

discordance,  
ILS & the coalescent

**Species tree**

Represents the evolutionary history of species - cladogenesis or the sequence of speciation events.

**Gene tree(s)**

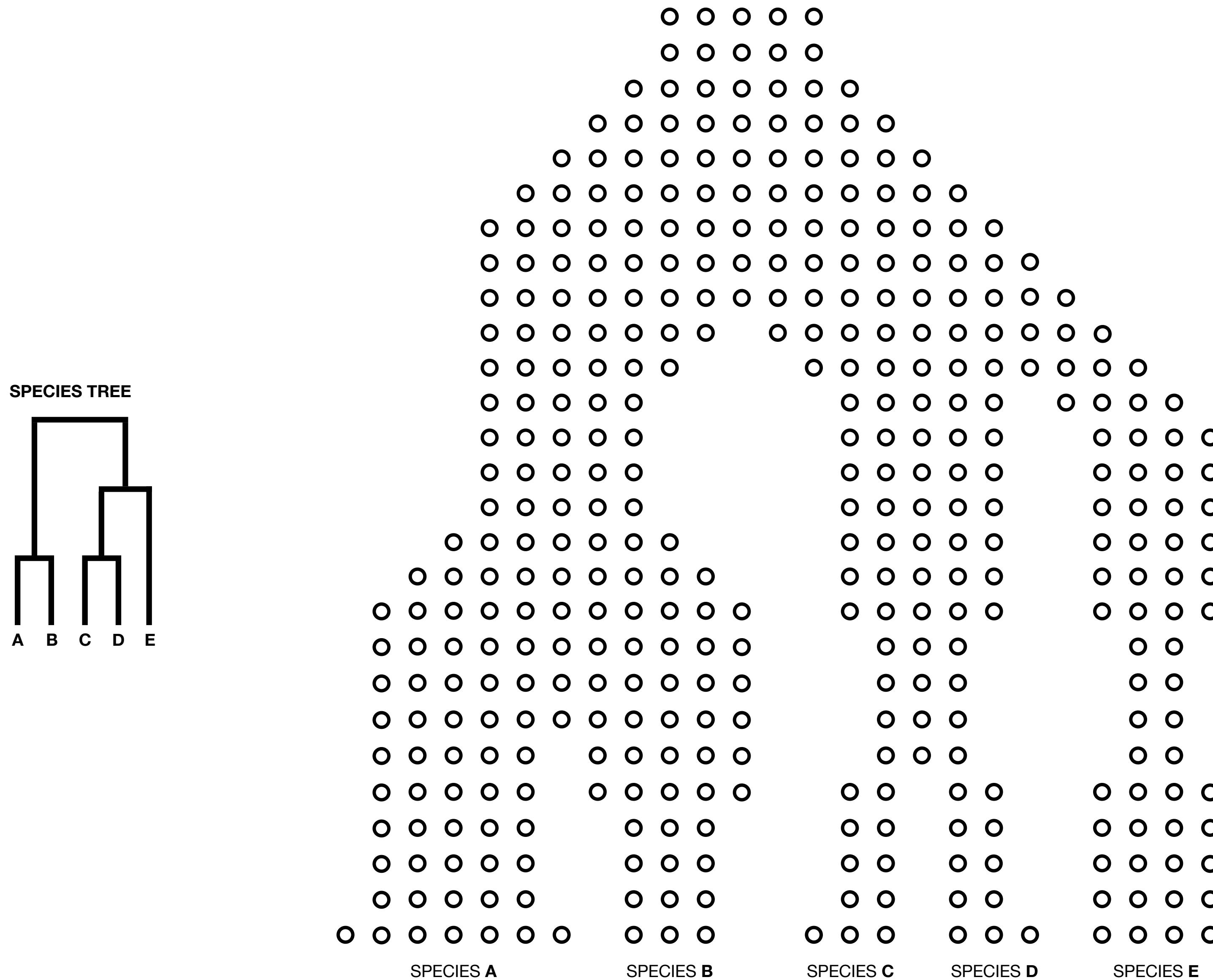
Traces the evolutionary history of a single locus and may differ biologically from the species tree.

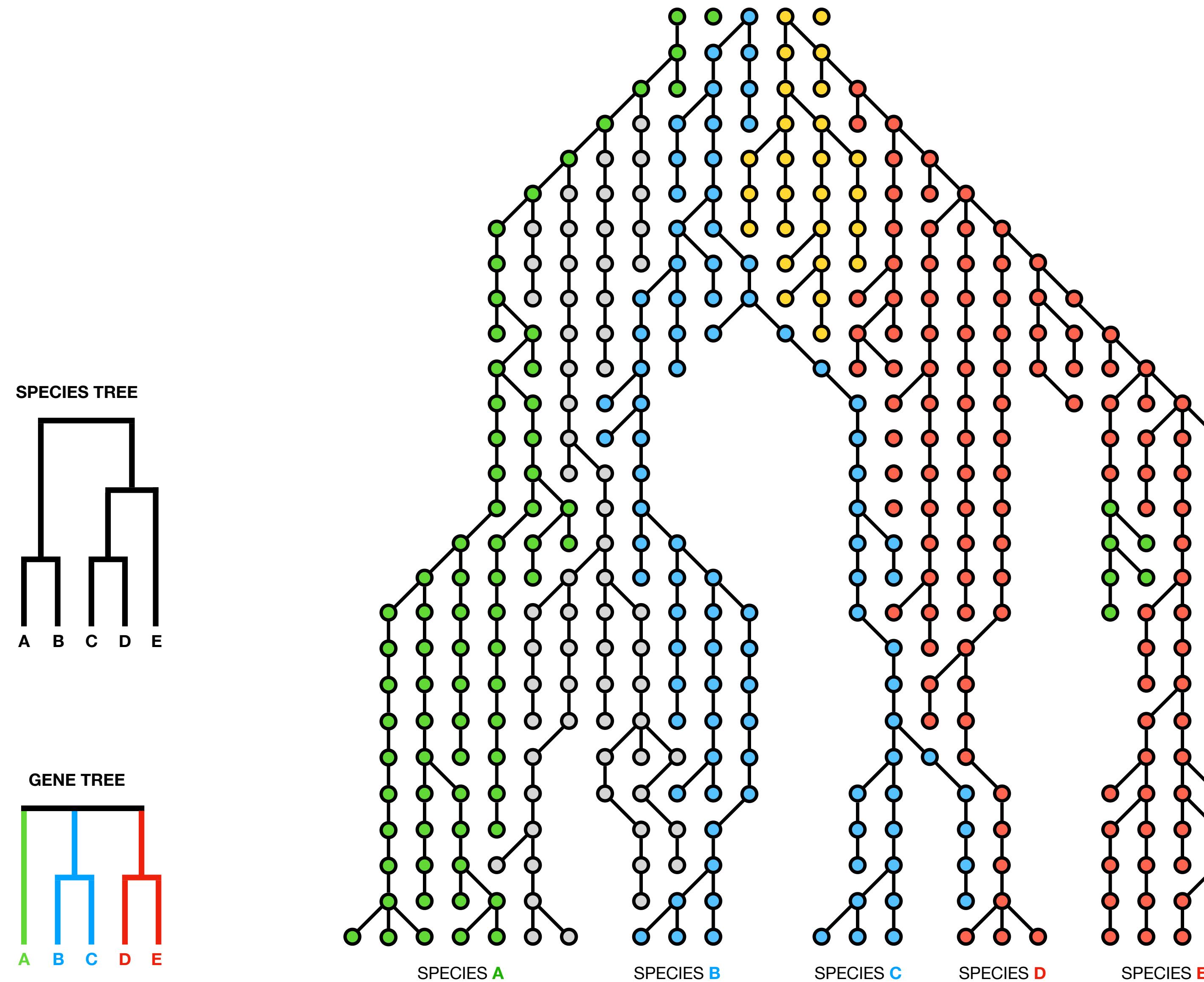
A common cause of gene tree discordance is **Incomplete Lineage Sorting (ILS)**, which is especially likely when:

- effective population sizes are large
- speciation events are closely spaced in time

You can **think of ILS as**:

- ancestral polymorphisms that persist across speciation events,
- alleles sorting into descendant lineages *erratically*, not reflecting the species splits.





While ILS is a major cause of gene tree conflict ...

**... also other biological processes can result in gene tree discordance:**

- **Horizontal Gene Transfer (HGT)**

Transfer of genetic material between unrelated lineages, common in microbes.

- **Hybrid Introgression (HI)**

Transfer of genetic material between species through hybridization followed by backcrossing.

These processes can result in **true biological conflict** between gene trees and the species tree!

In Lesson 14, we'll explore the **technical artifacts and sources of error** that can create the **illusion of conflict** between gene trees and the species tree — even when no true biological discordance exists.

By the way ... **Robinson-Foulds (RF) distance** ...

... are a commonly used metric to **compare two phylogenetic trees** with the same set of taxa.

### How it works:

It measures the symmetric difference between two trees. Specifically, it counts the number of bipartitions (splits) that are present in Tree A but not in Tree B and present in Tree B but not in Tree A.

- Higher RF distance = greater disagreement between the trees.
- Lower RF distance = smaller disagreement between the trees.
- Often normalized to range from 0 (identical) to 1 (completely different)

While RF distance captures topological differences, it treats all differences equally ... regardless of their biological relevance.

Other more nuanced metrics aim to assess the phylogenetic informativeness of the differences.

## Concatenation-based species tree inference

- a single data matrix containing aligned sequences (species × genes)
- assumes all genes share the same evolutionary history = one tree

## Coalescent-based species tree inference

- a set of unrooted, binary gene trees (each with leaves labeled by species) Models gene tree
- infers a species tree from multiple gene histories discordance due to incomplete lineage sorting

The **coalescent** is a framework that describes how genes sampled from an extant population trace back to a **common ancestor** over time.

The **coalescent theory** sits between phylo- and pop- genetics.

It's a **retrospective model**: the coalescent looks **backward in time** to understand the **genealogical relationships** between individuals or alleles.

The coalescent is based on the mathematical notion that mutations within genes - **new alleles** - can be **traced backwards in time**, to the point where the mutation initially occurred.

Given that this is a retrospective, instead of describing these mutation moments as **divergence** events, these appear as moments where mutations come back together as **coalescence** events.

There are a number of applications of coalescent theory, and it is particularly fitting process for understanding the **neutral demographic history of populations and species**.

The initial coalescent model was described in the 1980s, built upon by a number of different ecologists, geneticists and mathematicians. However, **John Kingman** is often attributed with the formalization of the original coalescent model.

Key **assumptions** of the coalescent:

- **constant population size** - effective population size ( $N_e$ ) does not change over time
- **random mating** - all genes have equal chance of being parent of any gene in the next gen
- **no selection** - all alleles are neutral and drift is the only force acting on allele frequencies
- **no migration** - no gene flow from other populations and no spatial or geographic structure
- **non-overlapping generations**
- **no recombination within loci**
- **infinite sites or infinite alleles model**
  - **infinite sites**: mutations occur at unique sites, no back or parallel mutations
  - **infinite alleles**: every mutation creates a completely new allele

**What's the chance two gene copies coalesce?**

Imagine two haploid gene copies sampled from a population. In each generation:

- each gene copy chooses a parent randomly from the previous generation
- in a population of effective size  $N_e$ , there are  $2N_e$  possible parental gene copies.

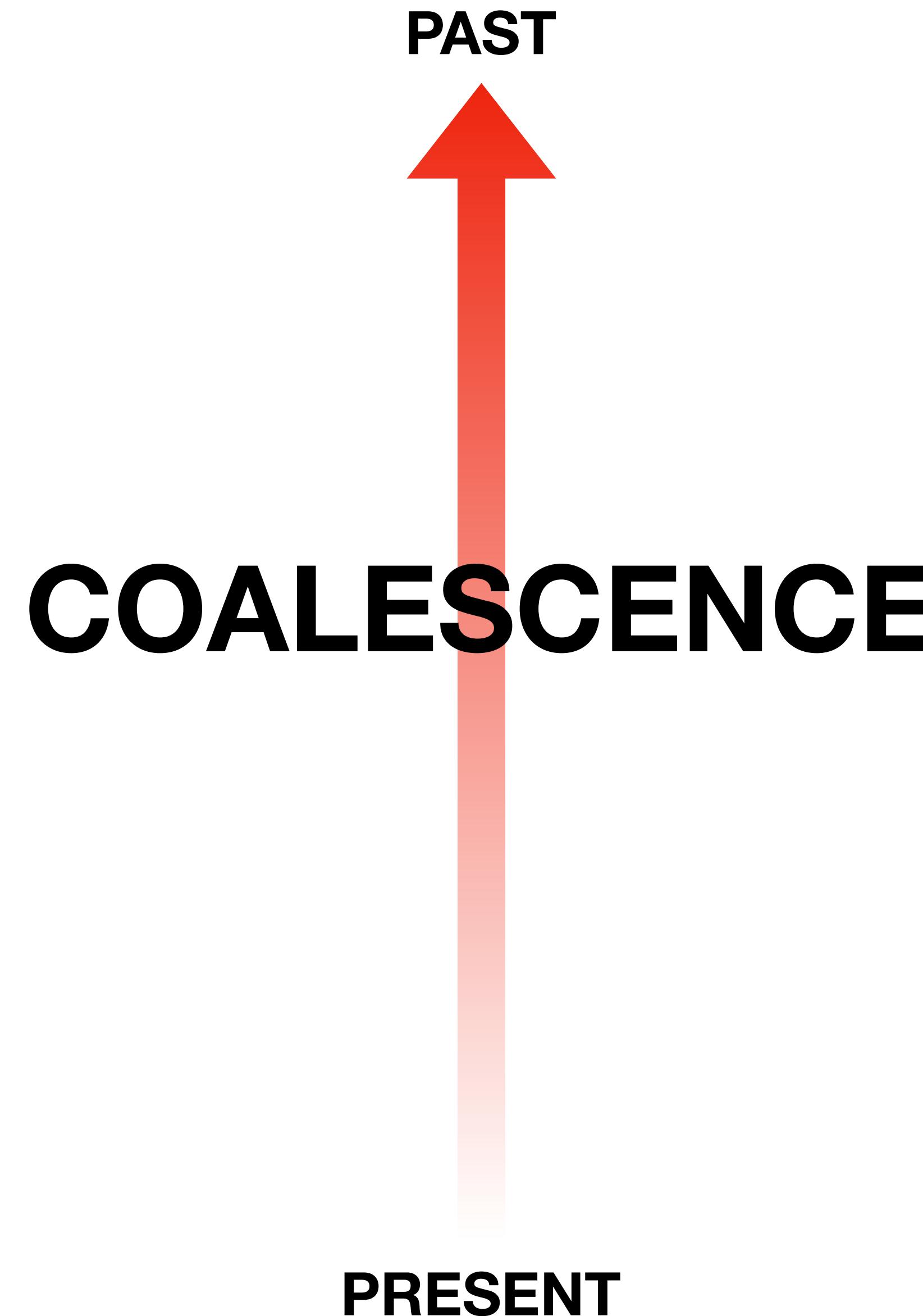
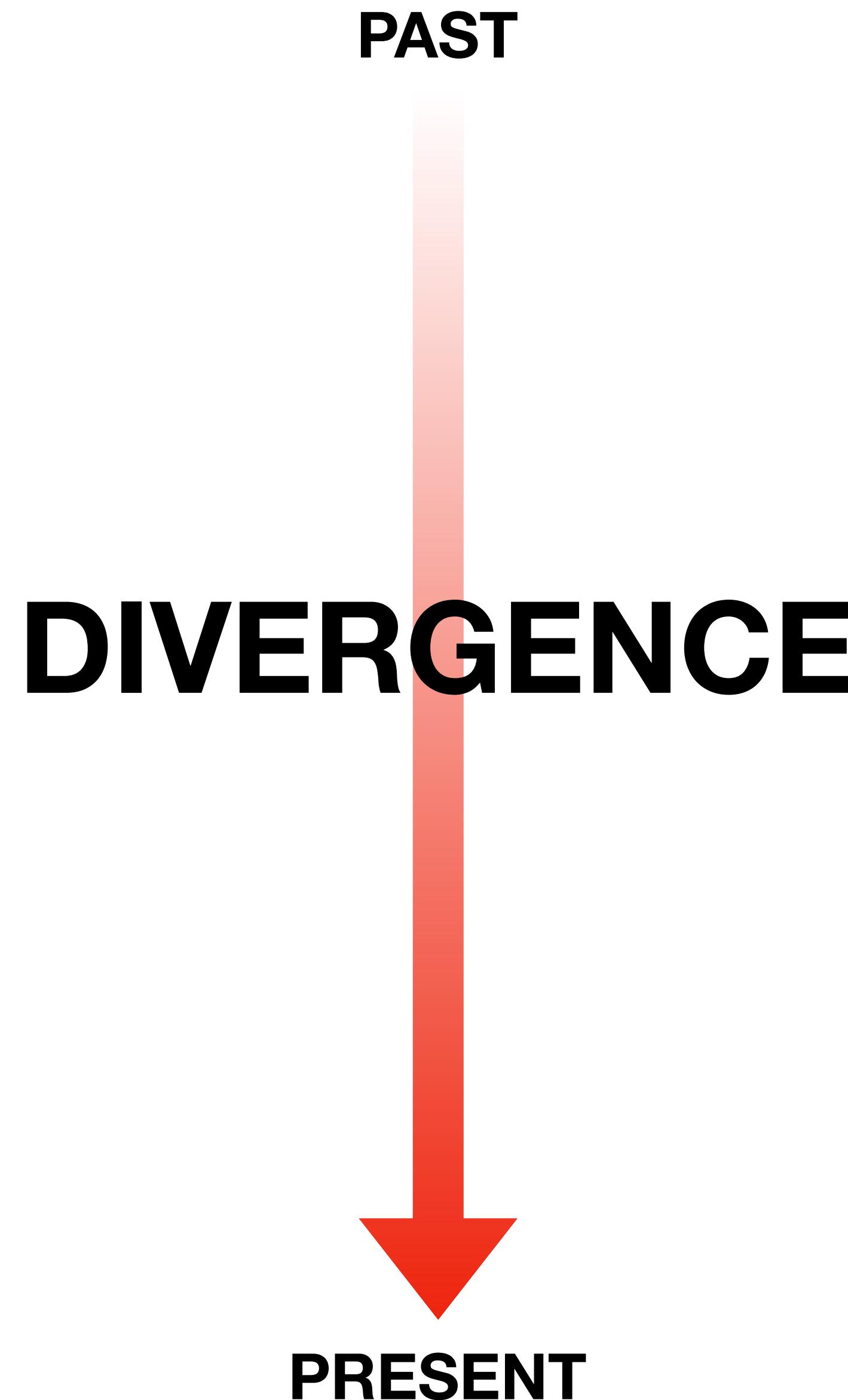
So...

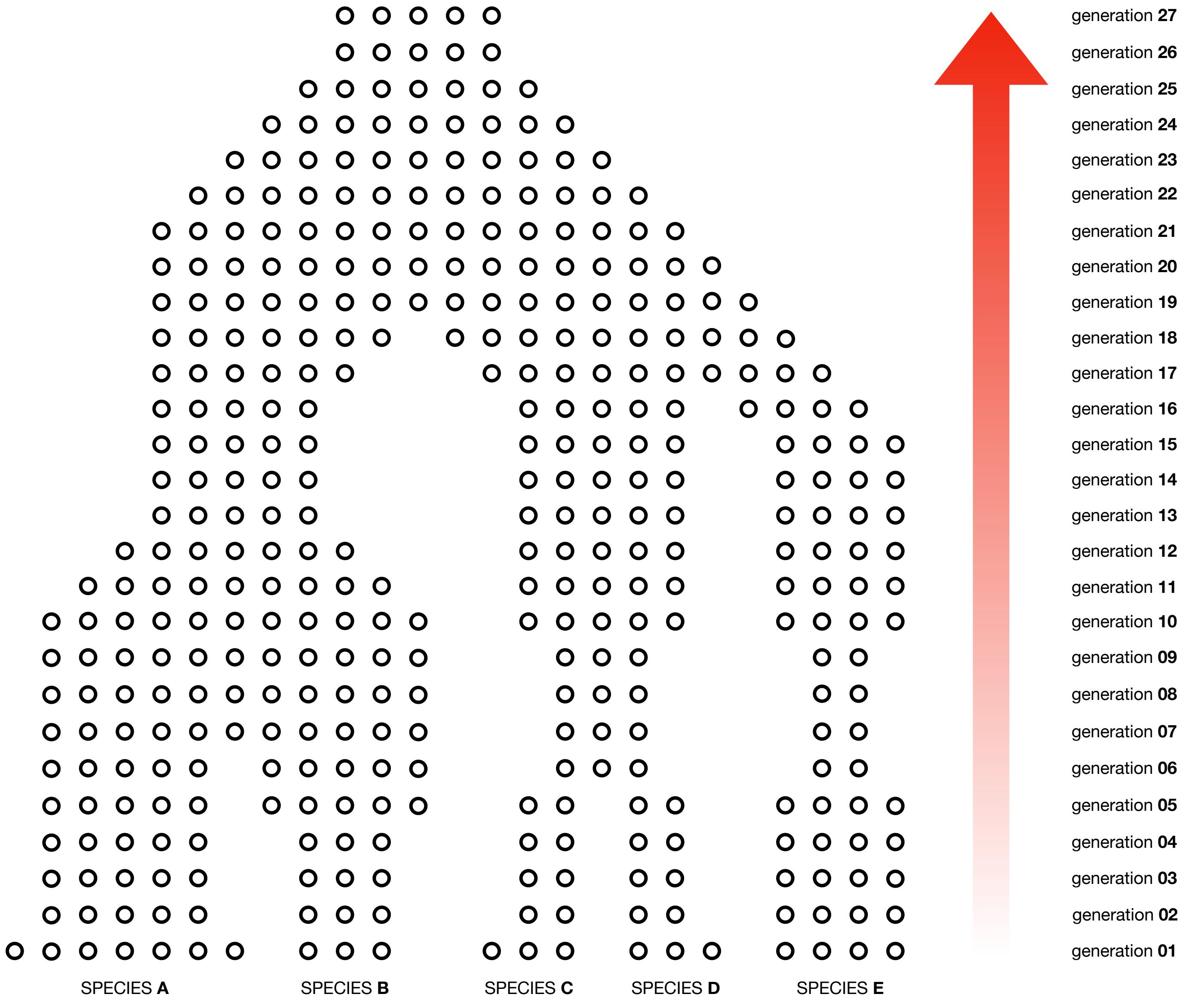
- the probability they coalesce in the previous generation:  $P(\text{coalescence}) = \frac{1}{2N_e}$
- the probability they don't coalesce in the previous generation:  $P(\text{no coalescence}) = 1 - \frac{1}{2N_e}$

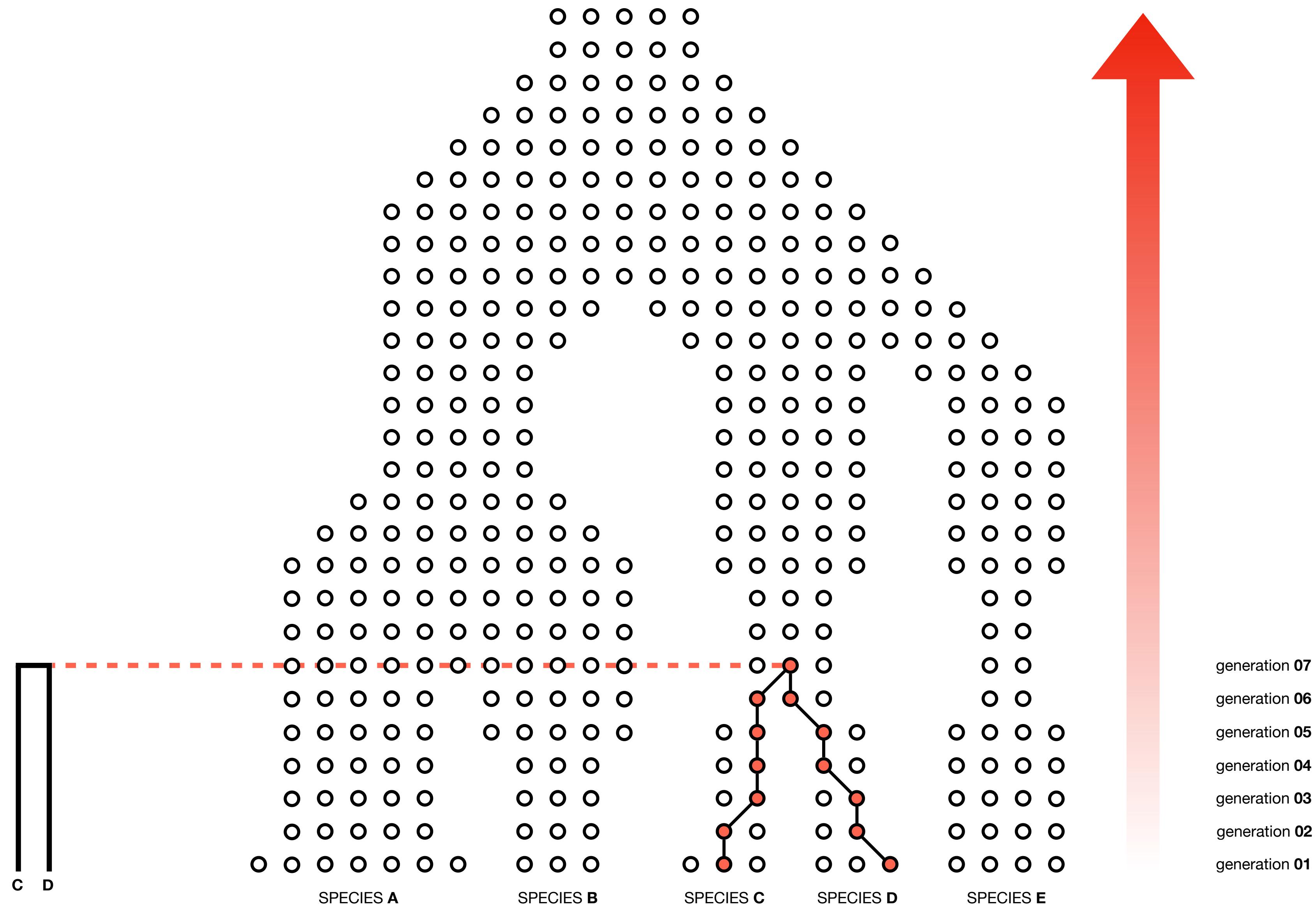
**What's the chance two gene copies share a common ancestor  $t$  generations ago?**

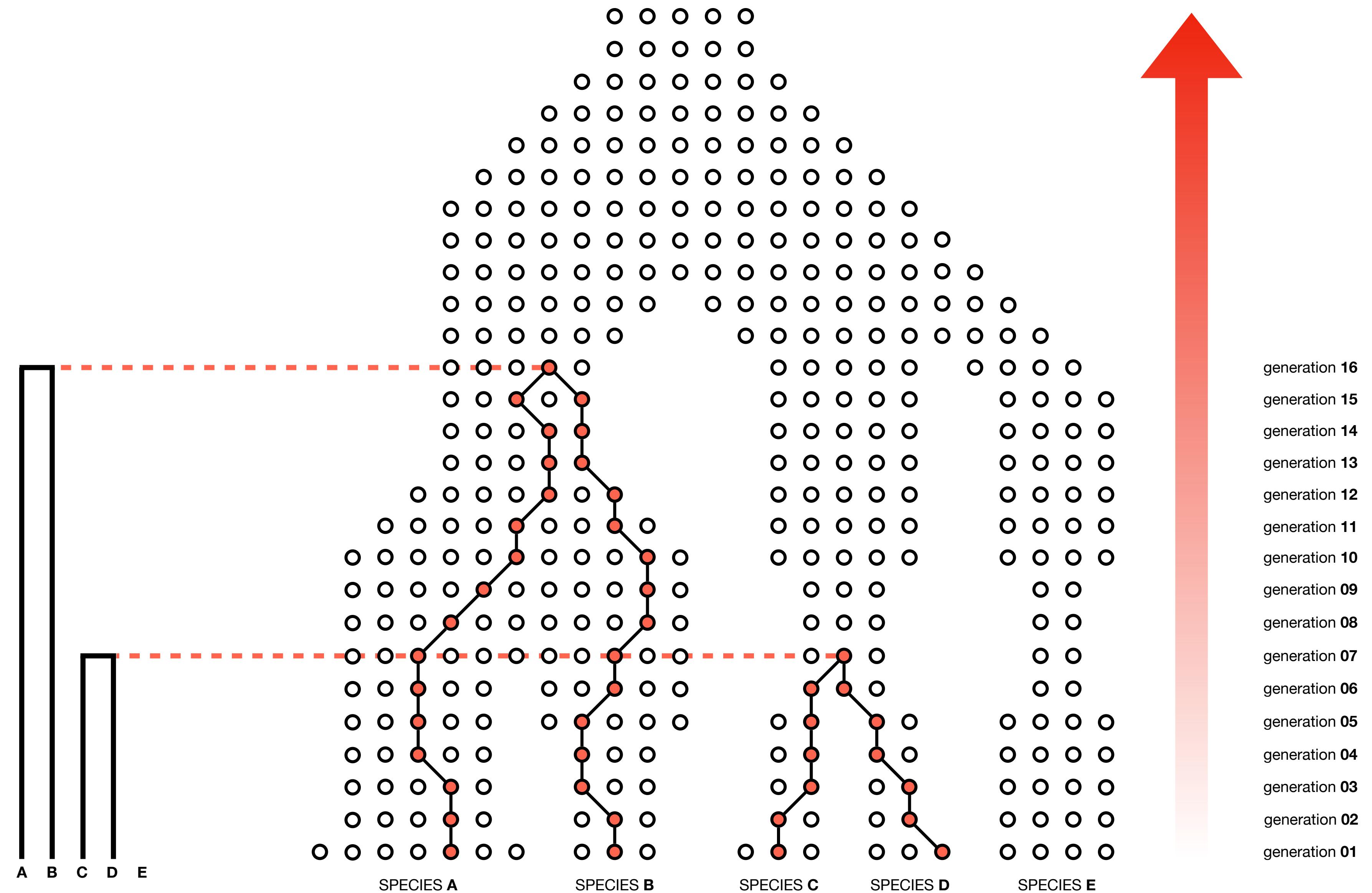
- the probability they coalesce a generation  $t$ :  $P(\text{coalescence at generation } t) = \left(1 - \frac{1}{2N_e}\right)^{t-1} \cdot \frac{1}{2N_e}$

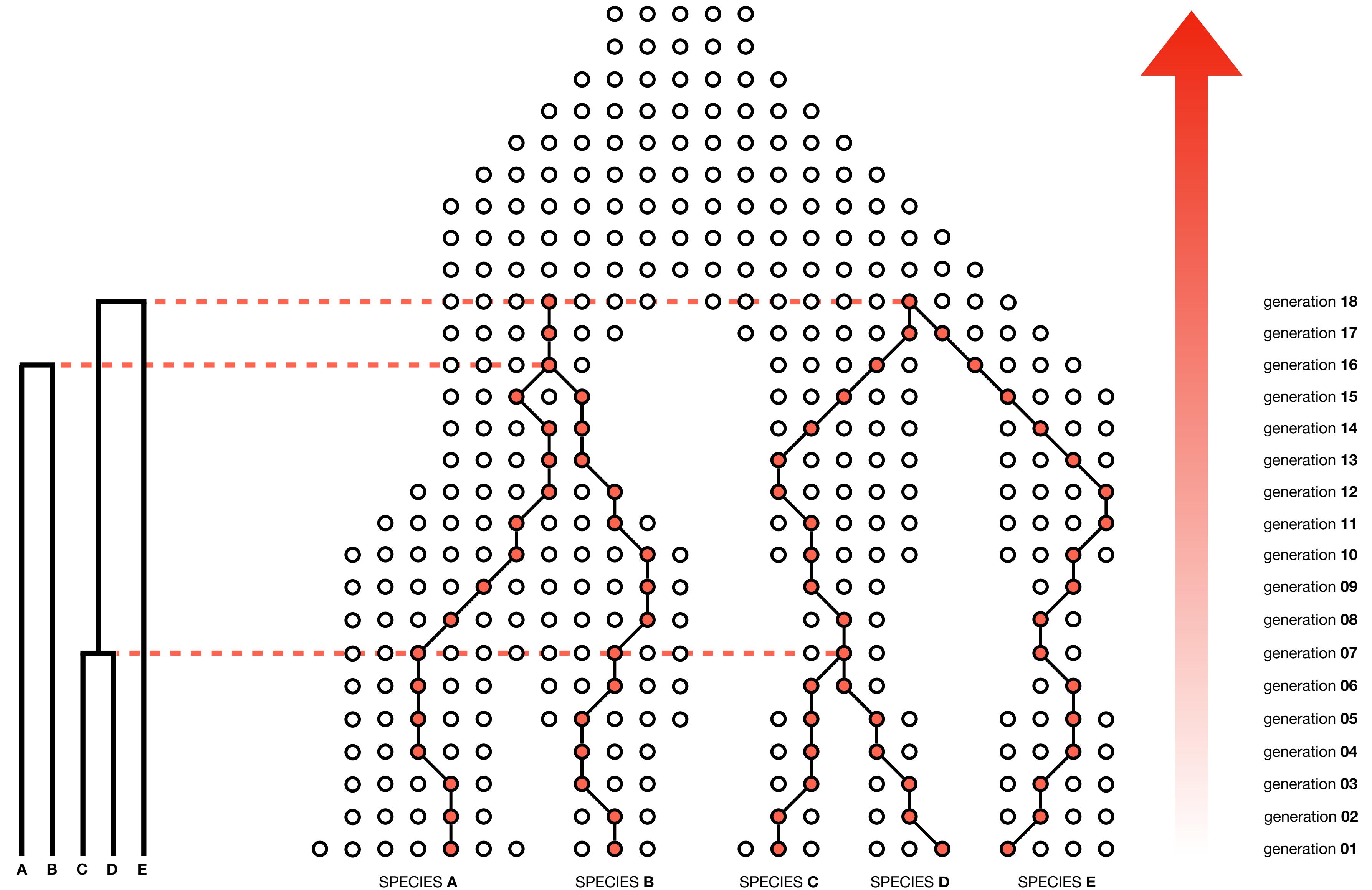
This forms the basis of the exponential distribution of coalescent times.

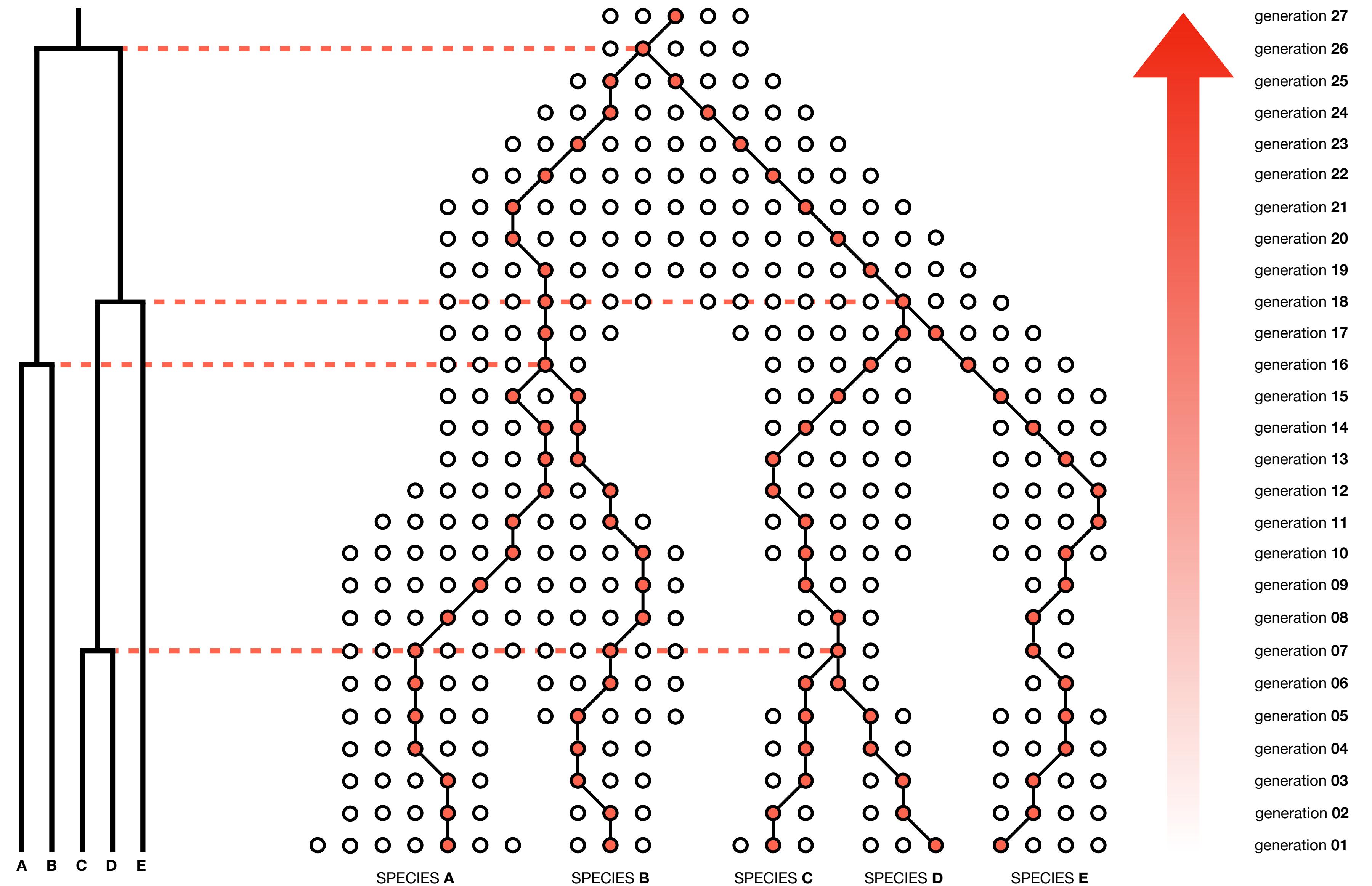












This is theoretically complicated ... 🤯 ... and we won't dive into the full mathematical depths of coalescent theory.

However, also consider that the coalescent gives us a **null model**  
- a baseline expectation - of how gene lineages and mutations would  
**coalesce back in time assuming ideal conditions.**

**Comparing observed data to predictions of the coalescent**, we can:

- detect population size changes
- infer population structure or migration
- identify signals of selection or recombination

## Gene tree summary methods for unrooted trees

- ASTRAL - gold standard for multi-locus species tree inference under ILS; very fast and accurate.
- ASTER - includes ASTRAL + CASTER + ASTRAL PRO!
- BUCKy - estimates concordance factors across genes; great for detecting gene tree conflict.
- ASTRID - fast; distance-based species tree inference using gene tree internode distances.

## Gene tree summary methods for rooted trees

- MP-EST - uses rooted gene trees and a pseudo-likelihood framework; sensitive to rooting errors.

## Site-based methods

- SVDQuartets - uses site patterns; no need for gene trees; efficient for medium datasets.
- SNAPP - bayesian method for biallelic SNP data; coalescent-based and very accurate, but slow.
- CASTER - coalescent-aware method for genome-wide data; uses site patterns, new.

## Bayesian co-estimation methods

- StarBEAST2 - full Bayesian co-estimation of gene and species trees.

In coalescent-based inference there are different:

## branch support metrics

- **Local Posterior Probability (LPP)**

Probability a branch is correct based on quartet frequencies.

Uses the Dirichlet prior - a pseudocount of 1 per topology:  $LPP_{T1} = \frac{(n_1 + 1)}{(n_1 + n_2 + n_3 + 3)}$   
Ranges from 0 to 1.

- **Multi-locus Bootstrap (MLBS)**

Branch stability across bootstrapped gene trees.

Ranges from 0 to 100%.

## branch length units

- **Coalescent Units**

The expected number of coalescent events per unit time, scaled by population size.

So, 1 coalescent unit =  $2N_e$  generations. This scaling allows comparisons of coalescent processes across populations of different sizes.

**FINISH**