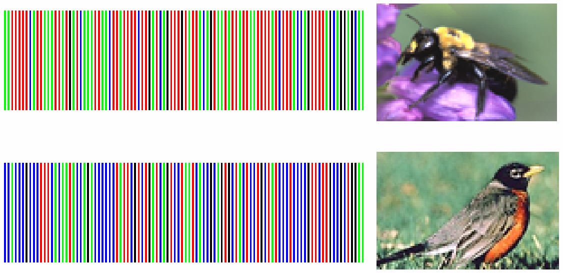
BARCODING LIFE, ILLUSTRATED

Goals, Rationale, Results

Mark Stoeckle The Rockefeller University
Paul E. Waggoner Connecticut Agricultural Experiment Station
Jesse H. Ausubel Alfred P. Sloan Foundation

v1.3 January 26, 2005

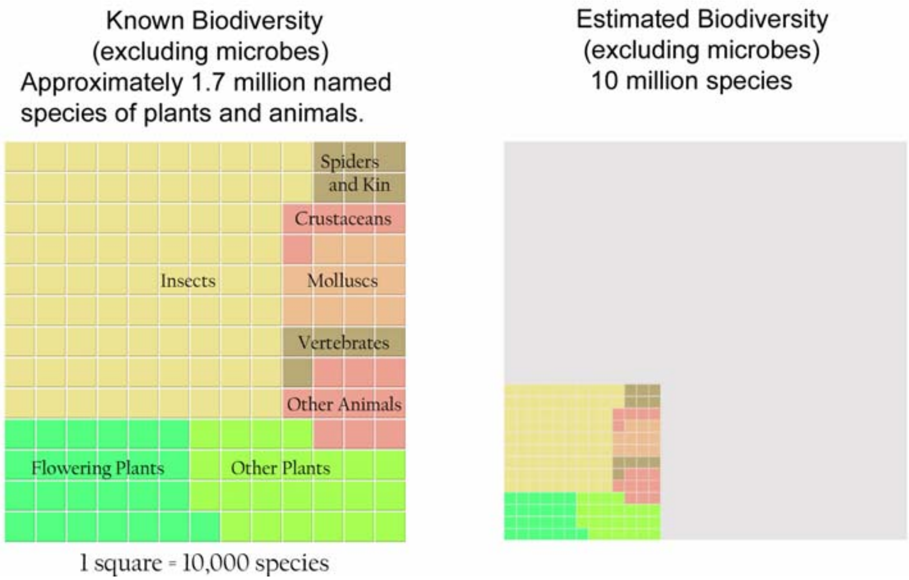


BARCODE OF LIFE: A short DNA sequence, from a uniform locality on the genome, used for identifying species.

Barcoding is a standardized approach to identifying animals and plants by minimal sequences of DNA.

1. Why barcode animal and plant species?

By harnessing advances in electronics and genetics, barcoding will help many people quickly and cheaply recognize known species and retrieve information about them, and will speed discovery of the millions of species yet to be named. Barcoding will provide vital new tools for appreciating and managing Earth's immense and changing biodiversity.

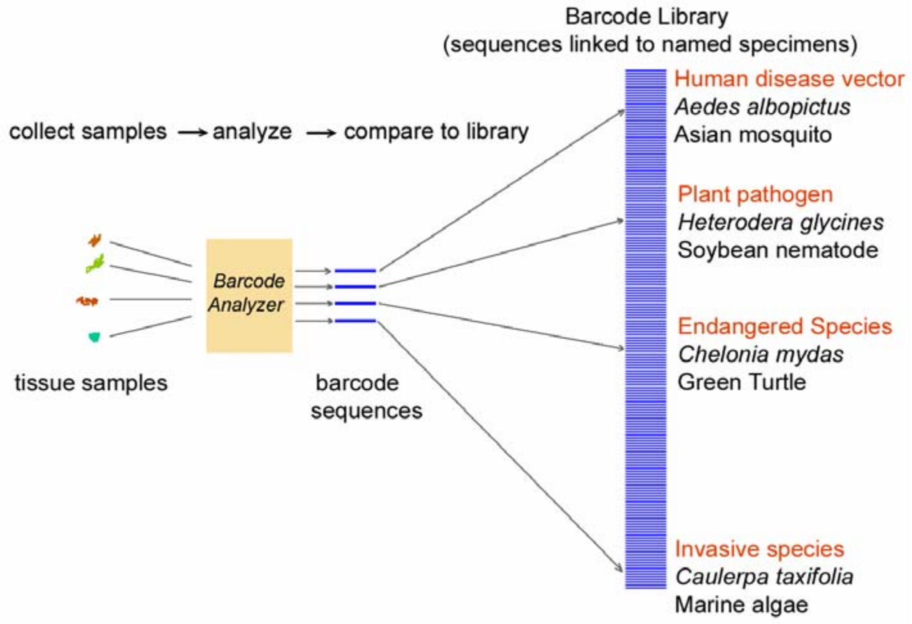


2. What are the benefits of standardization?

Researchers have developed numerous ways to identify species by DNA, typically tailoring the approach to answer a specific question in a limited set of species. Like convergence on one or a few railroad gauges, barcoding aims to capture the benefits of standardization. Standardization typically lowers costs and lifts reliability, and thus speeds diffusion and use.

For barcoding, standardization should help accelerate construction of a comprehensive, consistent reference library of DNA sequences and development of economical technologies for species identification. The goal is that anyone, anywhere, anytime be able to identify quickly and accurately the species of a specimen whatever its condition.

Results so far suggest that a mitochondrial gene barcode will enable identification of most animal species. For plants, mitochondrial genes do not differ sufficiently to distinguish among closely related species. Promising approaches to standardize plant identification using one or possibly two barcode regions are under development.



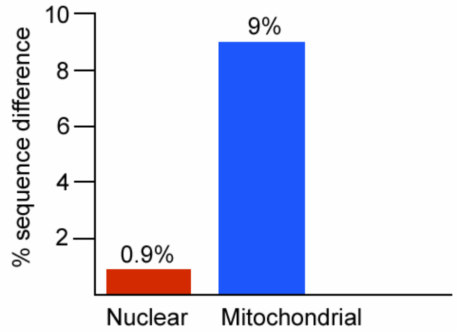
3. Why barcode animals with mitochondrial DNA?

Mitochondria, energy-producing organelles in plant and animal cells, have their own genome. Twenty years of research have established the utility of mitochondrial DNA sequences in differentiating among closely related animal species. Four properties make mitochondrial genomes especially suitable for identifying species.

Copy number. While each cell typically contains only 2 copies of nuclear DNA sequences, the same cell encompasses 100-10,000 mitochondrial genomes. Recovering mitochondrial DNA sequences succeeds much more often than nuclear sequences, especially from small or partially degraded samples. Greater success with smaller samples means lower processing costs.

Greater differences among species. Sequence differences among closely related animal species average 5- to 10-fold higher in mitochondrial than nuclear genes. Thus, shorter segments of mitochondrial DNA distinguish among species, and because they are shorter, less expensively.

Average sequence differences in nuclear and mitochondrial DNA between human and chimpanzee



Few differences within species. Intraspecific variation in mitochondrial DNA is low in most animal species. This may reflect rapid loss of ancestral polymorphisms due to maternal inheritance or selective sweeps following emergence of advantageous mutations. Regardless of cause, small intraspecific and large interspecific differences signal distinct genetic boundaries between most species, enabling precise identification with a mitochondrial barcode.

Absence of introns. In animals, mitochondrial genes rarely contain introns, which are non-coding sequences interspersed between the coding regions of a gene. Thus, amplification of mitochondrial DNA is usually straightforward. In contrast, amplification of coding regions of nuclear genes is often limited by introns, which may be long.

4. What are the main limits to barcoding encountered so far?

Groups with little sequence diversity. An example was found among a small number of corals and anemones from the marine phylum Cnidaria. The prevalence of such groups is not yet known, as researchers have analyzed only a few Cnidaria, and mitochondrial DNA sequences do distinguish some closely related species from this group. A comparison of mitochondrial sequences from 2238 species in 11 animal phyla showed 98% of closely related species pairs had more than 2% sequence difference, which is enough for successful identification of most species.

Resolution of recently diverged species. Collections of closely related organisms that have recently passed the threshold to win the status of species challenge separation by any method, including morphology. In some cases, a mitochondrial barcode may narrow identification to two (or more) closely related species and no further. The frequency of species with shared barcodes is low in groups studied so far.

Hybrids. Identification systems based on a single gene (nuclear or mitochondrial) will not allow the certain identification of hybrids, that is, individuals whose male and female parent are from different species. Such specimens may be misidentified morphologically as well.

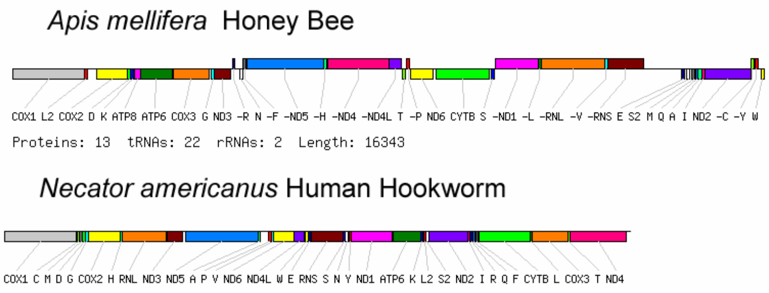
Nuclear pseudogenes. Pseudogenes, which are inactive copies of genes usually containing multiple mutations and/or deletions, can complicate identification by either mitochondrial or nuclear genes. Pseudogenes have proven a minor limitation to using a mitochondrial barcode in groups studied so far.

.

5. Why select the barcode sequence from within one gene?

With a few exceptions, animal mitochondria contain an identical set of genes: 13 protein-coding, 2 ribosomal RNA, and 22 transfer RNA genes. While the order of the genes and their polarity (location on plus or minus strand) differ markedly among animal phyla, sequences from diverse organisms can be easily compared as long as the barcode locality is limited to one gene. Staying within the boundaries of a single gene also eases development of broad range techniques for recovery of barcode sequences from diverse organisms.

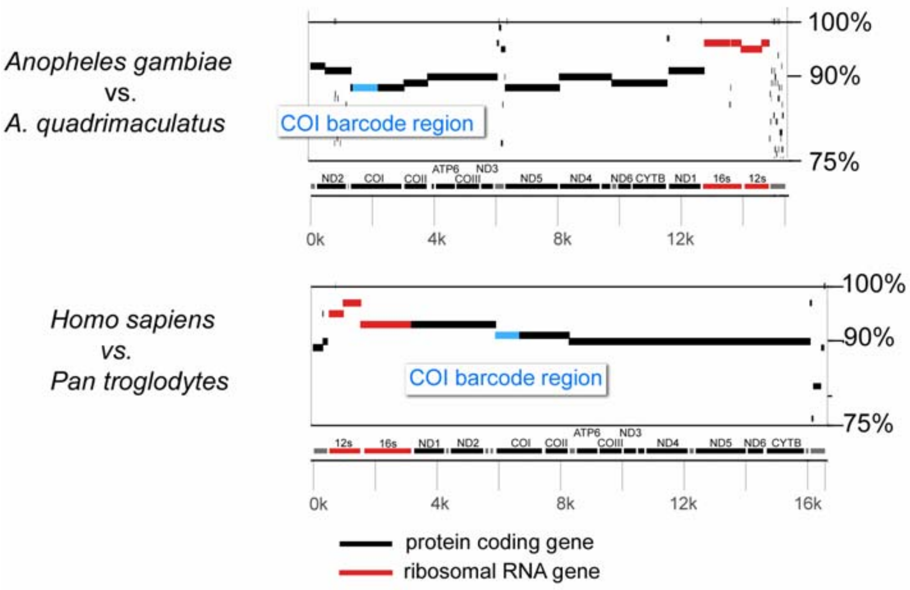
Mitochondrial genome organization differs among animals. As an example, genomes of bee and hookworm are shown. Their gene arrangements differ at 37 breakpoints. Thus, working with sequences that straddle genes poses problems.


Apis mellifera Honey Bee
COX1 L2 COX2 D K ATP8 ATP6 COX3 G ND3 -R N -F -ND5 -H -ND4 -ND4L T -P ND6 CYTB S -ND1 -L -RNL -V -RNS E S2 M Q A I ND2 -C -Y W
Proteins: 13 tRNAs: 22 rRNAs: 2 Length: 16343
Necator americanus Human Hookworm
COX1 C M D G COX2 H RNL ND3 ND5 A P V ND6 ND4L W E RNS S N Y ND1 ATP6 K L2 S2 ND2 I R Q F CYTB L COX3 T ND4

6. Why standardize on COI for animals?

The mitochondrial protein-coding genes generally contain more differences than the ribosomal genes and thus are more likely to distinguish effectively among closely related species. Sequence comparisons among protein-coding genes are easier because they generally lack insertions or deletions frequently present in ribosomal genes.

Percent identity plot (PIP) analysis of complete mitochondrial genomes. The protein-coding genes generally show more differences between species than the ribosomal genes.



Among candidate protein-coding gene regions, the cytochrome *c* oxidase I (COI) locality contains sequence differences representative of those in other mitochondrial protein-coding genes. Possible gains in accuracy or cost from using a different protein-coding domain would likely be small in light of the general similarity of these regions.

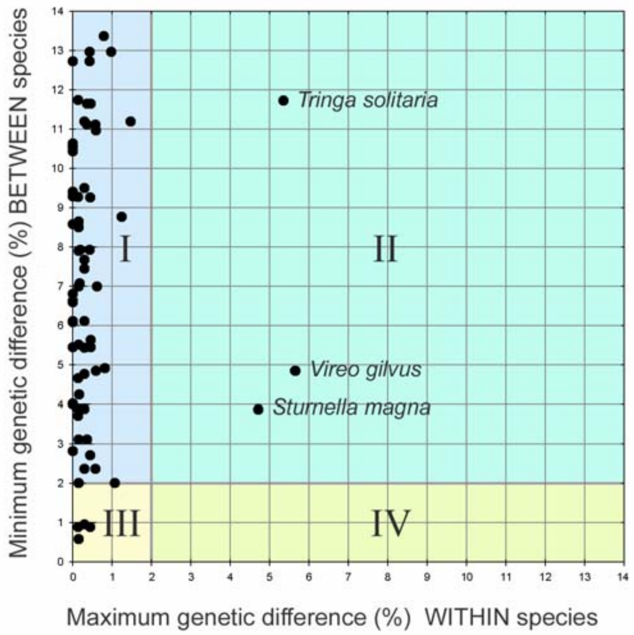
The COI region that is rapidly gaining currency represents approximately the first half of the gene and is 648 base pairs, a length easy to process in one "grab" with current technology and thus cheap. Results to date indicate that this COI barcode is:

1) easy to recover from diverse taxa, using a limited set of primers
2) readily aligned for sequence comparisons
3) effective in distinguishing among closely related animal species from a variety of invertebrate and vertebrate taxa

7. What do barcode differences among and within animal species studied so far suggest?

COI barcode sequences differ much more among than within species. For example, among 260 species of North American birds, differences between closely related species averaged 18-times higher than differences within species. Thus, a COI barcode alone should identify most bird species. Exceptions occur among some species that diverged very recently or hybridize regularly. Alternatively, low barcode differences between specimens attributed to different species may indicate synonymy, i.e., single species incorrectly split into separate taxa, or misidentified specimens. On the other hand, large barcode differences of specimens within a species may signal the presence of species mistakenly lumped together by current taxonomy.

Interspecific vs. intraspecific COI barcode differences



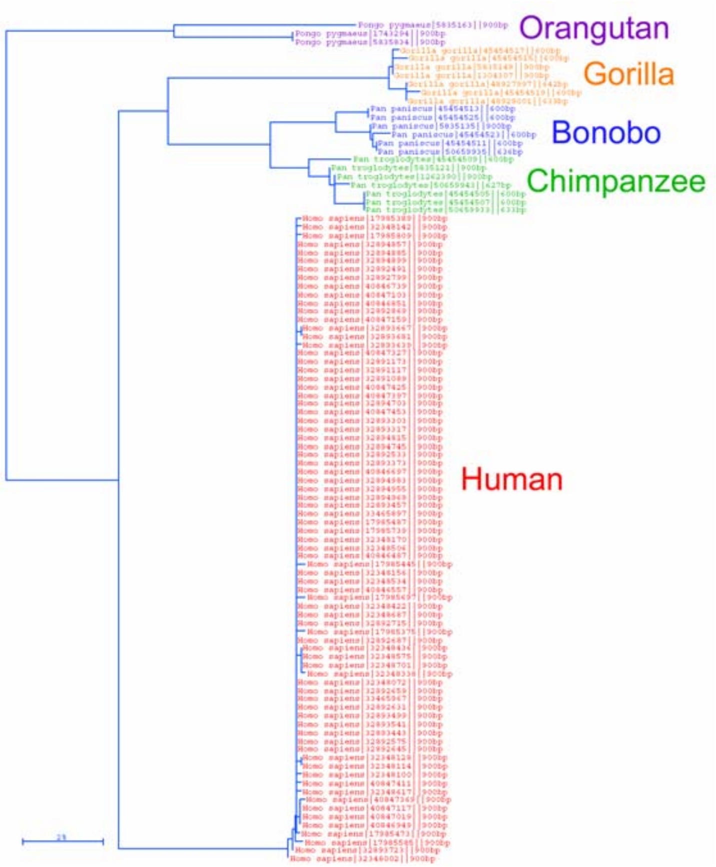
Results for 73 species of North American birds are shown. Quadrants represent different categories of species:

I. consistent with current taxonomy
II. possible lumped species (candidate for taxonomic split)
III. recent divergence, hybridization, or possible synonymy
IV. possible taxonomic misidentification

8. What about humans?

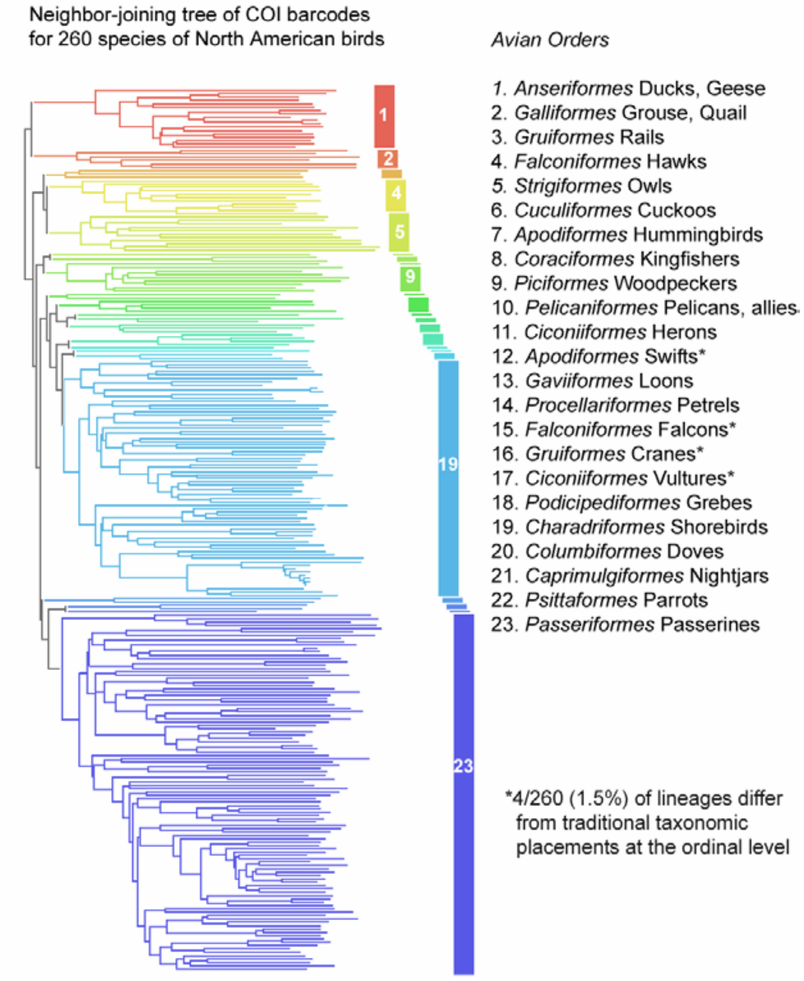
Barcodes affirm the unity of the species *Homo sapiens*. Comparison of COI barcode sequences shows we typically differ from one another by only one or two base pairs out of 648, while we differ from chimpanzees at about 60 locations and gorillas at about 70 locations. Large intraspecific differences may signal the presence of hidden species, as for example in the recent recognition of two species of orangutan.

Neighbor-joining tree of genetic distances in COI among and within 100 Hominidae.



9. Can barcodes aid understanding history of animal and plant species?

While barcoding’s goal is identification of specimens at the level of species, various rules also assemble groups of barcodes in “trees” suggesting evolutionary distances and relationships among species. For centuries biologists have worked to construct a tree of life or phylogeny showing the history of species. These efforts benefit from analysis of multiple characters, especially across long eras and varied groups. In the few cases examined so far, genetic distances among COI barcodes are largely congruent with understanding developed through traditional taxonomy, suggesting a library of barcodes will help evolutionary study.


Neighbor-joining tree of COI barcodes for 260 species of North American birds

Avian Orders
1. *Anseriformes* Ducks, Geese
2. *Galliformes* Grouse, Quail
3. *Gruiformes* Rails
4. *Falconiformes* Hawks
5. *Strigiformes* Owls
6. *Cuculiformes* Cuckoos
7. *Apodiformes* Hummingbirds
8. *Coraciformes* Kingfishers
9. *Piciformes* Woodpeckers
10. *Pelicaniformes* Pelicans, allies
11. *Ciconiiformes* Herons
12. *Apodiformes* Swifts*
13. *Gaviformes* Loons
14. *Procellariformes* Petrels
15. *Falconiformes* Falcons*
16. *Gruiformes* Cranes*
17. *Ciconiiformes* Vultures*
18. *Podicipediformes* Grebes
19. *Charadriformes* Shorebirds
20. *Columbiformes* Doves
21. *Caprimulgiformes* Nightjars
22. *Psittaformes* Parrots
23. *Passeriformes* Passerines
1
2
4
5
9
19
23

*4/260 (1.5%) of lineages differ from traditional taxonomic placements at the ordinal level

10. Who is advancing barcoding?

The Consortium for the Barcode of Life (CBOL) is an international collaboration of natural history museums, herbaria, biological repositories, and biodiversity inventory sites, together with academic and commercial experts in genomics, taxonomy, electronics, and computer science. The mission of CBOL is to speed compilation of DNA barcodes of known and newly discovered animal and plant species, establish a public library of sequences linked to named specimens, and promote development of portable devices for DNA barcoding. More information is available at:

http://barcoding.si.edu   http://www.barcodinglife.org
http:// phe.rockefeller.edu/BarcodeConference/index.html

SOURCES

Species known and unknown

Blaxter M. 2003. Counting angels with DNA. Nature 421: 122-124.

Tudge C. 2000. The Variety of Life. 684pp. Oxford University Press.

Why standardize?

Hebert PDN, Cywinska A, Ball SL, deWaard, JR. 2003. Biological identifications through DNA barcodes. Proc Royal Soc Lond B 270: 313-322.

Janzen DH. 2004. Now is the time. Phil Trans Royal Soc Lond B 359:731-732.

Mitochondrial DNA for identifying species

Avise JC, Walker D. 1999. Species realties and numbers in sexual vertebrates: perspectives from an asexually transmitted genome. Proc Natl Acad Sci 96:992-995.

Wildman DE, Uddin M, Liu G, Grossman LI, Goodman M. 2003. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus *Homo*. Proc Natl Acad Sci USA 100: 7181-7188.

Mitochondrial gene content and order

Jameson D, Gibson AP, Hudelot C, Higgs PG. 2003. ORGe: a relational database for comparative analysis of mitochondrial genomes. Nucl Acids Res 31: 202-206.

Why COI for animal species?

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. 1994. DNA primers for amplification of cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol 3: 294-299.

Hebert PDN, Ratnasingham S, deWaard JR. 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc Lond B 270: S596-S599.

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W. 2000. PipMaker—a web server for aligning two genomic DNA sequences. Genome Res 10: 577-586.

Differences among and within species

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM. 2004. Identification of birds through DNA barcodes. PLoS Biol 2:1657-1663.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astrapes fulgerator*. Proc Natl Acad Sci USA 101:14812-14817.

Non-human primate COI barcode sequences provided by Coriell Institute for Medical Research (http://www.IPBIR.org)

An earlier illustrated brochure offered
Ten Reasons for Barcoding Life:

1. Works with fragments. Barcoding can identify a species from bits and pieces, including undesirable animal or plant material in processed foodstuffs and morphologically unrecognizable products derived from protected or regulated species.

2. Works for all stages of life. Barcoding can identify a species in its many forms, from eggs and seed, through larvae and seedlings, to adults and flowers.

3. Unmasks look-alikes. Barcoding can distinguish among species that look alike, uncovering dangerous organisms masquerading as harmless ones and enabling a more accurate view of biodiversity.

4. Reduces ambiguity. A barcode provides an unambiguous digital identifying feature for identification of species, supplementing the more analog gradations of words, shapes and colors.

5. Makes expertise go further. Scientists can equip themselves with barcoding to speed identification of known organisms and facilitate rapid recognition of new species.

6. Democratizes access. A standardized library of barcodes will empower many more people to call by name the species around them.

7. Opens the way for an electronic handheld field guide. Barcoding links biological identification to advancing frontiers in DNA sequencing, electronics, and information science, paving the way for handheld devices for species identification.

8. Sprouts new leaves on the tree of life. Barcoding the similarities and differences among the estimated 10 million species of animals and plants will help show where their leaves belong on the tree of life.

9. Demonstrates value of collections. Compiling the library of barcodes begins with the multimillions of specimens in museums, herbaria, zoos, and gardens, and other biological repositories, thus highlighting their ongoing efforts to preserve and understand Earth's biodiversity.

10. Speeds writing the encyclopedia of life. A library of barcodes linked to named specimens will enhance public access to biological knowledge, helping to create an on-line encyclopedia of life on Earth.

http://phe.rockefeller.edu/barcode/docs/TenReasonsBarcoding.pdf