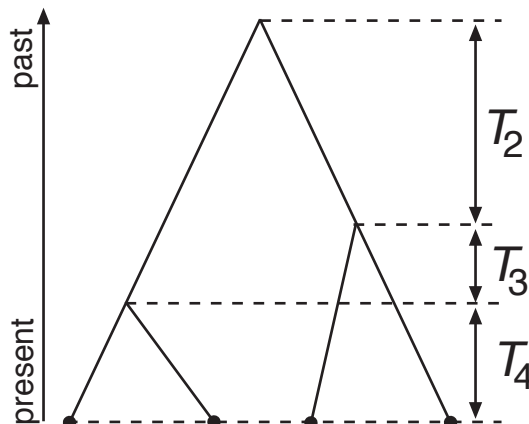


THE COALESCENT

INTRODUCTION TO THE COALESCENT

READING: Hedrick pp. 347-357

- Often want to use patterns of genetic variability to estimate parameters such as mutation and migration rates.
- e.g., Under infinite-alleles model saw that $\hat{H} = \frac{4Nu}{1 + 4Nu} = \frac{\theta}{1 + \theta}$, where \hat{H} is the expected equilibrium heterozygosity and $\theta = 4Nu$.
 - estimate θ by replacing \hat{H} with sample heterozygosity (H_{observed}) and solving equation for θ : $\theta_{\text{estimated}} = \frac{H_{\text{observed}}}{1 - H_{\text{observed}}}$
- This illustrates a **prospective** approach, since estimate is based on a forward-looking model
 - H_{observed} is also assumed representative of entire population’s true heterozygosity
- Alternatively, can develop estimates based on models that look backwards in time (i.e., are **retrospective**) and focus entirely on the set of sampled alleles (rather than entire population)
 - Called **coalescent** approaches.
- Main assumption behind coalescent: all alleles at a locus in a sample can be traced back to a single ancestral allele.
- **The coalescent** = lineage (genealogy) of sampled alleles traced back to their common ancestor.



- T_k = amount of time there are k distinct lineages.
 - each T_k is an independent random variable
- Are interested in T_{tot} , the total time in all branches of the genealogy until the entire set of alleles coalesces (i.e., can be traced back to a single common ancestor allele).
 - For above example, $T_{\text{tot}} = 4T_4 + 3T_3 + 2T_2$
 - Since the T_k 's are random variables, so is T_{tot}
- If u = mutation rate/generation, then $E[\text{different alleles in sample}] = E[\# \text{ mutations in genealogy from the common ancestor}] = uE[T_{\text{tot}}]$
- Coalescent approach often used to estimate $\theta = 4Nu$ based on the **infinite-sites model**.
 - assumes each allele is an infinitely long DNA or polypeptide sequence
 - every mutation occurs at a different site
 - Note: infinite-sites model like infinite-alleles model except in IAM, don't know *how* different the alleles are.
 - Will see that S = number of segregating sites (i.e., variable sites) can be used to estimate θ since $E[S] = uE[T_{\text{tot}}]$:
 - In fact, will show that $E[T_{\text{tot}}] = 4N \sum_{i=2}^n \frac{1}{i-1} = 4N \left(1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n-1} \right)$ where n = number of alleles in sample.
 - So $E[S] = u \left(4N \sum_{i=2}^n \frac{1}{i-1} \right) = 4Nu \left(\sum_{i=2}^n \frac{1}{i-1} \right) = \theta \sum_{i=2}^n \frac{1}{i-1}$
 - Suggests: $\theta_{\text{estimated}} = \frac{S_{\text{observed}}}{\sum_{i=2}^n \frac{1}{i-1}}$
 - Logic behind formula for $E[T_{\text{tot}}]$ = expected time to coalescence for a sample of n alleles:
 - Consider the probability of “no coalescence” in previous generation:

- * 1st allele’s ancestor in previous generation is one of $2N$ possible alleles
- * 2nd allele has *different* ancestor in previous generation with probability $1 - \frac{1}{2N}$
- * 3rd allele has different ancestor from first two alleles with probability $1 - \frac{2}{2N}$
- ⋮
- * n th allele has different ancestor from first $n-1$ alleles with probability $1 - \frac{n-1}{2N}$

– So... P(no coalescence in previous generation) = P(alleles have n distinct ancestors in previous generation) = $\left(1 - \frac{1}{2N}\right)\left(1 - \frac{2}{2N}\right) \cdots \left(1 - \frac{n-1}{2N}\right) \approx 1 - \frac{1}{2N} - \frac{2}{2N} - \cdots - \frac{n-1}{2N}$

– Finally: P(at least one coalescence in previous generation) = $1 - \text{P}(n \text{ distinct ancestors}) = \frac{1}{2N} + \frac{2}{2N} + \cdots + \frac{n-1}{2N} = \frac{1+2+\cdots+(n-1)}{2N} = \frac{n(n-1)/2}{2N} = \frac{n(n-1)}{4N}$

– Implies time to first coalescence in a sample of n alleles, T_n , is “geometrically distributed”:

* Geometric distribution is well studied. E.g., know that $E[T_n] = \frac{4N}{n(n-1)}$.

* By similar argument: $E[T_i] = \frac{4N}{i(i-1)}$.

– Know $T_{\text{tot}} = nT_n + (n-1)T_{n-1} + \cdots + 2T_2$ so

$$E[T_{\text{tot}}] = \sum_{i=2}^n iE[T_i] = \sum_{i=2}^n \frac{4N}{i-1} = 4N \left(1 + \frac{1}{2} + \cdots + \frac{1}{n-1}\right)$$

– Coalescence-based derivation of equilibrium homozygosity, \hat{f} under IAM:

- Consider P(two alleles not IBD) = $1 - \hat{f}$
- Two alleles will be IBD if they coalesce *before* a mutation occurs on either lineage.

– since P(two gene coalesce) = $1/2N$ per generation and P(mutation) = $2u$ per generation

– P(IBD) = $\frac{1/2N}{1/2N + 2u} = \frac{1}{1 + 4Nu} = \frac{1}{1 + \theta}$.

– Same result as before, but coalescent approach far easier!

– Example: (Aguadé et al. 1989)

- yellow-achaete-scute region of *D. melanogaster*.
- examined $n = 64$ chromosomes, found 9 polymorphic sites out of 2112 nucleotide sites
- $\theta_{\text{estimated}}(\text{entire region}) = \frac{9}{\sum_{i=2}^{64} (i-1)} = 1.9$
- Implies about 3 “effective alleles” segregating
- $\theta(\text{per site}) = \theta(\text{region})/2112 = 9 \times 10^{-4}$ per site.