Species Barcoding

For as long as scientists have measured DNA, it has been used to identify species, museum specimens, and other types of biological tissues. Recently, the use of shorter, standardized DNA sequences has increased the efficacy of this approach. But standardized genetic markers, such as DNA barcodes, depend on the existence of a library of reference sequences in order to be meaningful.

The number of named species—from plants and animals to fungi and protists—has been estimated at more than 1.7 million worldwide (IUCN 2011). The actual number of species, however, is likely to range from 3 million to 100 million (Gaston and Simon 2007).

Our ability to rapidly and accurately identify, measure, and monitor aspects of species population and diversity faces four challenges:

1. The actual number of species is not known, to within an order of magnitude (Gaston and Levin 2007).
2. Rates of species extinctions are higher than they have been at any time outside of the previous five mass extinction events (Barnosky et al. 2011).
3. Human-mediated species translocations will occur more frequently in a global economy (Lockwood, Cassey, and Blackburn 2005).
4. Taxonomists (the biologists who describe species diversity) are not being trained and replaced as quickly as they are retiring, which has resulted in a global phenomenon called the "taxonomic impediment." This term has been applied both to the shortage of taxonomic knowledge (as pertains to the undescribed or unrecognized portion of global diversity in nature) and also to the shortage of taxonomists and funding for taxonomy (House of Lords 2008).

Meeting these four challenges requires the integration of traditional and modern means of species identification. Using DNA as a tool for species and specimen identification is an important part of that modern tool kit.

DNA Barcode

A DNA barcode is a short, standardized DNA sequence used for the identification of specimens and species, as well as a tool for the discovery of previously unappreciated provisional species that are often morphologically cryptic (Floyd et al. 2002; Hebert et al. 2003). Where traditional sources of taxonomic knowledge exist, the DNA barcode can be used in concert with them (although it is not a replacement for such systems) and also as a transparent first-pass survey of a system or taxa where there is a paucity of other knowledge sources (Smith et al. 2009).

DNA barcode information, as characters or as distances, can help identify known species, perhaps from trace amounts of tissue, or a taxonomically nonlabile life history stage; also, it can be part of a suite of characters used for the discovery and description of new species, and the flagging of otherwise cryptic diversity.

Barcode Gap

Critical to successfully establishing a standardized marker for species identification was the selection of a gene or gene region that displayed low within-species (intraspecific) variation and larger between-species (interspecific) variation. Some have called the difference between these two types of variation the "barcoding gap." Calculating intraspecific variation requires the collection and analysis of multiple specimens from a single

species across the range of the species. Failure to account for this potential geographic structuring of intraspecific variation (phylogeography) can lead to misidentification of an artificially large barcode gap.

Animal DNA Barcode

The standardized barcode region for animals is part of the cytochrome *c* oxidase 1 (CO1) mitochondrial gene. The efficacy of this gene was initially tested using diverse arthropod groups (Hebert et al. 2003). The CO1 barcode region fit the requirements of containing a region of sufficient variability (to discriminate even recently evolved species) that was flanked by relatively conserved regions (to enable the design of polymerase chain reaction primers that can be used across a wide taxonomic breadth). Selecting CO1 from among the other thirteen protein-coding mitochondrial genes was at least partially pragmatic, since many of the other markers would work; however, at the time (2002–2003) there were more arthropod CO1 sequences in GenBank than any of the other thirteen protein-coding mitochondrial genes. (See Library Creation and Analysis, below, for more about GenBank.)

Early thinking suggested that this arthropod CO1 stockpile would accelerate the completion of the library in one of the most diverse groups of metazoans. It was learned very quickly that the agreed-upon data standards by the barcoding community were high enough to prevent the inclusion of most of this early parcel of data.

The use of mitochondrial DNA (mtDNA) as a DNA barcode has led to some criticisms based on the mode of inheritance. Mitochondrial DNA is maternally inherited, and therefore if the species hybridizes, or has experienced introgression or incomplete lineage sorting, a mtDNA barcode would produce an erroneous answer—in the former case, the maternal species, and in the latter, an alternate or ancestral species.

Uses

The animal DNA barcode has several practical applications:

- It prevents consumer market fraud that occurs when a product is labeled incorrectly and the consumer is charged a higher price. If the product is biological, DNA barcodes can be used to monitor and test identifications; as when, for instance, the labeled ID of more than 30 percent of fish samples was contradicted by DNA barcoding, which raised both economic and health concerns (Lowenstein, Amato, and Kolokotronis 2009).
- It monitors the commercial trade of endangered species, particularly when the products are processed (Eaton et al. 2010).
- It ensures that we understand who eats whom in a pragmatic and repeatable fashion, which allows moving toward easily identified food web units (Smith et al. 2011).

Plant DNA Barcode

For plants, the search for a standardized region took longer to complete. The mtDNA marker selected for animals, CO1, does not possess sufficient nucleotide variation to identify most species. Botanists therefore searched for another marker, or small number of markers, that would permit a similarly precise and accurate standardized system of species and specimen identification. The botanical community has converged on a two-gene solution based on recoverability, sequence quality, and capacity for species discrimination (Hollingsworth et al. 2009). Like the animal marker, the two-loci combination of *rbcL+matK* is non-nuclear—and so is exposed to the same theoretical concerns regarding inheritance, in that a species and a single gene have not necessarily experienced the same types of selection and do not necessarily reveal the same answer.

Library Creation and Analysis

To facilitate the creation, curation, and analysis of various DNA barcoding loci, the barcoding community uses both the Barcode of Life Datasystem (BOLD) (Ratnasingham and Hebert 2007) and the more general, public genetic repository GenBank. A DNA- barcode-specific database, BOLD is an online workbench for DNA barcoding that is designed to aid in the collection, management, analysis, and use of DNA barcodes and associated metadata (e.g., photographs, GPS). As of 2011, BOLD contained more than 1.3 million specimens representing nearly 110,000 species. As total global diversity likely exceeds this total by one or two orders of magnitude, the creation of this critical reference library is not at the beginning of the end, but it is, perhaps, the end of the beginning.

GenBank created a restricted keyword—Barcode—to be used when a sequence meets the minimum data standards associated with a DNA barcoding initiative, such as the International Barcode of Life project (iBOL 2011). These standards, established by the Consortium for the Barcode of Life (CBOL 2011) and the National Center for Biotechnology Information (NCBI 2011), require that the sequence be from a community-agreed-upon gene locus (e.g., CO1 in eukaryotic animals or *rbcL*

and *matK* in plants), that they trace files and specimen collection locality, and that they be of a minimum length (500 base pairs) and quality (fewer than 2 percent ambiguous bases).

Criticisms

Since the publication of the first DNA barcoding papers in the early twenty-first century, there have been criticisms of the efficacy of such an approach. Researchers using DNA barcodes were warned not to use the standard sequence typologically but rather as part of an integrative taxonomy. Concerns were also raised regarding the non-nuclear basis of both the animal and the plant barcodes—and how this would particularly bias species discovery by incorrectly labeling species that possess deep intraspecific variation as de novo species. In addition, worries about the effects of hybridization and introgression on these mitochondrial and plastid barcoding regions were raised. Hybridizing species and those possessing introgressed DNA are a challenge to nuclear or mitochondrial DNA-based identification systems—and indeed an integrative approach is necessary in order to capture the philosophical "fuzziness" (i.e., a non-Aristotelian definition) of what a "species" is. Others have asked why the apparently reductionist nature of the DNA barcoding endeavor (using one single gene or a small number of genes) would be pursued in an era when we expect the cost of sequencing entire genomes to continue to decline dramatically.

Future

Even though the cost of DNA sequencing will fall, it is likely that using a single, or a small number of, barcoding sequence(s) to identify specimens—and subsequently selecting from which of these to sequence entire genomes, when and if that is desirable—will stay the standard practice as sequencing costs for small fragments will remain comparatively more affordable. The key original elements remain critical: standardized selection of loci and the existence of a reference library of identified specimens.

M. Alexander SMITH
University of Guelph

See also Biological Indicators (*several articles*); Ecosystem Health Indicators; Fisheries Indicators, Freshwater; Fisheries Indicators, Marine; Global Strategy for Plant Conservation; Index of Biological Integrity (IBI); Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES); Land-Use and Land-Cover Change

Further Reading

Barnosky, Anthony D., et al. (2011). Has the Earth's sixth mass extinction already arrived? *Nature, 471*(7336), 51–57.

Consortium for the Barcode of Life (CBOL). (2011). What is CBOL? Retrieved August 28, 2011, from http://www.barcodeoflife.org/content/about/what-cbol

Eaton, Mitchell J., et al. (2010). Barcoding bushmeat: Molecular identification of central African and South American harvested vertebrates. *Conservation Genetics, 11*(4), 1389–1404.

Floyd, Robin; Abebe, Eyualem; Papert, Artemis; & Blaxter, Mark. (2002). Molecular barcodes for soil nematode identification. *Molecular Ecology, 11*(4), 839–850.

Gaston, Kevin J., & Levin, Simon Asher. (2007). Global species richness. In Simon Asher Levin (Ed.), *Encyclopedia of biodiversity* (pp. 1–7). New York: Elsevier.

Hebert, Paul D. N.; Cywinska, Alina; Ball, Shelley L.; & deWaard, Jeremy R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society—Biological Sciences, 270*(1512), 313–321.

Hollingsworth, Peter M., et al. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America, 106*(31), 12794–12797.

House of Lords. (2008). Systematics and taxonomy: Follow-up (HL Paper 162). Science and Technology Committee, 5th Report of Session 2007–08. London: The Stationery Office Limited.

International Barcode of Life (iBOL). (2011). What is iBOL? Retrieved August 28, 2011, from http://ibol.org/about-us/what-is-ibol/

International Union for Conservation of Nature and Natural Resources (IUCN). (2011). The IUCN red list of threatened species (Version 2011.1). Retrieved August 4, 2011, from http://www.iucnredlist.org

Lockwood, Julie L.; Cassey, Phillip; & Blackburn, Tim. (2005). The role of propagule pressure in explaining species invasions. *Trends in Ecology & Evolution, 20*(5), 223–228.

Lowenstein, Jacob H.; Amato, George; & Kolokotronis, Sergios-Orestis. (2009). The real *maccoyii:* Identifying tuna sushi with DNA barcodes—contrasting characteristic attributes and genetic distances. *PLoS ONE, 4*(11), Article e7866. Retrieved August 4, 2011, from http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0007866

National Center for Biotechnology Information (NCBI). (2011). About NCBI. Retrieved August 28, 2011, from http://www.ncbi.nlm.nih.gov/About/index.html

Ratnasingham, Sujeevan, & Hebert, Paul D. N. (2007). BOLD: The barcode of life data system (www.barcodinglife.org). *Molecular Ecology Notes, 7*(3), 355–364.

Smith, M. Alex, et al. (2011). Barcoding a quantified food web: Crypsis, concepts, ecology and hypotheses. *PLoS ONE, 6*(7), Article e14424. Retrieved August 4, 2011, from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0014424

Smith, M. Alex; Fernandez-Triana, Jose; Roughley, Rob; & Hebert, Paul D. N. (2009). DNA barcode accumulation curves for understudied taxa and areas. *Molecular Ecology Resources, 9*(s1), 208–216.