Integrative Biology 200A "PRINCIPLES OF PHYLOGENETICS" Spring 2014
University of California, Berkeley Kipling Will- 21 March

DNA Taxonomy:
Maintains that DNA –not morphology- be the main or exclusive data used for taxonomic decisions. Intends to function as the universal reference system for biology using sequences as the handles and in some sense as the names. Previous discussion in class on the use of multiple lines of evidence, the need to connect our phylogenies and taxonomic classifications to fossil and rare taxa and the many issues regarding the evolution and analysis of DNA sequence data, should be enough to make you critical of hardcore DNA taxonomy.

By some, it has been proposed as a realistic, but flawed, heuristic:

"*To be clear that what is being estimated for a specimen is not necessarily its membership to a 'species', however defined, we call the taxa yielded by grouping of specimens through a set of markers OTU. We have coined the term MOTU (Floyd et al. 2002); MOTU have also been called 'phylotypes' and 'genospecies'. MOTU can be simply defined by sequence identity: if two specimens yield sequences that are identical within some defined cut-off, they are assigned to the same MOTU*." (Blaxter 2004)

But in practice it tends to be multi-gene, phylogenetic analyses focusing on the species-level or faunal genetic sampling and analyses. The former is pretty typical and may even be most appropriate in some cases. Most sample the "obvious" morphological diversity and presumed populations and, not surprisingly, result largely with distinct clades consistent with the morphology and population structure. The latter tends to be studies that lack sufficient sampling as they don't focus on possible clades, but rather on a fauna.

DNA barcodes:
"*We are convinced that the sole prospect for a sustainable identification capability lies in the construction of systems that employ DNA sequences as taxon 'barcodes'.*" (Hebert et al. 2003)

DNA barcodes are to taxonomy as Twitter is to news reporting. How it is proposed to work:
A short sequence, ~650bp (about 4% of a typical mt genome, less than 50% of the COI gene) from the Folmer region of COI is used. Hypothetically this contains enough information to resolve 10-100million species. ("micro barcodes" of 100bp have also been proposed for more degraded materials). To do so depends on having randomized arrangement of the four nucleotides over the 650bp.

A good DNA barcode sequence is conserved enough to be amplified with "universal" primers while divergent enough to resolve closely related species. COI is asserted to have these properties.

The method (see first figure):
1. Gather this short sequence from all samples.
2. Build "profile" trees. Generally NJ is used. (overall similarity, diagnostic similarity are used, but see second figure and Meier & Zhang's paper)
3. Match taxonomic names to terms.
4. For unknowns their identity can be read from the resultant topology, typically, but not always by grouping with a cluster that is 98% or more similar, or are "nearby" in the NJ tree, or are more similar than the mean divergence between pairs (see second figure and Meier & Zhang's paper).

Purported good properties and possible applications (The top 10 list on the Barcode Website)
1. Works with fragments.
2. Works for all stages of life.
3. Unmasks look-alikes.
4. Reduces ambiguity.
5. Makes expertise go further.
6. Democratizes access.
7. Opens the way for an electronic handheld field guide, the Life Barcoder.
8. Sprouts new leaves on the tree of life.
9. Demonstrates value of collections.
10. Speeds writing the encyclopedia of life.

Problems:
-Resolving recently diverged species and hybrids may be impossible for COI. There is no way to know when the answer is wrong except in well-known and well sampled groups. However, often the "wrong" is shifted to non-barcode evidence without justification. Effectively limits what species are by setting a range of divergence based on this segment of a gene.

-No single gene is conserved across all life. So it will take a few, at least.

-Must be able to distinguish between interspecific and intraspecific variation and many papers refute the notion of a "barcode gap" (see third figure). However, reliance on a gap is necessary… "*BOLD ID engine (www.barcodinglife.org),...uses a 2% cutoff for assigning specimens to species*". Or … a 1% cutoff, that is then reported as identification with a confidence of 100% (Ratnasingham & Hebert 2007)… or the use of the mean distance, etc.

-Reference sequences must be from "taxonomically confirmed" specimens or one must accept unique COI haplotype clusters as the "important" units. These are Hebert's gene-species and Blaxter's MOTUs.

-Identification does not equal Science: Often integrative taxonomy and applied taxonomy are confused.

-"dark taxa". Numerous barcode sequences that have limited identification. (see Rod Page's blog "iPhylo")

-In general the whole enterprise hasn't amounted to much (see the readings for today). Good science tends to prevail and the meme "barcode" has shifted meaning except for its use by ecologists.

What should we be doing?

Integrative Taxonomy. The use of multiple independent or nearly independent lines of evidence and appropriate tests to establish taxonomic entities.
Applied Taxonomy. The implementation of techniques and technology to identify a semaphorant or sample as a token of a taxonomic entity.

Identification is placing individuals into circumscribed groups or recognizing that they fit none that exist. The science of taxonomy- determining those groups- would have to be nearly finished for this to even be partially useful beyond tentatively identifying samples. The reference database would need to be nearly complete (see Meier & Zhang's paper) and this is unrealistic.

-An appropriate analysis gives the taxa, a taxon gives the characters, the characters do NOT give the taxon, but characters can identify a token of a taxon. After systematic and taxonomic study (integrative taxonomy) the best tools to identify important biological units can be used. It might be DNA. Might even be COI.

-How is divergence in COI related to species (or smallest names clades) and what is its relationship to units for which we have interesting biological questions? There is no evidence of a connection between mtDNA and speciation (despite vague claims that there should be). Fundamentally, barcoding ignores difficult issues of species concepts and the dynamics of biological classifications and taxonomic hypotheses. In the BOLD system…"Reference barcodes are a validated subset of the full database containing only those species represented by three or more individuals showing less than 2% sequence divergence." Is this a barcode species concept?

-Problems with mtDNA largely ignored include mitochondrial heteroplasmy, identical sequences in different species, introgression and hybrid speciation, incomplete lineage sorting, NUMTs- Nuclear Pseudogenes.

Where does it succeed? When "almost" is close enough (e.g. like horseshoes and hand grenades) and possible error can be ignored or greatly reduced.

In well studied groups. To reduce error, just do the science first. In a group that is well studied and sampled, especially if it is of economic and/or human health concern (like ticks and

mosquitoes) having a broad sample and well done taxonomy is important for many reasons. Using DNA identification tools then makes good sense (see last figure).

In limited systems. An ecologists could make a first pass, sorting of samples from a restricted fauna, e.g. insects in a stream system, to make a contained database that subsequent samples would be compared against. Of course if the taxonomy of the groups sampled is not done they would only be able to identify unique haplotypes.

See also:

Will, K. W., and Rubinoff, D. 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics 20: 47-55.

Hebert, P. D. and Gregory, T. R. (2005). “The promise of DNA barcoding for taxonomy.” Systematic Biology 54(5): 852-859.

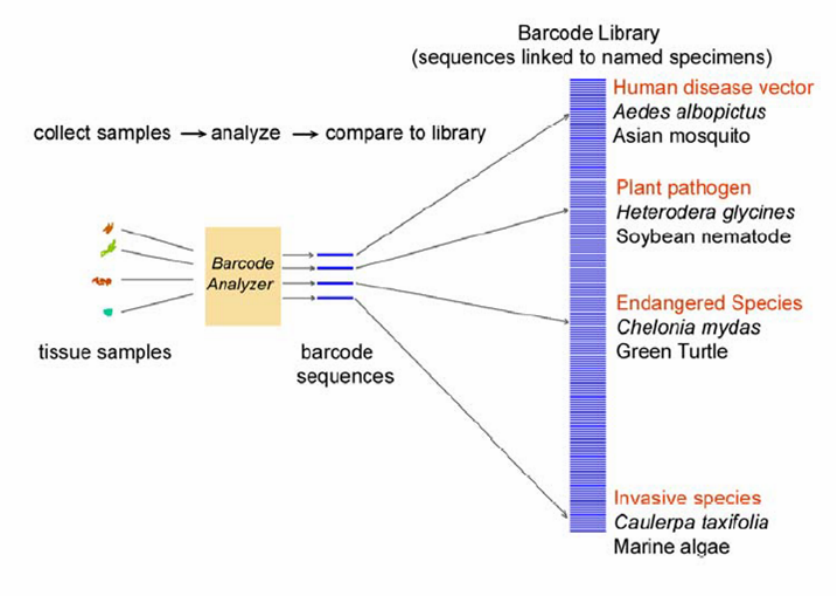
Smith, V. S. (2005). “DNA barcoding: perspectives from a "Partnerships for Enhancing Expertise in Taxonomy" (PEET) debate.” Systematic Biology 54(5): 841-844.

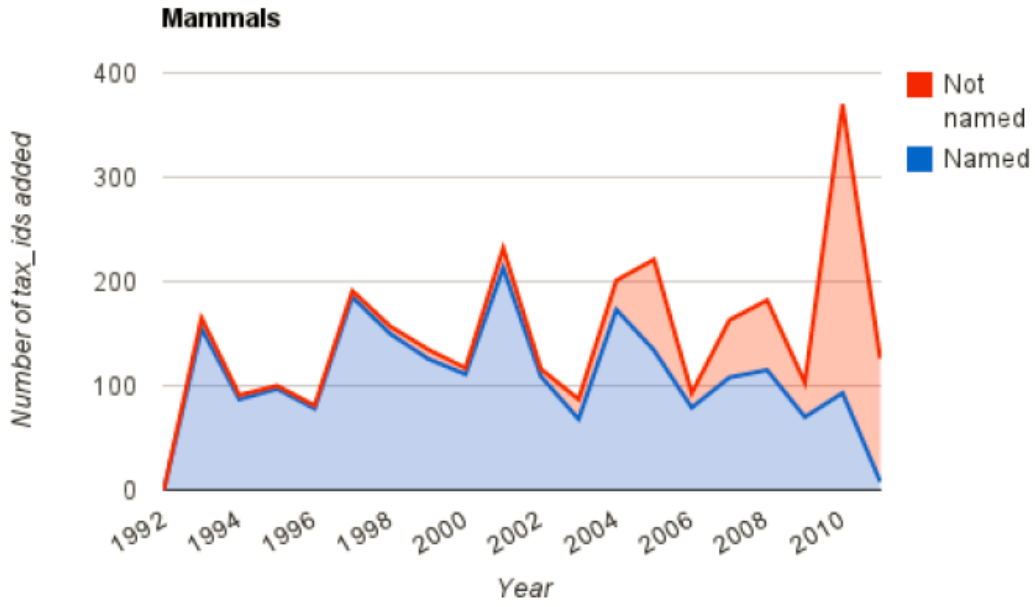
Will, K. W., Mishler, B. D. and Wheeler, Q. D. (2005). “The perils of DNA barcoding and the need for integrative taxonomy.” Systematic Biology 54(5): 844-851.

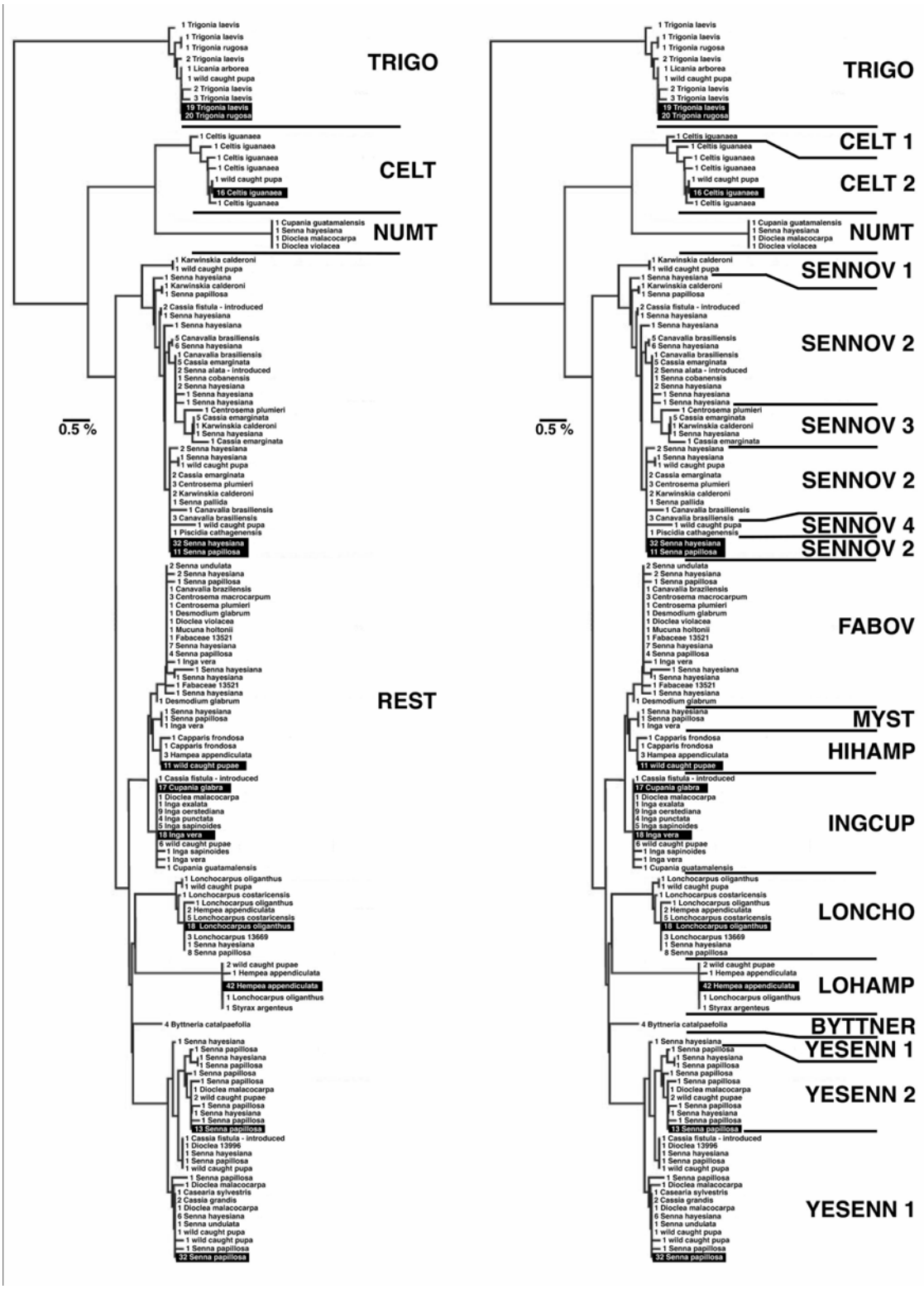
Stoeckle, M. Y. and Hebert, P. D. N. (2008). “Bar Code of Life: DNA Tags Help Classify Animals.” Scientific American 299(4): 82-88.

Meier, R. and Zhang, G. (2009). DNA barcoding and DNA taxonomy in Diptera: An assessment based on 4261 COI sequences for 1001 species. Diptera Diversity: Status, Challenges, and Tools. Pape, T., Bickel, D. and Meier, R. New York, Brill Academic Publishers, 349-380.

Marakeby et al. (2014) A System to Automatically Classify and Name Any Individual Genome-Sequenced Organism Independently of Current Biological Classification and Nomenclature Plos1.


Barcode Library
(sequences linked to named specimens)
collect samples → analyze → compare to library
Barcode Analyzer
tissue samples
barcode sequences
Human disease vector
Aedes albopictus
Asian mosquito
Plant pathogen
Heterodera glycines
Soybean nematode
Endangered Species
Chelonia mydas
Green Turtle
Invasive species
Caulerpa taxifolia
Marine algae


Mammals
Number of tax_ids added
400
300
200
100
0
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
Year
Not named
Named


TRIGO
CELT
NUMT
REST
0.5 %
TRIGO
CELT 1
CELT 2
NUMT
SENNOV 1
SENNOV 2
SENNOV 3
SENNOV 2
SENNOV 4
SENNOV 2
FABOV
MYST
HIHAMP
INGCUP
LONCHO
LOHAMP
BYTTNER
YESENN 1
YESENN 2
YESENN 1
0.5 %

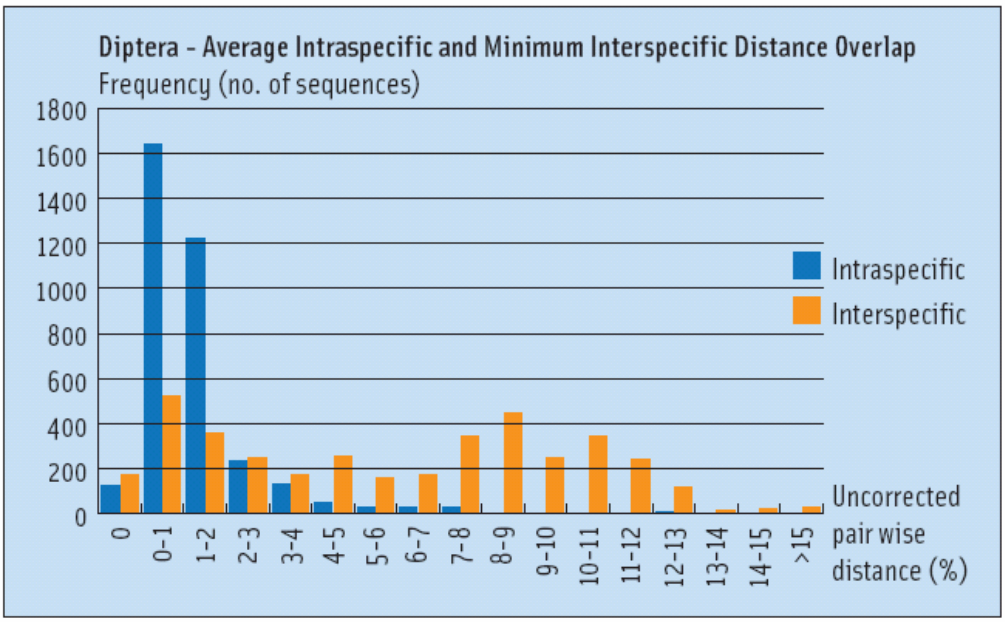
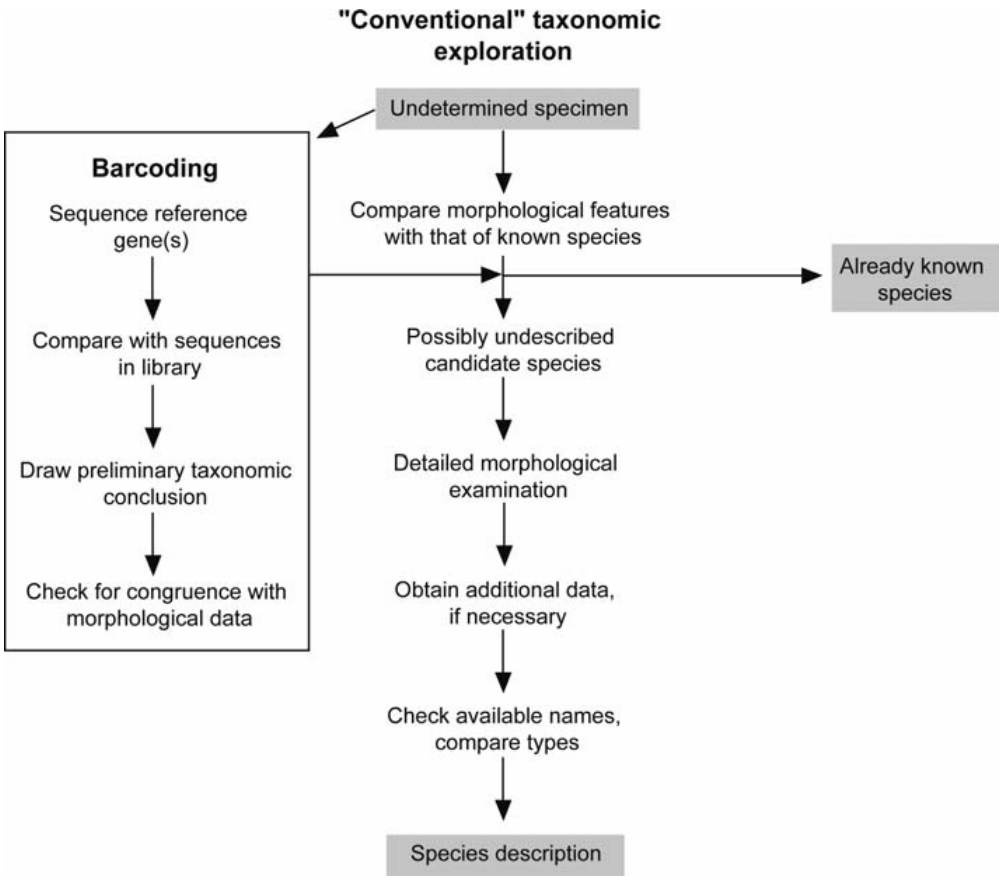


Figure 13.1. Barcoding gap for Diptera.

From Meier & Zhang. 2009. 4,261 COI sequences for 1001 species of Diptera



If freed from naïve expectations and the claim of exclusivity, barcoding is a useful tool in the hands of skilled systematists, especially in connection with classical methods.
Frank Köhler 2007