§2.6: The Coalescent

- Let $\theta = 4Nu$

Recall, when drift = mutation, $H = \frac{4Nu}{1+4Nu}$

As $\theta \uparrow$, we get more heterozygosity (approaching 1)
(large pop or large mutation rate)

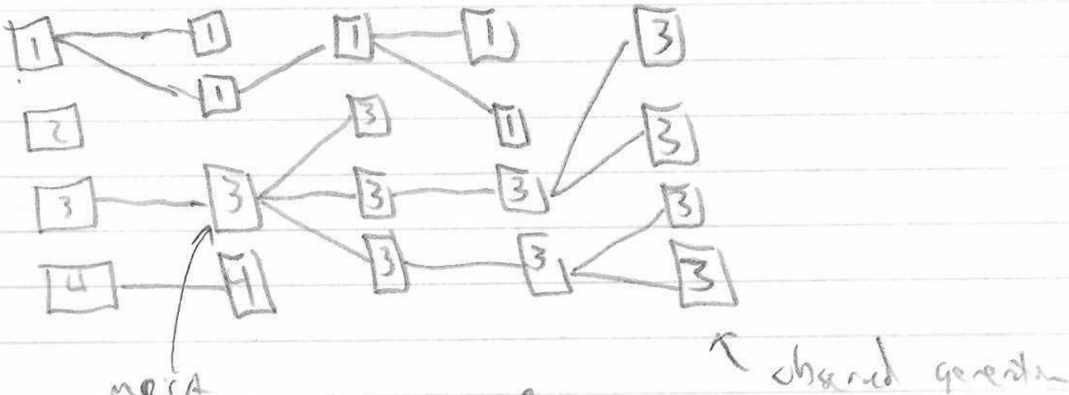
As $\theta \downarrow$, we get less heterozygosity
(small pop or small mutation rate)

- Before, we estimated H from observations to get estimate of θ

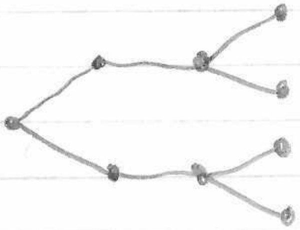
- Here, we will estimate θ by coalescent theory.

- Coalescent: Lineage of alleles traced back in time to most recent common ancestor

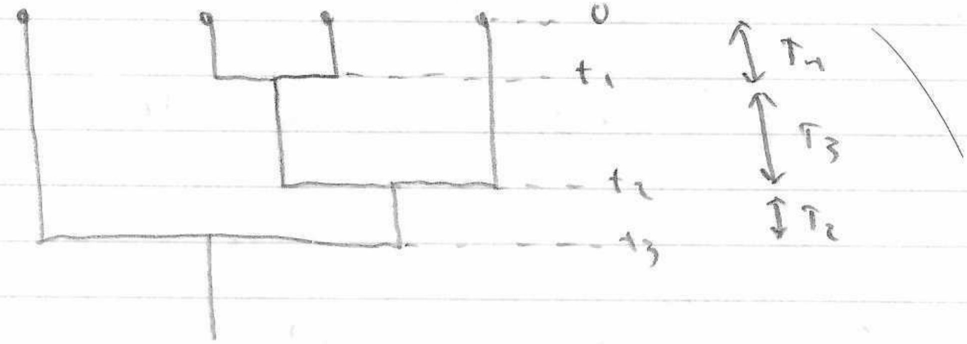
Ex)



Coalescent



- For typical N, the coalescent will look like this



t_i is in units of generations under Wright-Fisher model.

- We use mutations in the lineage to
  1. Estimate coalescent (called phylogenetics)
  2. Estimate θ

- Let S_n = # segregating (polymorphic) sites in a population
- Goal: use S_n to estimate θ.
- Let T_c = total time in coalescent

↑ in above figure, $T_c = 4t_1 + 3(t_2 - t_1) + 2(t_3 - t_2)$

Let T_i = time with i alleles in pop

$$T_c = 4T_4 + 3T_3 + 2T_2$$

- Expected # mutations in whole coalescent is $T_c \underbrace{u}_{\frac{\text{mutations}}{\text{generation}}}$

- We will show that (for a sample for 4 alleles)

$$E[T_c] = 4N\left(1 + \frac{1}{2} + \frac{1}{3}\right) = \frac{44N}{6} \quad \text{so} \quad uT_c = \theta\frac{11}{6}$$

↑ $\theta = 4Nu$

- Thus $\overbrace{E[S_4]}^{= u E[T_c]} = \theta\, 11/6$

  ↳ easy way to estimate θ

- Goal: Get $E[T_c]$ for any coalescent

- First interval T_n, second T_{n-1}, ..., T_2

- Consider n alleles (pop size is N)

Pr(allele 1 and 2 have different parents) $= 1 - \underbrace{\frac{1}{2N}}_{Pr(\text{same parent})} = \frac{2N-}{2}$

Pr(allele 3 diff | 1 and 2 diff) $= 1 - \frac{2}{2N} = \frac{2N-2}{2N}$

etc...

$$Pr(\text{all diff}) = \left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\cdots\left(1-\frac{n-1}{2N}\right)$$

can't be one of the n-1 already chosen

$$\approx 1 - \frac{1}{2N} - \frac{2}{2N} - \cdots - \frac{n-1}{2N} + \mathcal{O}\left(\frac{1}{N^2}\right)$$

$$Pr(\text{coalescent event}) = 1 - Pr(\text{all diff})$$

$$\approx \frac{1}{2N} + \frac{2}{2N} + \cdots + \frac{n-1}{2N} = \frac{1}{2N}(1+2+\cdots+n-1)$$

$$= \frac{1}{2N}\,\frac{n(n-1)}{2}$$

$$= \frac{n(n-1)}{4N}$$

- Each generation back, $Pr(\text{coalesce}) = \frac{n(n-1)}{4N}$

So time to coalescence $\sim \text{Geometric}\left(\frac{n(n-1)}{4N}\right)$

So $E(T_n) = \frac{4N}{n(n-1)}$ (Property of geometric distribution)

So if $T_c = \sum_{i=2}^{n} i\, T_i$

$$E(T_c) = \sum_{i=2}^{n} i\, E[T_i]$$

$$= \sum_{i=2}^{n} i \frac{4N}{i(i-1)}$$

$$= 4N \sum_{i=2}^{n} \frac{1}{i-1}$$

$$= 4N \sum_{i=1}^{n-1} \frac{1}{i}$$

$$E[S_n] = u\,E[T_c] = 4Nu \sum_{i=1}^{n-1} \frac{1}{i} = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

$$\Rightarrow \hat{\theta} = \frac{S_n}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

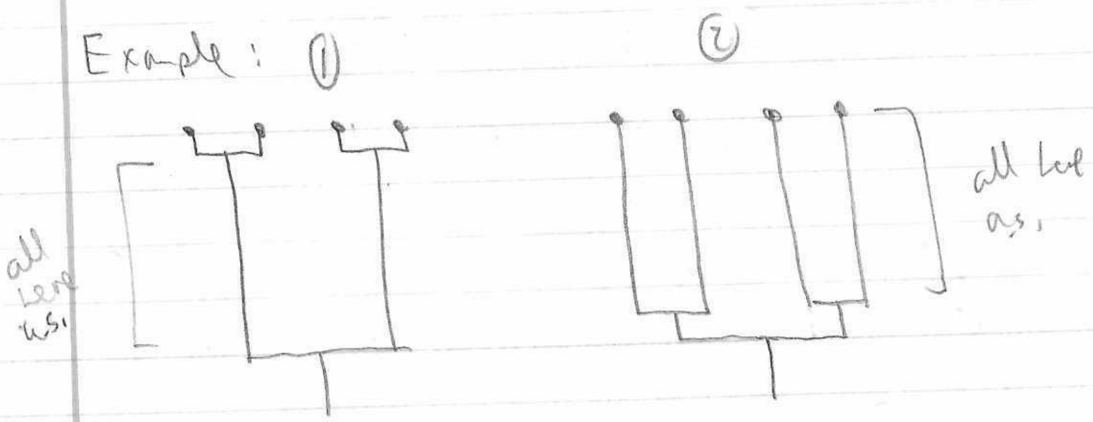
- Tajima's D: Compare observed heterozygosity to that given by $\hat{\theta}$

↳ tests for neutral model

↳ If observed heterozygosity = that from $\hat{\theta}$, then no evidence against neutral model

$$\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2 = \frac{1}{16} + \frac{9}{16} = \frac{5}{8}$$

Example:



Put 5 random mutation

Case ①: mutation is either in left or right
$p = 1/2$ for all mutations $\quad H = 2(\frac{1}{2})(1-\frac{1}{2}) = \frac{1}{2}$
$\uparrow$ allele freq

Case ②: mutation is in one branch
$p = \frac{1}{4}$ $\quad$ heterozygosity $= 2(\frac{1}{4})(1-\frac{1}{4}) = \frac{3}{8}$

$H = \Pr(\text{differ by state} \mid \text{drawn w/ replacement})$

- We will show later that $E\left[\sum_{i=1}^{\infty} 2p_i(1-p_i)\right] = \theta$ $\quad$ ↖ under neutral loci

let $\pi = \sum_{i=1}^{\infty} 2\,p_i(1-p_i)$
$\uparrow$ unknown

$$\hat{\pi} = \frac{n}{n-1} \sum_{i=1}^{S_n} 2\,\hat{p}_i(1-\hat{p}_i)$$

$\uparrow$ sum over segregating sites.

Tajima D: $D_T = \frac{\hat{\pi} - \hat{\theta}}{C}$

C chosen s.t. $\frac{\hat{\pi} - \hat{\theta}}{c} \sim$ Normal

$\hat{\pi}$ is estimate of θ from heterozygosity
$\hat{\theta}$ is estimate from coalescent

- $D_T > 0 \Rightarrow$ more heterozygosity than expected
  $D_T < 0 \Rightarrow$ less heterozygosity than expected

- A proof that $E\left[\sum_{i=1}^{\infty} 2p_i(1-p_i)\right] = \theta$:

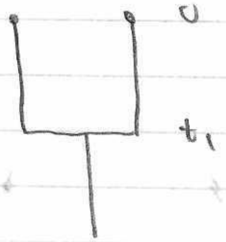
π is expected # of nucleotide differences between randomly selected pair of alleles.

$Z_i :=$ pair of alleles differ at nucleotide i

$E[Z_i|p_i] = 2p_i(1-p_i)$

$\pi = \sum_{i=1}^{\infty} Z_i$ = # differences b/t pair of alleles

$E[\pi|p] = E\left[\sum_{i=1}^{\infty} Z_i|p_i\right] = \sum_{i=1}^{\infty} 2p_i(1-p_i)$



$E[t_1] = \frac{4N}{2(2-1)} = 2N$

differences $= 2t_1 u$

$\Rightarrow E[\# \text{ differences}] = 4Nu = \theta$ //

- A coalescent derivation of

$$H = \frac{4Nu}{1+4Nu}$$

- Previous derivation: found ΔH in terms of mutation and drift, set $= 0$

① $\underbrace{A'}_{\text{Next gen}} = \underbrace{(1-u)^2}_{\text{No mutation}} \left(\underbrace{\underbrace{\frac{1}{N}}_{\text{Same}} + \underbrace{\left(1-\frac{1}{N}\right)}_{\text{diff}} \underbrace{A}_{\text{prev gen}}}_{\text{Same gen No mutation}} \right) = P_r(\text{same state} \mid \text{diff origi}$

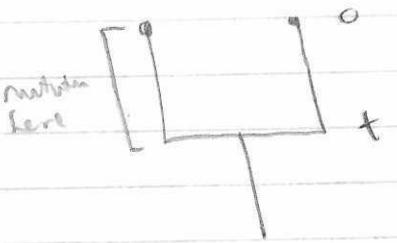
② $H' = 1 - A'$

③ use $u^2 \approx 0$, $\frac{u}{N} \approx 0$

④ find $\Delta H = H' - H$

⑤ set $\Delta H = 0$, solve for H

Two alleles differ by state $\Leftrightarrow$ mutation after common ance



$P_r(\text{Coalesce in 1 generation}) = \frac{1}{2N}$

$P_r(\text{Mutation in 1 generation}) = 1 - (1-u)^2$

Note: $1-(1-u)^2 = 1-(1-2u+u^2)$

$= 2u - u^2$

$\approx 2u$ since u^2 is small

Heterozygous if mutation occurs first

$$\Pr(\text{mutation} \mid \text{mutation or coalesce})$$

$$= \frac{\Pr[\text{mutation and } (\text{mutation or coalesce})]}{\Pr(\text{mutation or coalesce})}$$

$$= \frac{\Pr(\text{mutation})}{\Pr(\text{mutation or coalesce})}$$

$$= \frac{2u}{2u + \frac{1}{2N}}$$

$$= \frac{4Nu}{1+4Nu} \quad //$$