March 19, 2018. **Classification IV: DNA barcoding and DNA taxonomy**
and **Phylogenetic trees VIII: More on tree-to-tree comparisons; supertrees**

A. DNA barcoding and DNA taxonomy

"Your work, Sir, is both new and good, but what's new is not good and what's good is not new."
-- Samuel Johnson

1. DNA Taxonomy - the new but not good

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 taxonomies 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)"

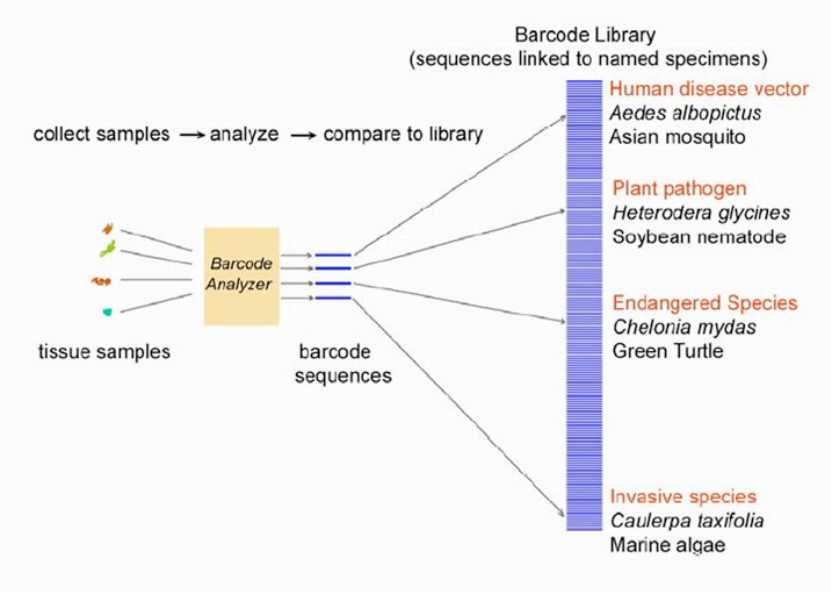
But in practice it tends to be multi-gene, phylogenetic analyses focusing on the species-level or faunal genetic sampling and analyses. The former is pretty typical and may even be most appropriate in some cases. Most sample the "obvious" morphological diversity and presumed populations and, not surprisingly, 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.

2. DNA barcoding - the good but not new

"*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) from the Folmer region of COI is use. Potentially this contains enough information to resolve 10-100million species. ("micro barcodes" of 100bp have also been proposed for more degraded materials). This depends on having any 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 figure above):

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 "near by" in the NJ tree, or are more similar than the mean divergence between pairs.

<u>Purported good properties and possible applications (The top 10 list on the Barcode Website):</u>

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.

<u>Problems:</u>

-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.

-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". However, reliance on a gap is necessary… "BOLD ID engine (www.barcodinglife.org),...uses a 2% cutoff for assigning specimens to species" (recent blog entry). Or … a 1% cutoff, that will is then reported as identification with a confidence of 100% (Ratnasingham & Hebert 2007).

-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 (see section on DNA taxonomy above).

Solution: Integrative Taxonomy (Will, Mishler, & Wheeler, 2005). The use of multiple independent lines of evidence and appropriate tests to establish taxonomic entities.

Where does DNA identification 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.

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 do approximate identifications.

For more discussion, and references, see:

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.

B. More on tree-to-tree comparisons; supertrees

-- There are many reasons why one would want to compare trees, falling into three basic categories:

-- *Within an analysis of one clade, with the same OTUs*; e.g., equally or nearly equally parsimonious (or likely) trees, trees resulting from different character partitions, models of evolution, or methods of analysis, and comparisons with trees from the literature.

-- *Within an analysis of one clade, with different OTUs*; trying to come up with a general tree for all OTUs, e.g. super trees.

-- *Comparing analyses of different clades*, e.g., gene family evolution, migration between populations, vicariance biogeography, host/ parasite relationships, symbiosis, community evolution, or any long-term ecological association. We will cover these later in the class, but for now just be aware that most of the same principles obtain.

Methodology for comparing phylogenies:

(1) consensus techniques (strict, semi-strict, majority rule, Adams) -- for finding shared signal among trees. [This is a review of the Feb 16th lecture -- see those notes too.]

Strict consensus: Only monophyletic groups found in all source trees are found in the resultant tree. The tree excludes a subset of all possible trees and conversely includes a subset of possible trees, whether or not they are part of the source set, e.g. (A(B(CD))) + (A(C(BD))) = (A (BCD)) but this also implies (A(D(BC))). In some sense the most conservative consensus.

Semistrict consensus: Only monophyletic groups found in at least **one** of the source trees and compatible (not in conflict) with all other source trees are found in the resultant tree, i.e. if a clade is never contradicted, but not always supported, then it is still included in this compromise tree. E.g. (A(B(CD))) + (A (BCD)) = (A(B(CD)))

Majority-rule consensus: Shows groups that appear in more than a pre-specified percentage of source trees, usually >50%. Not recommended for summary of equally-optimal trees resulting from a search.

Adams Consensus: Inconsistently placed taxa are moved down to the first node that summarizes the possible topologies. N.B., groups can appear in Adams consensus that are not found in **any** source tree. Adams trees have no biological or phylogenetic interpretation, but they do point to "wildcard" taxa. Those taxa may be experimentally removed from the matrix and the resulting analysis compared to when they are included.

(2) tree-to-tree distance metrics. There are two types of approaches. One counts the number of steps needed to transform one tree into another (e.g., NNI interchange metric, partition metrics, agreement subtrees). The second represents two trees as sets of simpler structures and then measures similarity between these (e.g., quartet measures)

Transforming one tree into another

A good example of a measure defined in terms of transforming one tree into another is the nearest nearest neighbor interchange (NNI) metric (e.g., Waterman and Smith, 1978) which measures the minimum number of NNIs required to change T_1 into T_2. In the example below, one NNI is required to convert T_1 into T_2, so $d_{NNI}(T_1, T_2) = 1$.

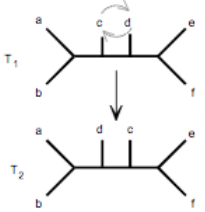


Figure 5.1 Transforming T_1 into T_2 by a single nearest neighbor interchange of leaves c nd d

from the Component User's Guide, by Rod Page (http://taxonomy.zoology.gla.ac.uk/rod/cplite/title.pdf)

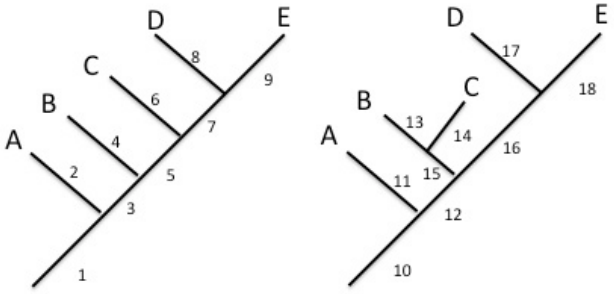
The Robinson Foulds metric is a commonly used metric; defined as the sum of [the number of partitions of data implied by the first tree but not the second tree] plus [the number of partitions of data implied by the second tree but not the first tree]. It is also called the symmetric difference metric (Robinson, D. R., Foulds, L. R. 1981. *Mathematical Biosciences*, **53**, 131-147).

(3) component analysis (more in the biogeography lectures) -- finding individual statements of relationship that are shared among trees, basically a node relating some taxa to the exclusion of others.

(4) Maximum likelihood approaches (e.g., parametric bootstrapping) -- comparing alternative trees or alternative models of evolution for your data (e.g., Efron, et al. 1996. Proceedings of the National Academy of Sciences 93: 13429).

(5) Brooks parsimony, i.e., representing the grouping information in separate trees as characters in a matrix (e.g., using Brooks parsimony, also called "matrix representation parsimony"). This might be used when comparing hosts and parasites, or phylogenies of different taxa that all live in the same areas of endemicity. [See Brooks & McLennan, 1991; Brooks 1981, Syst. Zool. 30:229; Wiley 1988, Syst. Zool. 37:271; and see Kluge 1988, Syst. Zool. 37:315 for some suggested modifications.]

Simple example: two trees being compared/combined for the same taxa



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	1	1	0	0	0													
B	1	0	1	1	0													
C	1	0	1	0	1													
D	1	0	1	0	1													
E	1	0	1	0	1													

In class exercise: fill in the rest of this character matrix. Then when you are all done, what topology would the matrix support?

(6) Supertrees are one of the frequent applications of tree comparisons, in this case attempting to combine different trees of the same larger clade that were developed from different sets of OTUs. In the simplest case, detailed phylogenies of individual genera are stitched together using a backbone phylogeny of a family that might have one representative of each genus. For a more analytical approach, Brooks parsimony can be used (branches in the separate trees are represented in a data matrix for analysis).

(7) Supermatrices are at the opposite end of the spectrum from the supertree approach -- these are so-called total evidence analyses that concatenate all the data types into one matrix. Note that different models can be used for different partitions (i.e., different genes, morphological data, etc.).