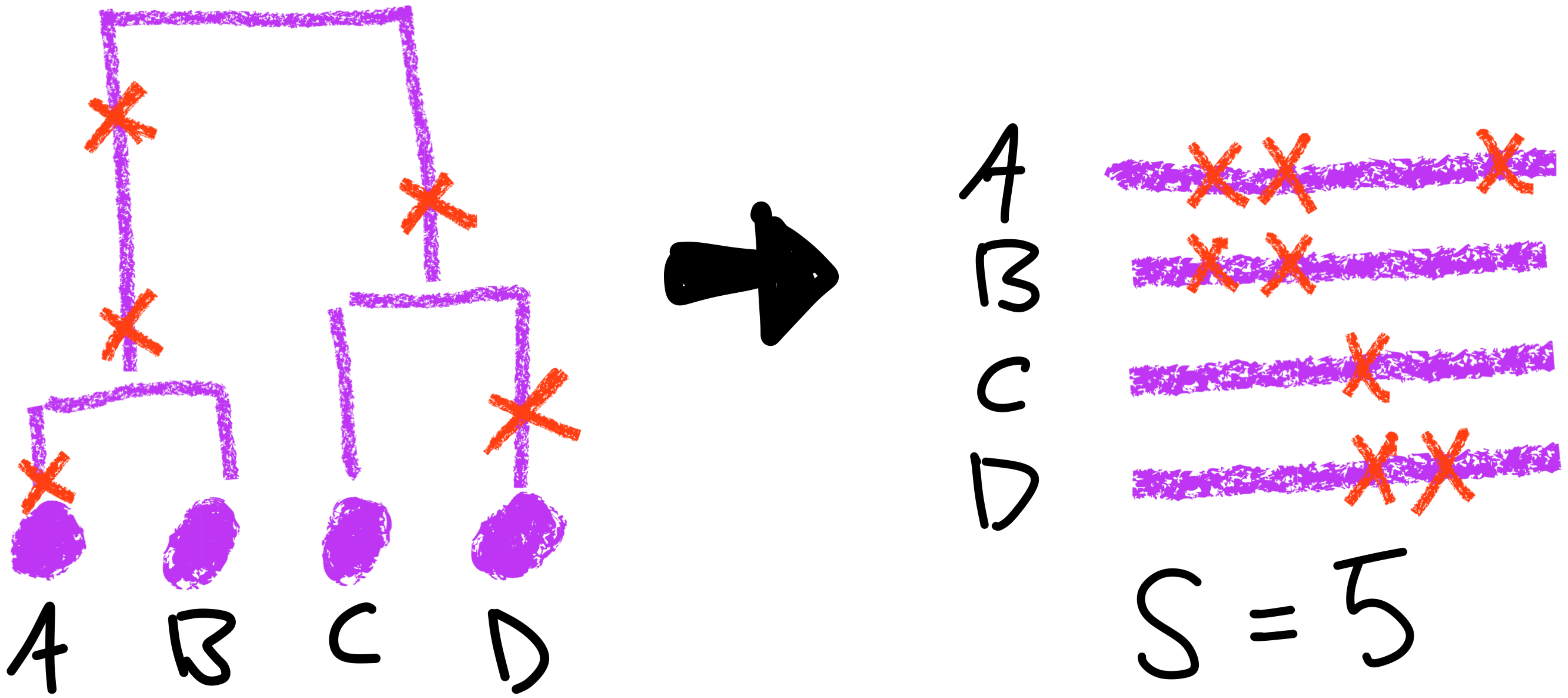
Previously on...

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



Constant N case :

total branch length

$$\mathbb{E}[S] = \mu\, \mathbb{E}[T_{total}]$$

$$= \mu \sum_{i=2}^{n} i\, \mathbb{E}[T_i]$$

$$= \mu \sum_{i=2}^{n} i \frac{2N}{\binom{i}{2}}$$

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

Genetic diversity stats

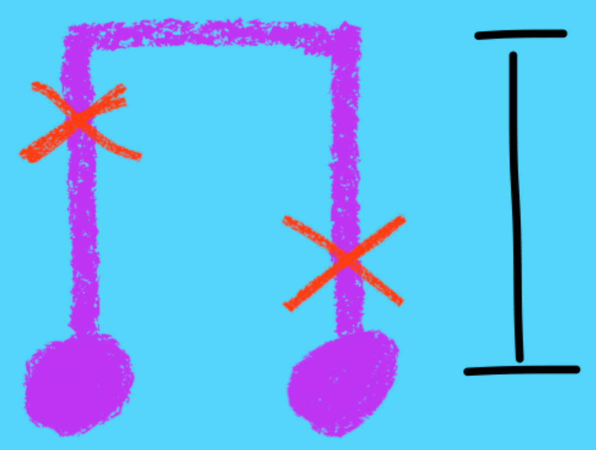
segregating sites, S, equals # mutations in the sample's history (infinite sites approximation):

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

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

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

Derivation:



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

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

Question: why 4?

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

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