The Neutral Theory of Molecular Evolution

- Kimura (1968)
  - initially viewed as a challenge to Darwinian evolution
    - e.g., King & Jukes 1969 *Science* 164:788-798.
- many genetic polymorphisms have no effect on fitness and are therefore selectively neutral
- neutral polymorphisms are maintained by the combined effects of mutation and drift
  - mutations introduce new alleles as others are lost through drift

Origins of the "Selectionist-Neutralist" Debate

- the only "mutations" early biologists saw were ones that had phenotypic effects
- 1953 - structure of DNA (Watson & Crick / *Franklin*)
- 1960-70's - protein electrophoresis
  - revealed allelic diversity for many genes
- 1968 - Motoo Kimura - the neutral theory
  - motivated by allozyme (amino acid) variation
- discovery of "junk DNA"
  - 98.5% of human genome is non-coding
- DNA sequencing has revealed substantial silent and non-coding variation, suggesting that much genetic variation is selectively neutral (or nearly so!)

Infinite Alleles/Sites Model

- what is the expected level of genetic diversity (heterozygosity) given mutation and drift in a finite population?
- suppose a gene is 900 base pairs long, coding for 300 amino acids
  - there are $4^{900} = 10^{542}$ possible sequences (sorta…)
- thus, we can reasonably assume that each new mutation generates a unique allele…

Infinite Alleles/Sites Model

- it follows that alleles with the same sequence are identical by descent
- autozygous - a genotype with two alleles that are identical by descent
- allozygous - a genotype with alleles that are not identical by descent (is this possible?)
  - arbitrarily declare all alleles unique at $t = 0$
- autozygous = homozygous under the infinite alleles model
  - thus, the level of heterozygosity can be predicted from the expected level of autozygosity

Infinite Alleles/Sites Model

- F_t = probability that two randomly chosen alleles are *IBD*
  - same as autozygosity if we randomly choose alleles to form genotypes

$$F_t = \left(\frac{1}{2N}\right)(1-\mu)^2 + \left(1-\frac{1}{2N}\right)(1-\mu)^2 F_{t-1}$$

  - in this model, mutations generate new alleles and "erase" *IBD*

Infinite Alleles Model

- in a random-breeding population of constant size, an equilibrium is reached where the increase in autozygosity (~*IBD*) due to loss of alleles by drift is exactly countered by the increase in heterozygosity produced by new mutations
- solving for $F_t = F_{t-1}$ yields:

$$\hat{F} = \frac{1}{1+4N\mu}$$

Infinite Alleles Model

$$\hat{F} = \frac{1}{1+4N\mu} = \frac{1}{1+4N_e\mu} = \frac{1}{1+\theta}$$

- given the assumption of infinite alleles, any genotype that is not autozygous is heterozygous, so

$$\hat{H} = 1 - \hat{F} = \frac{4N_e\mu}{1+4N_e\mu} = \frac{\theta}{1+\theta}$$

Neutral Expectations for Genetic Diversity

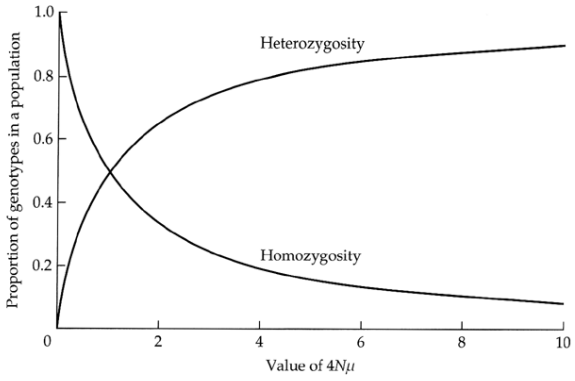


FIGURE 4.7 Plot of average homozygosity and average heterozygosity for the infinite-alleles model. Intermediate values of heterozygosity are maintained over only a small range of $\theta = 4N_e\mu$.

$$\hat{F} = \frac{1}{1+4N\mu} = \frac{1}{1+\theta}$$

- although an ideal population is expected to reach an "equilibrium" value of *F*, the population is not really *at equilibrium*, but rather in a "dynamic steady state" because there is a continual turnover of alleles
  - the most common allele is periodically replaced by another, other alleles are lost, and new alleles are produced by mutation

- DNA sequencing revealed surprisingly high levels of neutral genetic variation

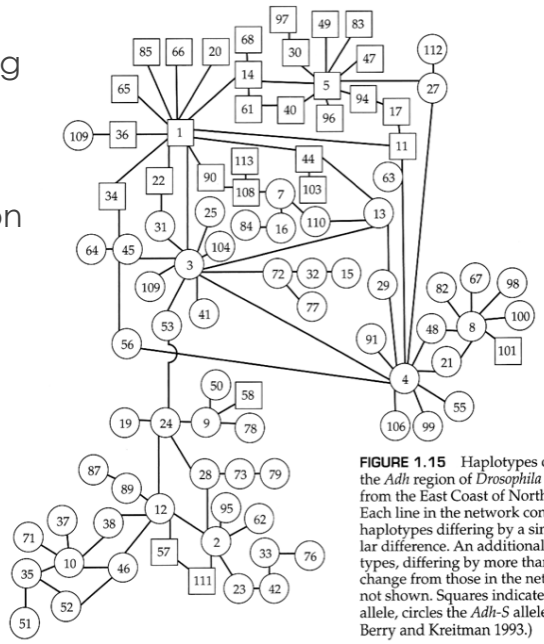


FIGURE 1.15 Haplotypes of alleles in the *Adh* region of *Drosophila melanogaster* from the East Coast of North America. Each line in the network connects two haplotypes differing by a single molecular difference. An additional 20 haplotypes, differing by more than one change from those in the network, are not shown. Squares indicate the *Adh-F* allele, circles the *Adh-S* allele. (From Berry and Kreitman 1993.)

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

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

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

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

$$E[total\_treelength] = \sum_{k=2}^{n} kE[t_k] = \sum_{k=2}^{n}\frac{2k}{k(k-1)} = 2\sum_{k=2}^{n}\frac{1}{k-1} = 2\sum_{k=1}^{n-1}\frac{1}{k}$$

$$E(S) = \left(2\sum_{k=1}^{n-1}\frac{1}{k}\right)\frac{\theta}{2} = \theta\sum_{k=1}^{n-1}\frac{1}{k}$$

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:

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

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

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

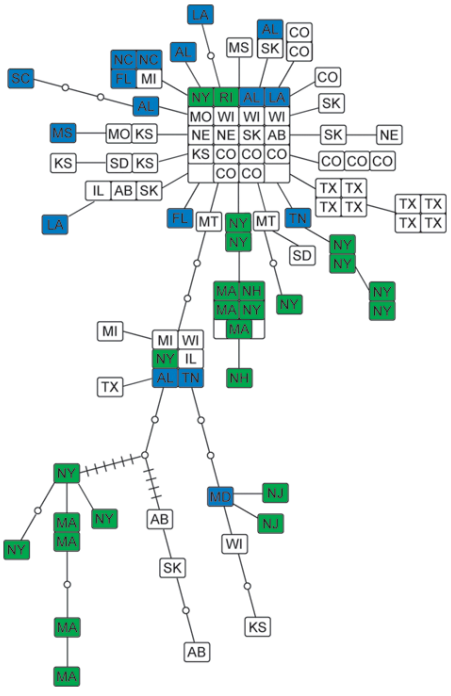
$$E(\theta) = \Pi = \left(\frac{(6 \times 6) + (4 \times 9) + (7 \times 1) + (0 \times 484)}{10}\right) = 7.90$$

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

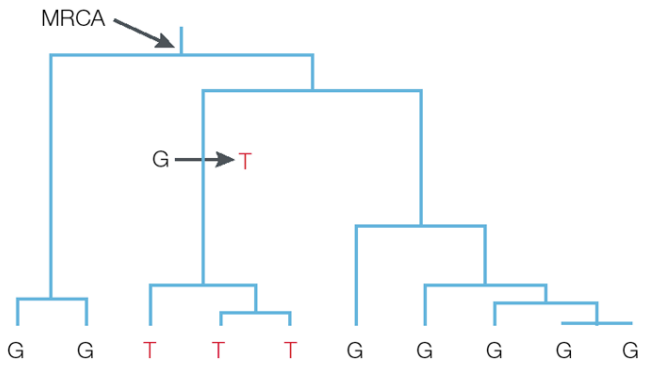
- θ (Π) = 5.35
- θ (S) = 10.42



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…

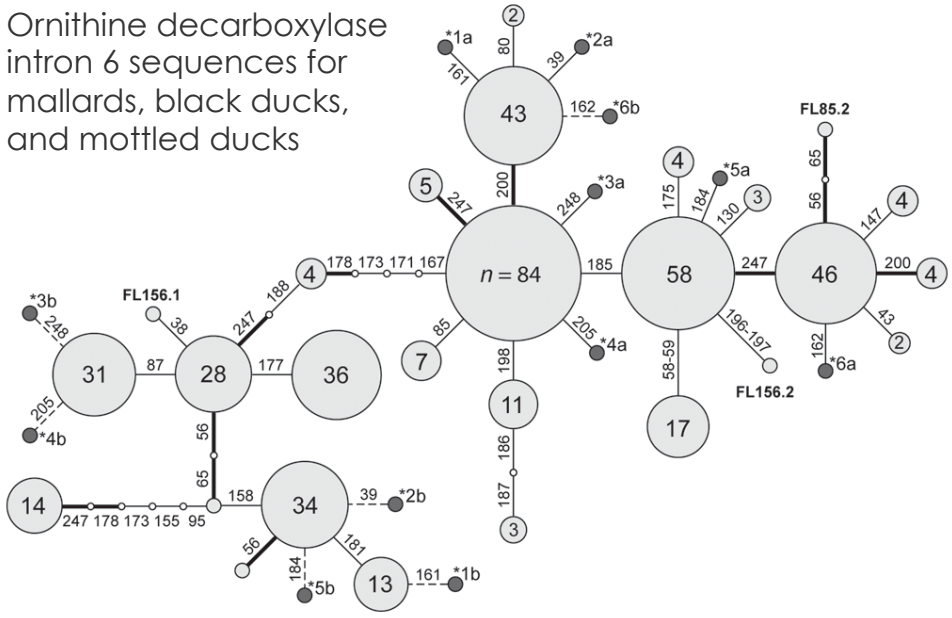


Rosenberg & Nordborg 2002 *Nat Rev Gen*

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*

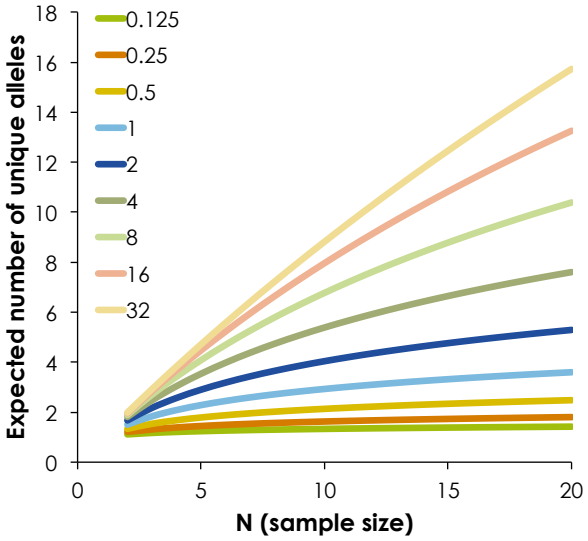
The Ewens Distribution

- beyond F, there is additional "information" available in the number of alleles present and the distribution of allele frequencies
  - "allelic configuration"
  - or "allele-frequency spectrum"
- Ewens (1972) - expected number of alleles k in a sample of size n, depends on θ

$$E(k) = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \ldots + \frac{\theta}{\theta+n-1}$$

Values of $\theta = 4N\mu$

$$E(k) = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \ldots + \frac{\theta}{\theta+n-1}$$



The Ewens Sampling Formula

- but there's more than just k (number of alleles)
- the Ewens Distribution specifies the probability distribution on the set of all ***partitions*** of the integer n
  - a.k.a. the "Chinese Restaurant" problem
  - broad applicability outside of population genetics

Partitions of 8

- 8
- 7 + 1
- 6 + 2
- 6 + 1 + 1
- 5 + 3
- 5 + 2 + 1
- 5 + 1 + 1 + 1
- 4 + 4
- 4 + 3 + 1
- 4 + 2 + 2
- 4 + 2 + 1 + 1
- 4 + 1 + 1 + 1 + 1
- 3 + 3 + 2
- 3 + 3 + 1 + 1
- 3 + 2 + 2 + 1
- 3 + 2 + 1 + 1 + 1
- 3 + 1 + 1 + 1 + 1 + 1
- 2 + 2 + 2 + 2
- 2 + 2 + 2 + 1 + 1
- 2 + 2 + 1 + 1 + 1 + 1
- 2 + 1 + 1 + 1 + 1 + 1 + 1
- 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1

Ewens' Sampling Formula (from Wikipedia!)

- Ewens' result provided the basis for a formula (Karlin & McGregor, 1972) giving the probability of a given allele frequency configuration (note: this is just one formulation)…

$$\Pr\{a_1,...,a_n\} = \frac{n!}{\theta(\theta+1)\ ...\ (\theta+n-1)}\prod_{j=1}^{n}\frac{\theta^{a_j}}{j^{a_j}a_j!}$$

where $a_1,...,a_n$ are counts of the number of alleles represented one, two,..., n times in the sample. $a_1,...,a_n$ are nonnegative integers that satisfy: $a_1 + 2a_2 + 3a_3 + ...+ na_n = n$

Karlin & McGregor 1972

- this equation works too…

$$\Pr(\theta\,;\, n_1,\, n_2,...,\, n_k) = \frac{r!}{n_1 n_2 ... n_k}\,\frac{1}{\alpha_1!\alpha_2!...\alpha_p!}\,\frac{\theta^k}{L_r(\theta)}$$

$$\text{where } L_r(\theta) = \theta(\theta+1)(\theta+1)\ ...\ (\theta+r-1)$$

Suppose the number of distinct integers in the set n_1, n_2,..., n_k is p and that there are exactly α_1 indices i such that $n_i = n_1$, exactly α_2 indices i such that $n_i = n_{\alpha_1+1}$, and so on, with exactly α_p indices i such that $n_i = n_k$.

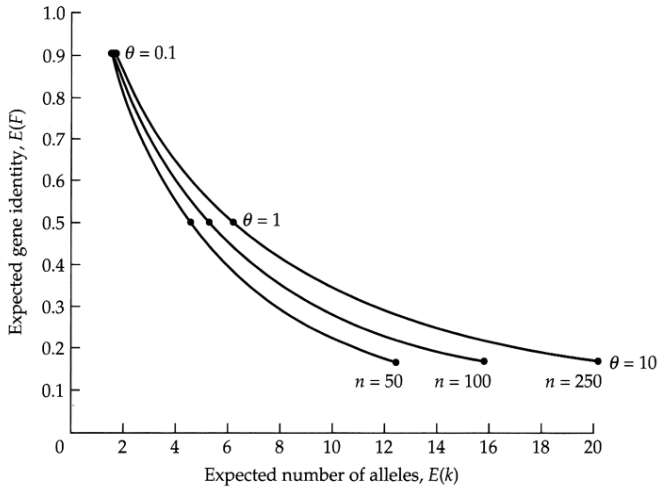


FIGURE 4.9 The infinite-alleles model prediction of the relation between the expected number of alleles $E(k)$ and the expected gene identity (homozygosity) F. The three curves represent a range of values of $\theta = 4N_e\mu$, starting at $\theta = 0.1$ in the upper left, and ending with $\theta = 10$ in the lower right. For the value of $\theta = 1$, the expected F, given by the relation $F = 1/(1 + \theta)$, is $\frac{1}{2}$, regardless of the sample size. Larger sample sizes always lead to larger expected numbers of alleles, but the difference is greater in more diverse populations (those with smaller F).

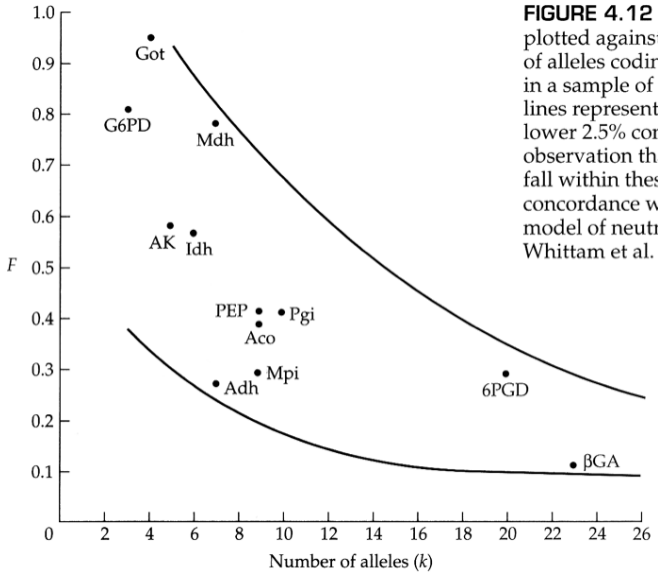
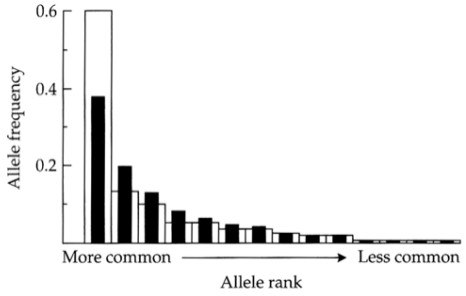


FIGURE 4.12 Gene identity (F) plotted against the observed number of alleles coding for various proteins in a sample of 279 *E. coli*. The solid lines represent the upper 97.5% and lower 2.5% confidence limits, and the observation that all of the tested loci fall within these limits suggests good concordance with the infinite-alleles model of neutral mutation. (From Whittam et al. 1983.)

Ewens' Sampling Formula

- *key point:* the Ewens distribution provides a basis for testing observed data against the neutral model

FIGURE 4.10 Observed (open columns) and expected (black bars) allele frequency spectrum of the *HRAS-1* gene in humans, identified by Southern blotting with the *pLM0.8* probe and *Taq*I digests. Observed data are from Baird et al. (1986). The expected distribution was generated using the Ewens sampling formula. In this sample of 490 genes there were 14 distinct alleles, four of which were present in just one individual. (From Clark 1988.)



Ewens' Sampling Formula

- *key point:* the Ewens distribution provides a basis for testing observed data against the neutral model
- and if the data fit neutral expectations, they can be used to estimate demographic and historical parameters

Site Frequency Spectrum

- applicable to DNA sequence data
- what is the distribution of frequencies for individual mutations (SNPs)
- number of derived "singletons" = θ
- why?
  - length of "external branches" = $4N$ generations
  - (note: $t = 2$ in Nielsen & Slatkin)
- frequency distribution for derived mutations

$$E\left[f_j\right] = \frac{1/j}{\sum_{k-1}^{n-1} \frac{1}{k}}, \text{ for } j = 1, 2, \ldots, n\text{-}1$$

Neutral Expectations with…

- …constant population size and mutation

nucleotide diversity (infinite sites)
 # segregating sites (infinite sites)
heterozygosity (infinite alleles)

$$E(\Pi) = \theta = 4N\mu$$

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

$$\hat{H} = \frac{\theta}{1+\theta}$$

 # unique alleles (infinite alleles)

$$E(k) = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \ldots + \frac{\theta}{\theta+n-1}$$

allele frequency distribution (infinite alleles)

$$\Pr\{a_1, \ldots, a_n\} = \frac{n!}{\theta(\theta+1)\ \ldots\ (\theta+n-1)} \prod_{j=1}^{n} \frac{\theta^{a_j}}{j^{a_j} a_j!}$$

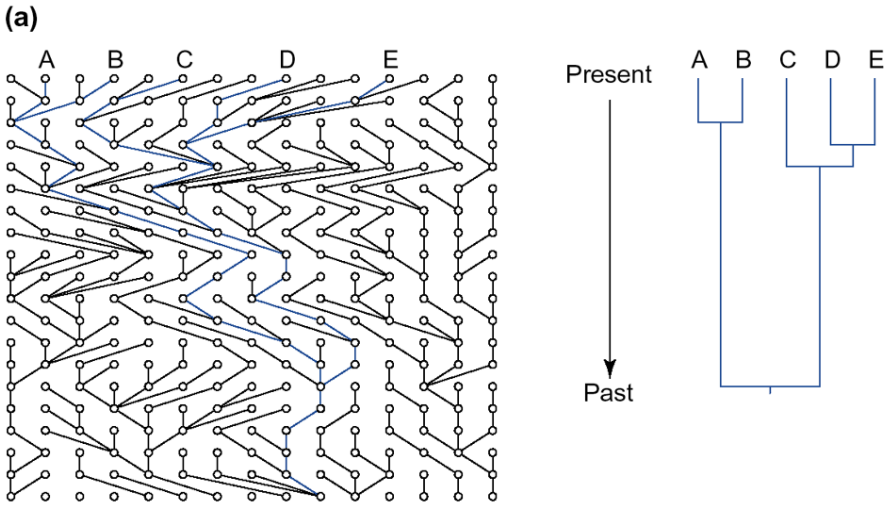
site frequency distribution (infinite sites)

$$E\left[f_j\right] = \frac{1/j}{\sum_{k-1}^{n-1} \frac{1}{k}}, \text{ for } j = 1, 2, \ldots, n\text{-}1$$

The coalescent with population growth

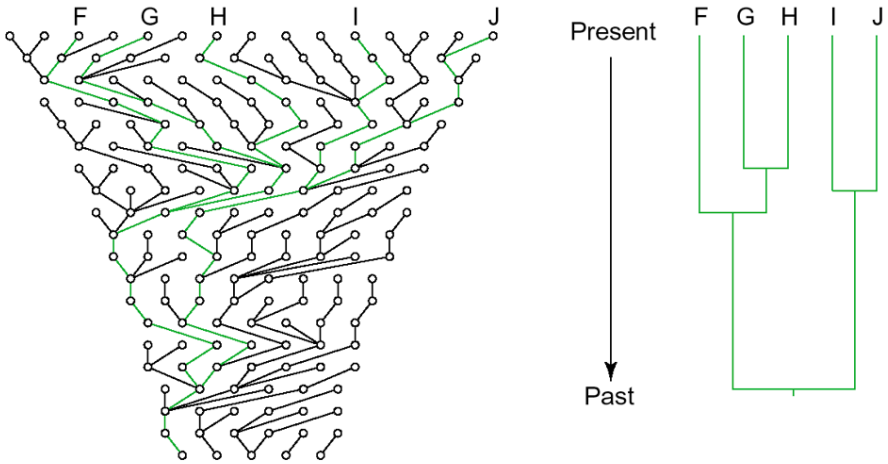
- coalescent trees are expected to be sparse (few lineages) near the root for populations of constant size
- in a growing or shrinking population, the distribution of coalescence times differs from expectations for the ideal population
- expanding populations have more nodes closer to the root of the tree
  - takes longer for alleles to "find each other" in a growing population

constant



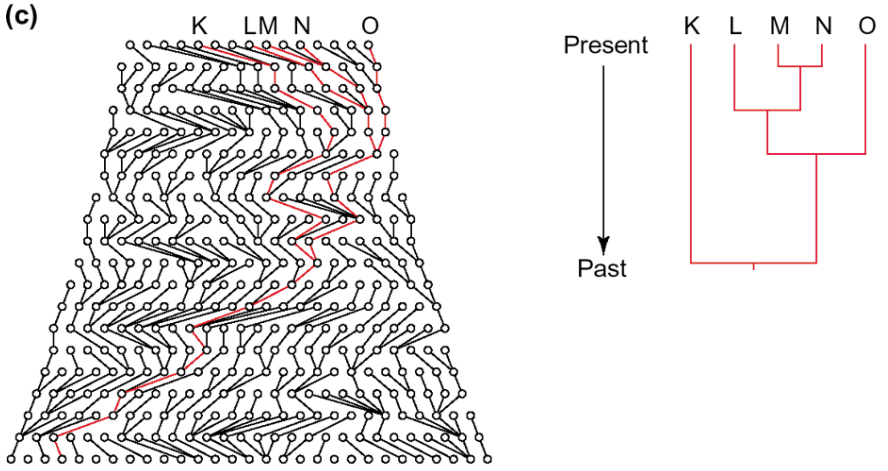
growing

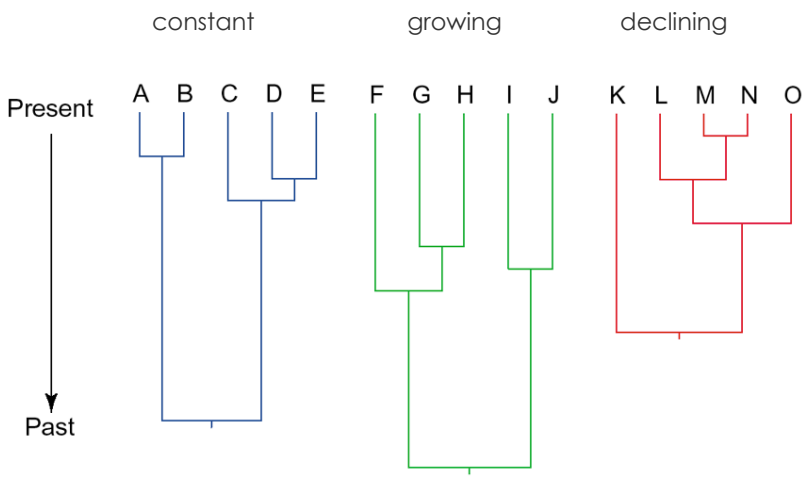
(b)



declining

(c)





So, what's the point?

- coalescent modeling
  - simulate genealogies under a given set of population parameters
  - "hang" random mutations on those trees in equal number (or at the same rate) as in the observed data
  - evaluate whether the observed data could have been produced by a random coalescent process (the null hypothesis)