February 3, 2014**. Molecular Data I**

Required reading: Maddison, W. P. and D.R. Maddison. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75 Chapter on Molecular Data: http://mesquiteproject.org/Mesquite_Folder/docs/mesquite/molecular/molecular.html

I. Techniques - kinds of data

-- intrinsically distance-based data:

-- immunology (cross reaction of antibodies)
-- DNA - DNA hybridization
-- AFLPs - RAPDs -DNA fingerprinting
-- microsatellites

-- character-based data:

-- allozymes
-- restriction enzyme sites
-- sequencing methods
-- direct
-- cloning
-- PCR
-- genomic data (gene arrangement)

Properties of a good marker, as compared between molecules (i.e., DNA sequence data) and morphology.

	molecules	morphology
1) COMPLEXITY AND COMPARABILITY	–	+
2) DISCRETE STATES	+	–
3) HERITABILITY	+	–
4) INDEPENDENCE	?	?
5) LOW RATE OF CHANGE (λ)	?	?
6) MANY POSSIBLE CHARACTER STATES	–	+

II. Special features of molecular data

-- purported advantages:

-- closer to (or equal to) the genetic information.
-- huge numbers of potential characters, especially useful in organisms with simple morphology.
-- ability to homologize across very broad groups.
-- independence from morphological characters which are perhaps more subject to adaptive convergence.
-- ability to model or weight, because of relatively simple models of change.
-- $$$.

-- purported weaknesses:

-- simplicity of characters (i.e., no ontogeny, few possible character states) leading to special problems with homoplasy.
-- sampling problems.
-- fossil taxa generally can't be included.

-- highly conserved regions, used to reconstruct deep branching points, are perhaps *more* subject to adaptive convergence.
-- $$$.

III. Methods of analysis (an overview for now -- more later in the class)

A. Phenetic

-- molecular systematics is the last hold-out of phenetic methods as used for phylogenetic reconstruction.
-- disadvantages:
-- usually assumes molecular clock.
-- many distance measures used are non-metric, therefore one can't interpret branch lengths.
-- hides homoplasy.
-- throws away the information on individual characters that was so laboriously obtained.
-- advantages:
-- ??? (at best able to mimic the results of a phylogenetic analysis)
-- Averaging across whole genome?
-- Avoiding problem of reticulation? (some argue phenetic methods are OK below species level, as in the field of "phylogeography").

B. Phylogenetic

-- many molecular systematists are deeply concerned with adapting standard character-based methods of phylogenetic analysis (e.g., parsimony and Maximum likelihood) to these data; most of the issues we have already discussed are involved:

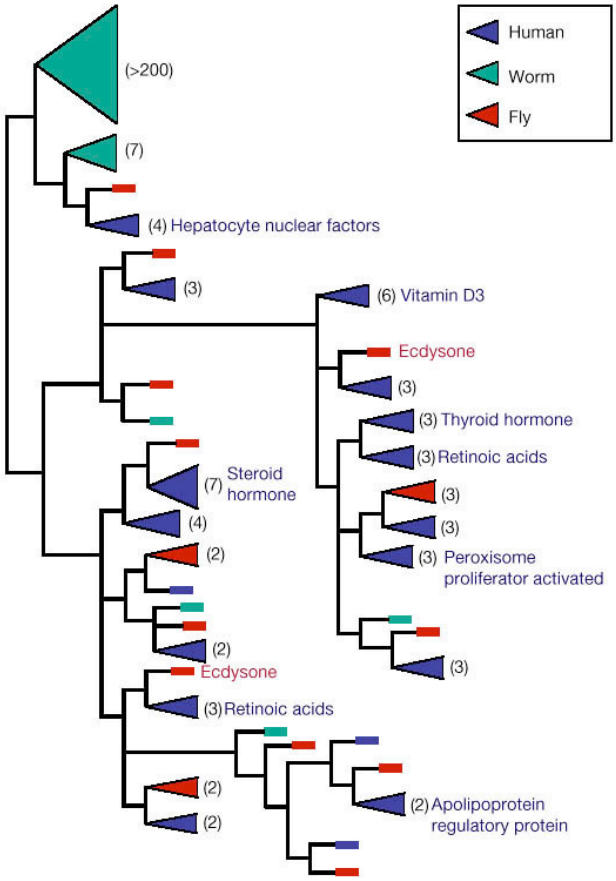
-- homology (including alignment problems)

-- what is a character?
-- nucelotide positions
-- character correlations
-- structural rearrangements (i.e., deletions, inversions) - more below
-- allozymes
-- restriction sites:
--RFLP's
--mapping
-- microsattelites

-- weighting/modeling issues:
-- gains versus losses
-- transitions versus transversions
-- purines A G
-- pyrimidines C T U
-- codon position bias
-- compensatory substitutions in RNA (due to secondary structure)
-- compatibility, "signature nucleotides" (i.e., the "true" synapomorphy approach in a new guise)

IV. Comparing genomes

-- synteny, rearrangements, insertions/deletions
-- exon shuffling
-- the gene "annotation" problem
-- multigene families
-- paralogy vs orthology
-- the fate of duplicated genes: ghost genes, subfunctionalization



Simplified cladogram of the 'many-to-many' relationships of classical nuclear receptors. Triangles indicate expansion within one lineage; bars represent single members. Numbers in parentheses indicate the number of paralogues in each group.

V. Recommendations (Mishler's Aphorisms):

-- treat these data as any other; if the object is phylogeny reconstruction, use phylogenetic methods.

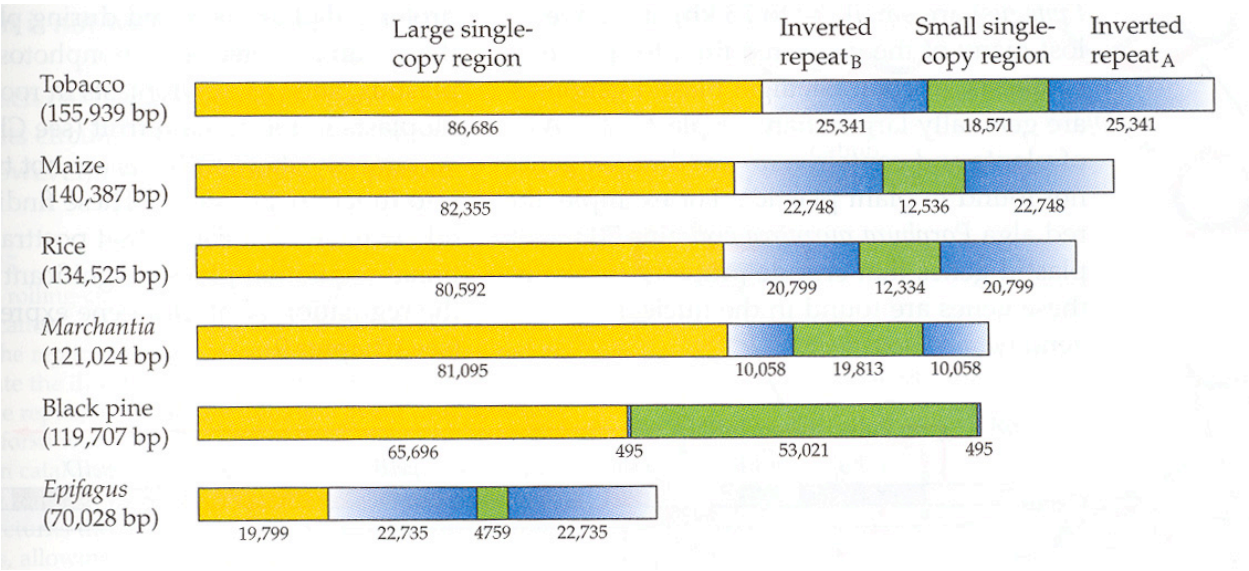
-- include all available data in an analysis, even if your own focus has been on molecules; it makes no sense to ignore older data just because newer data have been generated.

-- be wary of consensus tree approaches; they may be worthwhile as part of the analysis, but it is probably best to combine all putative homologies into one matrix (perhaps with weighting if this can be independently justified).

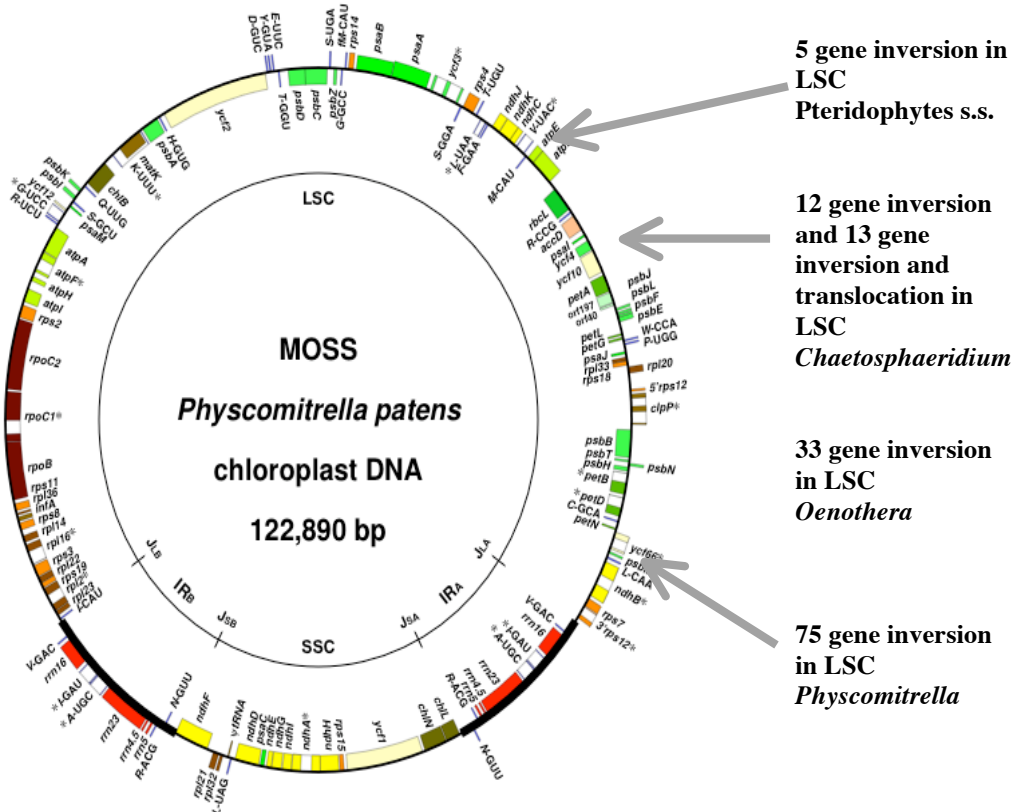
-- for reconstructing deep splits, it is much better to sequence portions of several different genes, scattered around the nuclear and organelle genomes, than it is to concentrate on extensive sequencing of a single gene (because of the problem of tight selective constraints on any one highly conserved region). Or for that matter, use morphology or genome structure.

-- it is probably better to break large surveys down into reasonable local analyses, to avoid spurious homoplasy (e.g., instead of putting all eukaryotes into one huge matrix, work on relationships within smaller, a priori justified monophyletic groups, and later link those groups together using archetypes: "compartmentalization" (Mishler, 2005).

-- molecular evolutionary studies and phylogeny reconstruction using molecules are two very different goals; for the former purpose, one should use phylogenies based on morphology (and other characters, perhaps including molecules -- but not the molecules that are being studied evolutionary).



Multiple gene inversion characters across Green Plants



D.G. Kelch, A. Driskell, and B.D. Mishler. 2004. Inferring phylogeny using genomic characters: a case study using land plant plastomes, *In* B. Goffinet, V. Hollowell, and R. Magill (eds.), Molecular Systematics of Bryophytes [Monographs in Systematic Botany 98], pp. 3-12. Missouri