plus faster and cheaper sequencing and better information technology, will make this new barcode key increasingly practical and powerful.

Panoramic Vision

DNA barcoding is also a boon for exploration. For more than three centuries biology has multiplied its power and reach with microscopes to see the small and fine. Some phenomena, however, are too large to see and grasp, and global explorers need *macroscopes*. Increasingly, this need challenges ecology, which flourished first in bell jars and at the scale of a meadow, but must now urgently probe patterns on a broad, even planetary, scale. These patterns may encompass, for example, the changing flora in biodiversity hotspots such as the Kruger region of South Africa (3), Western Ghats of India (4), or the entire Arctic. Such surveys require reliable identification of manifold specimens. DNA barcoding enables parataxonomists and others to identify samples that previously only very scarce experts could, and frees the keenest experts to work on exceptional difficulties.

For evolution, DNA barcoding opens new views of molecular diversification. Accumulating large libraries of aligned sequences opens formerly ungraspable territory. Already the Barcode of Life Database (www.barcodinglife.org) has collected cytochrome *c* oxidase I (*COI*) sequences from over 620,000 specimens from over 58,000 animal species. Goloboff et al. (5) show that comparable data sets for other genes, accumulated piecemeal over 25 years, are all much smaller, with the small subunit (SSU) ribosomal RNA gene topping the chart at 20,462 taxa and cytochrome *b* next at 13,766. For plants, the largest dataset of interest is *rbcL* with records for 13,533 taxa (6). With about 380,000 named plant species (7), the time has come for the standardization that speeds the building of libraries.

The CBOL Plant Working Group emphasized relative discriminatory power of different loci rather than overall discriminatory power with regard to the 380,000 or so potential subjects. Still, the good results of the group's strategy for angiosperms (flowering plants), which make up about 90% of named

Author contributions: J.H.A. wrote the paper.

Conflict of interest statement: Since 1994 I have served as a program manager for the Alfred P. Sloan Foundation, which provides funding to the Consortium for the Barcode of Life. Neither I nor my Rockefeller University laboratory participated in the concourse of the Plant Working Group.

See companion article on page 12794.

[1]To whom correspondence should be addressed at: Program for the Human Environment, The Rockefeller University, 1230 York Avenue, New York, NY 10065. E-mail: ausubel@rockefeller.edu.

land plants, imply a quick road to a useful library of sequences for *rbcL* and *matK* of 300,000 species or more. Laboratories should prepare to multiply the *rbcL* library by more than 20 times and *matK* even more steeply. Adding a macroscope that sharpens eyes 20-fold surely empowers watchers of the botanical heavens. One such heaven is beneath our feet. Barcoding may permit unprecedented appreciation of the 60% of plant biomass that lies underground and resists traditional means of identification.

A radically different and equally fascinating vista to survey is the changing base composition in DNA sequences themselves. To what extent are signals found in various codons and locations in various taxa? What biases may exist in mutation? Using the sequences as what mathematicians call indicator vectors (8), what might we discover with computational ease about the relatedness of barcoded plants at various taxonomic levels?

Horizontal Genomics

Building sequence libraries for multitudes of species and specimens is part of a larger revolution. Genomics began as a vertical endeavor to know the entire genome of a single species, such as the 3 billion base pairs of *Homo sapiens*. Now we are entering a companion era of horizontal genomics in which discovery comes from panoramic views of short sequences of many specimens and species. For environmental science, horizontal genomics allies with the emerging "e-Biosphere," embracing such endeavors as the Catalog of Life (www.catalogueoflife.org/search.php, which lists valid species names), Global Biodiversity Information Facility (www.gbif.org/, which among other things catalogs where specimens are archived), and Ocean Biogeographic Information System (www.iobis.org/, which already provides geospatial information for about 20 million observations of more than 140,000 species). The young integrator of all is the Encyclopedia of Life (EOL; www.eol.org/) opening a portal to images, text descriptions, and of course DNA sequences of every species. A happy sequel of the Plant Working Group strategy could be visitors to an EOL page linking barcode sequences seamlessly to, for example, the habitat, appearance, and behavior of the plant. Botanists anticipated the e-Biosphere in their heroic and beautiful volumes of the complete flora of nations and regions, linked to binomial names.

A person not specialized in molecular biology will want to know why plants differ so much in barcode-ability from animals (9), where *COI* succeeds in delivering a species identification in 95% or more cases. In plants, the mitochondrial *COI* sequence favored for animals suffers from low substitution rates, forcing a redirection of the barcode search to plastids, where sequences evolve relatively rapidly, and to a pair of loci that identify almost ¾ of the stunning diversity of the plant kingdom. Indeed, the strategy of two sequences may become hierarchical if a 3rd and even a 4th segment are needed to gain species identification in the recalcitrant cases. One option preferred by some researchers in the Plant Working Group was a 3-locus barcode of *matK* + *rbcL* + *trnH-psbA*. The 3rd region, a noncoding transgenic spacer, suffered from practical hardships in reading its base pairs. Also, a macroscope seems likely to find more significant insights scanning regions that code for proteins than those that do not.

But then keys for taxonomic identification have always been a sequence of questions, starting with queries about color or size or shape. While reaching a final answer after one or two questions is brilliant, reaching it in three or four is still clever. And developing more molecular discernment for the few thousand plant species among 380,000 that require it will be manageable and fascinating for researchers questing new knowledge and grants.

For many, DNA barcodes evoke grocery shelves, but among molecular biologists, the visual analogy to old black-and-white electrophoretic gels spurred interest. The Plant Working Group brings us full circle to jars of mysterious powders on shelves and websites of purveyors of health and nutrition. The crux is whether 72% is a good enough starting point for standardization of botanical barcodes. Recalling technology history, whether automobiles or typewriters or vaccines, I say yes. Cars became easier to drive as engineers added electric starters, windshield wipers, and now global positioning. In the few years of fish barcoding, initially hampered by lack of primers spanning a broad range of species, investigators not only discerned phony red snapper on your plate but also unified the morphologically diverse larval, male, and female versions of deep sea fish (10) and split look-alike species in the Indian Ocean (11). Already, nifty software is freely available to explore and display results. For example, web-based tools, available at www.ibarcode.org allow a user to manage barcode datasets, cull out nonunique sequences, identify haplotypes within a species, and examine the within- to between-species divergences. Barcode technology will relentlessly improve.

While the struggle of rash focus versus disabling diffusion will never end, humanity will enjoy the speed and economy of barcoding to learn quickly for what the present plant accuracy suffices. And, learning by doing, users will speed the technology forward. Let us accept the invitation of the 52 authors led by Hollingsworth to use the standard two-locus barcode of *matK* and *rbcL* to join in building a powerful botanical macroscope.

1. Stoeckle K, Strauss L (2008) High school students track down fish fraud. *Pacific Fishing* 29(9):34.
2. CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proc Natl Acad Sci USA* 106:12794–12797.
3. Lahaye R, et al. (2008) DNA barcoding the floras of biodiversity hotspots. *Proc Natl Acad Sci USA* 105:2923–2928.
4. Ragupathy S, Newmaster SG, Murugesan M, Balasubramaniam V (2009) DNA barcoding discriminates a new cryptic grass species revealed in an ethnobotany study by the hill tribes of the Western Ghats in southern India. *Mol Ecol Res* 9S1:164–171.
5. Goloboff PA, et al. (2009) Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. *Cladistics* 25:211–230.
6. Smith SA, Beaulieu JM, Donoghue MJ (2009) Mega-phylogeny approach for comparative biology: An alternative to supertree and supermatrix approaches. *BMC Evol Biol* 9:37.
7. Paton AJ, et al. (2008) Towards target 1 of the Global Strategy for Plant Conservation: A working list of all known plant species—progress and prospects. *Taxon* 57:602–611.
8. Sirovich L, Uglesich R (2004) The organization of orientation and spatial frequency in primary visual cortex. *Proc Natl Acad Sci USA* 101:16941–16946.
9. Fazekas AJ, et al. (2009) Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Mol Ecol Res* 9S1:130–139.
10. Johnson GD, et al. (2009) Deep-sea mystery solved: Astonishing larval transformations and extreme sexual dimorphism unite three fish families. *Biol Lett* 5:235–239.
11. Zemlak TS, Ward RD, Connell AD, Holmes BH, Hebert PDN (2009) DNA barcoding reveals overlooked marine fishes. *Mol Ecol Res* 9S1:237–242.