Genetic diversity stats

 # segregating sites, S, equals # mutations in the sample's history (infinite sites approximation):

$$\mathbb{E}[S] = 4N\mu \sum_{i=2}^{n-2} \frac{1}{i}$$

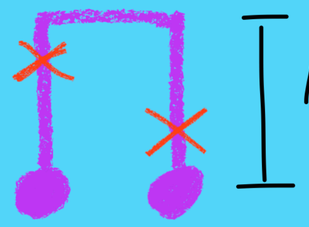
Pairwise divergence, π, # mutations in the history of two sampled haplotypes:

$$\mathbb{E}[\pi] = 4N\mu$$

Tajima's D: null hypothesis, standard neutral coalescent (constant N)

$$\frac{\mathbb{E}[S]}{\sum_{i=2}^{n-2} 1/i} - \mathbb{E}[\pi] = 0$$

Derivation:

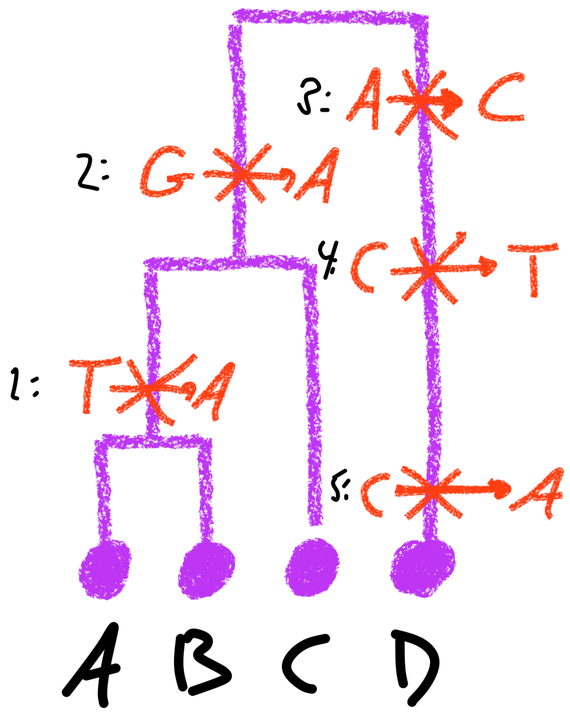


$$\mathbb{E}[T_2] = \frac{1}{1/2N} = 2N$$

$$\Rightarrow \mathbb{E}[\pi] = 4N\mu$$

Question: why 4?

Example



	1	2	3	4	5
A	A	A	A	C	C
B	A	A	A	C	C
C	T	A	A	C	C
D	T	G	C	T	A

$$S = 5$$

$$\pi = \frac{1}{\binom{4}{2}}\left(0+2+5+2+5+4\right) = \frac{8}{3}$$

$$D = \frac{5}{1+\frac{1}{2}+\frac{1}{3}} - \frac{8}{3} = \cdots \frac{2}{33}$$