Taxonomy, DNA, and the Barcode of Life **(aka *Banbury II*)**
Cold Spring Harbor Laboratory
September 10–12, 2003

Notes from Friday morning, session 1

1. Criteria for selecting pilot studies

Pilot studies provide an opportunity to test (validate) the barcoding approach to identifying and detecting species for various taxa, "i.e., proof of principle." This may include exploration of which genes work best for a particular taxon and/or particular methods (and ages) of preservation, but the emphasis should be on few and easily sequenced genes and detecting those with high resolving power. The overall goal is to develop a rapid and simple tool for species-level discrimination and identification. The following criteria should be considered when selecting appropriate pilot studies (sequence does not imply ranking in order of importance):

1. Likely discovery of new (i.e., undescribed or unrecognized) species, for the purpose of showing how barcoding will deal with them and facilitate their recognition, description, and/or detection.
2. Addresses research questions that are likely to yield new scientific insights.
3. Project rationale and logistics are cast in a "hypothesis-testing" style.
4. Demonstrates the potential for practical applications of the barcoding approach, such as biomedical importance (e.g., disease vectors), or agricultural/commercial importance (e.g., pest species), or educational benefits (e.g., ecotourism, pre-university school science).
5. Focus is either taxon-specific or biogeographic, or explicitly both.
6. Addresses taxa or environments that are highly relevant to conservation considerations.
7. Can readily link to existing large-scale research initiatives, e.g., ATOL, PBI, MaNIS and other "distributed" databases (e.g., the various "NET"s).
8. Can readily link to existing large-scale biodiversity inventory initiatives (e.g., ATBI of Great Smoky Mountains National Park; ATBI of Area de Conservacion Guanacaste, Costa Rica; national biodiversity inventory by INBio, Costa Rica; Terry Erwin/SI South American plots, Papua New Guinea Lepidoptera/SE Asia inventory, etc.).
9. Likely to yield new technologies and methods (hard- and software) for barcoding (i.e., deriving DNA sequence data) and for the use of barcoding data in species identification.
10. Likely to add significantly to the barcode sequence database.
11. Practicality—Is the project doable? Can it be executed within the year 2004?
12. Is there a willing and motivated professional community ready to carry out the pilot project?

13. Are the necessary specimens available politically and logistically (in hand or museum)? Are they already databased (or "databasable" as a simple byproduct of the project)?
14. High level of international interest or involvement.
15. High level of inter-institutional collaboration.
16. Explicit commitment of matching funds or in-kind support, in addition to direct funds required for the project.
17. Demonstrates and develops the long- and short-term integration of species-level taxonomic activity with the "gathering the sequence library" and "developing the delivery technology" (hard- and software) processes.
18. Demonstrates and develops the management and delivery of collateral information for every barcode and every voucher.

2. Specific suggestions of possible taxa, biotas, etc.

Primates of the world (xx spp.)

advantages

- popular appeal; "warm and fuzzy"
- high conservation priority
- rapid and convenient means of identifying individual animals (or animal parts) to species has broad practical application, e.g., for customs officials, wildlife inspectors, etc.
- accessible collections, including frozen tissues
- human genome is sequenced
- IPBIR database

disadvantages

- unlikely to yield (many) new species

Birds of the world (12,000 spp.)

advantages

- popular appeal
- barcoding effort underway by FAA/NMNH
- there already is a known enthusiast (Carla Dove and SI)
- chicken genome will be sequenced soon
- much of the needed raw material is already available from a few major collections
- may reveal cryptic species, or substantial genetic substructure within single species
- rapid and convenient means of identifying individual animals (or animal parts) to species has broad practical application, e.g., for customs officials, wildlife inspectors, etc.

disadvantages

- unlikely to yield many new species

Turtles of the world (300 spp.)

advantages

- circumscribed taxonomically
- several species are amenable to detailed population sampling, which provide an opportunity to test the impact of intraspecific variation on species identification via barcoding.
- high conservation priority in some regions (e.g., southeast Asia)
- accessible collections, including frozen tissues (virtually all currently recognized species are available, as well as hybrid individuals which can be used to assess the ability of barcoding to detect hybrids)
- digital database exists
- bacterial artificial chromosomes (BAC) library will be available soon
- popular appeal
- potential for novel applications, including the rapid identification of commercially harvested species (there was a really interesting case of turtle soup not long ago where the meat in the can actually came from several species, including one that was protected)
- the relevant scientific community of expert taxonomists, geneticists, and field biologists are eager to participate and are already organized (e.g., they recently jointly submitted a PBI project proposal to NSF)
- likely to yield new species; preliminary molecular studies have shown that each species examined seems to be composed of > 1 species

disadvantages

- some species are difficult to obtain in large numbers, which will limit ability to assess levels of intraspecific variation. However, this is mostly restricted to endangered taxa

Sphingid moths of the world (Sphingidae; 2000+ spp.)

advantages

- specimens of essentially all species are in two collections, one of which (TNHM) is nearly completely databased
- extant collections hold large series of many species from broad geographic areas
- willing taxonomic expertise exists (e.g., Kitching, Cadiou, Janzen), and would pull in the amateur community
- global distribution, and publicly well-known
- well-known taxonomically, including a picture guide to the world's species, and a modern taxonomic checklist; will not require major taxonomic revision or alpha-taxonomy before sense can be made of sequence data
- will likely reveal previously overlooked cryptic species and help resolve variation/subspecies arguments
- genomic resources exist for *Manduca*
- no one is doing global sphingid sequences at the moment
- mini-pilot sequencing the 137 species of Costa Rica is underway (Hebert, Janzen, Chacon), to be "finished" by the end 2003

disadvantages

- unlikely to yield MANY new species
- to complete the project within one year, a significant proportion (30%?) of the sequences will need to come from dry museum specimens collected before 1950. This would require corresponding forensic sequence development. This is not a problem for specimens collected more recently. A 2-year project duration would allow the distributed network of sphingid aficionados to get additional fresh material from far-flung parts of the world.
- the need to assess both local and geographic variation within species dictates that this project would probably require sequence data from 20 specimens per species, or a total of 20,000 sequences (tubes).

Mosquitoes (Diptera: Culicidae; 3500 spp.)

advantages

- human disease vectors
- representative specimens in two museums (NHML, NMNH)
- *Anopheles* genome sequenced (one other species approved, another five or six proposed)
- willing research community
- world taxonomic digital catalog linked to online pdf files of most taxonomic literature [www.mosquitocatalog.org]
- keys
- likely to yield new species
- high level of sampling density and biological knowledge, geographically well known
- taxa collectively display broad range of evolutionary histories, ages
- likely funding opportunities related to disease, including possible interest in developing new diagnostic technology for species recognition and identification

disadvantages

- perceived low conservation priority—"save the mosquitoes" unlikely to be embraced by the public—although they might prove to be useful indicators of environmental quality
- may be difficult to extract DNA from older museum specimens. (However, given the high epidemiological interest and the well-established network of scientists among several continents, it should be possible to quickly obtain fresh material, as needed.)

Skipper butterflies of the Area de Conservacion Guanacaste (ACG), Costa Rica (Hesperiidae; 400+ spp.)

advantages

- specimens of essentially all species are present in one collection (SI)
- large series exist for each species (mostly reared from caterpillars); more are readily obtained
- most specimens are less than 30 years old and will yield COI sequence data from a single dry leg
- many species range from Mexico to Paraguay; location covers the three major tropical vegetation types—rain forest, dry forest, cloud forest
- ACG includes 60% of species of Costa Rica and at least 50% of those of Central America
- integral part of conservation of the ACG
- integral part of the ATBI of the ACG (e.g., http://janzen.sas.upenn.edu) and the Costa Rican national biodiversity inventory (INBio)
- well-known taxonomically, yet rich in undescribed sibling species
- will not require major taxonomic revision or alpha-taxonomy before sense can be made of sequence data (but will require sufficient alpha-taxonomic revision to show how that will integrate with barcoding); results likely will resolve many subspecies and variation arguments
- well-known publicly, some are agricultural pests (e.g., *Urbanus*, on beans)
- global genus-level sequencing for phylogeny already underway (Andy Warren)
- mini-pilot (NSF, Smithsonian and Hebert) already in underway
- willing taxonomic expertise (John Burns) and enthusiasts (Janzen, Hebert) already involved, in association with multiple collaborators and institutions (USNM/SI, INBio, TNHM)
- likely integration with TOL and PBI efforts—phylogeny of the world's butterflies by a large and growing collection of international collaborators (not yet funded)
- all species and specimens are databased (with photos) either in the field as collected or as part of museum accession process (see http://janzen.sas.upenn.edu)
- could be done in one year (though two would be better); with a one-year push it could be expanded to be the entire Hesperiidae fauna of Costa Rica by using the parataxonomist team at INBio (with all the political advantages associated)
- would provide proof-of-concept for the use of sequences to associate conspecific adults and immatures (caterpillars, pupae)

disadvantages

- perceived low conservation priority, but might be useful as indicators of environmental quality
- to complete the project within one year, a significant proportion (10%?) of the sequences would need to come from dry museum specimens collected before 1950. This would require corresponding forensic sequence development. This problem does not apply to specimens collected more recently. If the project were to include the entire Hesperiidae of Costa Rica, sequences from older museum specimens (Smithsonian, TNHM) would need to be used.

- to assess both local and geographic variation within species would require data from an average of 20 specimens per species, which means 8,000 sequences (tubes). If expanded to the entire country, this could require as many as 14,000 sequences.

tephritid fruitflies (Diptera: Tephritidae; 4400 spp.)

advantages

- well known taxonomically, well sampled
- agricultural importance (pests)
- likely to yield new species
- willing expertise
- keys
- funding related to agriculture
- Med-fly genetics
- Digital names catalog soon [other information already available at www.sel.barc.usda.gov/diptera/tephriti/tephriti.htm/
- Possible interest in developing new technology for recognition
- potential as environmental quality monitoring tools because many species are associated with native plants (i.e., they are not pests)
- some species used as biological control agents for weeds

disadvantages

- Possibly limited availability of material. (Although not everyone sees this as a problem, given the large numbers of specimens accumulating from agriculture-funded surveys. Many non-agriculture species come to the chemical lures that are used for surveys.]

Costa Rica/INBIO national inventory (200,000+ spp.)

Caveat: A pilot project would not address all 200,000 species held at INBio. Instead, it would take a two-year slice through the INBio inventory, surgically doing ca. 10 taxa (200–500 species each), each spread over two years, selected for 1) good national-level representation of specimens and species in INBio, 2) detailed and enthusiastic participation of INBio curators, and 3) synergism with INBio goals of both national and Central American inventory and public biodiversity education (bioliteracy).

advantages

- enthusiastic institutional and staff participation, with the consequent "proof of concept" based on the collective efforts of a group people (the INBio curators, parataxonomists and their international collaborators as a whole).
- Execution of a very biodiverse tropical barcoding pilot project by self-motivated tropical citizens in a politically friendly environment.
- Basal involvement by a tropical institution in a process that potentially will very strongly impact human interaction with tropical biodiversity; makes INBio a stakeholder in a very direct way from the very beginning.
- The focal taxa can be easily chosen by the INBio curators and staff with a few days of discussion; potential examples—Scarabaeidae, Cerambycidae, spiders, Syrphidae, Tachinidae, terrestrial/freshwater molluscs, mycorrhizal fungi,

Tettigoniidae, *Enicospilus* (Ichneumonidae), microgastrine Braconidae, Hesperiidae (see above pilot), Poaceae, Rubiaceae, etc.

- Positive results easily displayed to other tropical countries as part of the established INBio international interaction with tropical biodiversity conservation and management.
- Positive results easily expanded to include the entire Costa Rican biota, which is at least 4% of the world.
- INBio would be happy to be the geographic-based effort to contrast with global taxon efforts.
- INBio curators, parataxonomists and collections already have an extensive global network of international taxonomists working with them (at least 400 taxonomists in tens of countries); pilot projects would instantly interact with this network, both technically and biopolitically.
- All INBio specimens have been collected since 1989, and the great majority pinned and oven-dried in the field (except for spiders, etc., which are stored in alcohol).
- Entire INBio collection is specimen-based databased at the time of collection (many millions of specimens now; http://www.inbio.ac.cr)
- All that is done with Costa Rica and INBio occurs in a social environment in which national parks are a legitimate and important part of the national socio-economic fabric, and therefore an ideal testing ground for the relevance of barcoding to the biodiversity-humanity interaction as well as for positive feedback from society to taxonomists.

disadvantages

- INBio was not present at Banbury I or Banbury II (though Janzen has been maintaining INBio staff moderately updated)
- Taxonomically very diverse.
- Not driven by a single individual, or PI (at present), so will have the fuzziness that comes through a committee and bureaucratic process (but there are obvious advantages to this as well).
- Will require many tens of thousands of sequences, and substantial involvement from the international taxasphere.

nematodes (a lot of species)

advantages

- *C. elegans* model
- barcoding projects underway
- disease-related, agricultural pests
- barcoding could prove to be one of the most effective ways to tell them apart

disadvantages

- retaining voucher specimens could be difficult
- Comprehensive study not possible in short time (1 yr)
- likely would need to recruit/organize taxonomic expertise to spearhead effort

Macro-parasites of the 940 species of vertebrates of the Area de Conservacion Guanacaste, Costa Rica.

advantages

- willing expertise: Dan Brooks (University of Toronto) has coordinated this inventory since 1995 and has strong interest in barcoding
- thousands of samples representing all classes (flukes, tapeworms, acanthocephalans, nematodes, bot flies, lice, fleas, mites, etc.) already collected from several hundred vertebrate species
- existing network of > 30 parasite taxonomists is already involved and working up material in classical and DNA ways
- team of three well trained parataxonomists already on site (both taxonomic collecting/sampling and collecting material for sequencing).
- another intensive collecting session is planned for May–July 2004.
- All specimens and data collected to date (including images) are databased and on the web (http://brooksweb.zoo.utoronto.ca/index.html). Web site integrates inventory database with species home pages and with database of all published phylogenetic trees for helminth parasites. Running list of publications from the inventory is also included in the web site.
- very diverse array of taxa, with obvious medical and educational applications
- project can begin immediately
- ideal for proof-of-concept for associating larvae and juveniles in intermediate hosts with adults in final hosts, and consequently quickly working out complex life histories cycles
- would accelerate ongoing taxonomic efforts (many new genera and species), including identification of cryptic species groups
- would demonstrate feasibility and logistics of cross-taxon yet site-specific discovery and characterization of biodiversity.
- would facilitate protocol development for tropical regions
- obvious linkage to the Costa Rican national conservation and biodiversity development process.

disadvantages

- would require at least two years to complete, although a huge amount, and maybe enough, could be done in just one year

salamanders

advantages

- ca. 70% of named species represented in frozen tissue collections; many are already sequenced
- many new (unnamed) species
- existing research infrastructure: AmphibiaTree, HerpNET, AmphibiaWeb
- high conservation priority, e.g., CI's Global Amphibian Assessment

disadvantages

- DNA sequence data already is used extensively in alpha taxonomy, and large levels of intraspecific variation may not support barcoding approach with these animals.

Gulf of Maine megafauna (ca. 1000 sp.)

advantages

- part of existing CML
- doable number of species
- socio-economically important ecosystem
- existing collections
- all-taxa approach

disadvantages

- unlikely to yield new species, except among benthic taxa

3. The following additional groups were suggested as appropriate or compelling subjects of pilot studies, but little additional supporting evidence was provided.

Catfishes: advantages—PBI recipient; disadvantages—probably couldn't be completed within a year or so (e.g., new PBI will run for 5 yr)

freshwater gastropods/Australo-Papuan caen…ids: advantages—existing databases, conservation priority

marine mammals/vertebrates (xx spp.)

mirid bugs: advantages—PBI recipient

Hawaiian terrestrial arthropods (xx spp.)

Australian marsupials (xx spp.)

Notes from Friday morning, session 2

1. ***Definition/description of the barcoding initiative***

A. What is the "barcode of life"?

Definition: A short DNA sequence that provides an aid to species recognition and identification in a particular domain of life.

Alternate definitions:

- A short DNA sequence(s) that serves as a unique identifier of each species.
- A DNA-based method that, through appropriate technology, enables rapid species recognition and identification by non-specialists, educators, health professionals, and in a variety of additional applications, e.g., a molecular key, or field guide.

Supplementary comments:

1) Barcode data should be derived by the professional community, which would establish protocols, standards, etc., for quality control.
2) Each DNA sequence should be linked to a voucher specimen accessioned in an institutional repository (e.g., museum collection).
3) DNA barcodes and related information should be made widely available through electronic databases, and linked with other complementary and efforts, such as MorphoBank, GenBank, GBIF.
4) The barcoding initiative is intended to capture and adopt rapidly evolving molecular technology to develop useful, field-based applications of barcode data.
5) Initial implementation of DNA barcoding should make use of museum collections.

B. What isn't it?

1) It is hoped (and intended) that DNA sequence data will contribute to the discovery and formal recognition of new species. However, DNA barcodes should not be used as the sole criterion for description of new species, which instead should employ diverse data, from morphology, to behavior, to genetics.
2) The goal of DNA barcoding is not the discovery of new species, per se, although it is expected that such discoveries will emerge from the enterprise.
3) A DNA barcoding approach may not be effective or appropriate in all instances.
4) DNA barcoding is intended to complement, facilitate, and enhance—not supplant or invalidate—existing taxonomic practice. It also is not intended to duplicate or compete with efforts to resolve the phylogenetic history of life on earth, e.g., ATOL.

2. ***Possible actions and duties of DNA barcoding steering/advisory group***

- assess proposals
- participate in monthly conference calls
- prepare documents, position papers, applications
- represent and advocate for initiative in Washington, DC, and elsewhere

3. ***Potential applications and products of barcoding, and corresponding sponsors.***

Application or "product"	***Potential sponsor(s)***[1]
Rapid ID of disease vector, e.g., malaria, West-Nile virus	NIH, DOD, Audubon Society, WHO (TDR),Gates Foundation, CDC, AID
Rapid ID of agricultural pests, e.g., Medfly; includes quarantine applications	DOA[2]
ID of illegal trade in endangered species	DOA, DOI, USFWS, CS
Environmental assay	EPA
Invasive species, e.g., ID larvae in ballast water	DOI, DOA
Assessing productivity in marine protected areas	NOAA
Education	NSF, DOEd
Bio-threats	DHS
Bio-forensics	FBI, CIA
Biodiversity inventory	CI, World Bank, TNC, local and regional organizations
Biodiversity prospecting	Merck
Technology R&D	venture capital
middle America	average Joe; typical homeowner
bioinformatics	IBM, Microsoft

[1] AID, U.S. Agency for International Development; CDC, U.S. Centers for Disease Control; CI, Conservation International; CIA, U.S. Central Intelligence Agency; CS, U.S. Customs Service; DHS, U.S. Department of Homeland Security; DOA, U.S. Department of Agriculture; DOD, Department of Defense; DOEd, U.S. Department of Education; DOI, U.S. Department of the Interior; EPA, U.S. Environmental Protection Agency; FBI, U.S. Federal Bureau of Investigation; IBM, International Business Machines; NIH, U.S. National Institutes of Health; NOAA, U.S. National Oceanographic and Atmospheric Administration; TCN, The Nature Conservancy; USFWS, U.S. Fish and Wildlife Service; WHO, World Health Organization.

[2] Also likely applies to departments of agriculture at all levels, from the FAO internationally, through many nations (especially USA, Australia, New Zealand, Latin America, Africa, Japan, etc.), through states such as California and Florida, through counties, etc.