Lecture Handout 25: MOLECULAR EVOLUTION AND PHYLOGENETICS 3

Final exam: 12/17/09 Thursday 8:00 AM,107 ASL

Comprehensive: will cover entire class. All previous plus current formula sheets will be attached to the Final.

For earlier sections, focus on (1) 2008/2009 exams, (2) problem sets; (3) lectures

For current section, focus on problem sets 9, 10, 11; 2008 Final; and lectures. ALSO: **chapter problems in Hedricks:**
Ch 10, Q 2 (D and D' only); Ch. 11, Q 4 and 5

Questions? Q/A in 404 ASL (or 432 ASL) on:
Tuesday 12/15, 3:30-5:30
Wednesday 12/16, 10:30-12:30
Wednesday 12/16, 3:30-5:30

The goal of **tree building methods** is to convert information in sequences into trees

Type of method I, distance or discrete

- **Distance methods** first convert aligned sequences into a pairwise distance matrix, then use the matrix to build the tree
- **Discrete methods** consider each nucleotide site (or some function of each site) directly on trees

Distance matrix

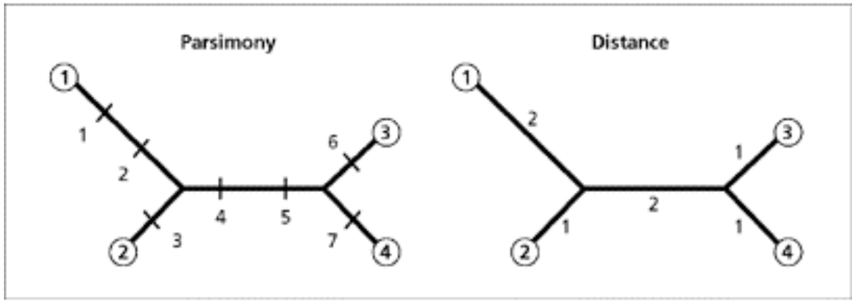
Alignment

1	T	T	A	T	T	A	A
2	A	A	T	T	T	A	A
3	A	A	A	A	A	T	A
4	A	A	A	A	A	A	T

Distance Matrix

$$\begin{array}{cccc}
1 & & & \\
2 & \mathbf{3} & & \\
3 & \mathbf{5} & \mathbf{4} & \\
4 & \mathbf{5} & \mathbf{4} & \mathbf{2} \\
 & 1 & 2 & 3
\end{array}$$

Parsimony vs. distance



Page, Holmes
Molecular Evolution

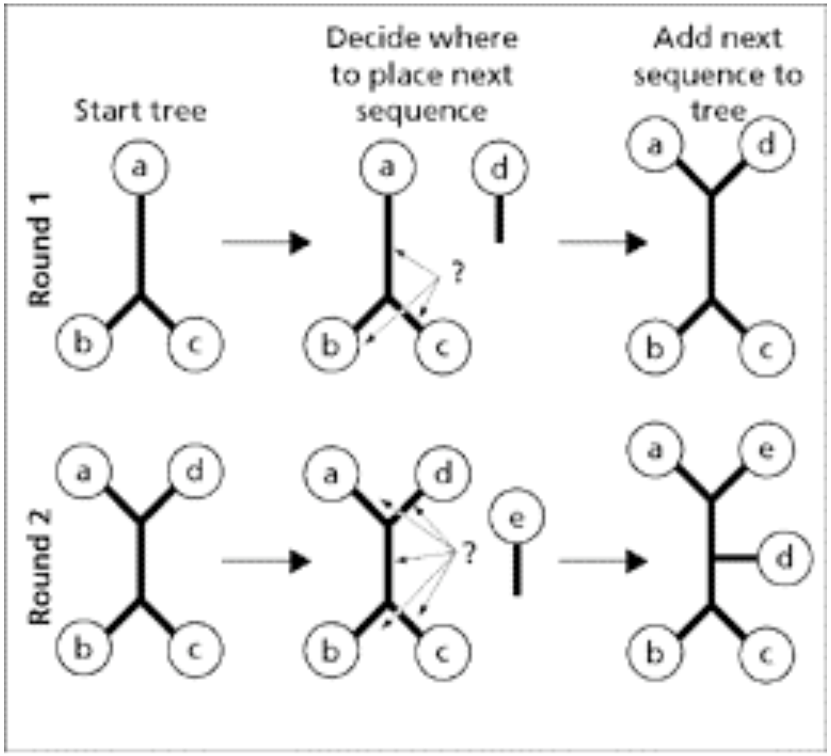
- Both trees have the same topology and branch lengths
- **However:** parsimony identifies which sites have contributed to lengths
  - Once we convert data into distance matrix, we lose information

Kinds of methods II:
By tree building algorithms

- **Clustering algorithms:** follow a set of steps until arriving to a tree

- **Optimality criteria:** choose among the set of all possible trees

Clustering methods
Example


Start tree
Decide where to place next sequence
Add next sequence to tree
Round 1
Round 2
a
b
c
d
e
?

Page, Holmes
Molecular Evolution

- **Round 1:** The tree is constructed starting with a tree for 3 sequences
- **Round 2:** Follow by addition of each of the remaining sequences to the most similar node

Clustering methods
Evaluation

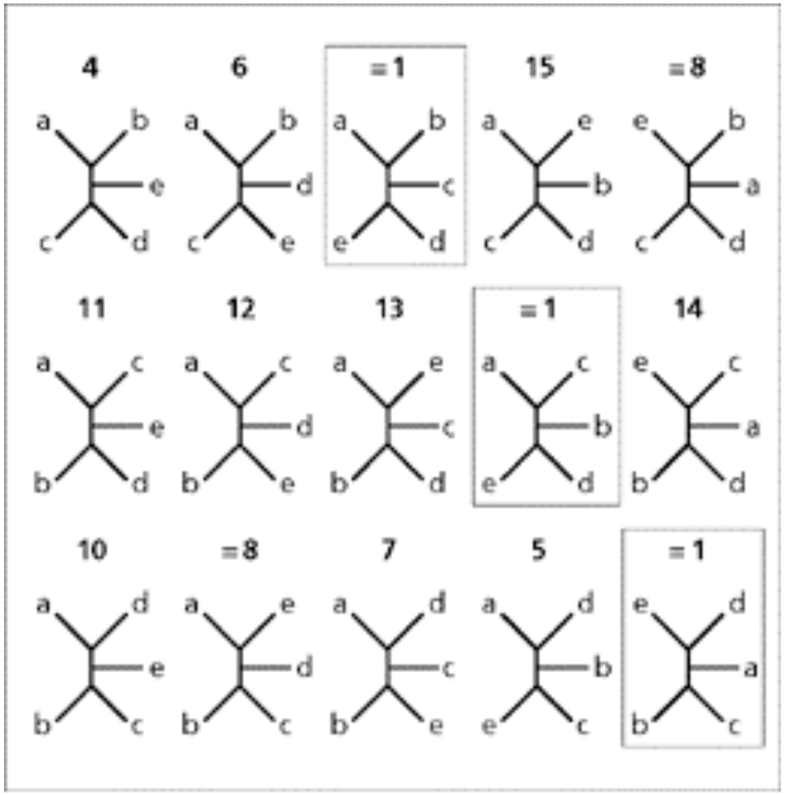
- **Advantages:**

1. Easy to implement, result in very fast computer programs
2. Always produce a single tree

- **Limitations:**

1. Result may depend on the order in which we add sequences to the tree
2. When the multiple hypotheses are possible, there is hardly any way to evaluate them

Optimality criteria methods
Example



b Page, Holmes
Molecular Evolution

- Based on evaluating all possible trees
- Each tree is assigned a score based on a specific measure
- Trees are ranked and the best tree is used to estimate the phylogeny
  - In the current example there are three optimal trees

Optimality criteria methods
Evaluation

- **Advantages:**
  1. Require an explicit function that relates data & tree (a model)
  2. Can evaluate the quality of any tree to discriminate between the competing hypotheses

- **Limitations:**
  1. Computationally very expensive: requires a search among all possible trees (or the use of a heuristic method--a short cut for searching tree space)
  2. Requires an optimality criterion by which to judge trees

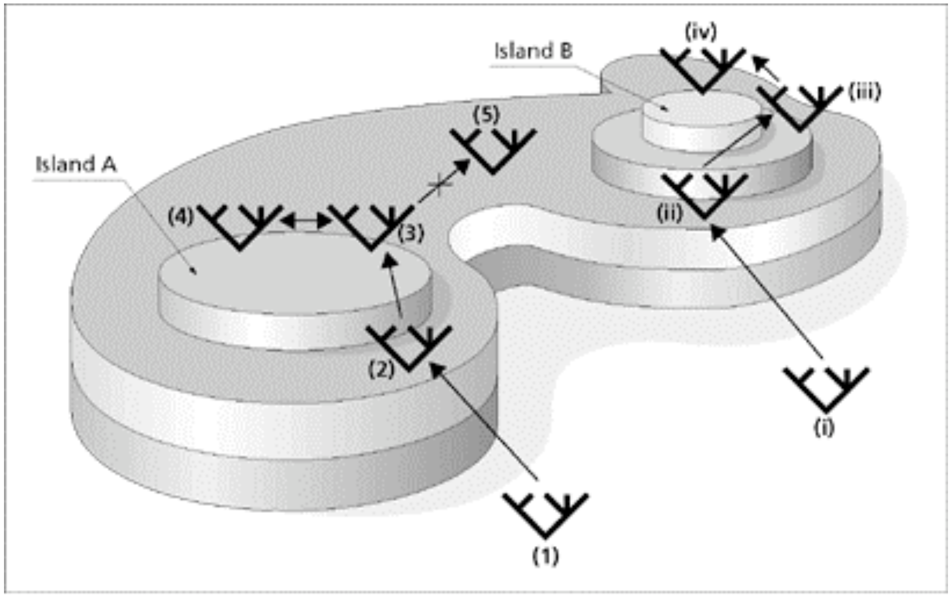
Heuristic methods

- Quick-and-dirty methods to deal with the computational slowness of calculation

- **Typical heuristic strategy**:
  - Start with a tree from a clustering algorithm (or at random)
  - Rearrange the tree keeping any rearrangements that produce a better tree
- Resembles "hill climbing" – the algorithm that will climb the highest hill => the most optimal tree

Heuristic methods

The problem of multiple islands

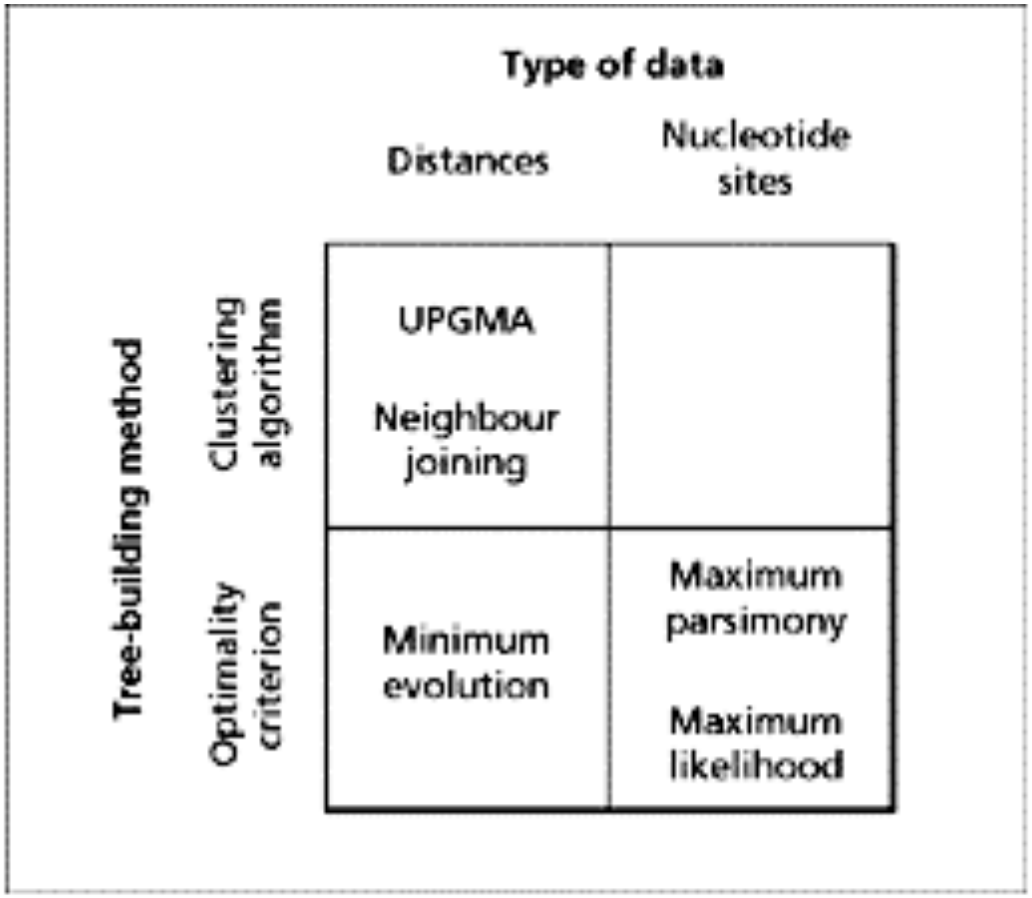


Page, Holmes
Molecular Evolution

If the set of possible trees contains more than one island, heuristic methods may land on an suboptimal island:

- A tree algorithm that started at tree **1** will succeed in finding trees **3** and **4**
- A tree algorithm starting at tree **i** will find tree **iv** which is the optimal tree
- Avoid getting stuck either by branch-swapping, or using multiple runs each with different starting trees.

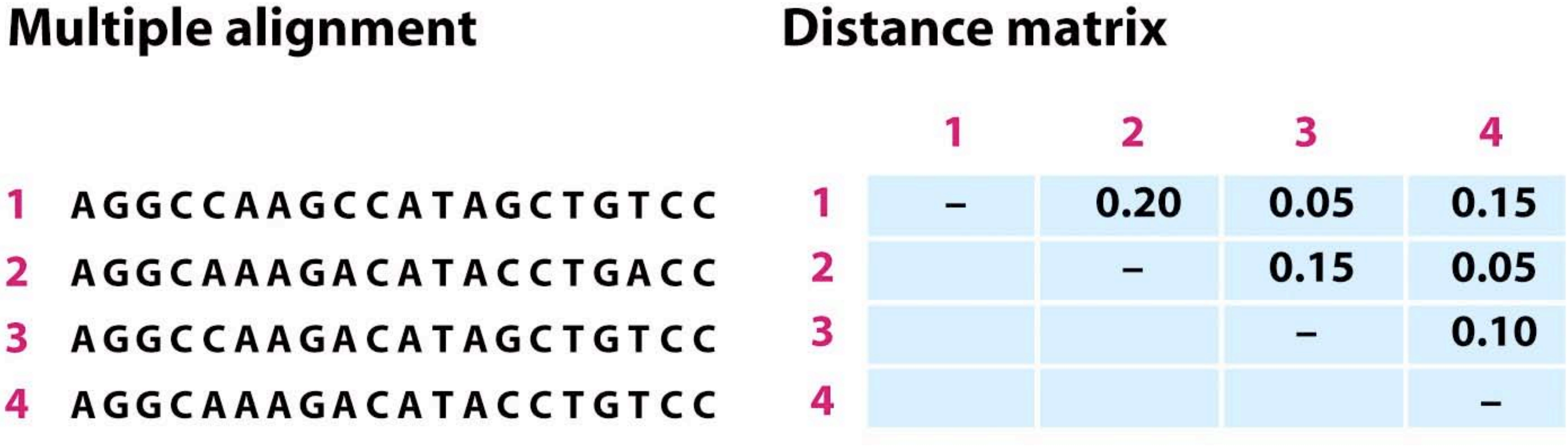
The ideal method?



		Type of data	
		Distances	Nucleotide sites
Tree-building method	Clustering algorithm	UPGMA Neighbour joining	
	Optimality criterion	Minimum evolution	Maximum parsimony Maximum likelihood

Page, Holmes
Molecular Evolution

Distance method: matrix



Multiple alignment

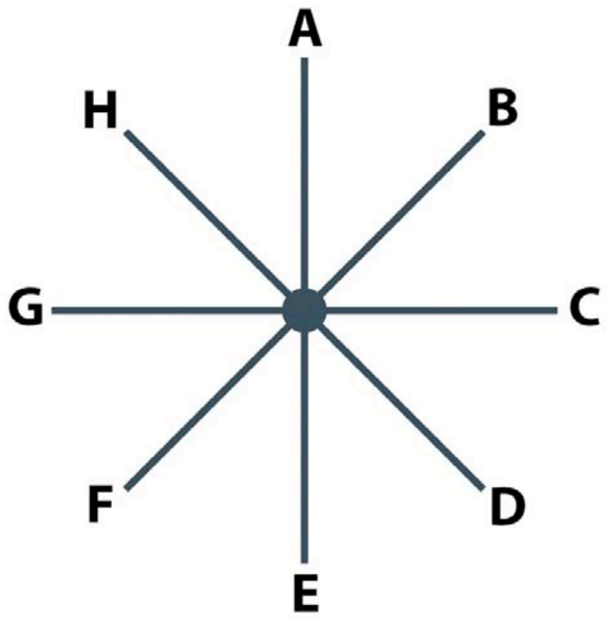
```
1  AGGCCAAGCCATAGCTGTCC
2  AGGCAAAGACATACCTGACC
3  AGGCCAAGACATAGCTGTCC
4  AGGCAAAGACATACCTGTCC
```

Distance matrix

	1	2	3	4
1	–	0.20	0.05	0.15
2		–	0.15	0.05
3			–	0.10
4				–

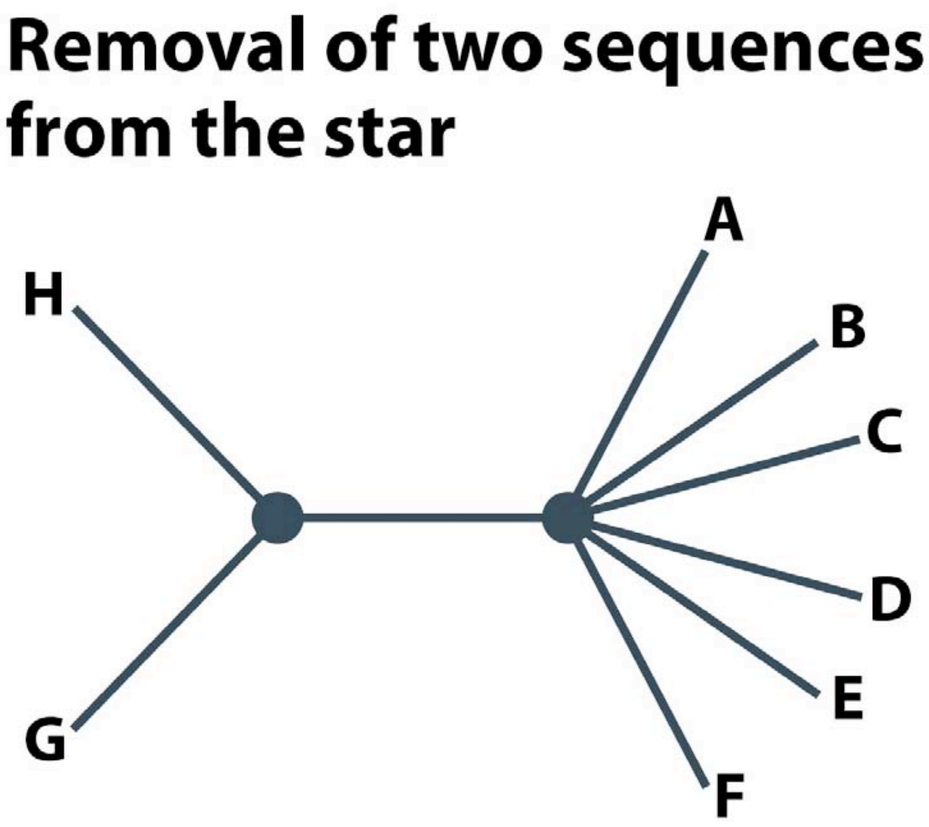
Distance methods
Neighbor-Joining Method

The starting point for the neighbor-joining method



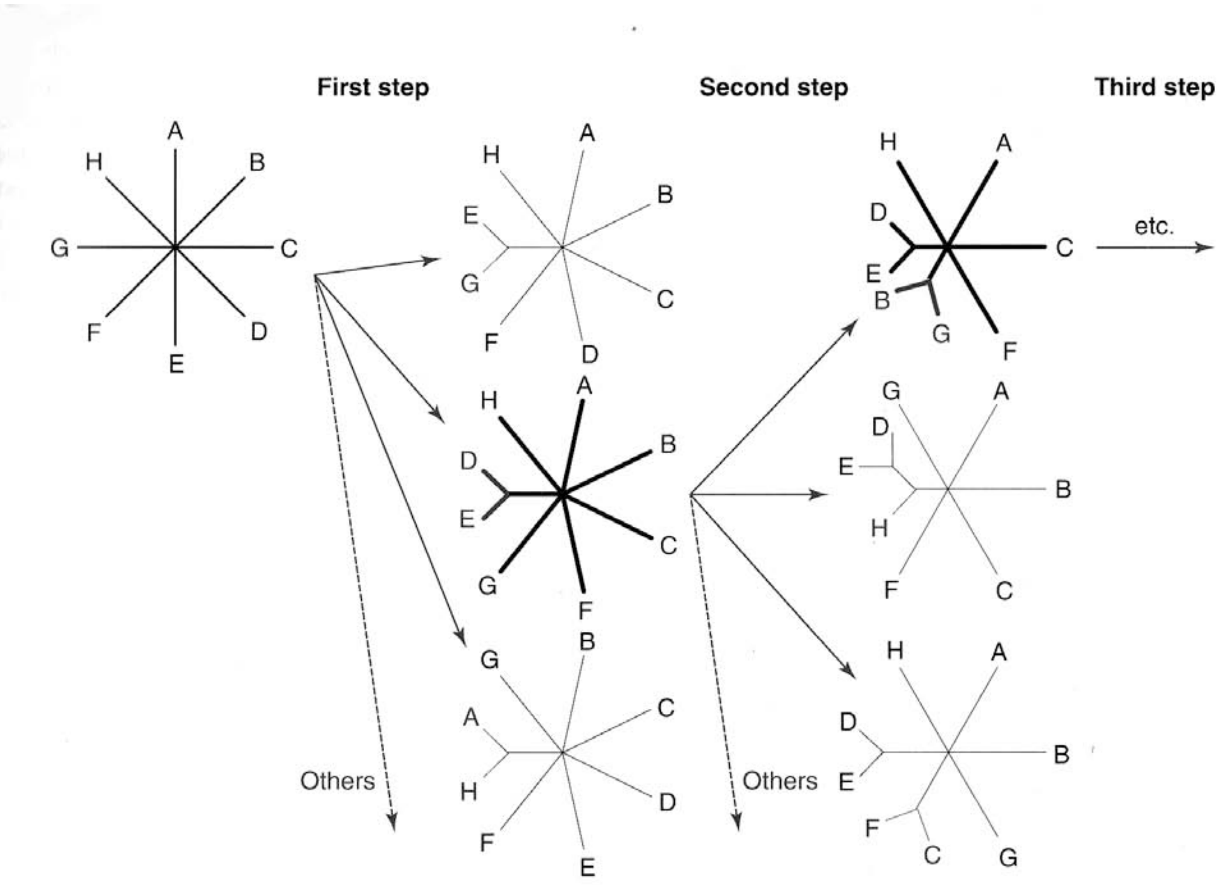
- Assume that there is only one internal node as a starting point
- Next, a pair of sequences is chosen at random and attached to a 2nd internal node

Distance methods
Neighbor-Joining Method



- The distance matrix is now used to calculate the total branch length in this new tree
- This is repeated until all possible pairs are examined and **optimal** tree is chosen

Distance methods
Neighbor-Joining Method



- This process of pair selection and matrix calculation is now repeated, each step with one less branch

Minimum evolution

- A distance method (like NJ), but does not use a clustering algorithm
- Rather, ME searches across tree space (all possible trees or using a heuristic approach)
- The sum of all branches is the length **L** of the tree, which is used as the optimality criterion
- **ME tree is the tree that minimizes L**

Distance methods
Objections

- Summarizing a set of sequences by a pairwise distance matrix leads to information loss
  - Can’t track individual sites

- Branch lengths estimated may not be evolutionary meaningful
  - Sometimes calculates fewer substitutions than are biologically possible (such as a negative branch length)

Discrete methods

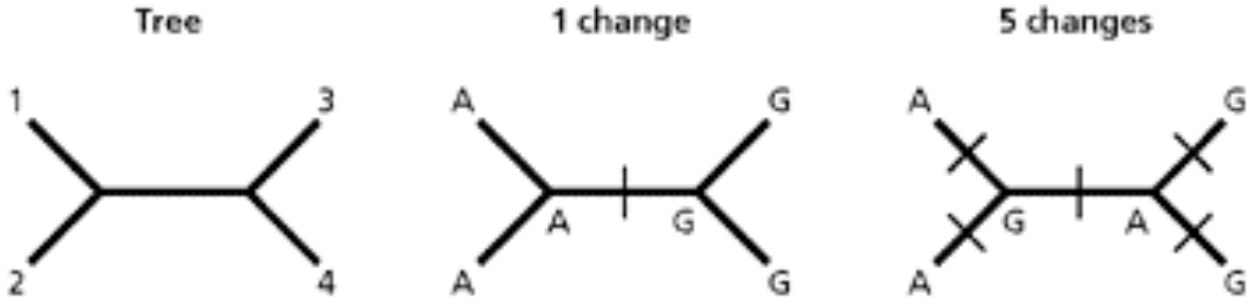
- Operate directly on sequences, not on distance matrices
- Avoid loss of information that occurs
- Two major methods:
  - **Maximum parsimony (MP)**
  - **Maximum likelihood (ML)**

Maximum parsimony

- Data for MP are the individual nucleotide sets
- For each site, the goal is to reconstruct an evolutionary tree with the smallest number of changes
- Note: MP does not generally use a correction model

1	A	T	A	T	T
2	A	T	C	G	T
3	G	C	A	G	T
4	G	C	C	G	T

Maximum parsimony
Example



- The unrooted tree **((1,2),(3,4))** and two possible reconstructions of the evolution of the first site

- In each reconstruction, two internal nodes are postulated
- One reconstruction requires one change, the other 5 changes

Parsimony
Length of a tree

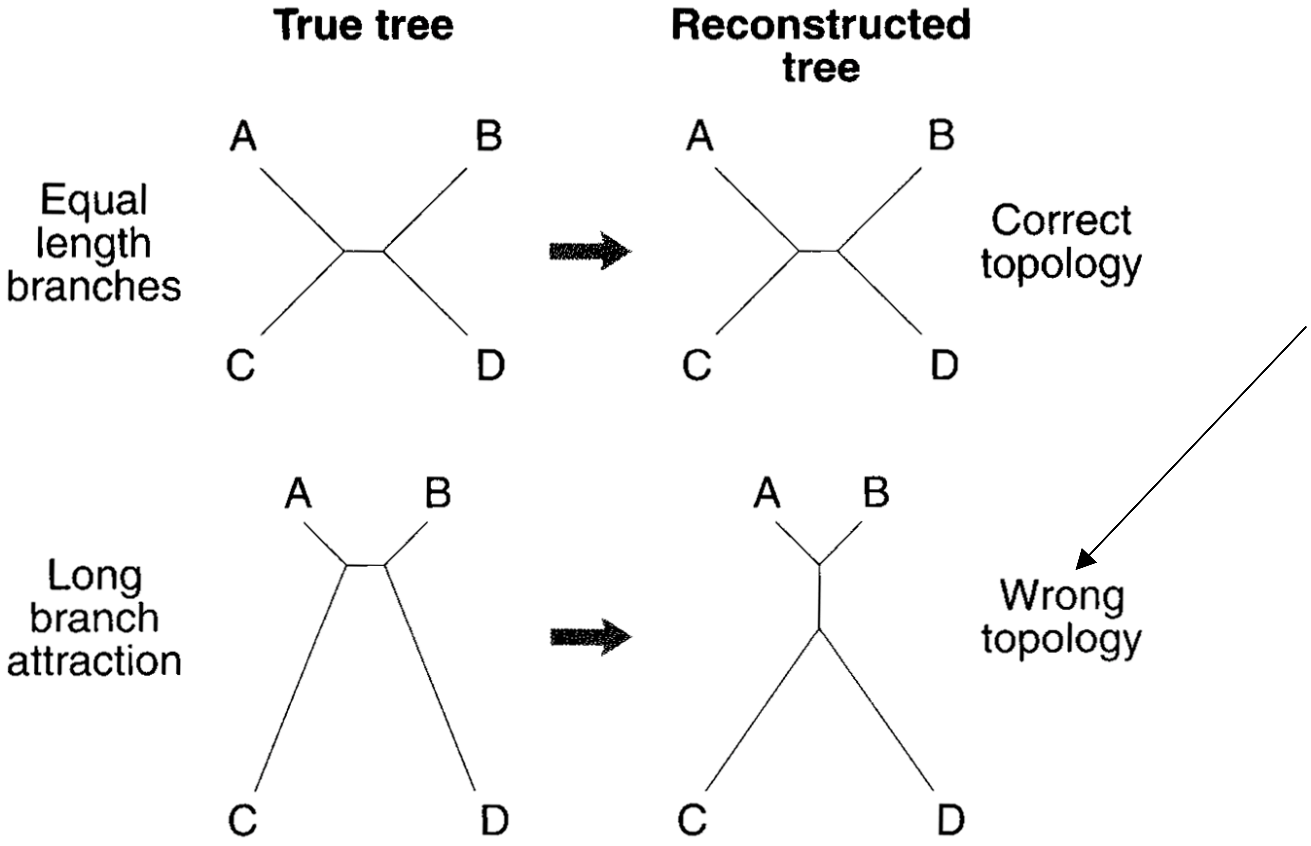
- The total number of evolutionary changes on a tree (tree length) L is the sum of the number of changes at all branches (both internal and terminal branches):

Maximum parsimony
Justifications

- Two arguments for **why** it can choose the best tree:

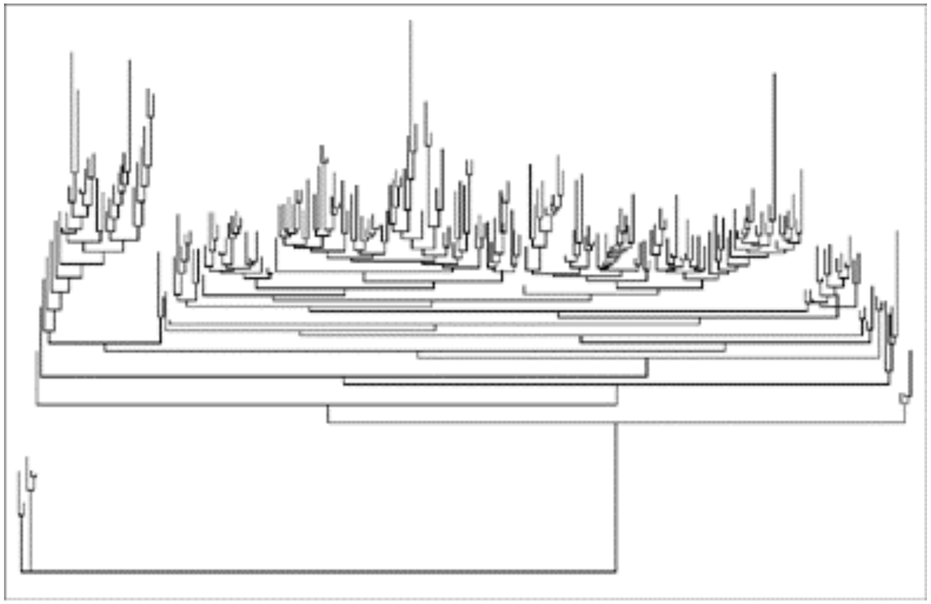
1. Maximizes the similarity that can be interpreted as homology
2. Based on the assumption that evolutionary change is rare, so that minimum change should represent the most likely phylogeny

Maximum parsimony
Long branch attraction problem



Maximum parsimony
Long branch attraction problem

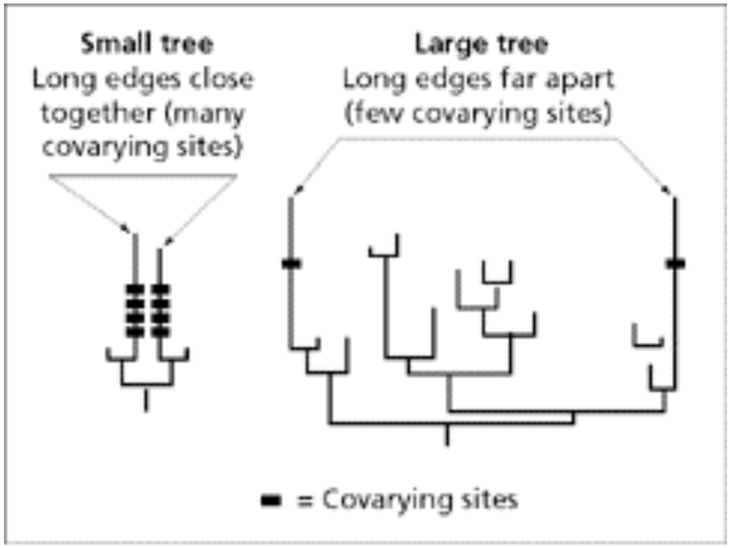
- In the case of large phylogenies, there is less of a problem
- Therefore, one strategy is to add more sequences to break the long branches …



Page, Holmes
Molecular Evolution
Blackwell Science

Maximum parsimony
Long branch attraction problem

- The problem seems to be not in the length, but in the similarity of substitutions between the branches
  - "fooling" the program into joining them together
- By adding sequences, we separate long branches so fewer of the changes appear homoplasious


Small tree
Long edges close together (many covarying sites)
Large tree
Long edges far apart (few covarying sites)
▬ = Covarying sites

Page, Holmes
Molecular Evolution

Blackwell Science

Maximum likelihood

- **Reminder:** the principle of likelihood is that the explanation that makes the observed outcome more likely is one to be preferred
- Given data (D) & hypothesis (H)

$$\text{Likelihood}=\Pr(D|H)$$

Obtaining maximum likelihood

- **Requires three elements:**
  1. Tree: topology and branch lengths
  2. Model of sequence evolution
  3. Observed data
- **Aims to find:**
  1. For a given tree topology, the branch length that makes data most likely
  2. Which tree of all possible trees is the most likely

Maximum likelihood
Justifications

1. Can readily exploit information present in the dataset (Ts:Tv ratios, nucleotide frequencies, etc.)
2. Gives exact probability for each potential tree, which makes comparisons across them straightforward

Maximum likelihood
Objections

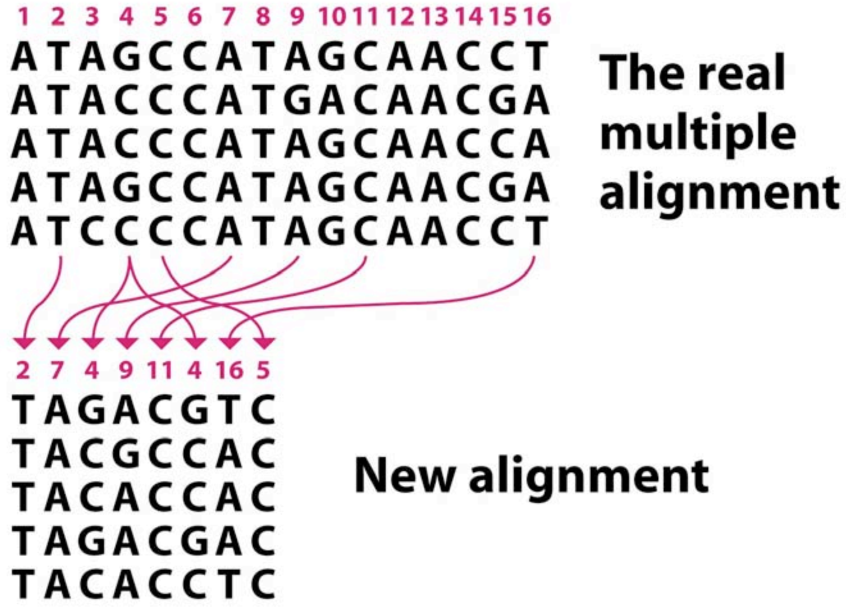
1. Requires an explicit model of evolution (both strength and weakness)
   - Where did the probabilities in the model come from?
2. Computationally time consuming:
   1. Has to work though all possible phylogenies
   2. Has to make detailed estimate and calculations for each

Was little used until recently

The bootstrap: uses pseudoreplicates to provide an estimate of the robustness of the relationships in the inferred phylogeny (for NJ, ME, MP, ML etc.)

- Given the 896 nucleotide sites, we could generate a pseudoreplicate by sampling at random and with replacement from the original dataset
  - Some of the sites will be represented more than once
  - Some will not be presented at all

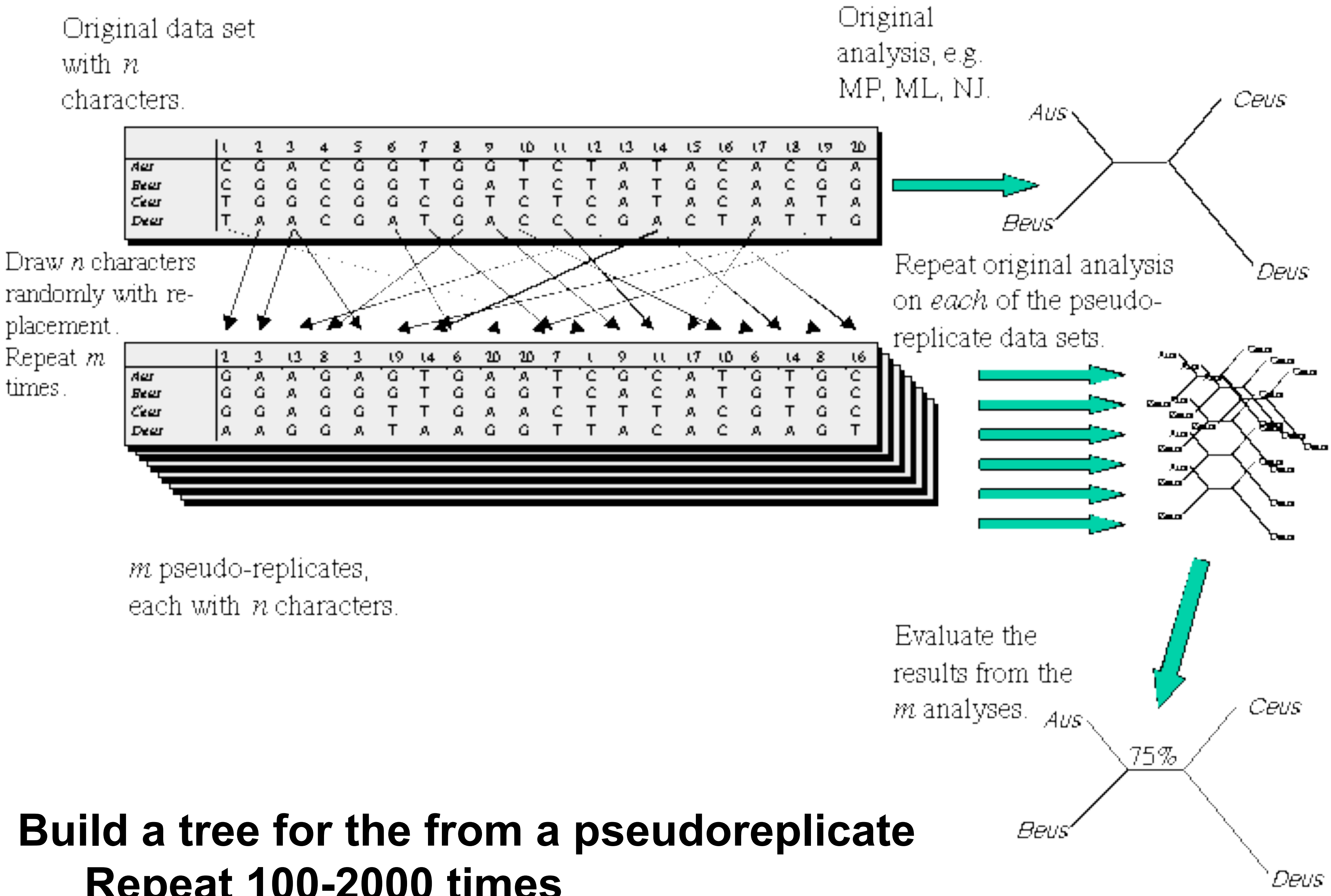
Example:



ATAGCCATAGCAACCT
ATACCCATGACAACGA
ATACCCATAGCAACCA
ATAGCCATAGCAACGA
ATCCCCATAGCAACCT

TAGACGTC
TACGCCAC
TACACCAC
TAGACGAC
TACACCTC

Figure 19-14 Genomes 3 (© Garland Science 2007)



Original data set with *n* characters.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Aer	C	G	A	C	G	G	T	G	G	T	C	T	A	T	A	C	A	C	G	A
Beer	C	G	G	C	G	G	T	G	A	T	C	T	A	T	G	C	A	C	G	G
Ceer	T	G	G	C	G	G	C	G	T	C	T	C	A	T	A	C	A	A	T	A
Deer	T	A	A	C	G	A	T	G	A	C	C	C	G	A	C	T	A	T	T	G

Original analysis, e.g. MP, ML, NJ.

Draw *n* characters randomly with re-placement. Repeat *m* times.

	2	3	13	8	3	19	14	6	20	20	7	1	9	11	17	10	6	14	8	16
Aer	G	A	A	G	A	G	T	G	A	A	T	C	G	C	A	T	G	T	G	C
Beer	G	G	A	G	G	G	T	G	G	G	T	C	A	C	A	T	G	T	G	C
Ceer	G	G	A	G	G	T	T	G	A	A	C	T	T	T	A	C	G	T	G	C
Deer	A	A	G	G	A	T	A	A	G	G	T	T	A	C	A	C	A	A	G	T

Repeat original analysis on *each* of the pseudo-replicate data sets.

m pseudo-replicates, each with *n* characters.

Evaluate the results from the *m* analyses.

Build a tree for the from a pseudoreplicate

Repeat 100-2000 times

Bootstrap value

- In practice, **bootstrap values** can be assigned to each internal node in the original tree:
  - This value being the **number of times that the branch pattern seen at that node was reproduced in the pseudoreplicate trees**
  - If the bootstrap value is **greater than 70/100,** then we can assign a reasonable degree of confidence to the topology at that particular internal node (maybe)

FINAL TOPICS: Quantitative Trait Loci (QTLs) and other topics.

Quantitative trait: phenotype determined by many genes (polygenic), e.g., body size in livestock; percentage of oil in maize. Quantitative phenotypes are often distributed on a *continuous* scale.

Heritability: the proportion of phenotypic variance that is genetically determined.

Quantitative trait loci, QTLs: genes or loci that affect quantitative traits, e.g. in agricultural species or human medicine. A **candidate gene** approach may be useful in some cases.

Quantitative trait loci, QTLs, example at UI: Oil and protein content in corn (maize) are affected by 56 and 123 genes, respectively. But this estimate assumes: no linkage disequilibrium, no epistasis, additivity and equal effect of each gene

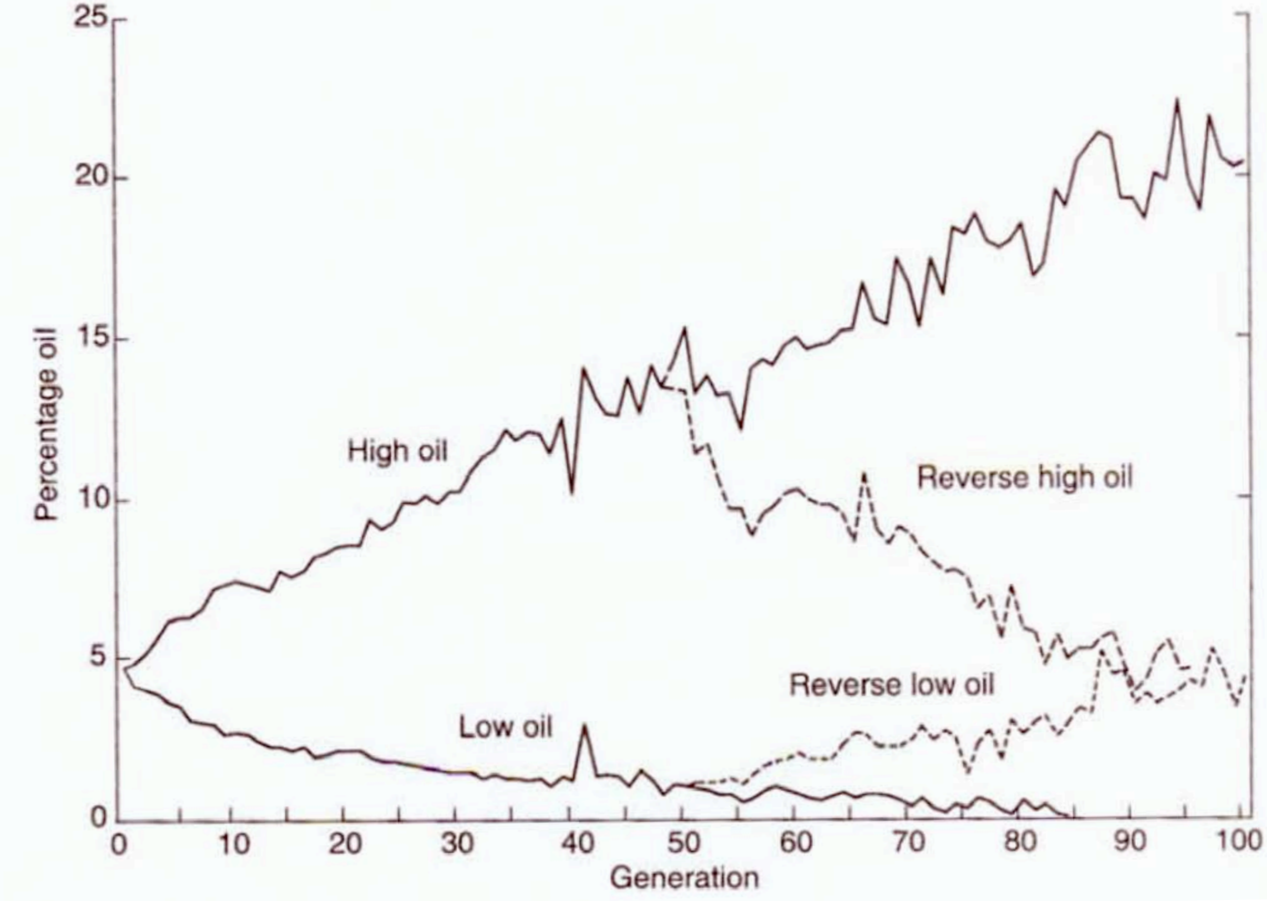


Figure 1.21. The percentage of oil in maize populations selected for 100 generations for either high oil content or low oil content (from Dudley and Lambert, 2004). Also given are selection responses in reverse selection lines for both the high and the low selection lines (broken lines).

Quantitative Trait Loci (QTLs)

May be identified by crossing two lines that differ in phenotype, examine segregation among descendants of a marker locus linked to a QTL

Interval mapping: multiple pairs of adjacent markers of known position are examined for association with QTL. A $\log_{10}$ likelihood score is determined: **LOD score (logarithm of odds)**, an estimate of the linkage between a trait and a marker. LOD score of 3 means the likelihood of observing result if *not* linked is less than 1 in 1000. But: requires correction for large number of tests (Bonferroni correction or permutation test).

Quantitative Trait Loci (QTLs): LOD score for loci influencing height in humans

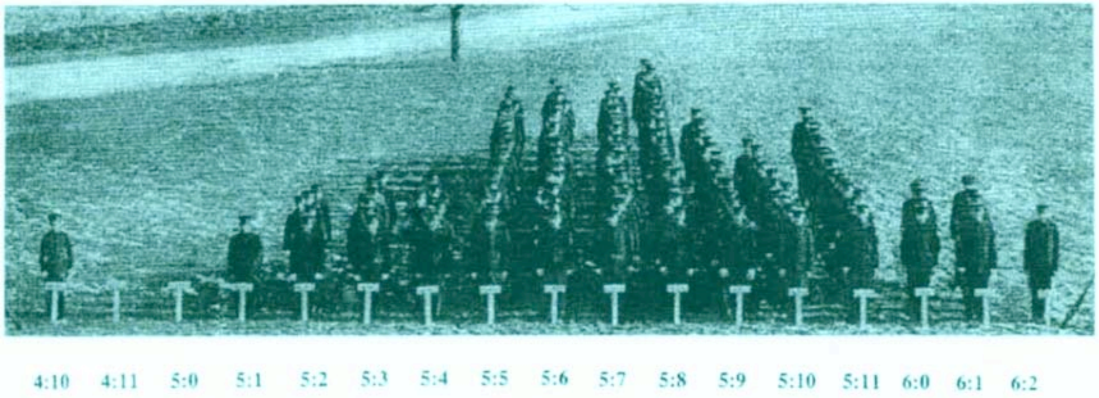


Figure 1.4. The distribution of heights, in inch categories, of male students at Connecticut Agricultural College (from Crow, 1997). (Courtesy of Albert Blakeslee, 1914.)

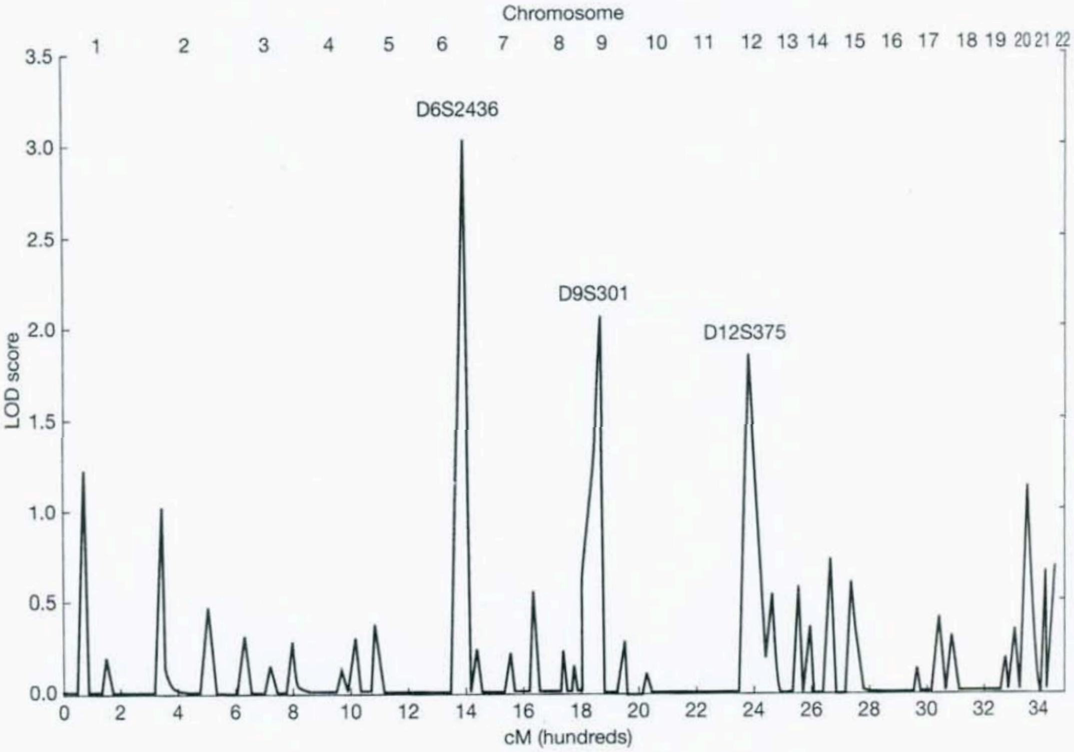


Figure 11.19. The results of a genome-wide scan for genes influencing height in humans as indicated by the LOD score over the 22 different autosomal chromosomes (across top) and over the cumulative genetic map (across bottom) (Xu *et al.*, 2002). The three peaks are identified by the microsatellite markers with highest LOD scores.

Quantitative Trait Loci (QTLs): **Haplotype Mapping**

- Hap Map Project identified SNPs across human individuals and populations

- Identified the "**tag SNPs"** that are especially useful for differentiating among different haplotypes

- Parts of the genome exist as **haplotype blocks**, regions with little evidence of recombination, separated by regions of low linkage disequilibrium or recombination hot spots. "Mosaic genome"

- **Association mapping**: tag SNPs identify haplotype blocks associated with disease phenotypes

OTHER TOPICS (unrelated to QTLs): use of molecular markers

Extra-pair paternity: fertilization of females by a male other than her pair-bond are more common in bird species than predicted by behavioral observations

Paternity exclusion: candidate males can be excluded as fathers when alleles present in the mother and offspring are known for polymorphic loci

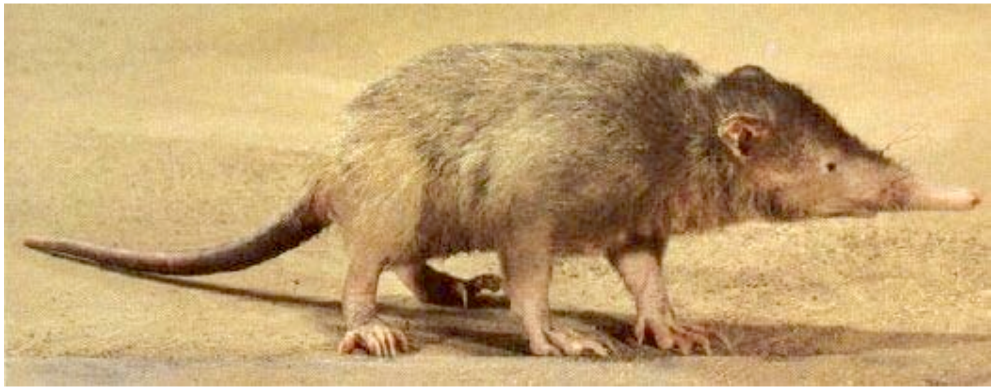
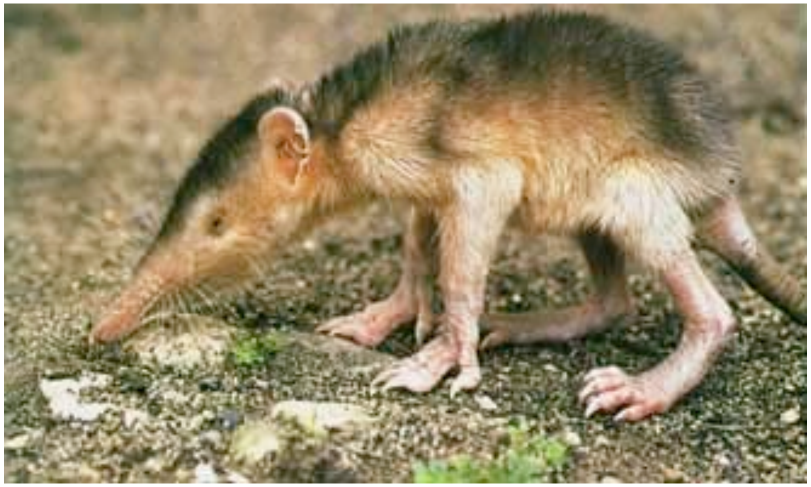
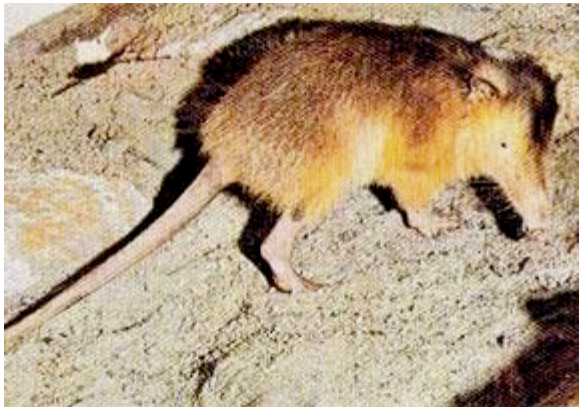
Both individual identity (eg FBI’s 13 CODIS highly polymorphic STR loci) and paternity may be assigned using likelihood and Bayesian approaches

OTHER TOPICS (unrelated to QTLs)

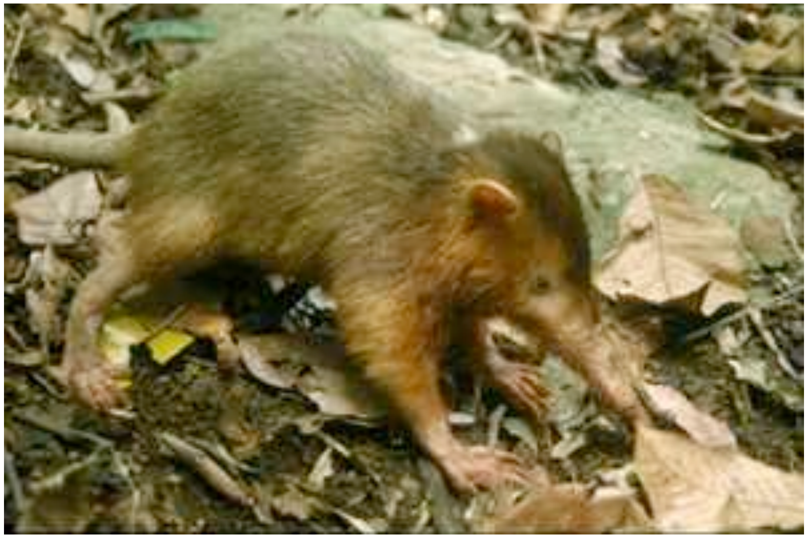
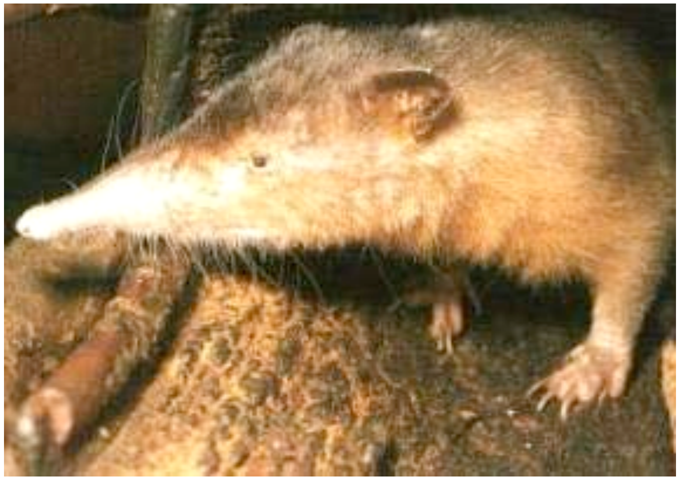
Phylogeography: study of the principles and processes governing the geographical distribution of genealogical (phylogenetic) lineages

Dispersal: active or passive movement of organisms from an ancestral origin to a new geographic area

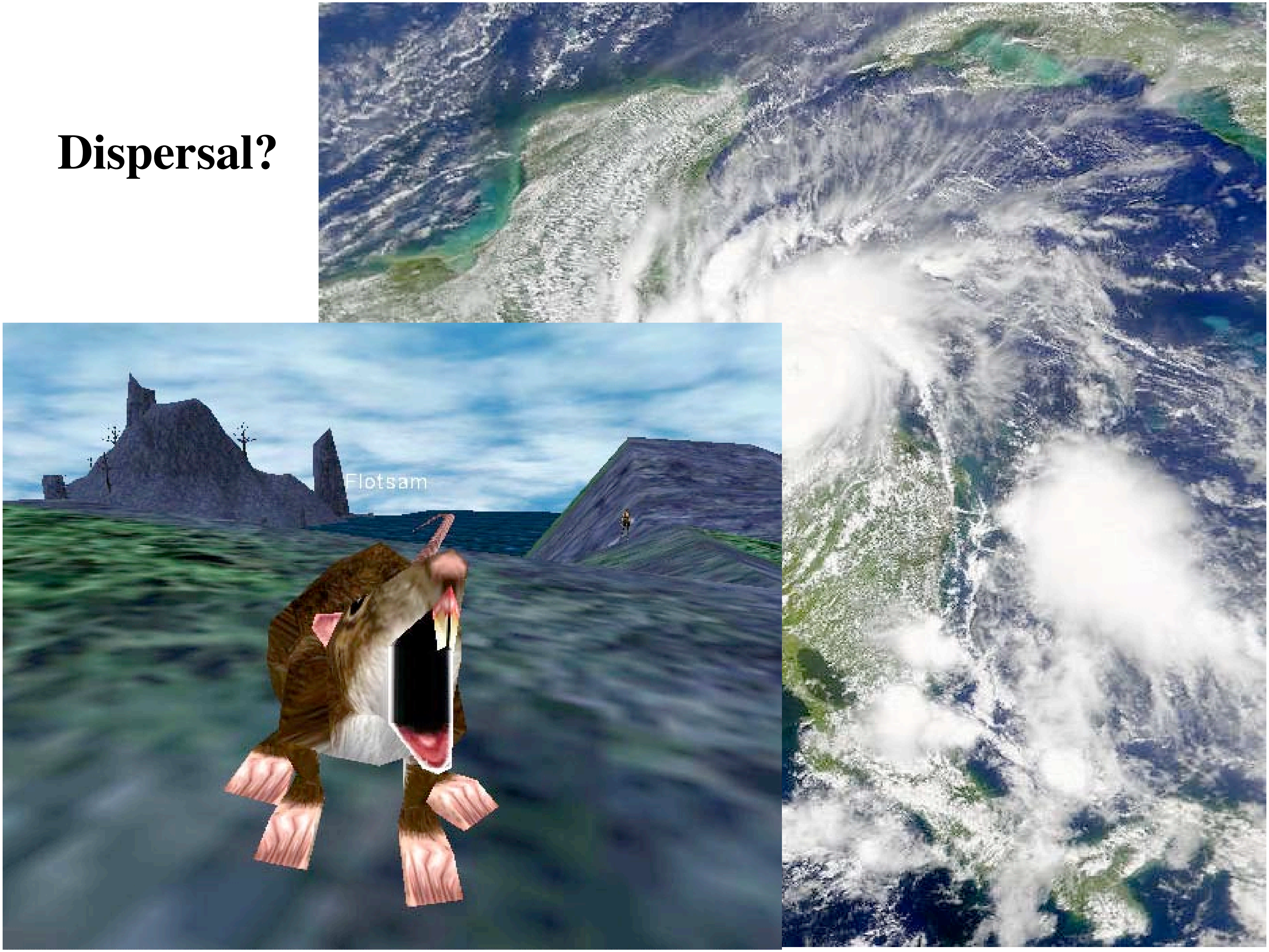
Vicariance: the separation of a previously continuous organismal range by past geological or environmental events

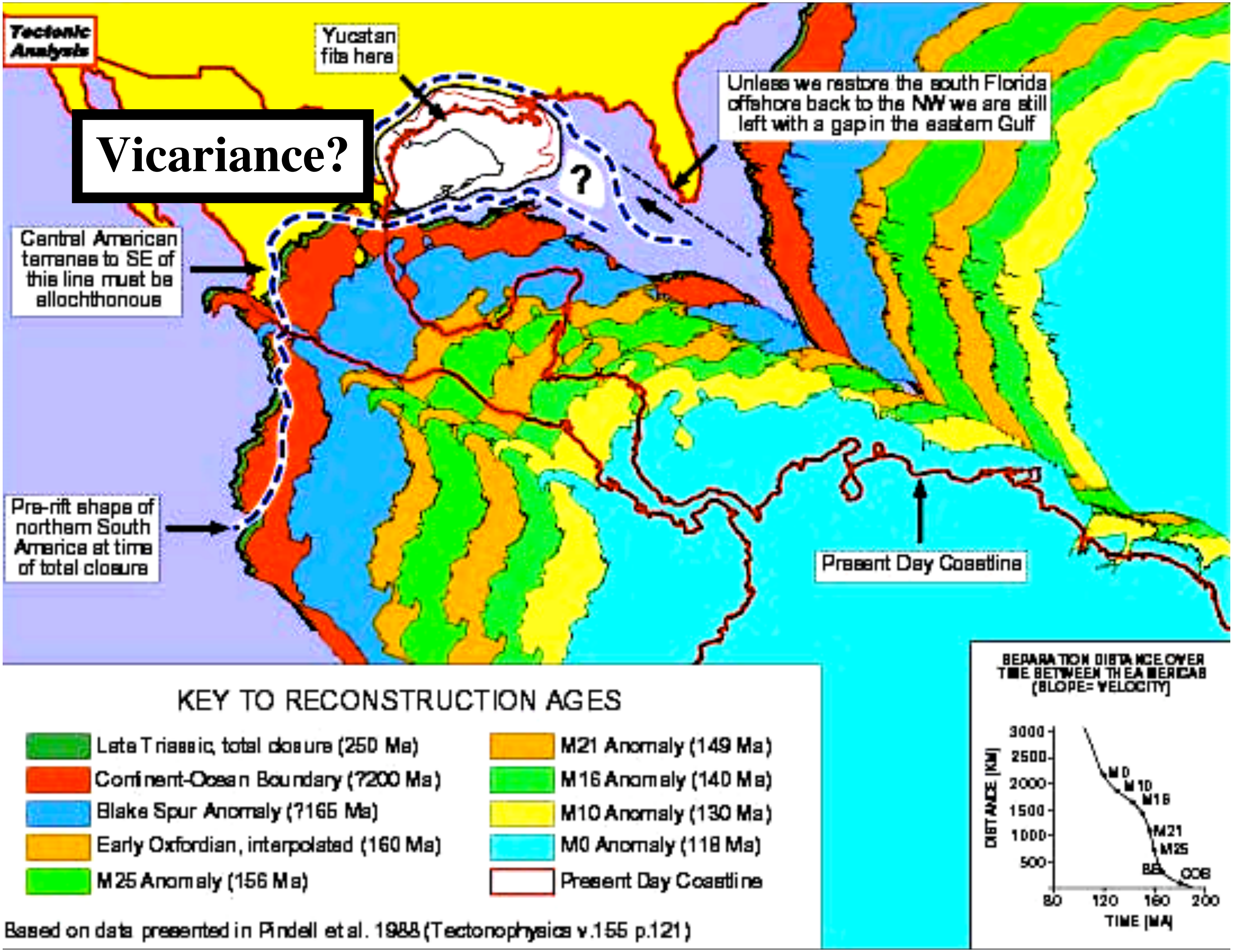
Solenodon paradoxus

Dispersal?



Vicariance?

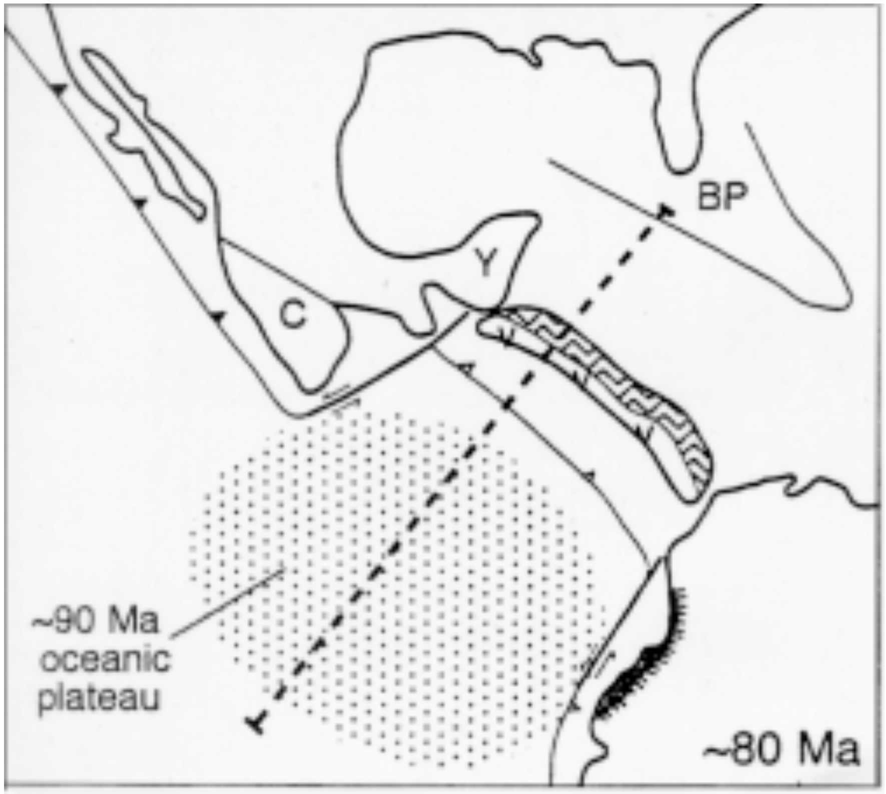

Tectonic Analysis
Yucatan fits here
Unless we restore the south Florida offshore back to the NW we are still left with a gap in the eastern Gulf
?
Central American terranes to SE of this line must be allochthonous
Pre-rift shape of northern South America at time of total closure
Present Day Coastline
SEPARATION DISTANCE OVER TIME BETWEEN THE AMERICAS (SLOPE= VELOCITY)
DISTANCE (KM)
3000
2500
2000
1500
1000
500
M0
M10
M16
M21
M25
BS
COB
80
120
160
200
TIME (MA)

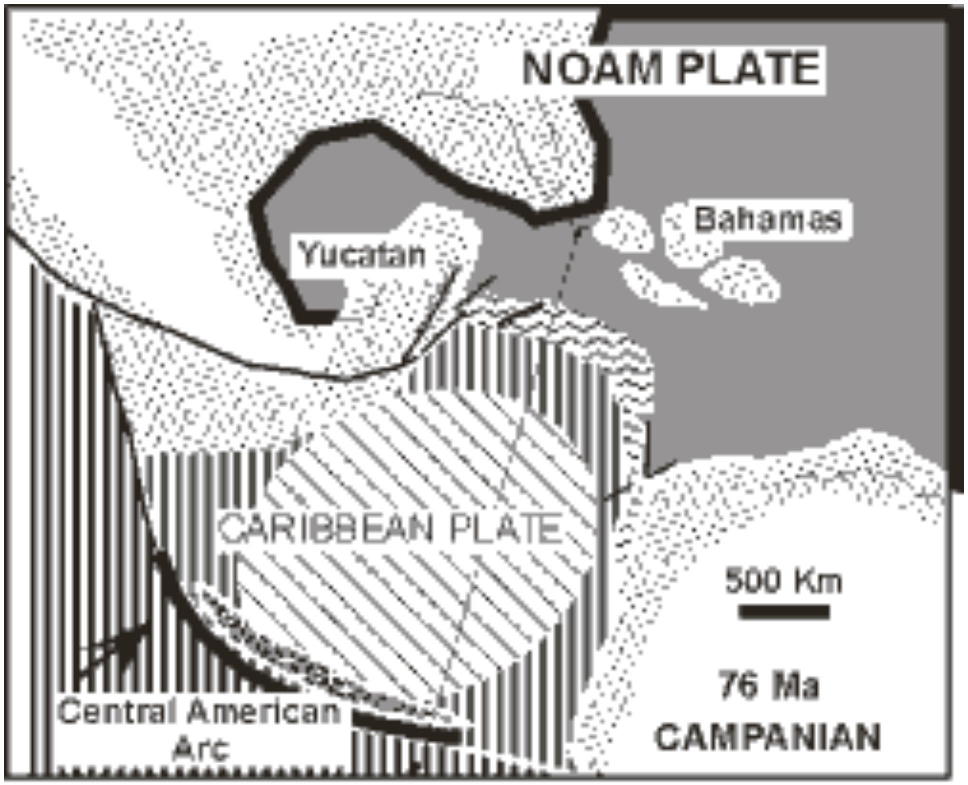
KEY TO RECONSTRUCTION AGES

🟩	Late Triassic, total closure (250 Ma)	🟧	M21 Anomaly (149 Ma)
🟥	Continent-Ocean Boundary (?200 Ma)	🟩	M16 Anomaly (140 Ma)
🟦	Blake Spur Anomaly (?165 Ma)	🟨	M10 Anomaly (130 Ma)
🟧	Early Oxfordian, interpolated (160 Ma)	🟦	M0 Anomaly (118 Ma)
🟩	M25 Anomaly (156 Ma)	□	Present Day Coastline

Based on data presented in Pindell et al. 1988 (Tectonophysics v.155 p.121)

Vicariance?


BP
Y
C
~90 Ma
oceanic
plateau
~80 Ma

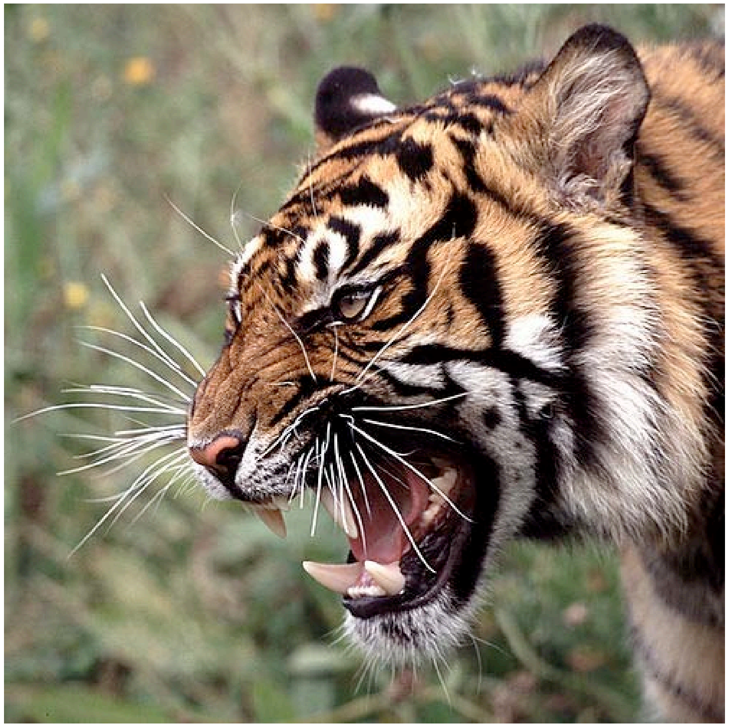

NOAM PLATE
Bahamas
Yucatan
CARIBBEAN PLATE
500 Km
Central American
Arc
76 Ma
CAMPANIAN

Evolutionarily significant units (ESUs): populations that are ecologically, historically or **genetically** unique enough to receive special protection.

Hedrick text cites Moritz in suggesting that **reciprocal monophyly** for mtDNA as a necessary criterion for ESUs but…

ESUs comprise a different concept than **subspecies**.

In at least one case--tigers, ESUs and subspecies appear to correspond

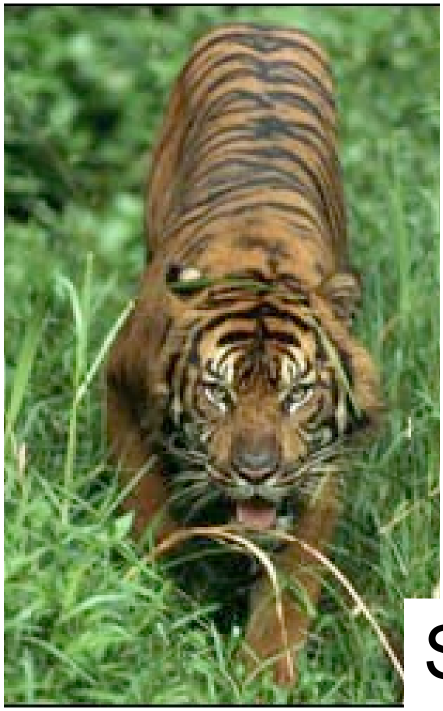


FIVE EXTANT TIGER SUBSPECIES

BENGAL

AMUR ("SIBERIAN")

INDOCHINESE

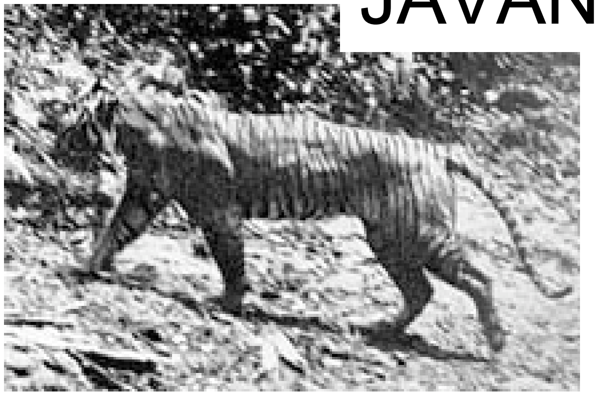


SUMATRAN

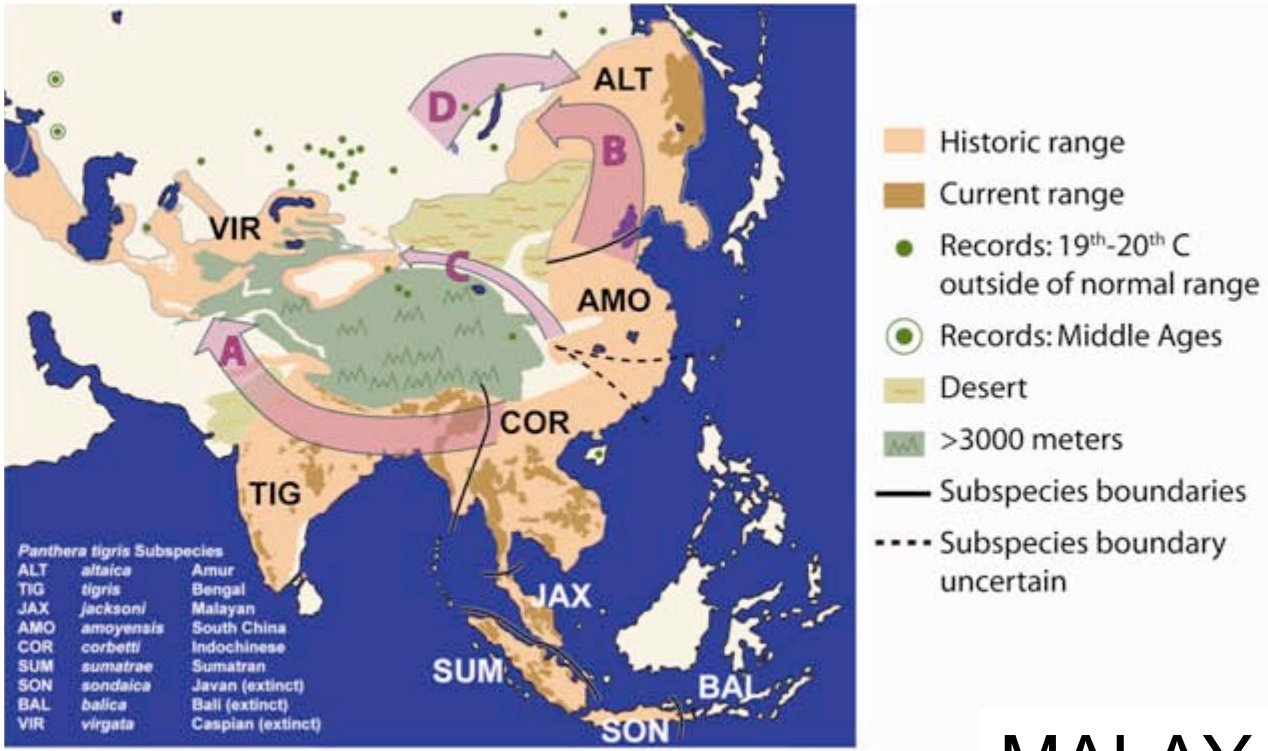
SOUTH CHINA

CASPIAN

JAVAN

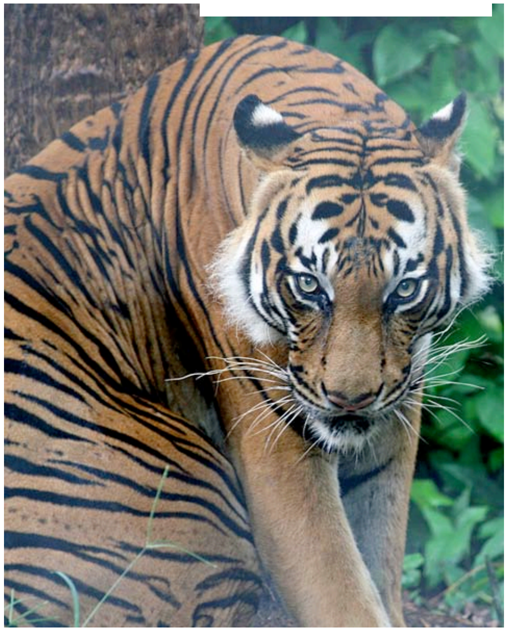


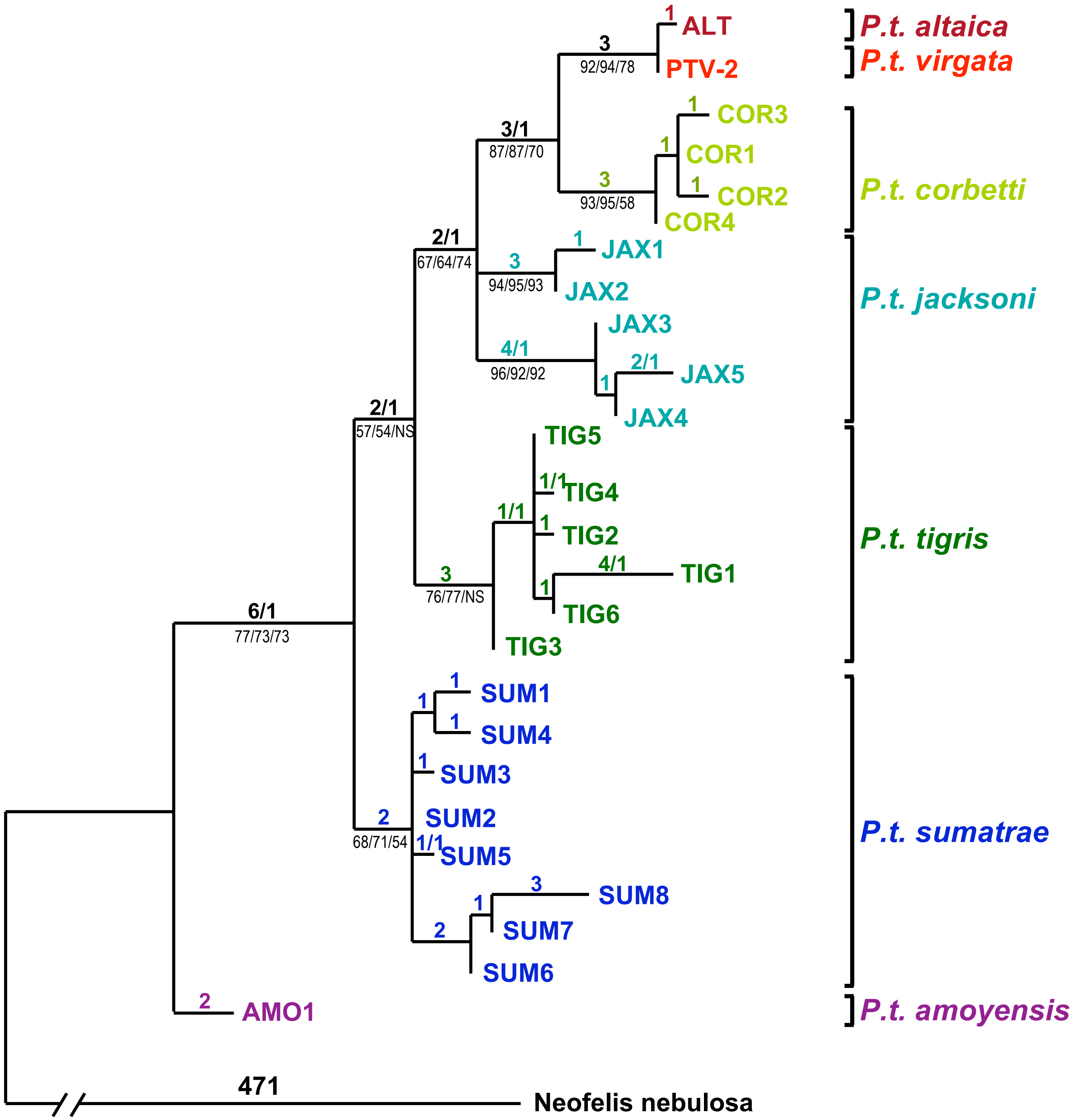
BALI



THREE EXTINCT TIGER SUBSPECIES PLUS A NEW EXTANT SUBSPECIES?

MALAY




1
ALT
3
92/94/78
PTV-2
1
COR3
3/1
87/87/70
1
COR1
3
93/95/58
1
COR2
COR4
2/1
67/64/74
1
JAX1
3
94/95/93
JAX2
JAX3
4/1
96/92/92
2/1
JAX5
1
JAX4
2/1
57/54/NS
TIG5
1/1
TIG4
1/1
1
TIG2
4/1
TIG1
3
76/77/NS
1
TIG6
6/1
77/73/73
TIG3
1
SUM1
1
1
SUM4
1
SUM3
2
68/71/54
SUM2
1/1
SUM5
3
SUM8
1
2
SUM7
SUM6
2
AMO1
471
Neofelis nebulosa
P.t. altaica
P.t. virgata
P.t. corbetti
P.t. jacksoni
P.t. tigris
P.t. sumatrae
P.t. amoyensis