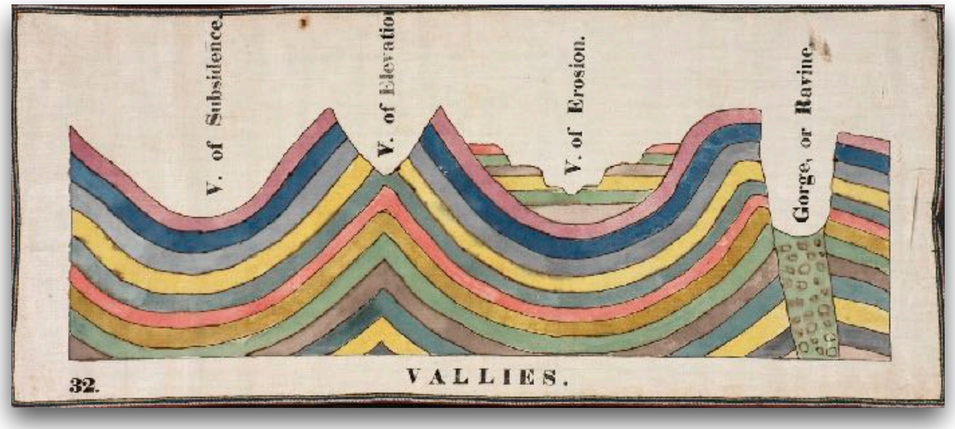
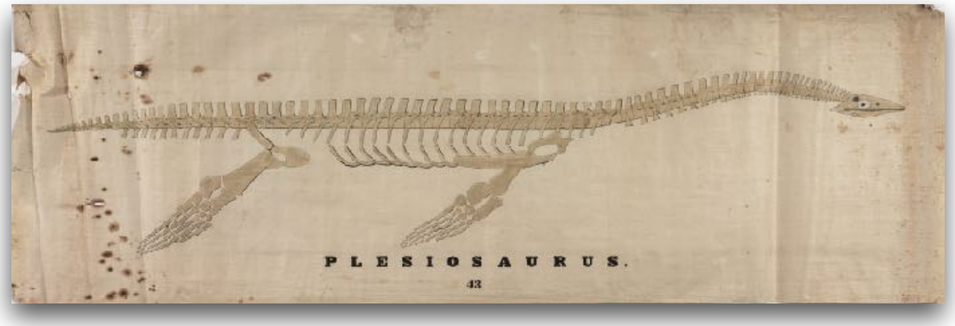




**discordance,**
**ILS & the coalescent**

## SPECIES TREE

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

## GENE TREE(S)

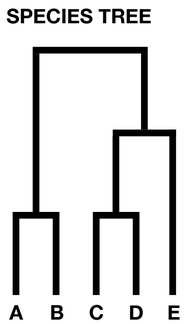
Traces the evolutionary history of a single locus and may differ biologically from the species tree.

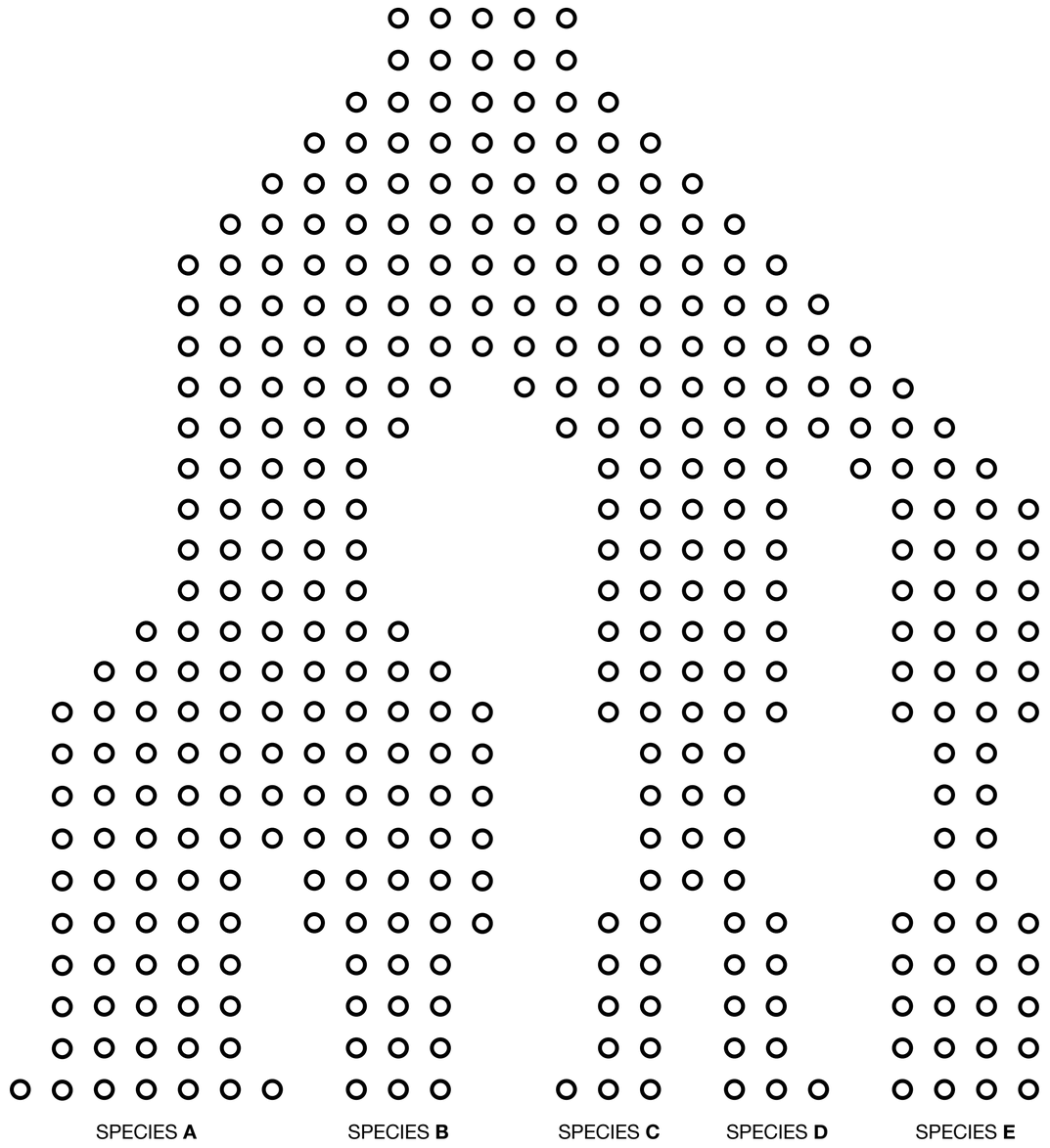
A common cause of gene tree discordance is Incomplete Lineage Sorting **(ILS)**,
which is especially likely when:

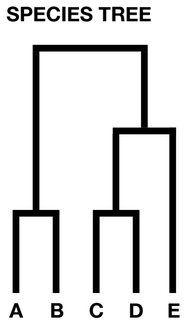
- effective population sizes are large
- speciation events are closely spaced in time

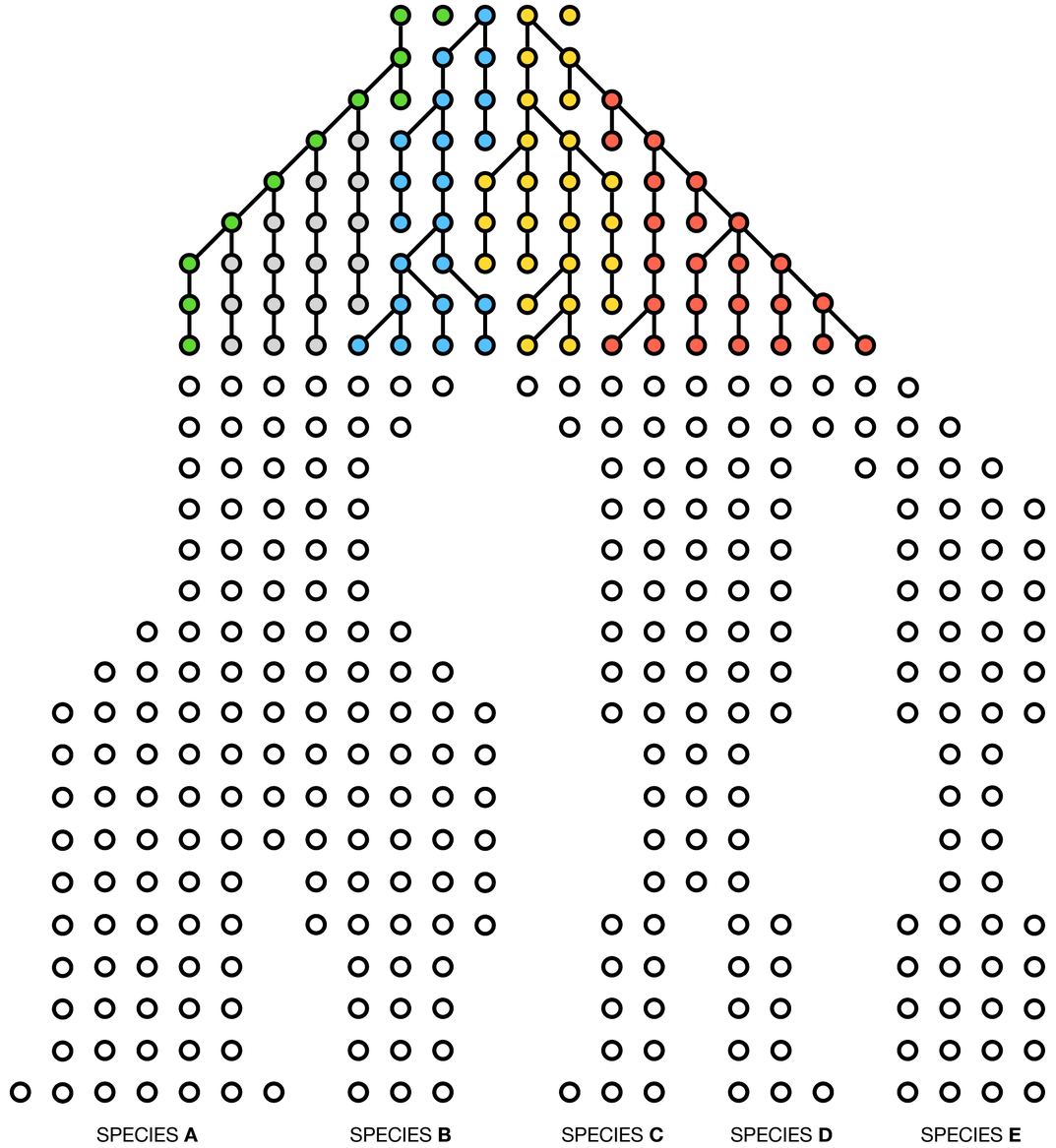
You can think of **ILS** as:

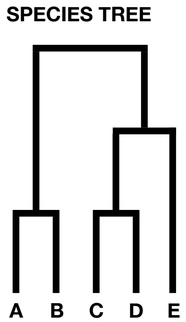
- ancestral polymorphisms that persist across speciation events,
- alleles sorting into descendant lineages *erratically*, not reflecting the species splits.

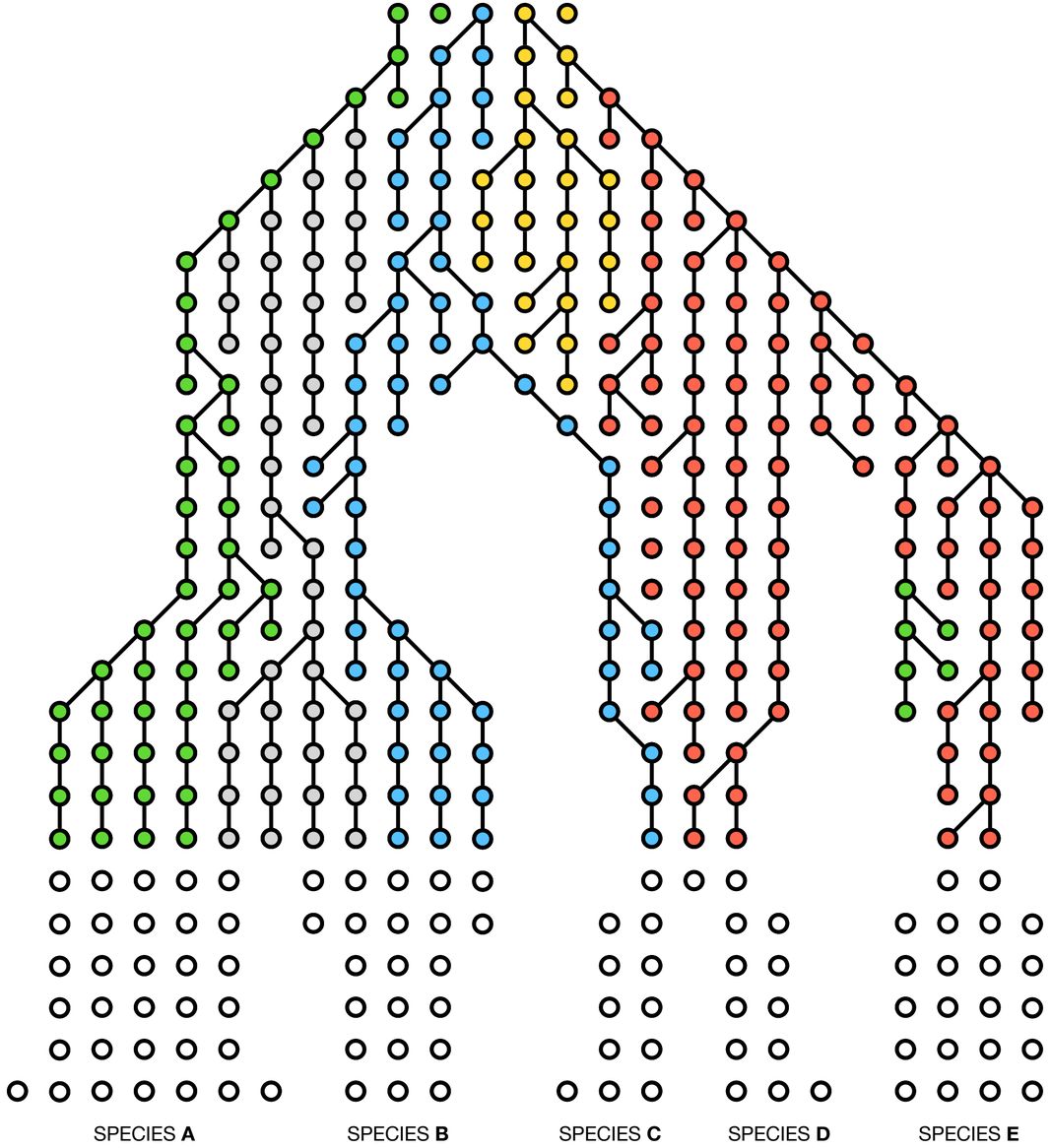

SPECIES TREE
A B C D E


SPECIES A
SPECIES B
SPECIES C
SPECIES D
SPECIES E

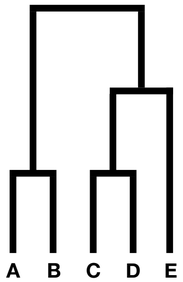

SPECIES TREE
A B C D E


SPECIES A
SPECIES B
SPECIES C
SPECIES D
SPECIES E

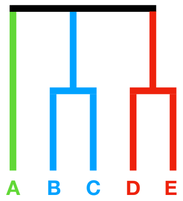

SPECIES TREE
A B C D E

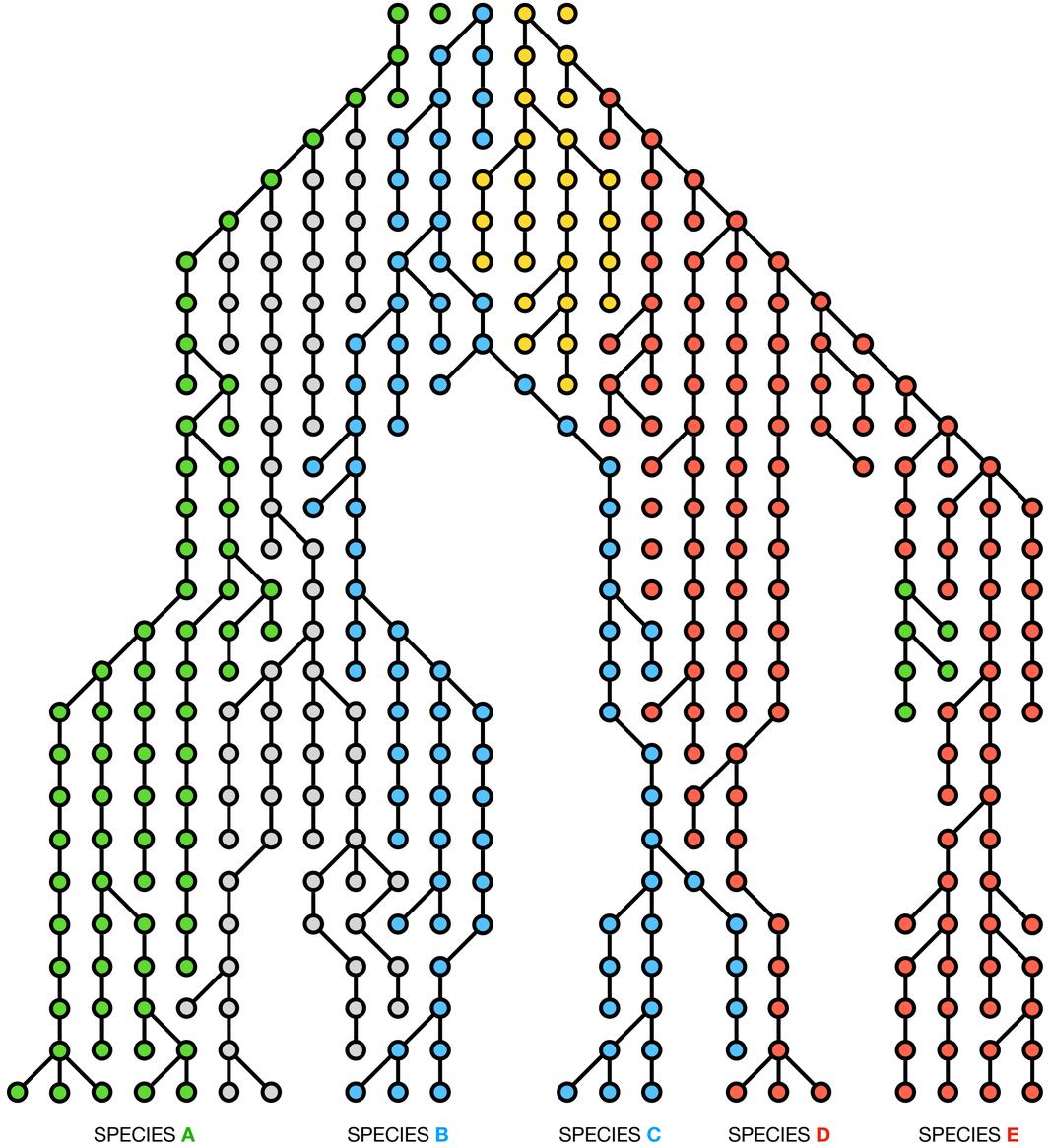

SPECIES A
SPECIES B
SPECIES C
SPECIES D
SPECIES E

SPECIES TREE




GENE TREE



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**)
  Vertical transfer between species through hybridisation 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

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

## COALESCENT-BASED SPECIES TREE INFERENCE

- takes as input a set of unrooted, binary gene trees (each with leaves labeled by species)
- infers a species tree from multiple discordant gene histories due to e.g. 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.

Coalescent theory sits between phylogenomics and population genomics.

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 demographichistory 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 2 GENE COPIES COALESCE?

Imagine two gene copies sampled from a population of diploid individuals. 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.

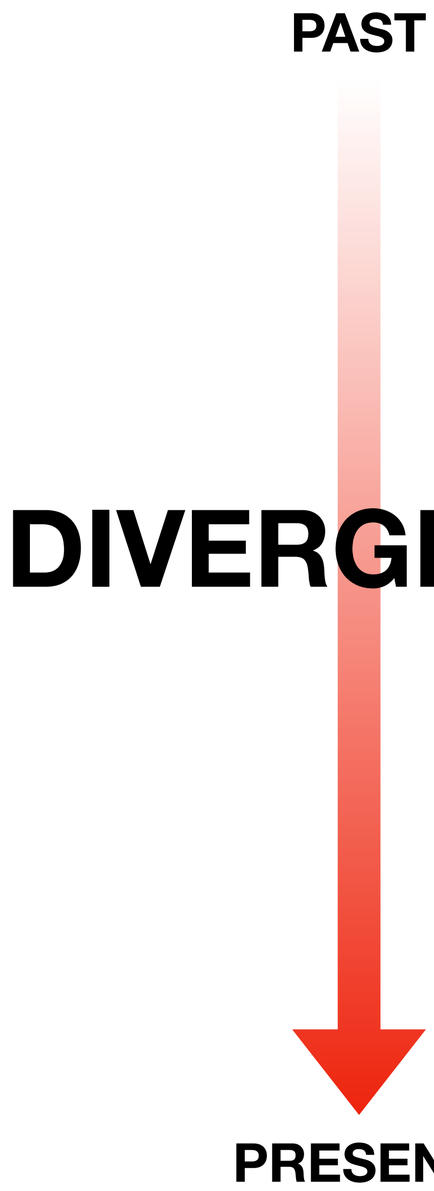
So…

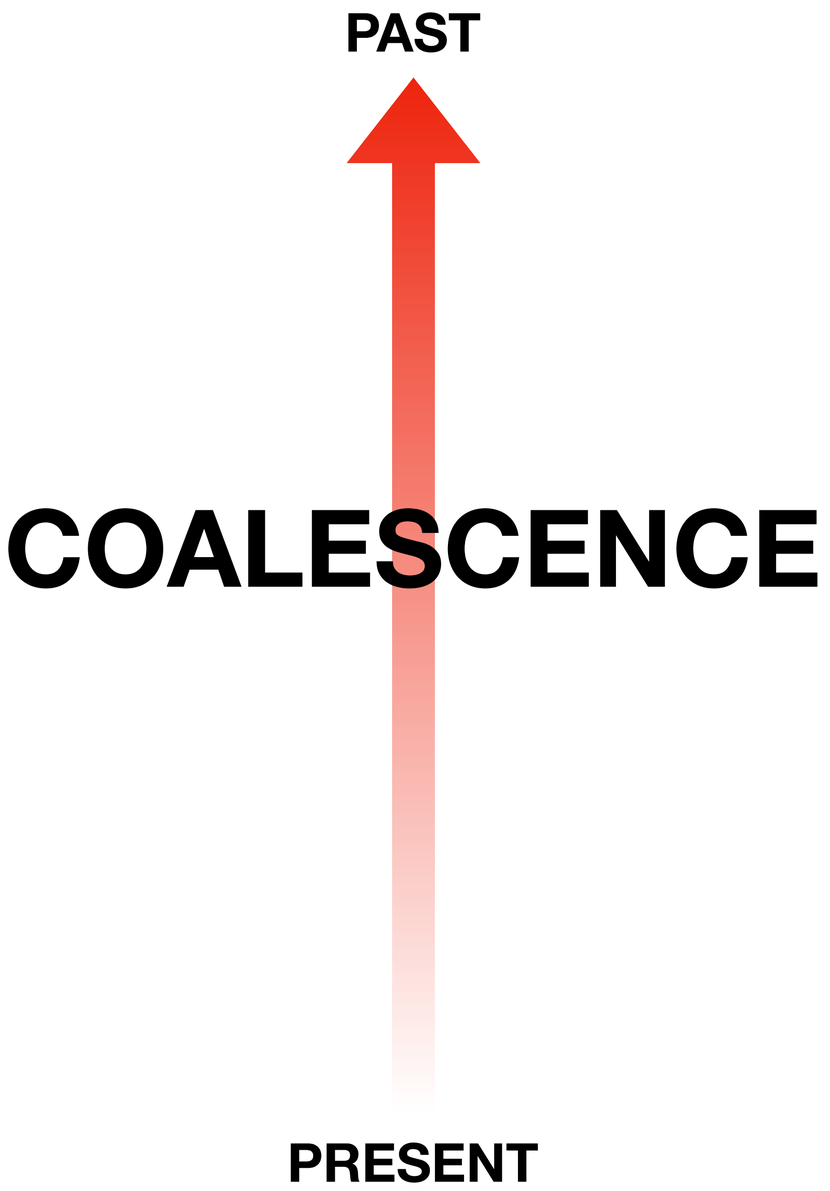
- the probability they coalesce in the previous generation $P(\text{coalescence}) = \frac{1}{2N_e}$
- the probability they don't coalesce in the previous generation: $P(\text{no coalescence}) = 1 - \frac{1}{2N_e}$

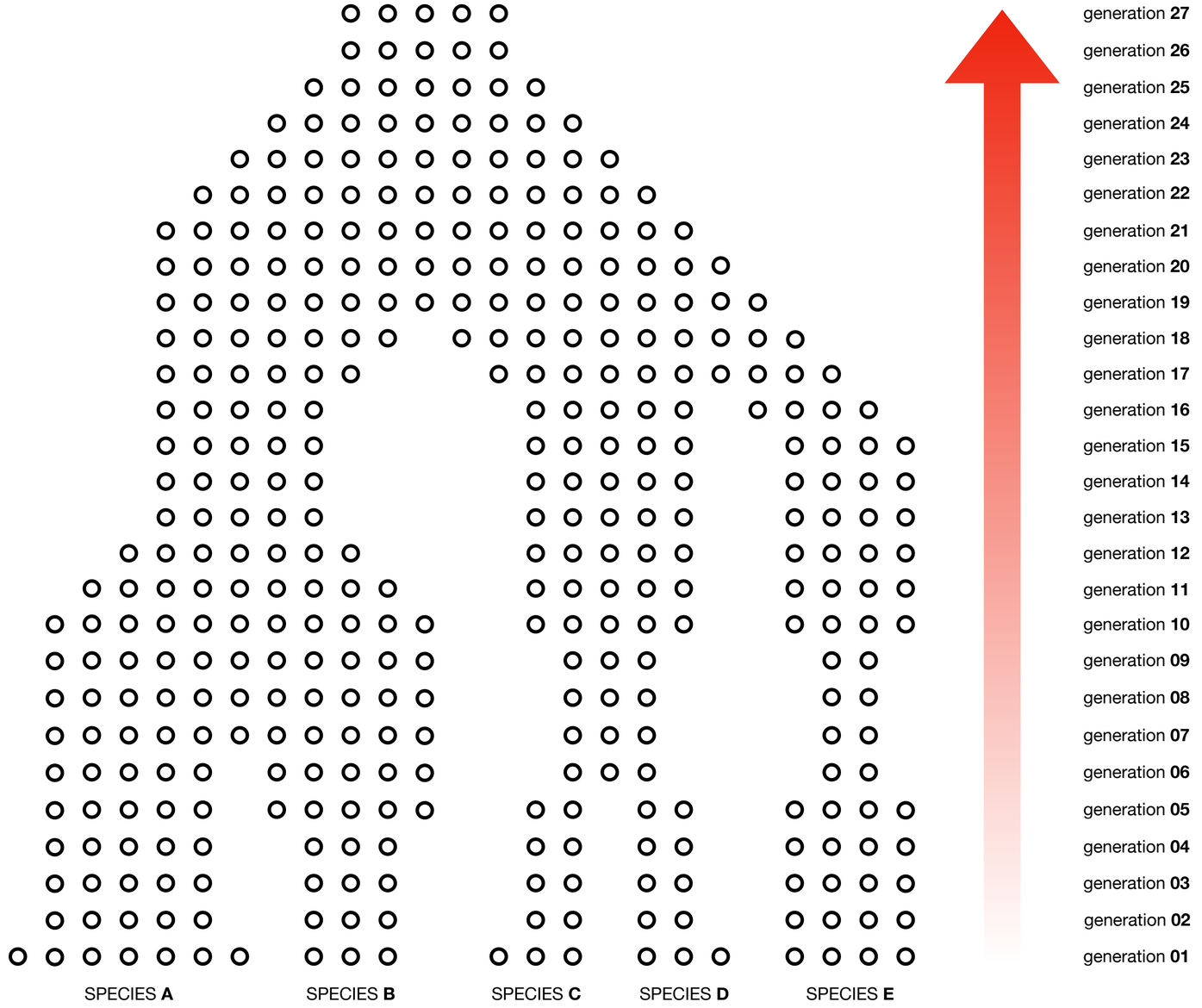
## WHAT'S THE CHANCE 2 GENE COPIES SHARE A COMMON ANCESTOR *T* GENERATIONS AGO?

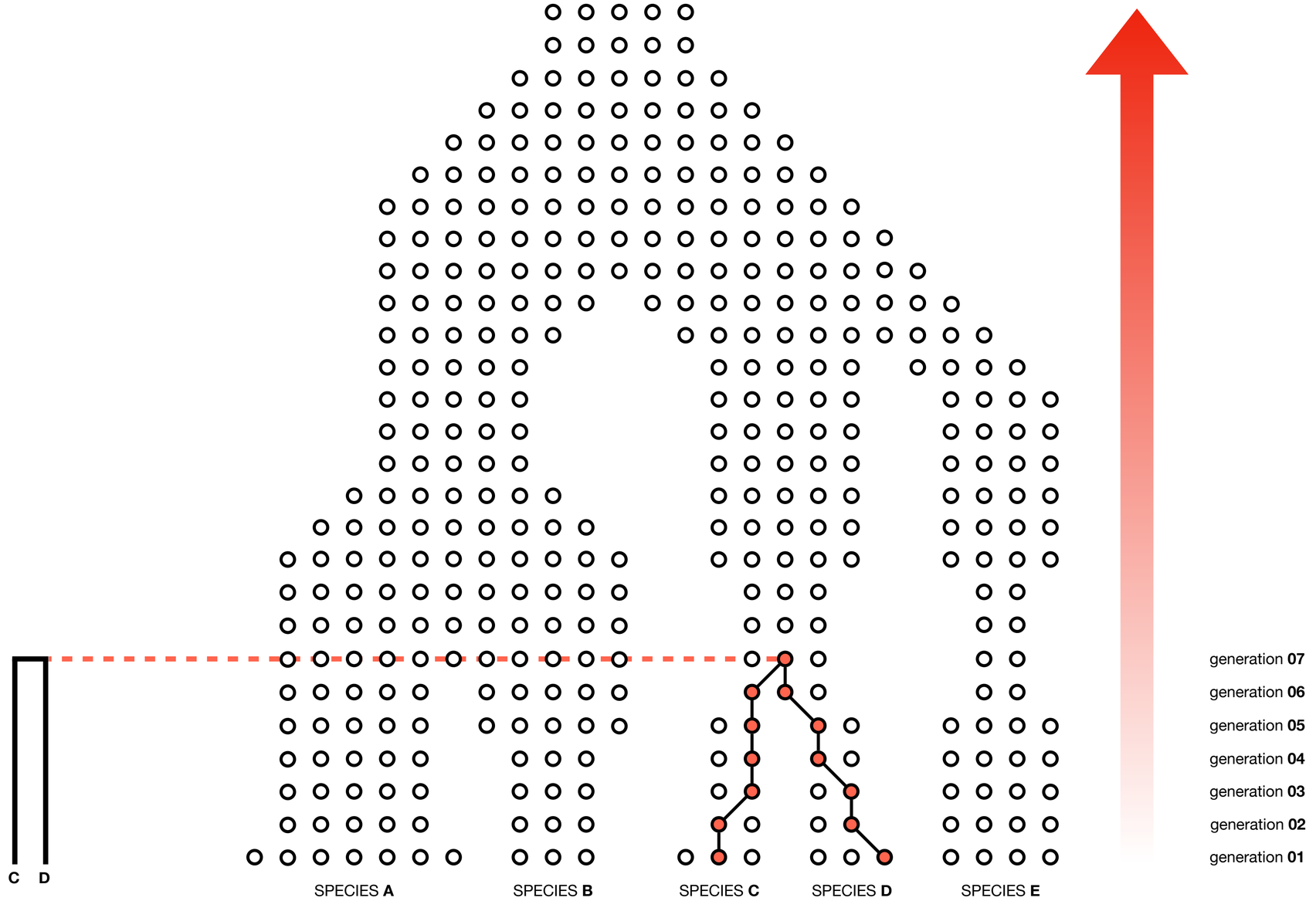
- the probability they coalesce a generation t: $P(\text{coalescence at generation } t) = \left(1 - \frac{1}{2N_e}\right)^{t-1} \cdot \frac{1}{2N_e}$

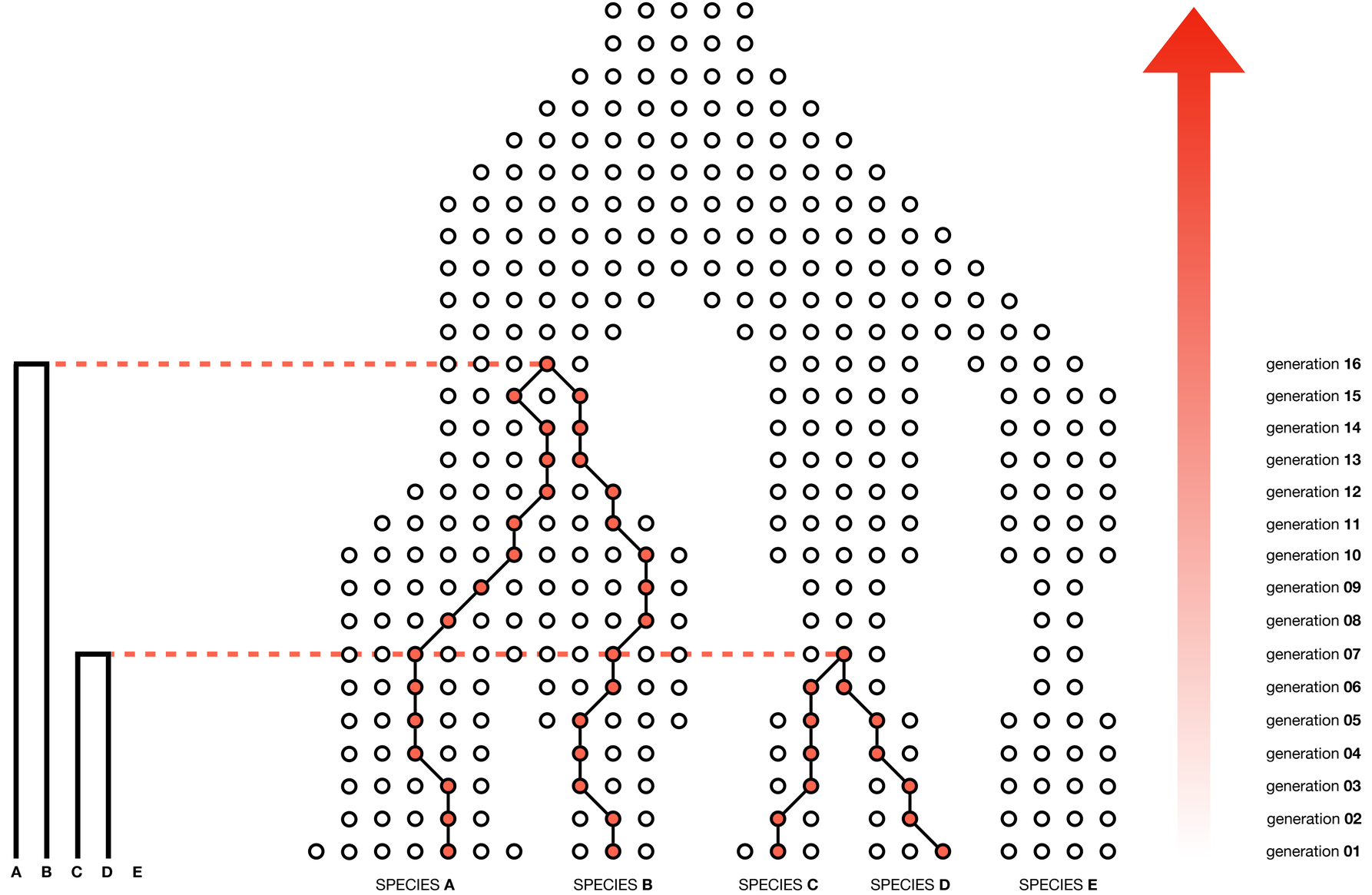
This forms the basis of the exponential distribution of coalescent times.

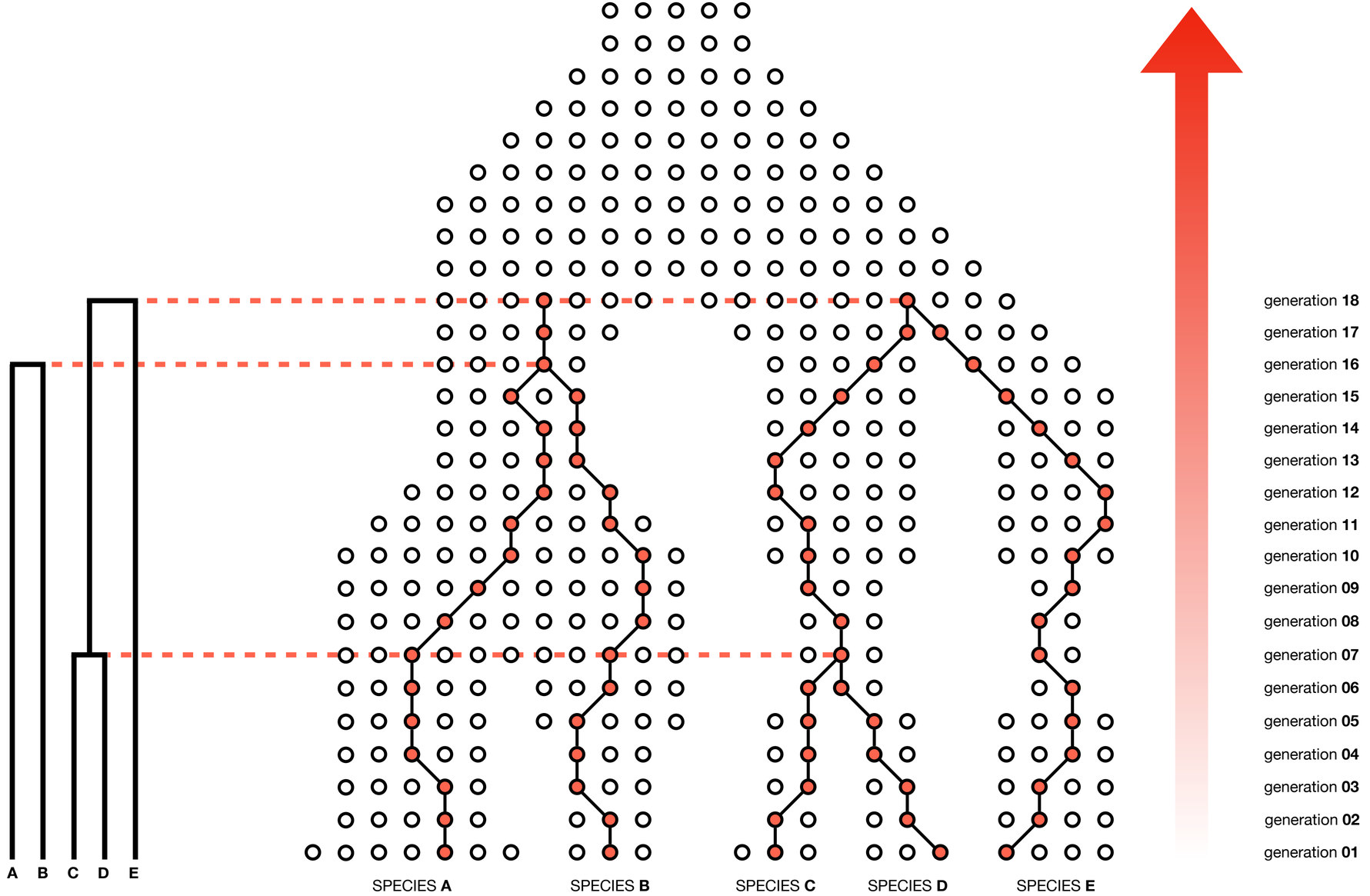

PAST
DIVERGENCE
PRESENT

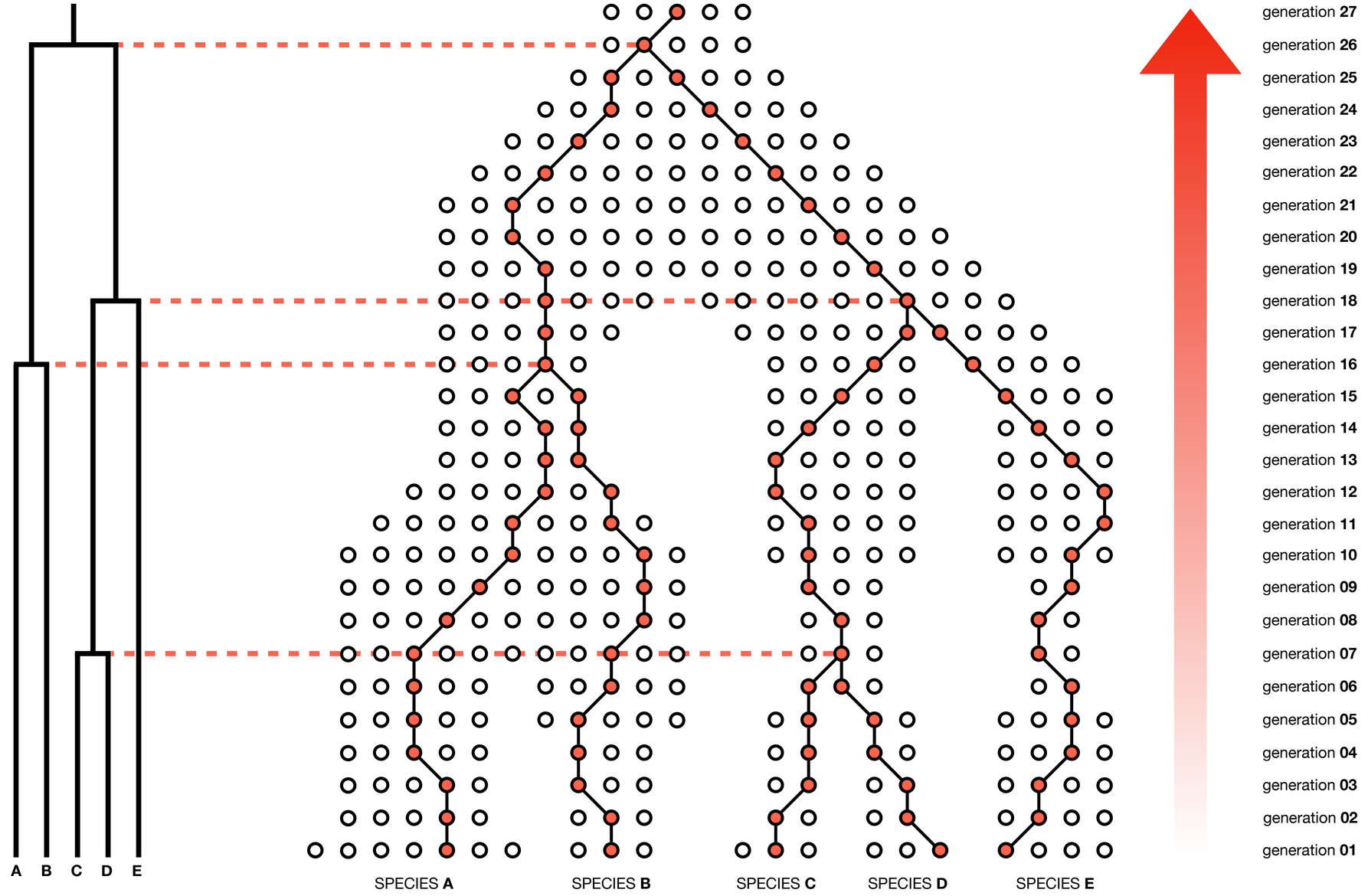

PAST
COALESCENCE
PRESENT


generation 27
generation 26
generation 25
generation 24
generation 23
generation 22
generation 21
generation 20
generation 19
generation 18
generation 17
generation 16
generation 15
generation 14
generation 13
generation 12
generation 11
generation 10
generation 09
generation 08
generation 07
generation 06
generation 05
generation 04
generation 03
generation 02
generation 01
SPECIES A
SPECIES B
SPECIES C
SPECIES D
SPECIES E


C
D
SPECIES A
SPECIES B
SPECIES C
SPECIES D
SPECIES E
generation 07
generation 06
generation 05
generation 04
generation 03
generation 02
generation 01


A B C D E
SPECIES A
SPECIES B
SPECIES C
SPECIES D
SPECIES E
generation 16
generation 15
generation 14
generation 13
generation 12
generation 11
generation 10
generation 09
generation 08
generation 07
generation 06
generation 05
generation 04
generation 03
generation 02
generation 01


A B C D E
SPECIES A
SPECIES B
SPECIES C
SPECIES D
SPECIES E
generation 18
generation 17
generation 16
generation 15
generation 14
generation 13
generation 12
generation 11
generation 10
generation 09
generation 08
generation 07
generation 06
generation 05
generation 04
generation 03
generation 02
generation 01


A
B
C
D
E
SPECIES A
SPECIES B
SPECIES C
SPECIES D
SPECIES E
generation 27
generation 26
generation 25
generation 24
generation 23
generation 22
generation 21
generation 20
generation 19
generation 18
generation 17
generation 16
generation 15
generation 14
generation 13
generation 12
generation 11
generation 10
generation 09
generation 08
generation 07
generation 06
generation 05
generation 04
generation 03
generation 02
generation 01

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

# Coalescent-based species tree inference:

- ## Gene tree summary methods

  - first estimate individual gene trees (e.g. with ML or Bayesian methods)
  - summarize them into a species tree using properties predicted by the coalescent.

These methods do not model the coalescent process directly to statistically infer the correct species tree. **Examples**: ASTRAL, ASTRID, MP-EST, BUCKy

- ## Site-based methods

Skip gene tree estimation entirely and work directly from sequence alignments or SNPs. Use site patterns predicted under the coalescent to infer the species tree. **Examples**: SVDQuartets, SNAPP

- ## Bayesian co-estimation methods

Jointly estimate gene trees and the species tree under a full coalescent model. Most statistically rigorous but computationally expensive, thus limited to smaller datasets. **Examples**: StarBEAST2

## How do summary methods rely on the coalescent?

The multispecies coalescent predicts that for any a quartet, the most frequent gene tree topology matches the species tree.

The two discordant topologies are equally probable and always less frequent.

ASTRAL Breaks gene trees into quartets, counts which topology is most frequent for each, and assembles the species tree consistent with the largest number of quartet topologies.

ASTRAL does not simulate coalescence or estimate $N_e$ directly. Its statistical correctness depends entirely on the MSC being true: most common quartet topology = true species tree topology.

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: $\mathrm{LPP}_{T1} = \frac{(n_1 + 1)}{(n_1 + n_2 + n_3 + 3)}$
  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