DNA Sequence-based Measures of Genetic Variation

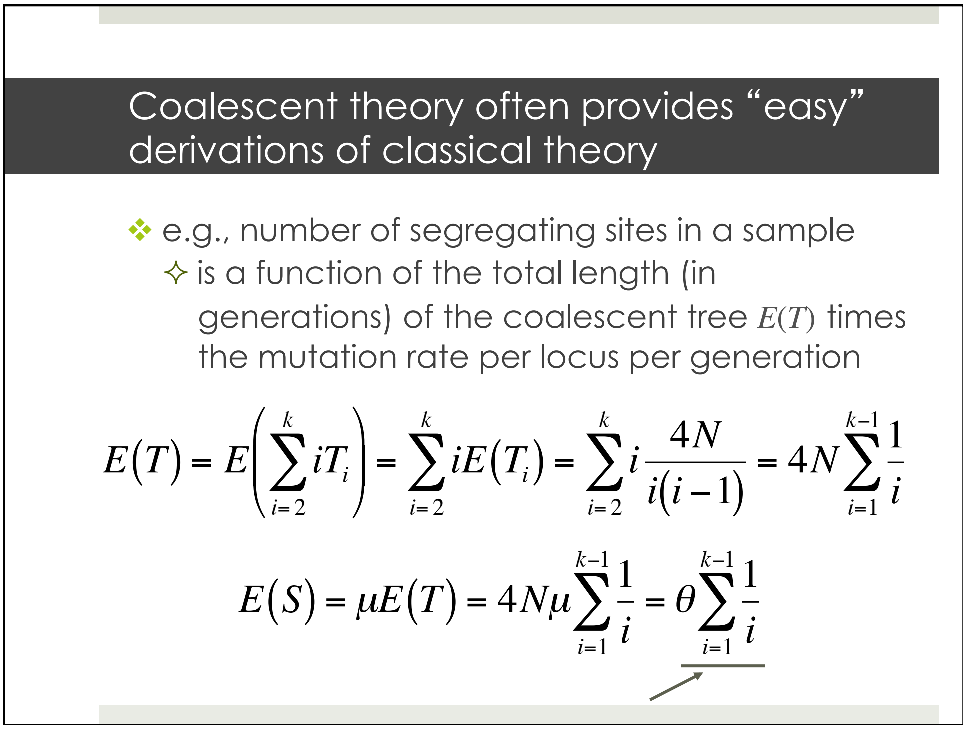
- S = number of segregating sites
- Π = average number of pairwise differences between sequences
- Π analogous to heterozygosity
- can derive theoretical expectations for both measures for an idealized, random breeding population (and also assuming an "infinite sites" model)…

Segregating sites

- expected number of segregating sites:

$$E(S)=\theta\left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots+\frac{1}{k-1}\right)$$

- where $\theta = 4N\mu$ and k = the number of sequences in the sample
- μ ("mu") is the per locus mutation rate = mutation rate per site per generation x length of sequence

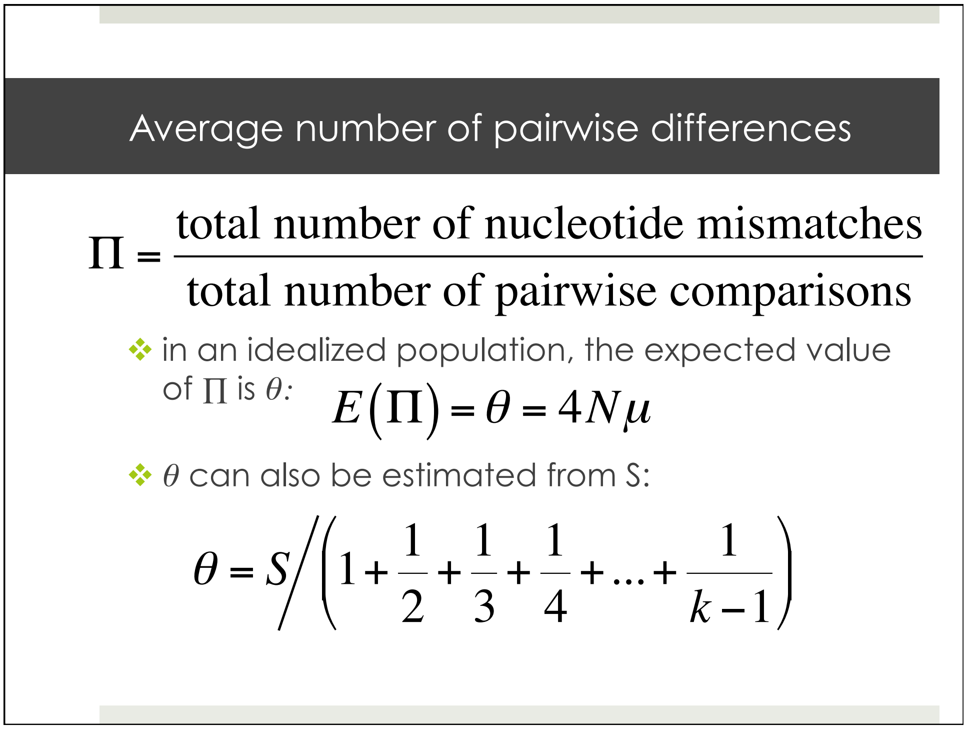


Coalescent theory often provides "easy" derivations of classical theory

- e.g., number of segregating sites in a sample
  - is a function of the total length (in generations) of the coalescent tree $E(T)$ times the mutation rate per locus per generation

$$E(T) = E\left(\sum_{i=2}^{k} iT_i\right) = \sum_{i=2}^{k} iE(T_i) = \sum_{i=2}^{k} i\frac{4N}{i(i-1)} = 4N\sum_{i=1}^{k-1}\frac{1}{i}$$

$$E(S) = \mu E(T) = 4N\mu\sum_{i=1}^{k-1}\frac{1}{i} = \theta\underline{\sum_{i=1}^{k-1}\frac{1}{i}}$$

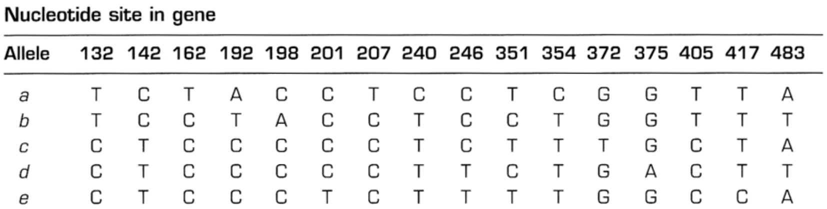


Average number of pairwise differences

$$\Pi = \frac{\text{total number of nucleotide mismatches}}{\text{total number of pairwise comparisons}}$$

- in an idealized population, the expected value of Π is θ: $E(\Pi) = \theta = 4N\mu$
- θ can also be estimated from S:

$$\theta = S\Big/\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots + \frac{1}{k-1}\right)$$



Nucleotide site in gene

Allele	132	142	162	192	198	201	207	240	246	351	354	372	375	405	417	483
a	T	C	T	A	C	C	T	C	C	T	C	G	G	T	T	A
b	T	C	C	T	A	C	C	T	C	C	T	G	G	T	T	T
c	C	T	C	C	C	C	C	T	C	T	T	T	G	C	T	A
d	C	T	C	C	C	C	C	T	T	C	T	G	A	C	T	T
e	C	T	C	C	C	T	C	T	T	T	T	G	G	C	C	A

$$\theta = 16 \Big/ \left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}\right) = 7.68$$

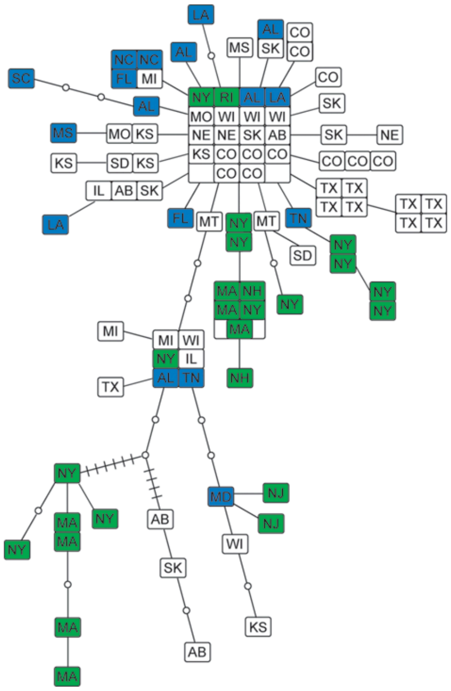
$$E(\theta) = \Pi = \left(\frac{(6 \times 6) + (4 \times 9) + (7 \times 1) + (0 \times 484)}{10}\right) = 7.90$$

Key point!

- differences in the values for number of segregating sites and average pairwise differences lead to the inference that the gene(s) or the population departs in one or more ways from the ideal "null model" (i.e., constant population size, no selection, etc..)

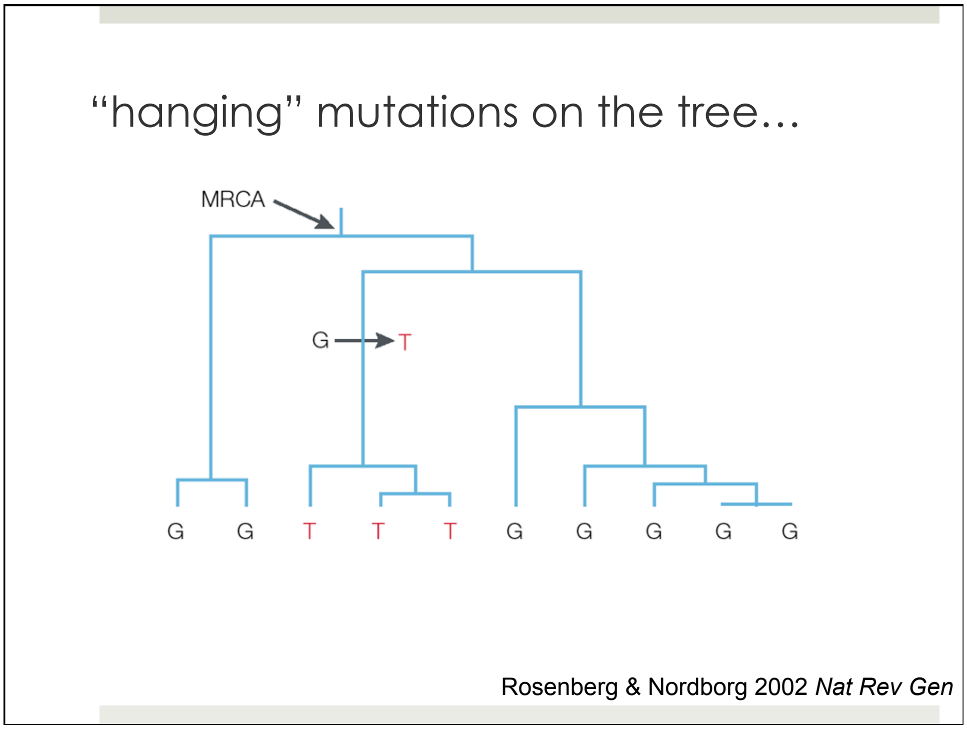
mtDNA haplotypes for big brown bats (*Eptesicus fuscus*) east of the Rockies

- θ (∏) = 5.35
- θ (S) = 10.42



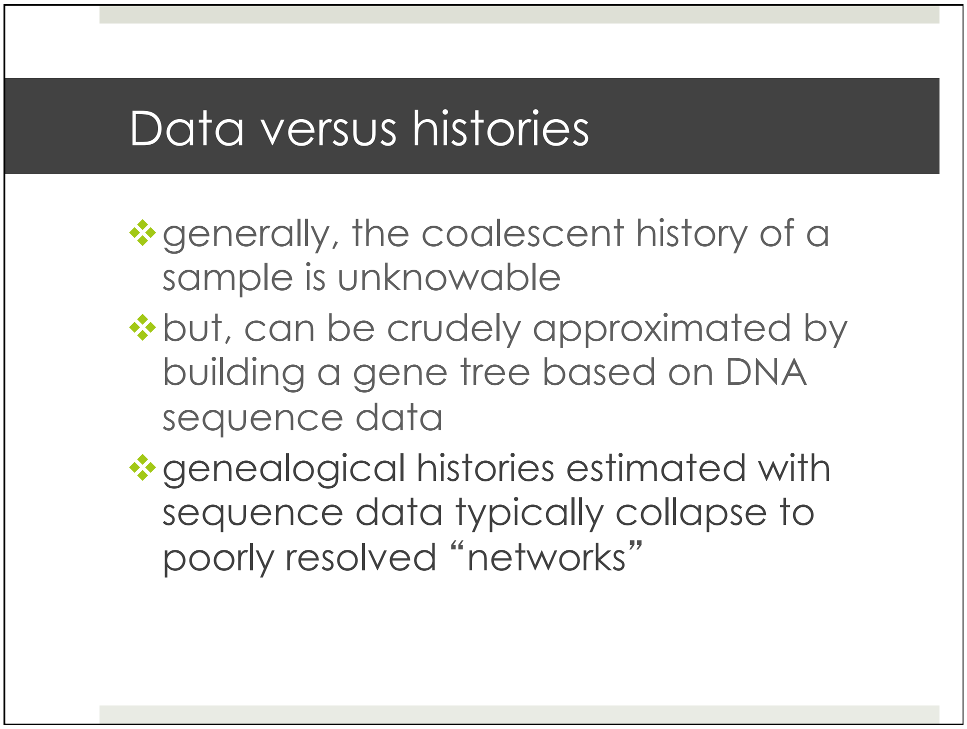
Data versus histories

- generally, the coalescent history of a sample is unknowable
- but, can be crudely approximated by building a gene tree based on DNA sequence data



“hanging” mutations on the tree…

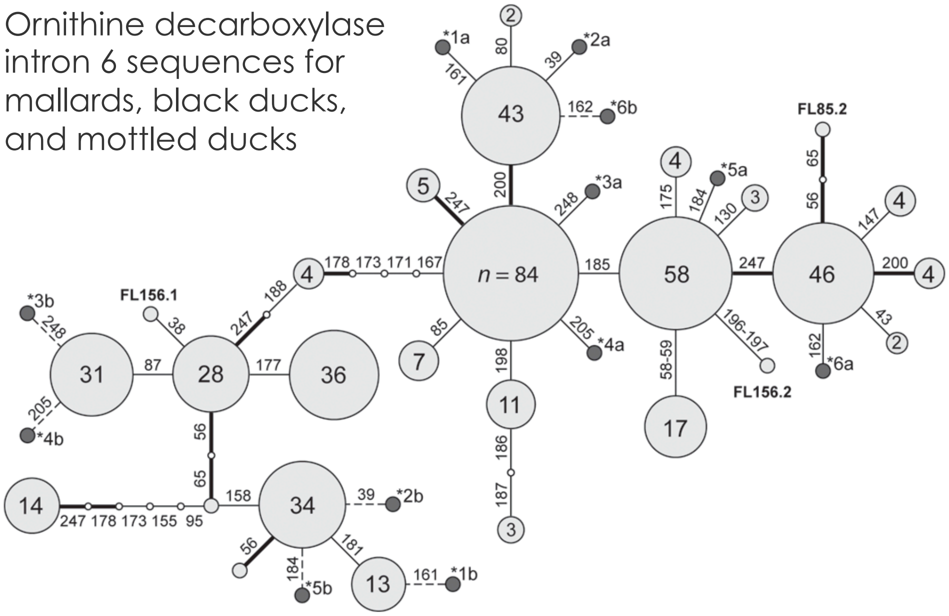
Rosenberg & Nordborg 2002 *Nat Rev Gen*



Data versus histories

- generally, the coalescent history of a sample is unknowable
- but, can be crudely approximated by building a gene tree based on DNA sequence data
- genealogical histories estimated with sequence data typically collapse to poorly resolved “networks”

Ornithine decarboxylase intron 6 sequences for mallards, black ducks, and mottled ducks



Harrigan *et al*. 2008 *Mol. Ecol. Resources*

Pybus *et al*. 1999 The mid-depth method and HIV-1: A practical approach for testing hypotheses of viral epidemic history. *Mol. Biol. Evol*.

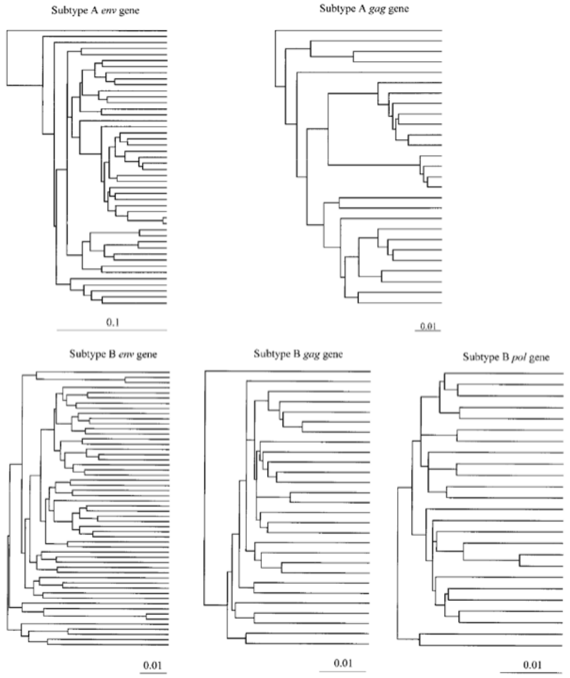
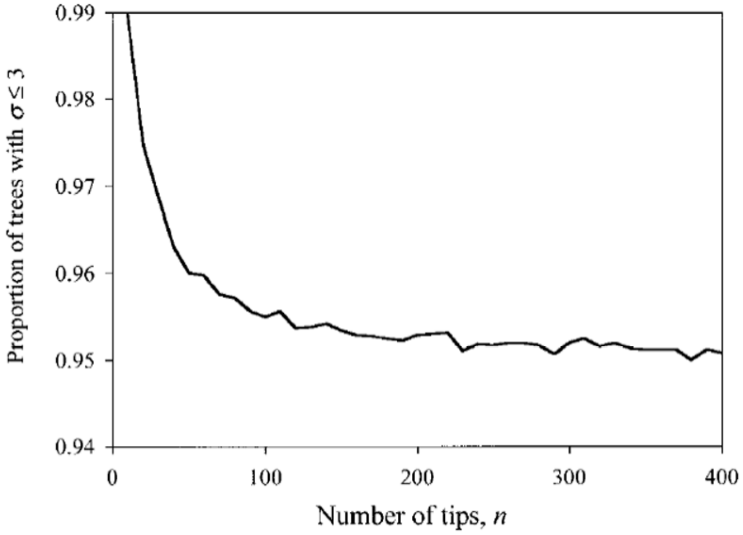


FIG. 4.—Estimated maximum-likelihood phylogenies for each of the five HIV-1 data sets. The scale bars represent expected number of substitutions per nucleotide.



Pybus *et al*. 1999 *Mol. Biol. Evol.*

FIG. 1.—Testing the hypothesis of constant population size. For each value of *n*, 10,000 endemic coalescent trees were simulated using the algorithm of Hudson (1990). The ordinate shows the proportion of simulated trees with *n* tips which have σ ≤ 3. More than 95% of all simulated trees with *n* < ≈200 tips have σ ≤ 3.

Still better…

- what is the likelihood L (=probability) of the observed data?

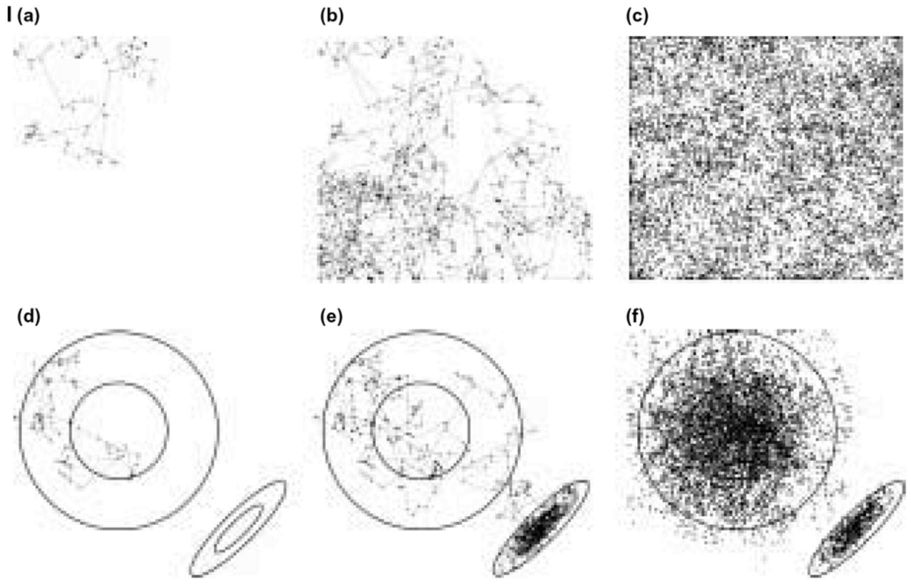
$$L = \sum_{G} \Pr\{D \mid G, \underline{\mu}\} \Pr\{G \mid \underline{\alpha}\}$$

- L = the sum across all possible genealogies (G) of the probability of the data given the genealogy and a model of the mutation process (μ) times the probability of the genealogy given a set of parameters (α) that characterize the population process

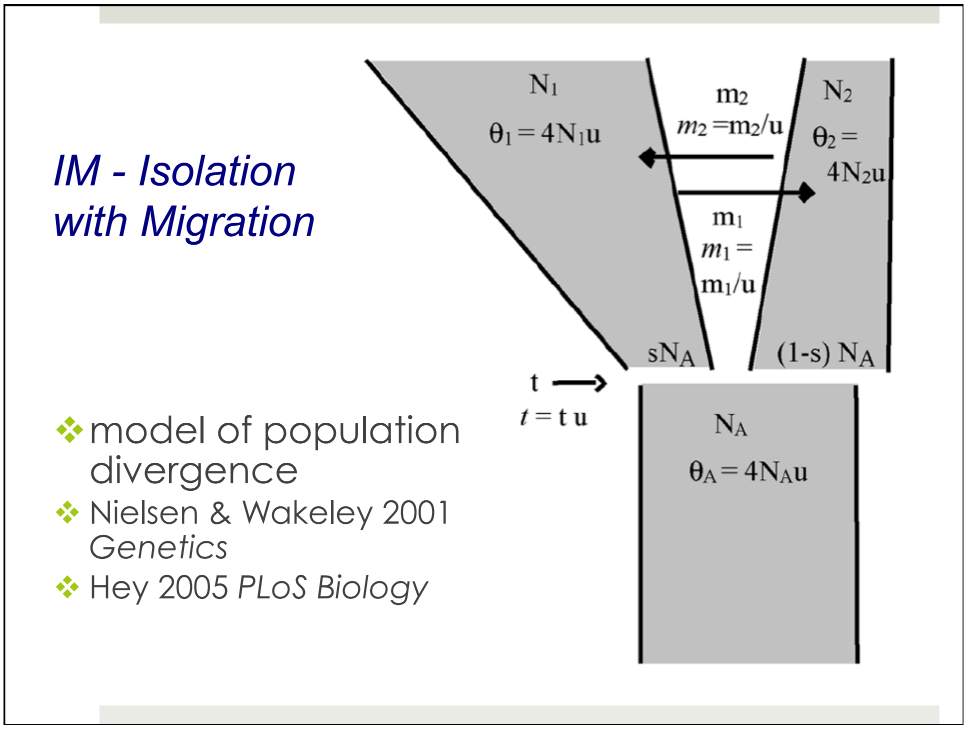
In practice…

- for a sample of k alleles, draw random coalescence times according to the exponential distribution
- estimate the likelihood of observing the actual data on that genealogy
- change a parameter, generate a new genealogy, calculate likelihood, **repeat** millions of times

Markov Chain Monte Carlo methods



TRENDS in Ecology & Evolution



IM - Isolation with Migration

- model of population divergence
- Nielsen & Wakeley 2001 *Genetics*
- Hey 2005 *PLoS Biology*