Gene trees, species trees and the coalescent

Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu, twitter:snacktavish

(With thanks to Peter Beerli, Laura Kubatko, Mick Elliot and Arne Mooers for slides)

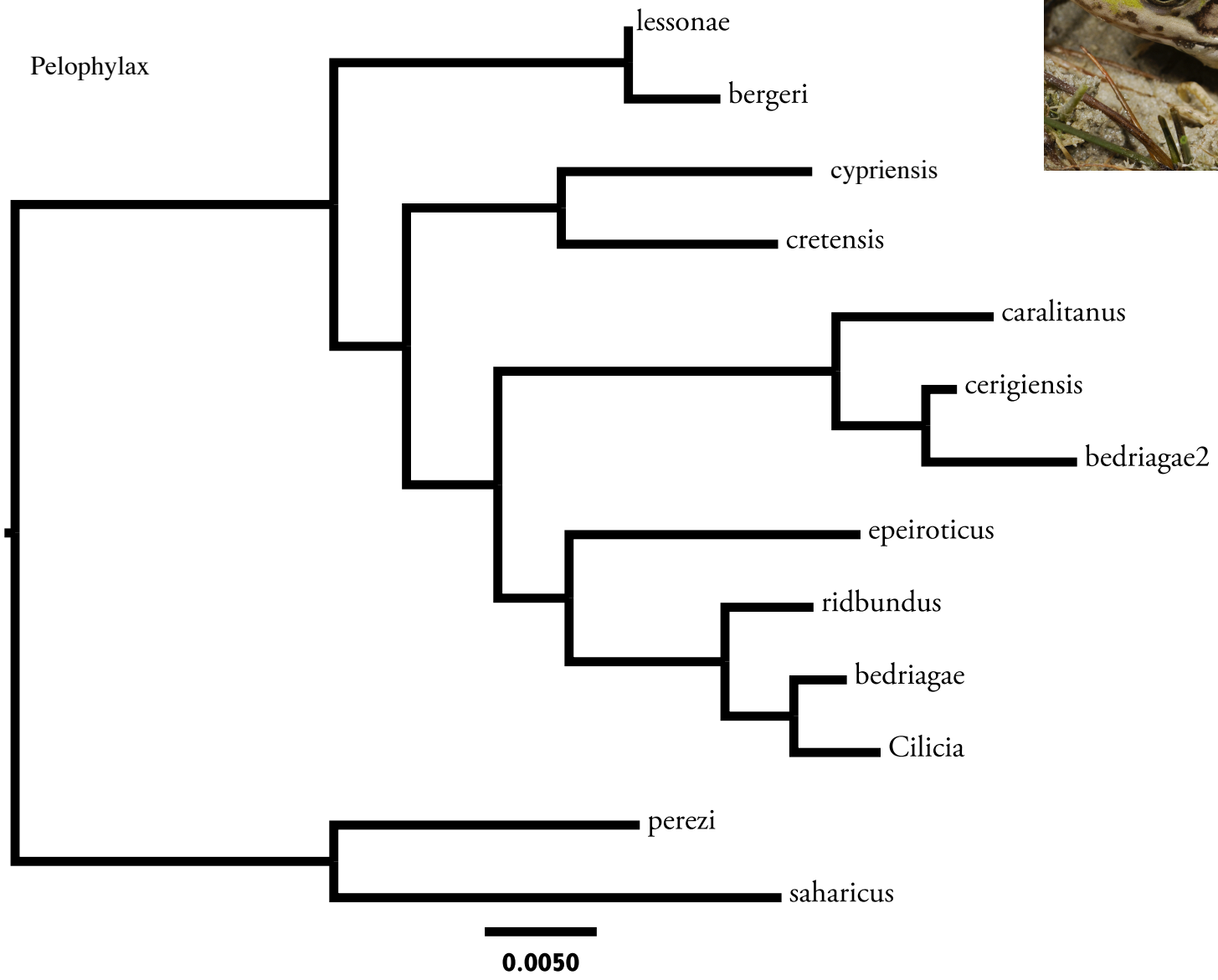
Relationship between phylogenetics and Population Genetics

- Population genetics: Study of genetic variation within a population
- Phylogenetics: Use genetic variation between taxa (species, populations) to infer evolutionary relationship

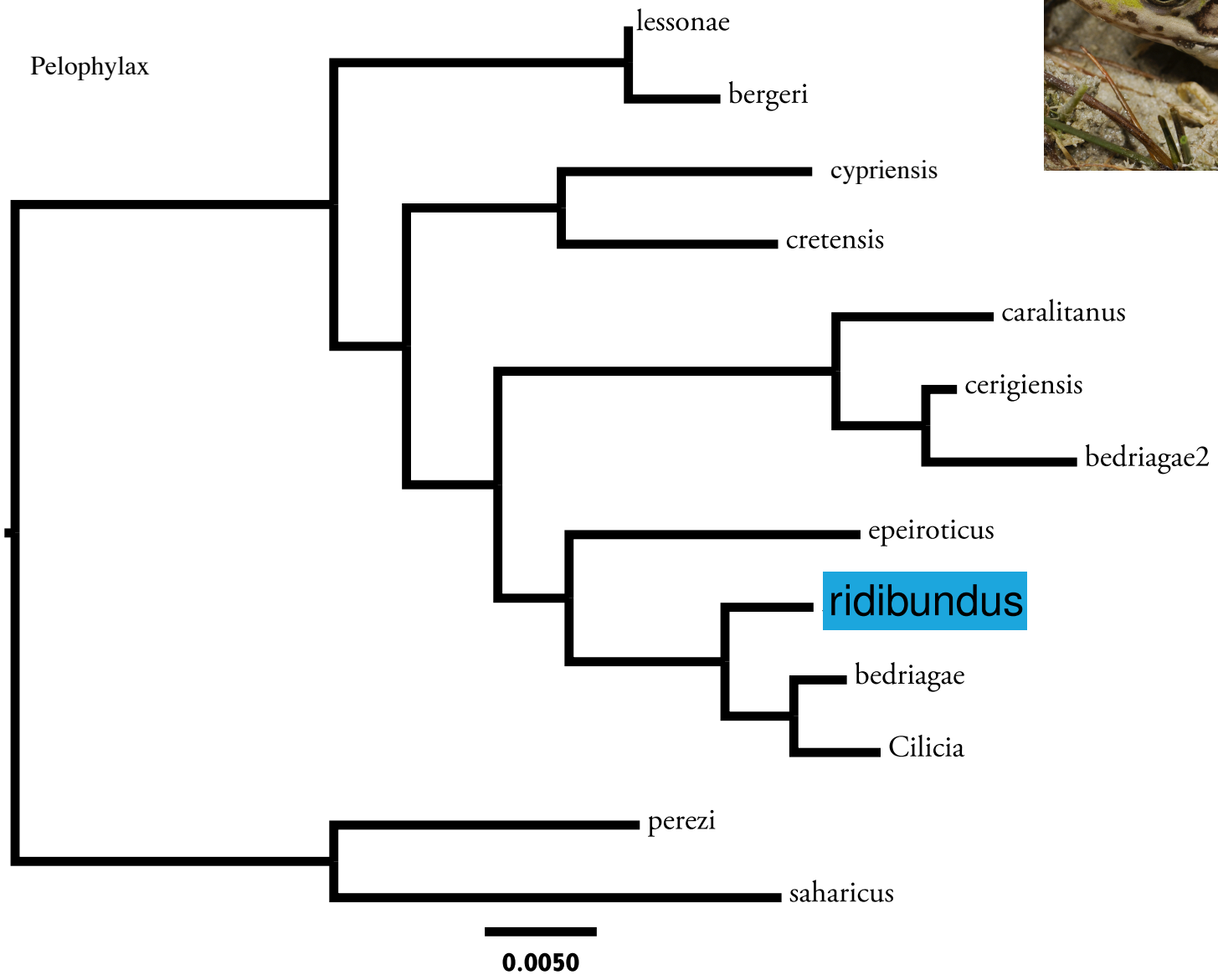
We expected tree like relationships among alleles within species.
The expectations for these trees is described by ‘coalescent theory’
(Kingman 1982)

What are reasons that a 'gene tree' may not show the same relationships as the species tree?

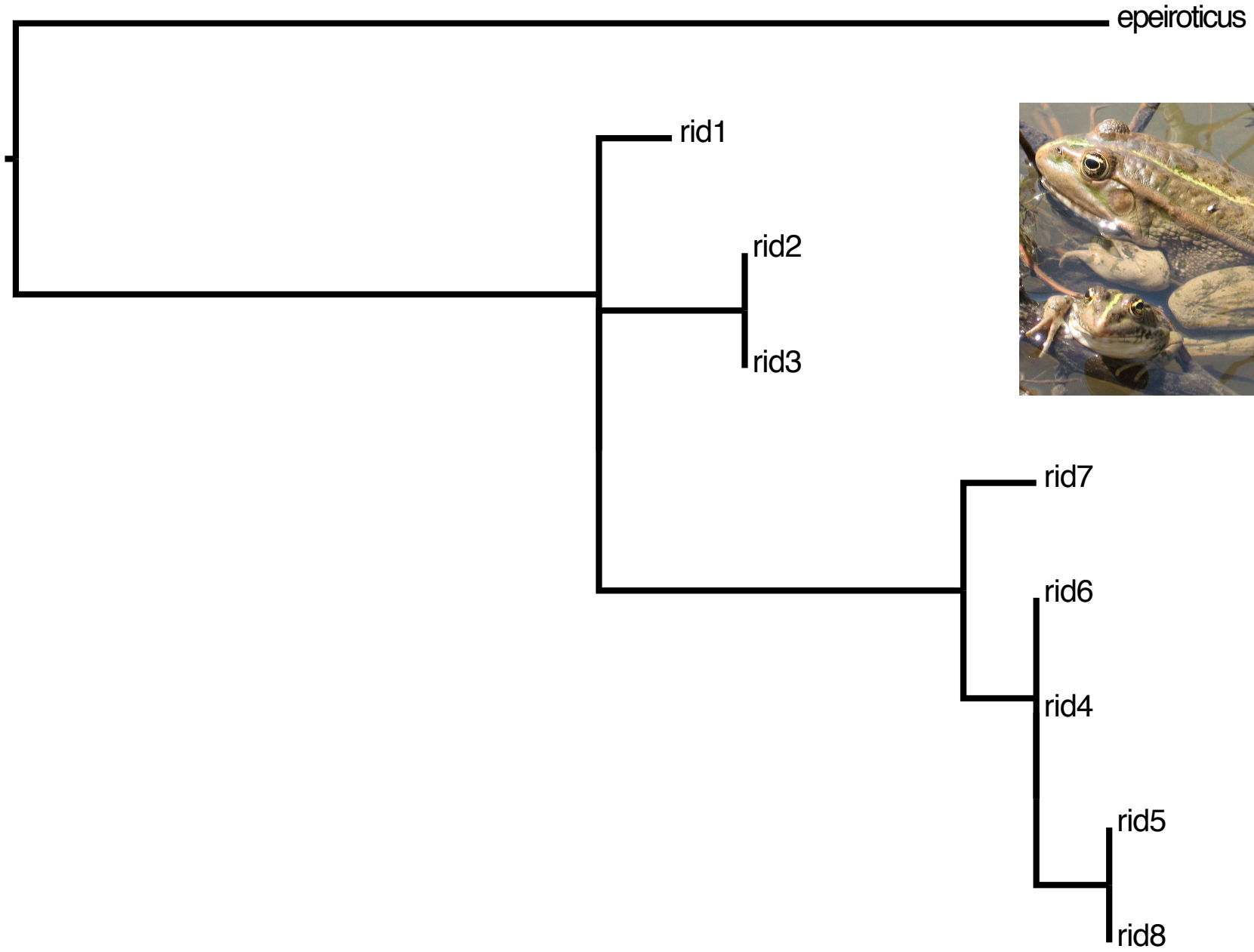
Species trees



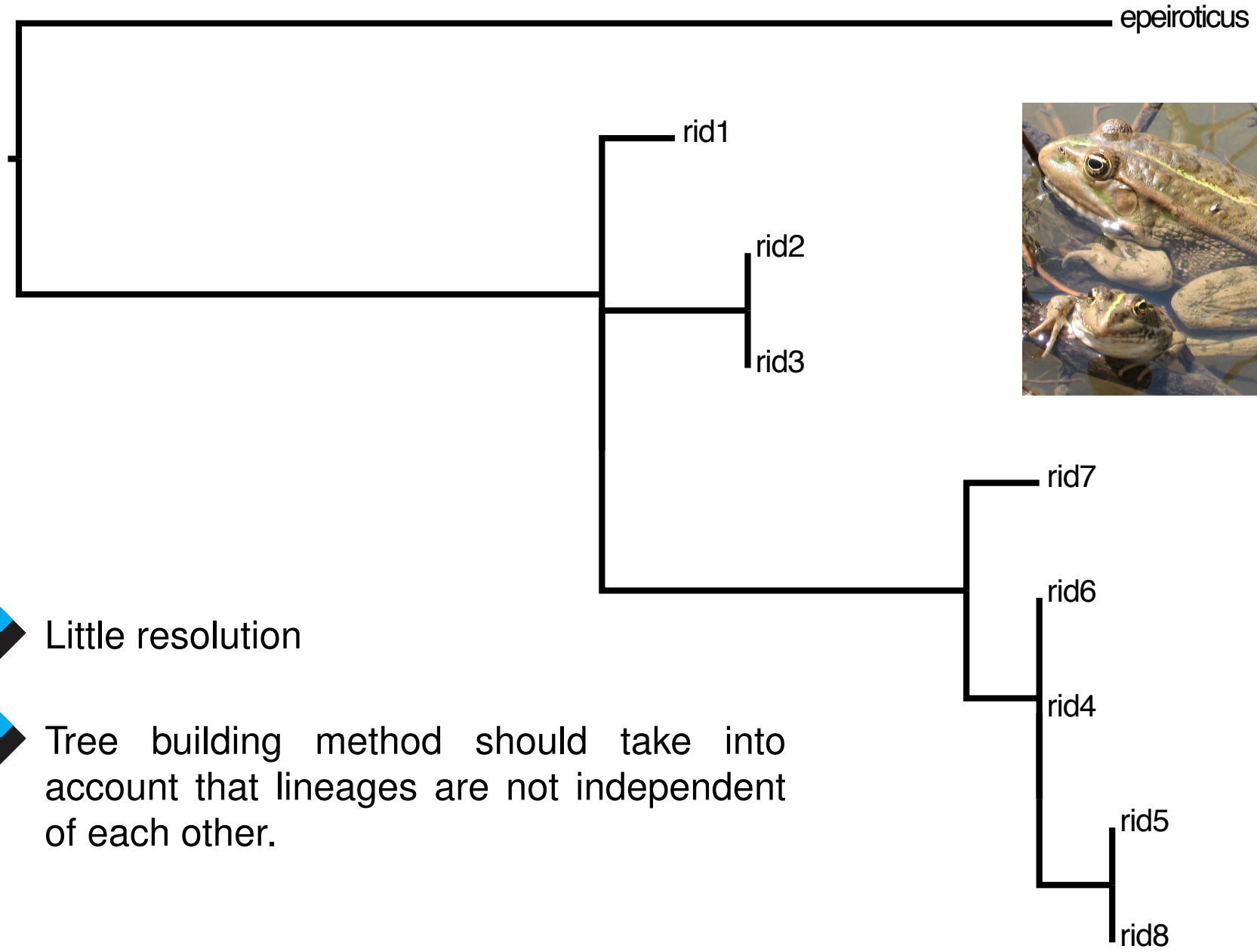
Species trees



Tree of individuals of same species

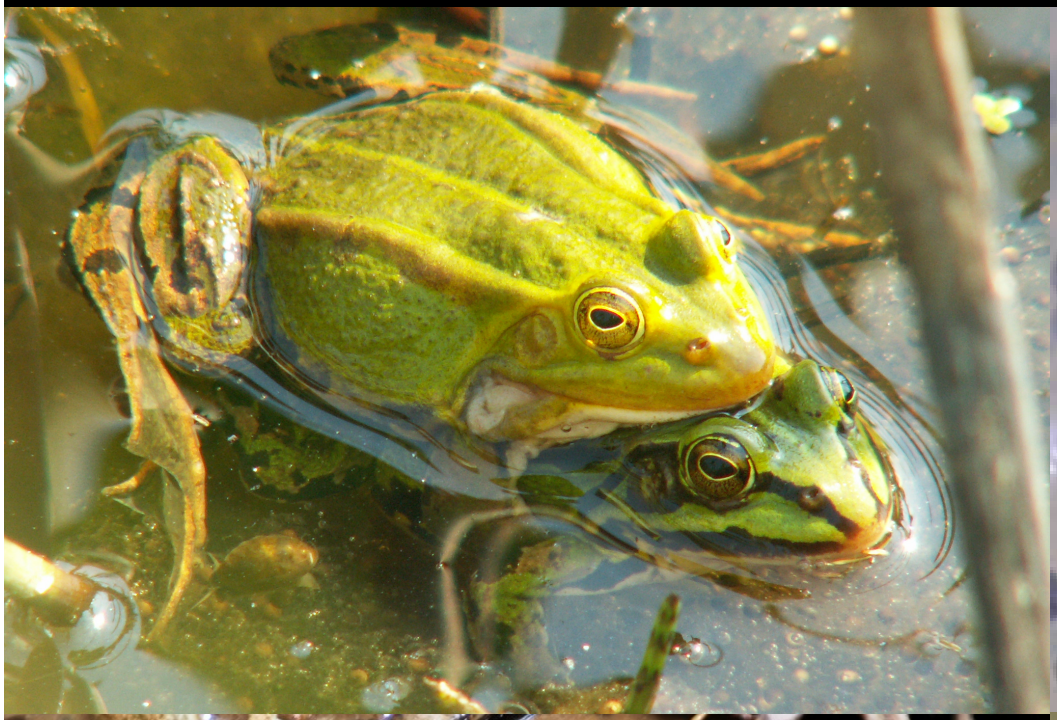


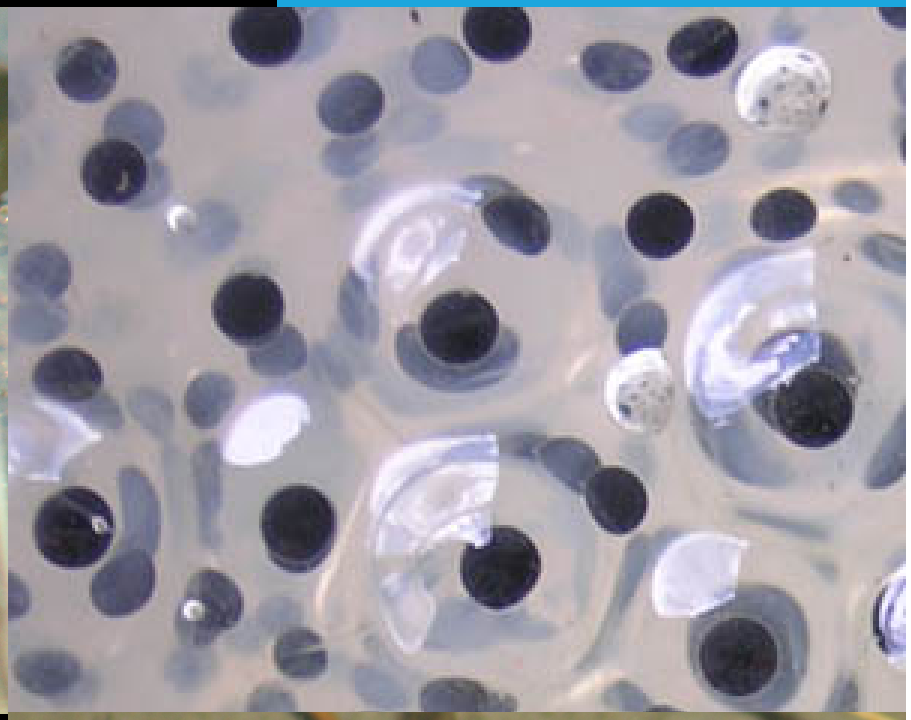
Tree of individuals of same species



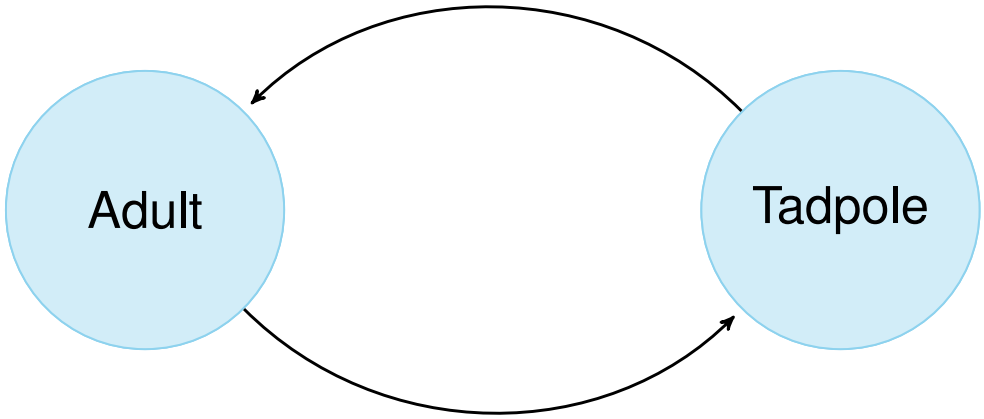
- Little resolution
- Tree building method should take into account that lineages are not independent of each other.

Interaction among individuals





Interaction among individuals



Wright-Fisher population model

- All individuals live one generation and get replaced by their offspring
- All have same chance to reproduce, all are equally fit
- The number of individuals in the population is constant

Population model

Past

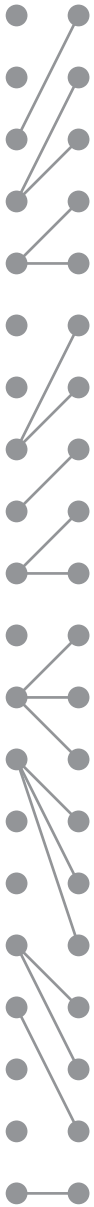
Present

Population model

Past

Present

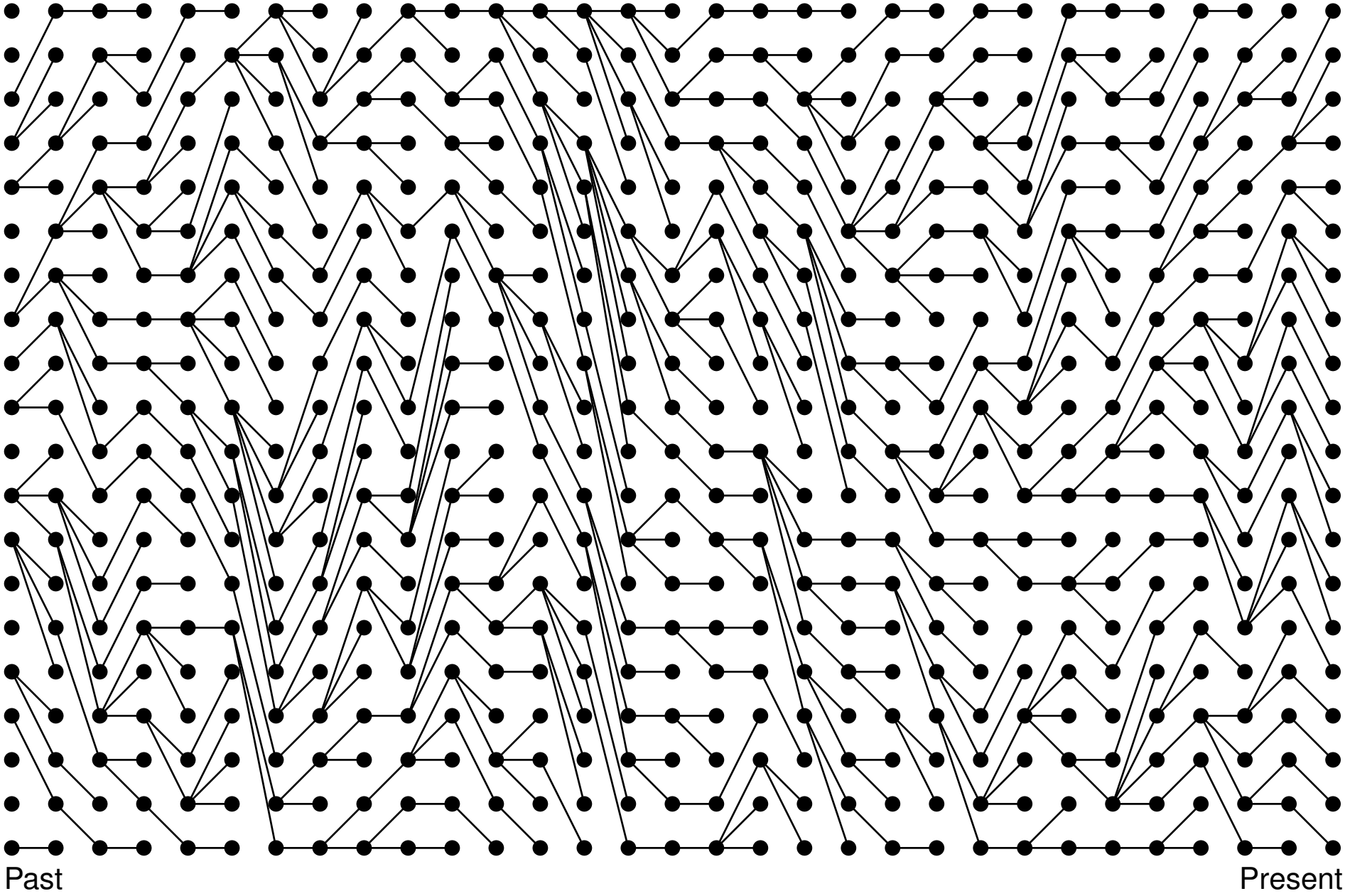
Population model



Past

Present

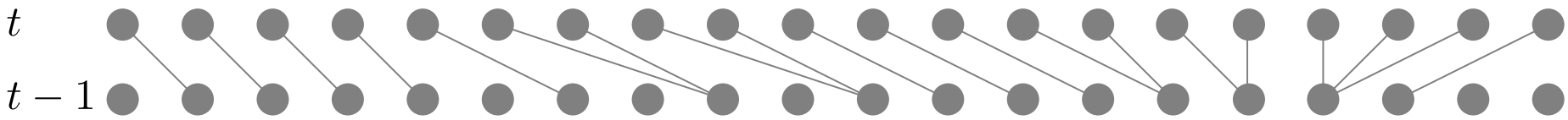
Population model



Population model

Wright

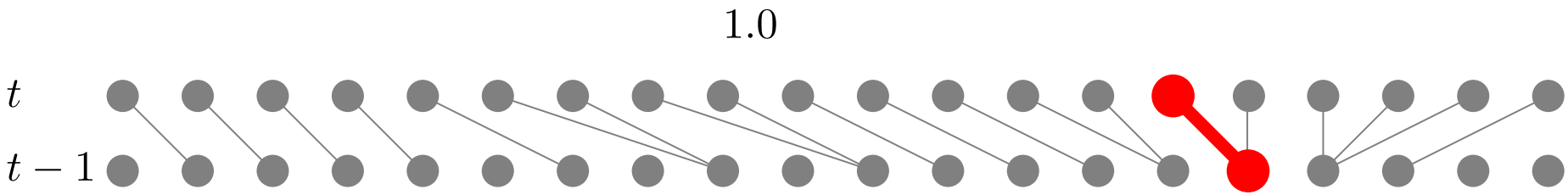
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is



Population model

Wright

Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is

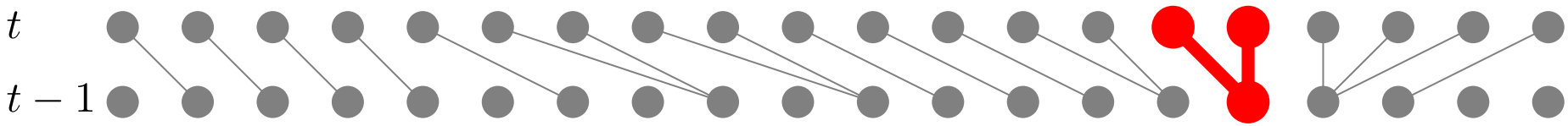


Population model

Wright

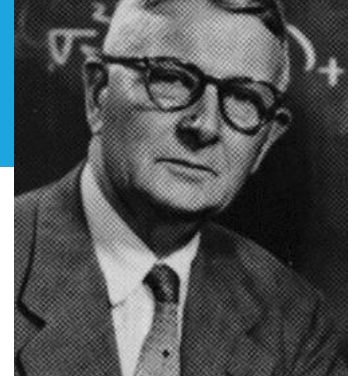
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is

$$1.0 \times \frac{1}{2N}$$



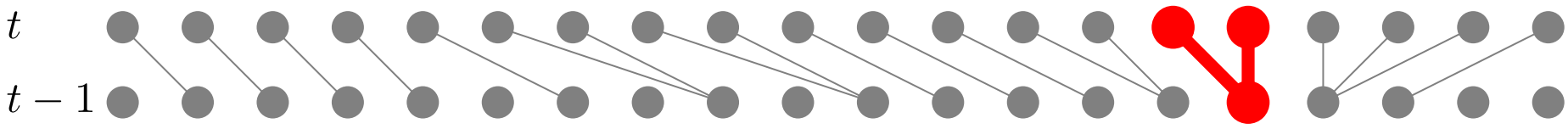
Population model

Wright



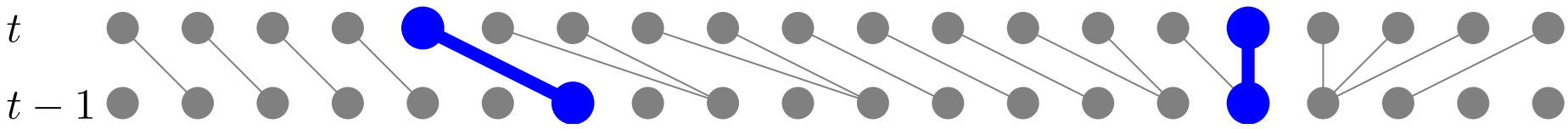
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in last generation is

$$\frac{1}{2N}$$



The probability that two randomly picked chromosome do not have a common ancestor is

$$1-\frac{1}{2N}$$



Population model

Wright

If we know the genealogy of the two individuals then we can calculate the probability as

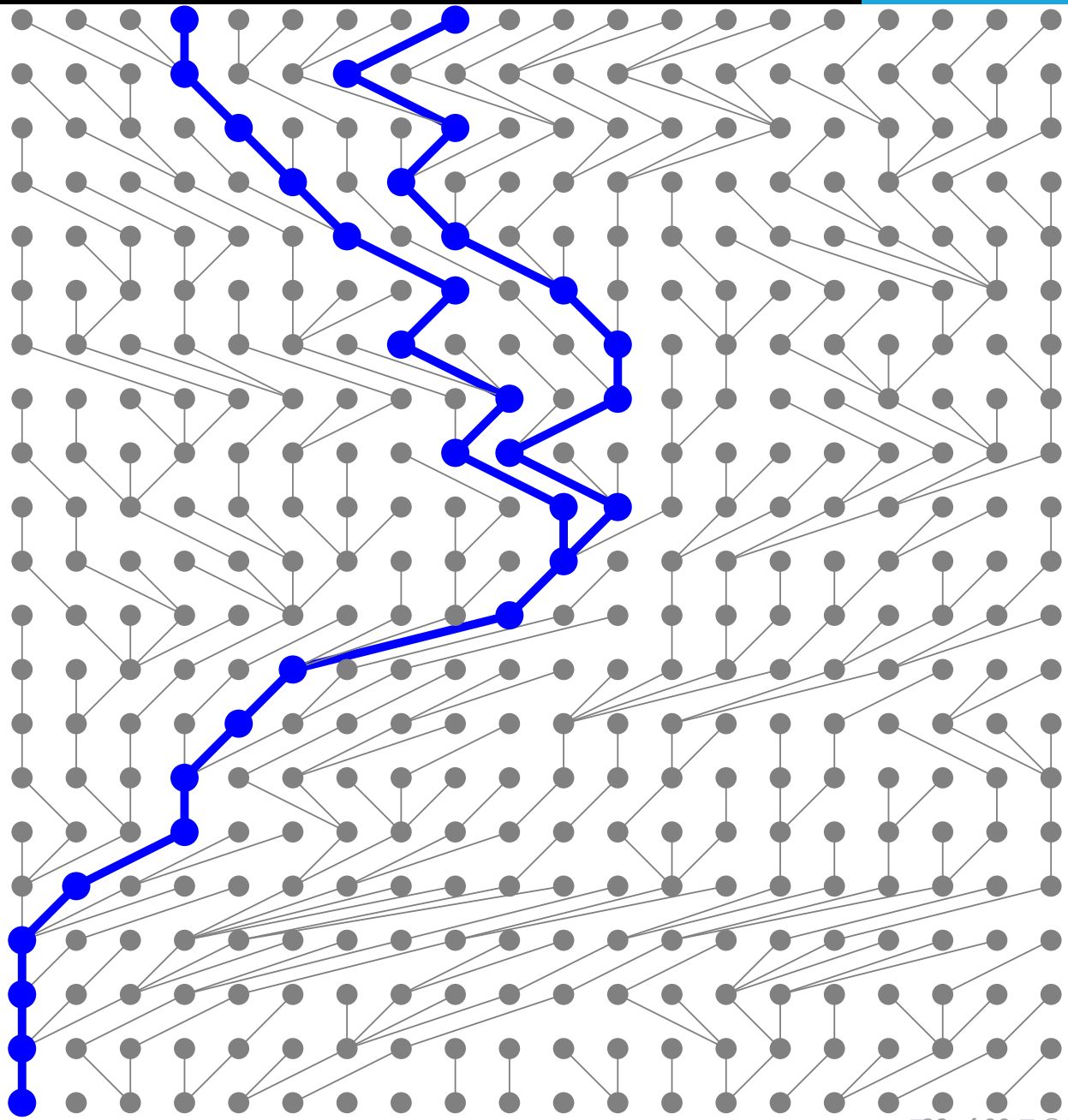
$$\mathrm{P}(\tau|N) = \left(1 - \frac{1}{2N}\right)^{\tau} \left(\frac{1}{2N}\right)$$

where τ is the number of generations with no coalescence. This formula is the Geometric Distribution and we can calculate the expectation of the waiting time until two random individuals coalesce:

$$\mathbb{E}(\tau) = 2N$$

Population model

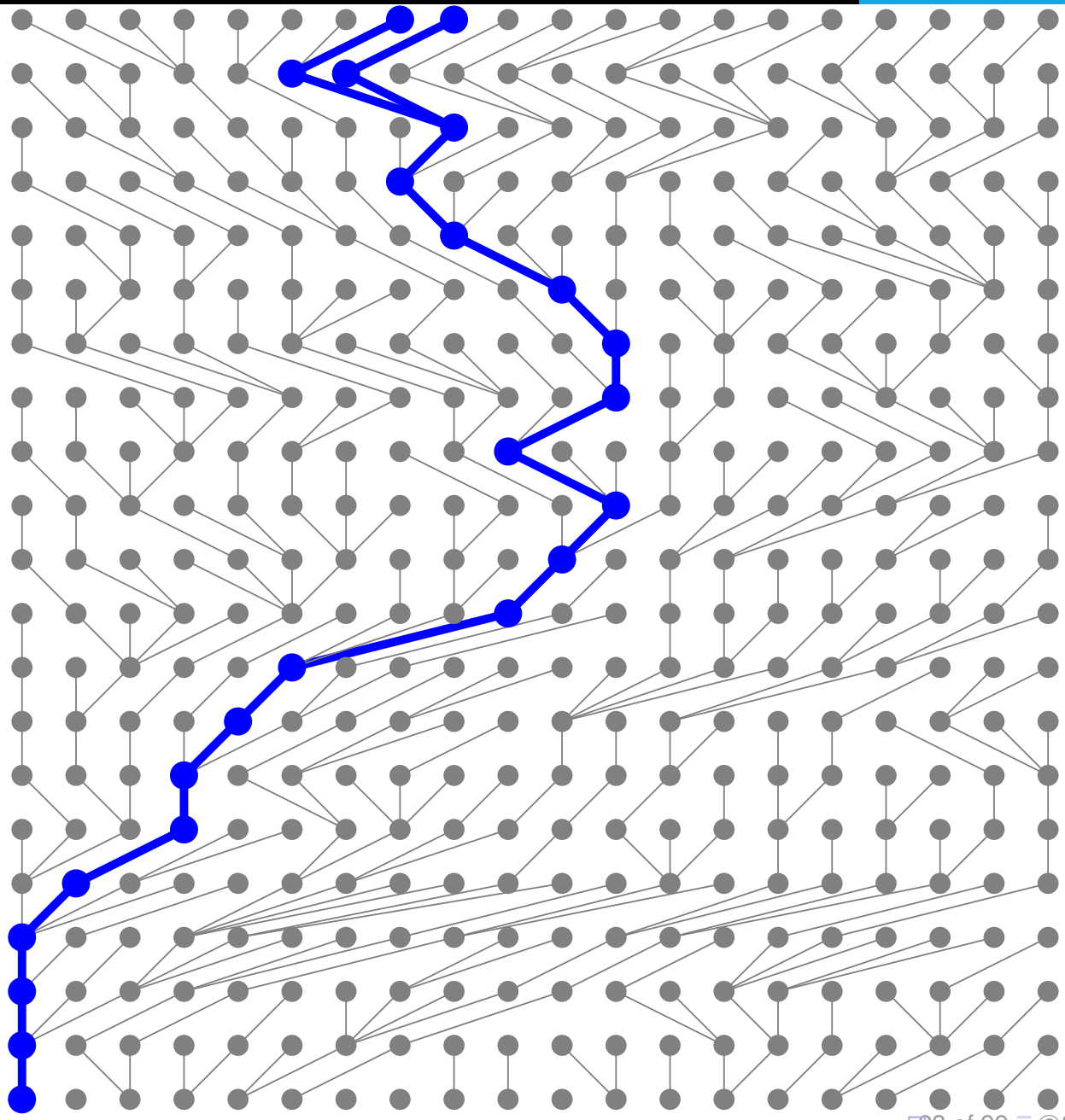
Present



Past

Population model

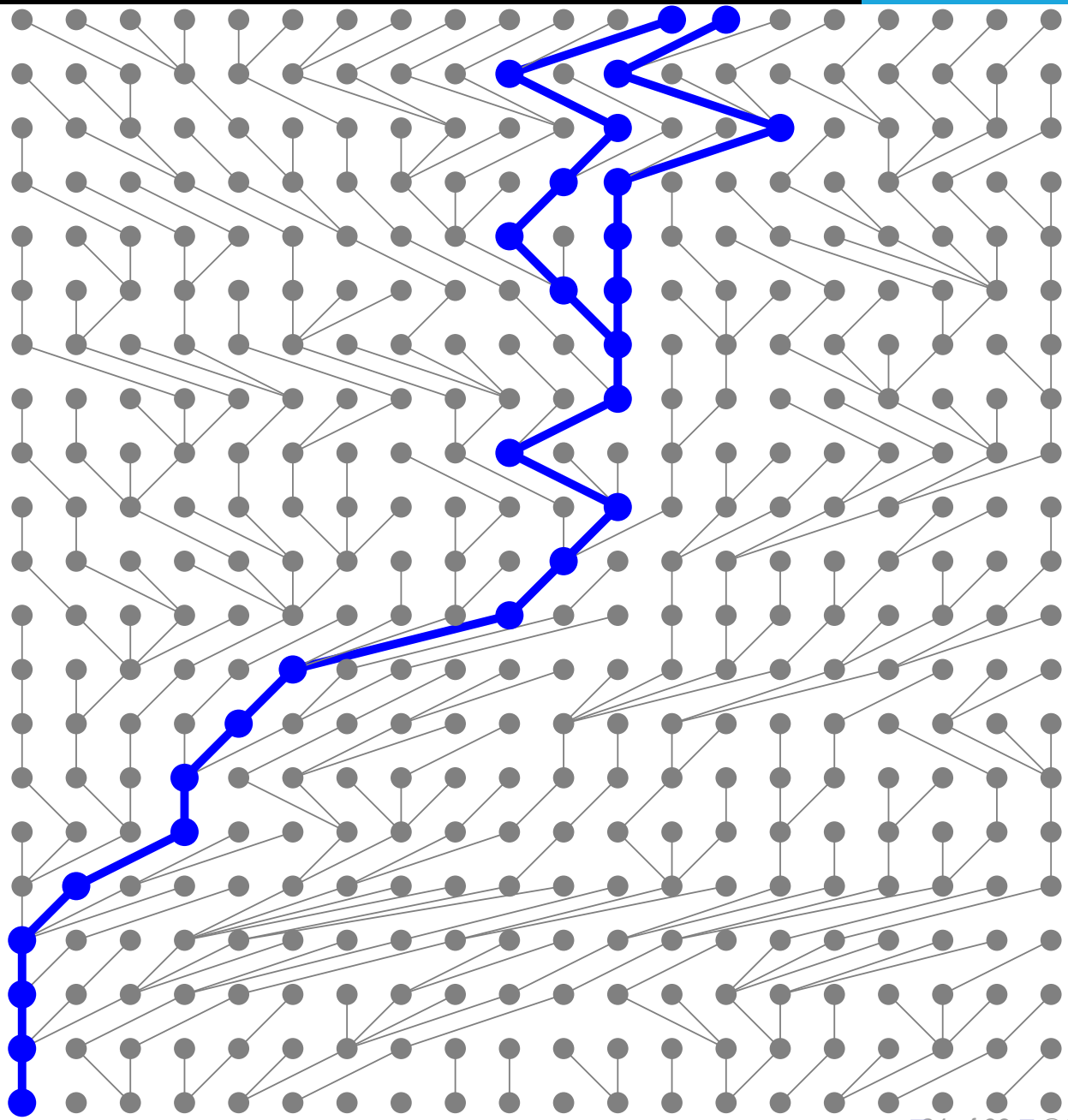
Present



Past

Population model

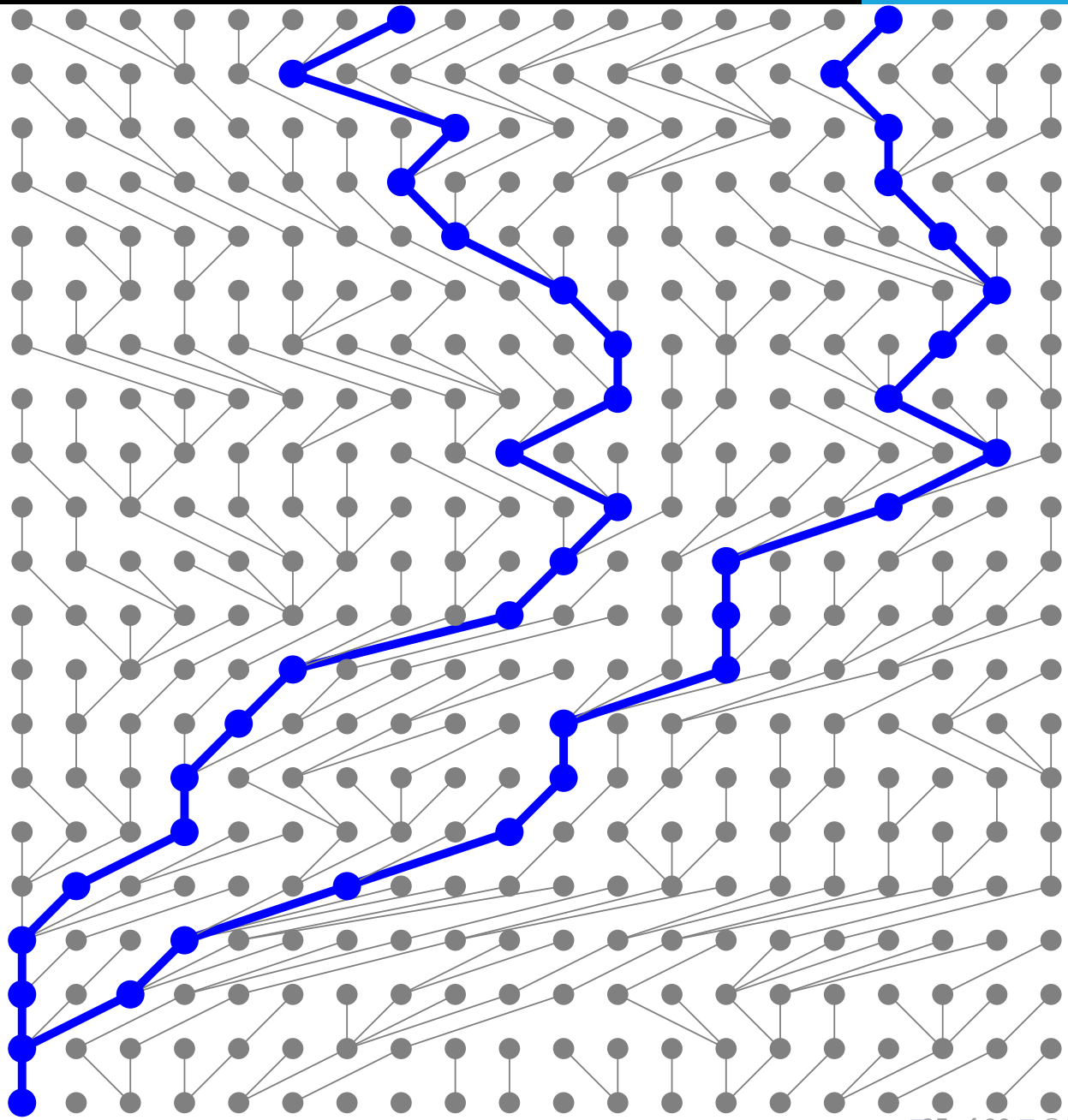
Present



Past

Population model

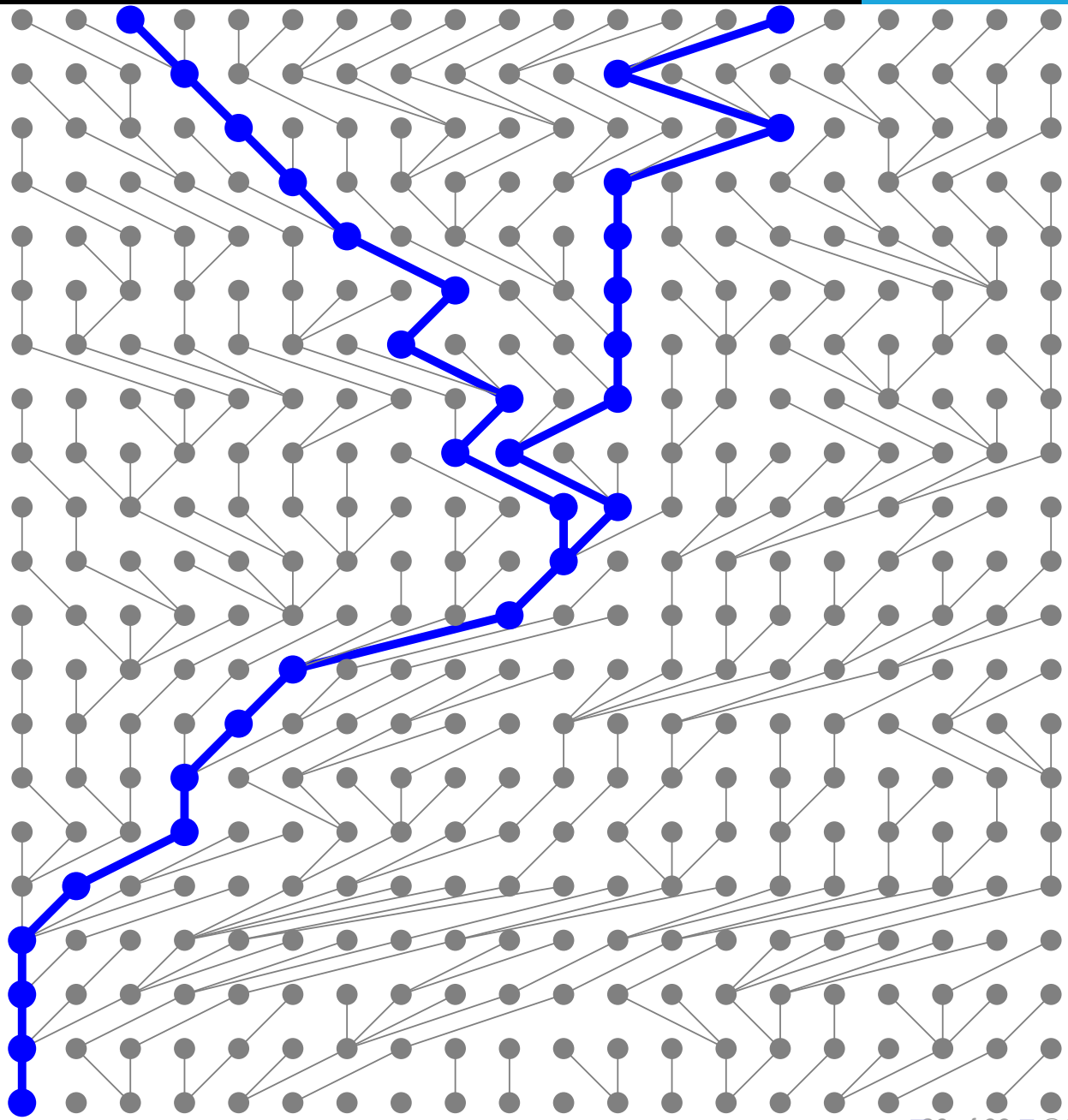
Present



Past

Population model

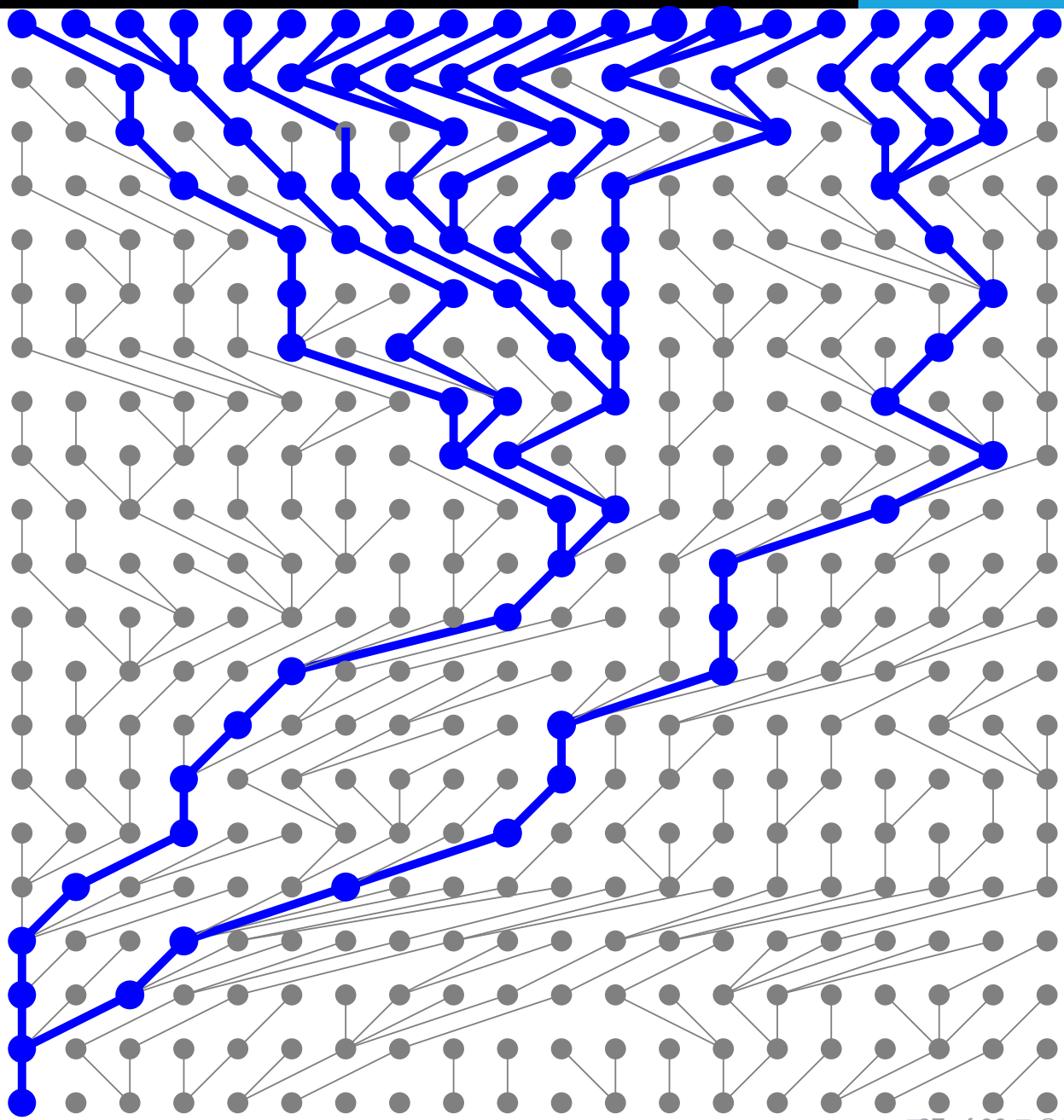
Present



Past

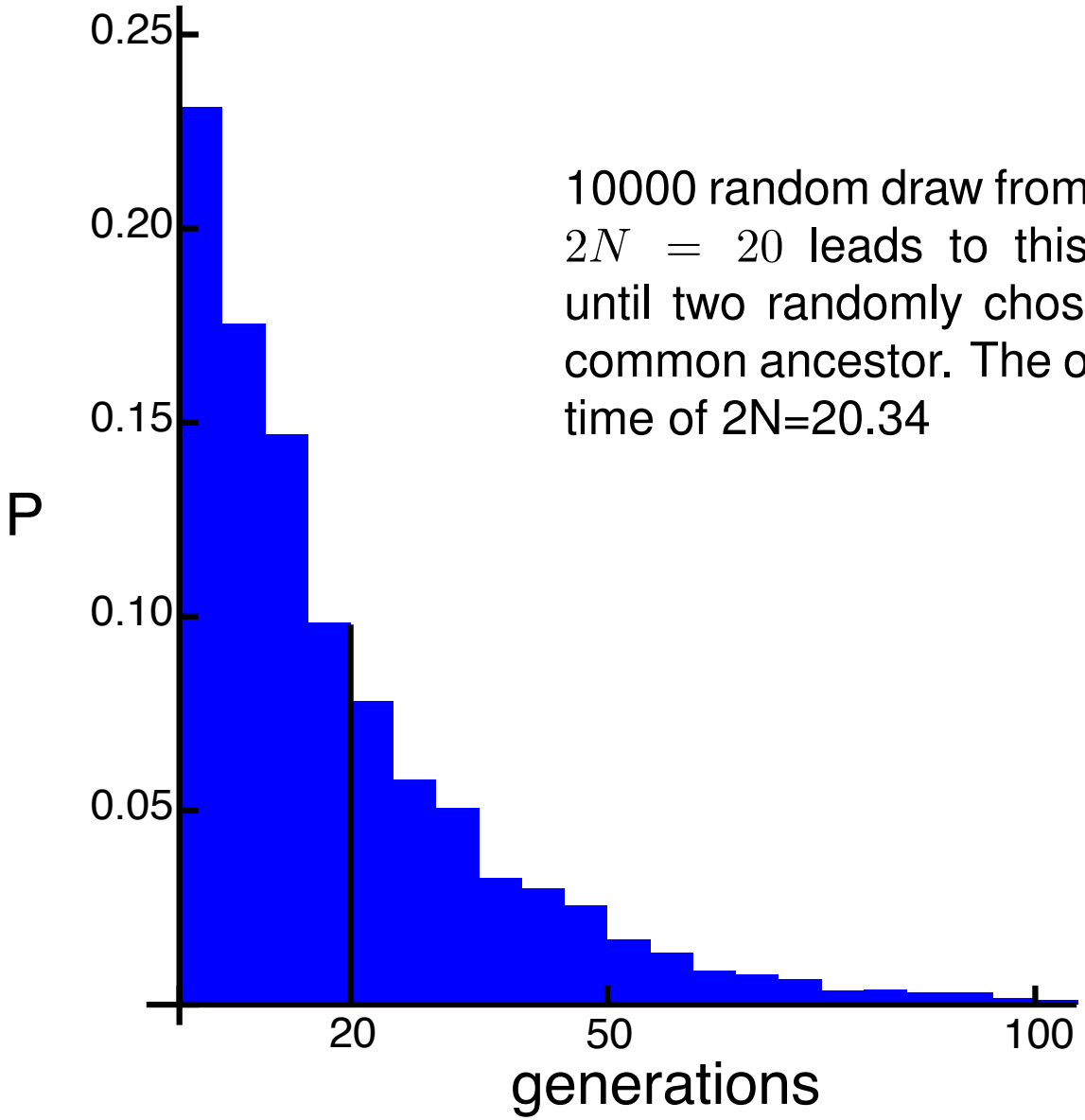
Population model

Present



Past

Probability Distribution



10000 random draw from a population with size $2N = 20$ leads to this distribution of times until two randomly chosen individuals have a common ancestor. The observed mean waiting time of 2N=20.34

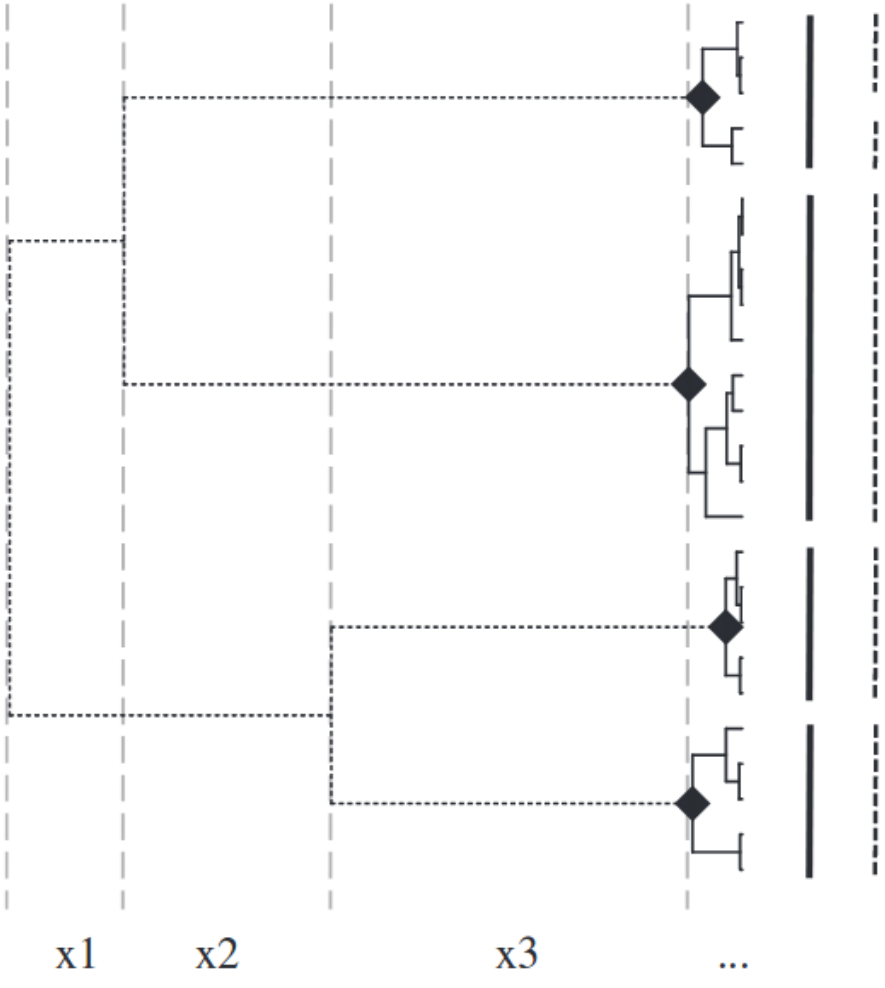
Observations

- For the time of coalescence in a sample of TWO , we will wait on average $2N$ generations assuming it is a Wright-Fisher population
- The model assumes that the generations are discrete and non-overlapping
- Real populations do not necessarily behave like a Wright-Fisher (the *'ideal' population)*
- *We assume that calculation using Wright-Fisher populations can be extrapolated to real populations.*

- What parameters drive expected timing of branching events of coalescent trees?

- What parameters drive expected timing of branching events of coalescent trees?
- What parameters drive expected timing of branching events of species trees?

Generalized Mixed Yule Coalescent (GMYC) method is a likelihood method for delimiting species by fitting within- and between-species branching models to reconstructed gene tree



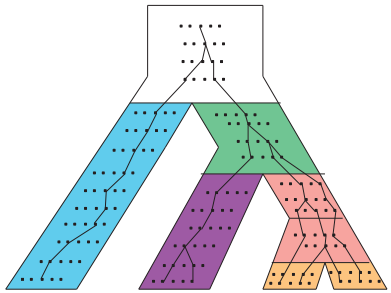
Fujisawa and Barraclough, 2013, Delimiting species using GMYC)

Do gene trees follow the species tree?

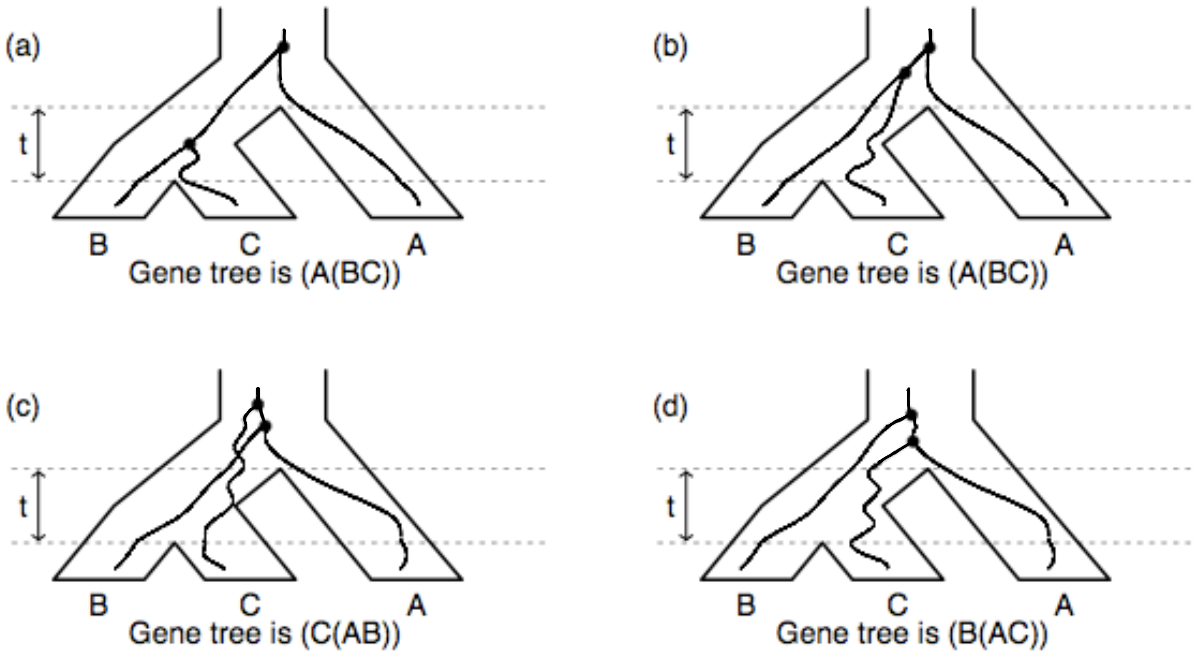
“ Phylogenomic analysis of 1,070orthologues from 23yeast genomes identified 1,070distinct gene trees, which were all incongruent with the phylogeny inferred from concatenation.” //
Salichos, L., Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497, 327–331 (2013). https://doi.org/10.1038/nature12130

Relationship between population genetics and phylogenetics

- Given current technology, we can do much more:
  - Sample many individuals within each taxon (species, population, etc.)
  - Sequence many genes for all individuals
- Need models at two levels:
  - Model what happens within each population
    [population genetics – coalescent model]
  - Link each within-population model on a phylogeny
    [phylogenetics]

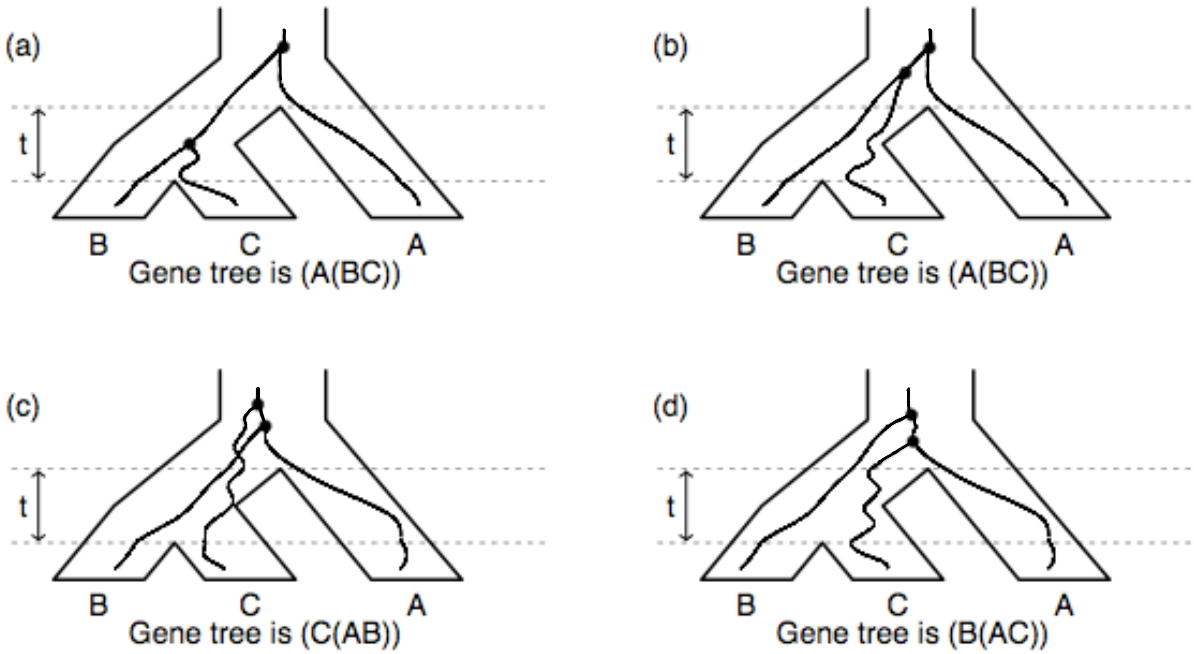


Phylogenetic coalescent model

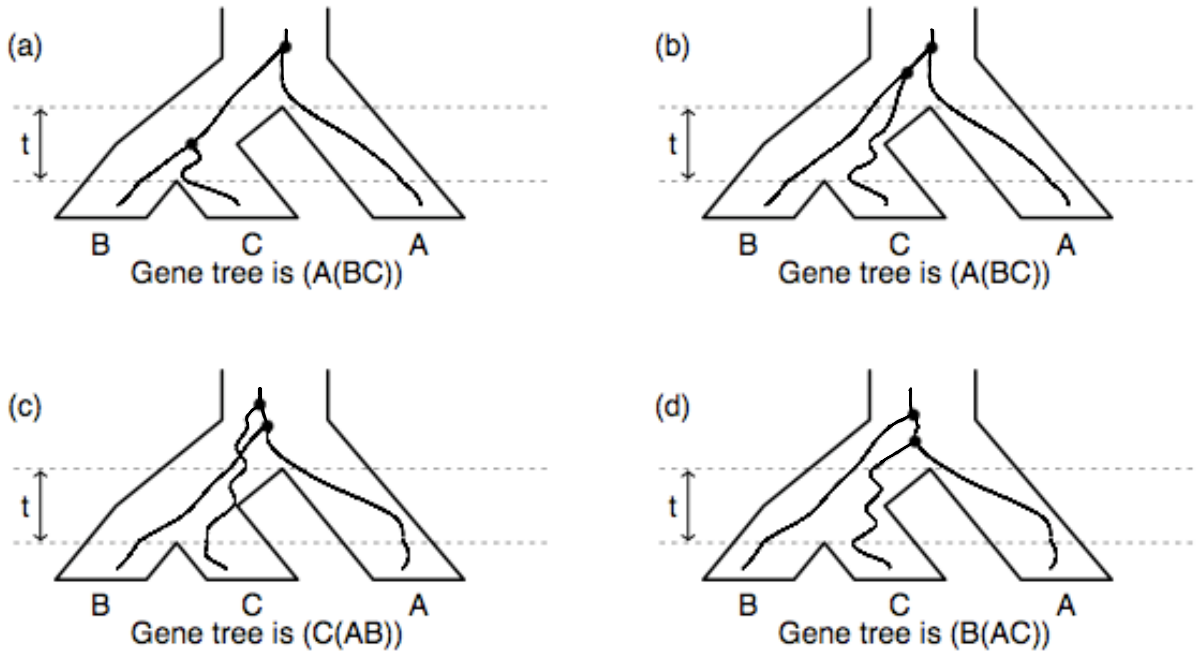


Which of these species trees is more likely to result in gene trees that match the species tree?

Phylogenetic coalescent model

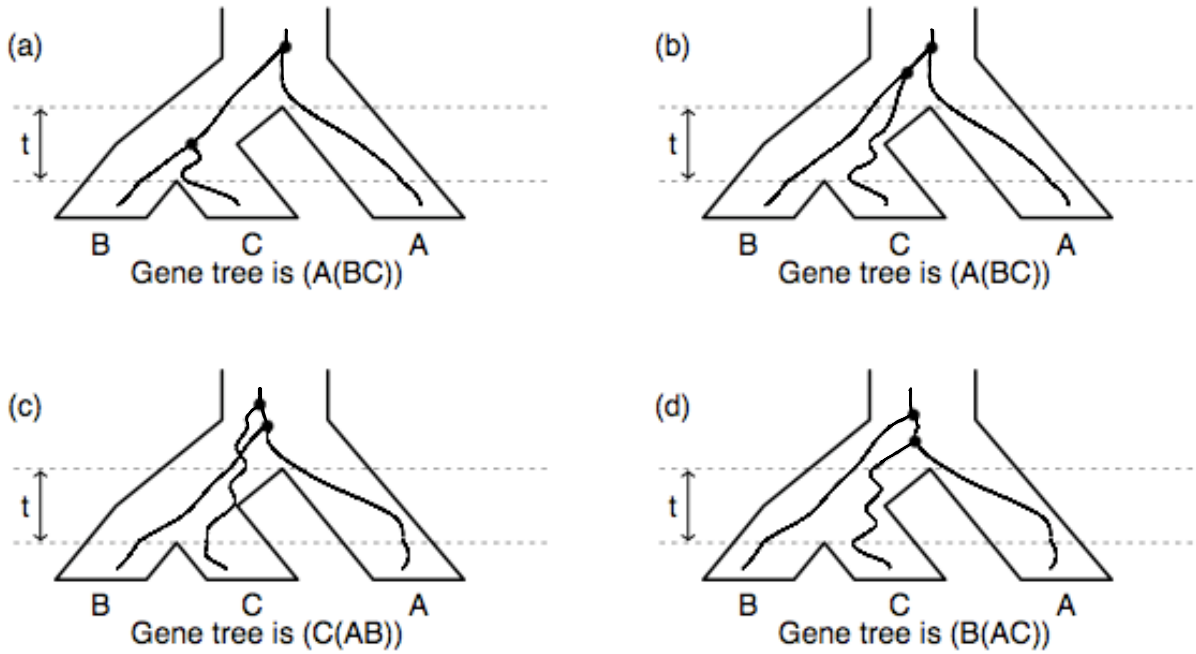


Phylogenetic coalescent model



t = length of interval between speciation events in **coalescent units**
= number of $2N$ generations

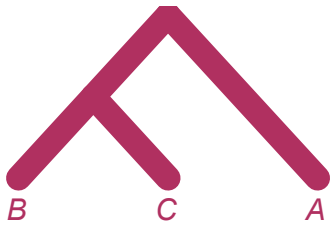
Phylogenetic coalescent model



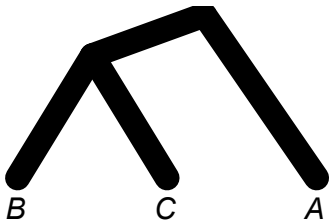
t = length of interval between speciation events in **coalescent units**
= number of $2N$ generations

Example: 1.2 coalescent units for an organisms with population size $N = 10,000$ and a generation time of 3 years $= 1.2 \times 20,000 \times 3 = 72,000$ years

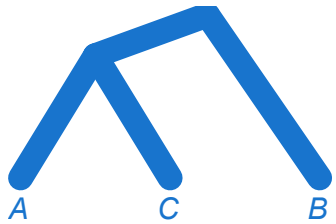
Probabilities of each gene tree history are shown below them
t = length of interval between speciation events



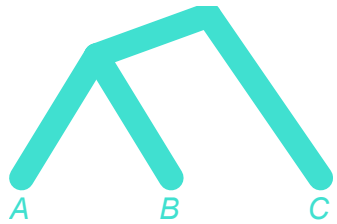
$1 - e^{-t}$



$\frac{1}{3}e^{-t}$



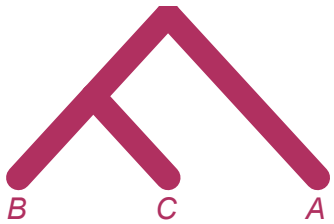
$\frac{1}{3}e^{-t}$



$\frac{1}{3}e^{-t}$

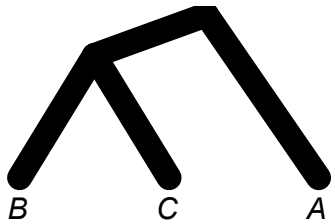
Phylogenetic coalescent model

t = length of interval between coalescent events $= 1.0$



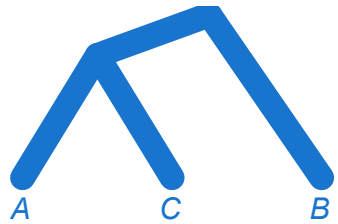
$1 - e^{-t}$

0.63



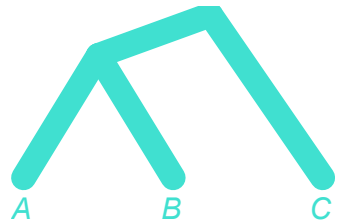
$\frac{1}{3}e^{-t}$

0.12



$\frac{1}{3}e^{-t}$

0.12

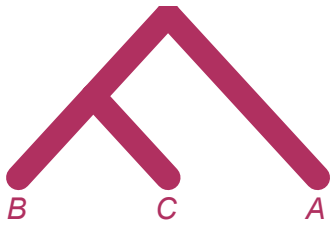


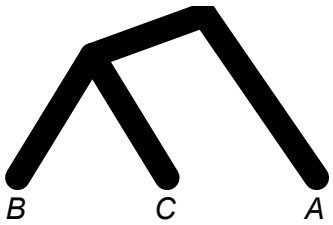
$\frac{1}{3}e^{-t}$

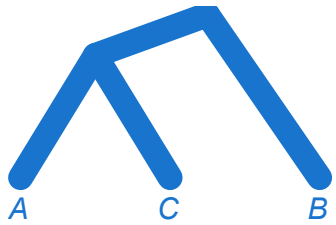
0.12

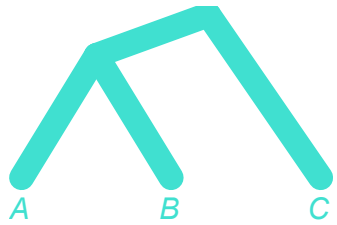
Phylogenetic coalescent model

t = length of interval between coalescent events $= 1.0 = 0.5$





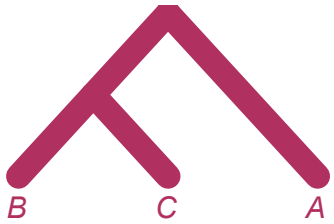


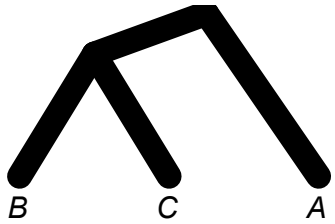


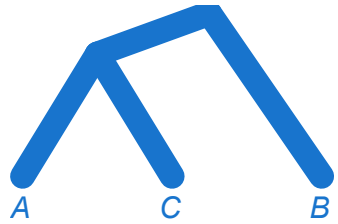
$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
0.40	0.20	0.20	0.20

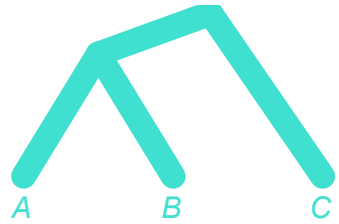
Phylogenetic coalescent model

t = length of interval between coalescent events $= 1.0 = 0.5 = 2.0$





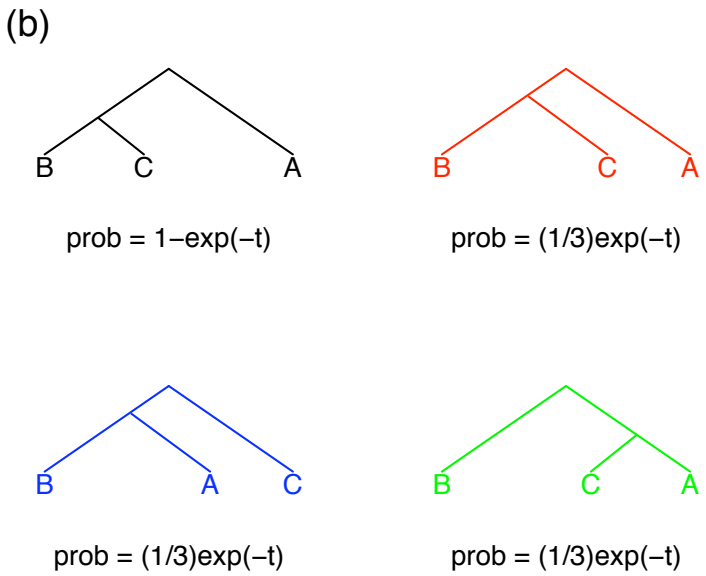


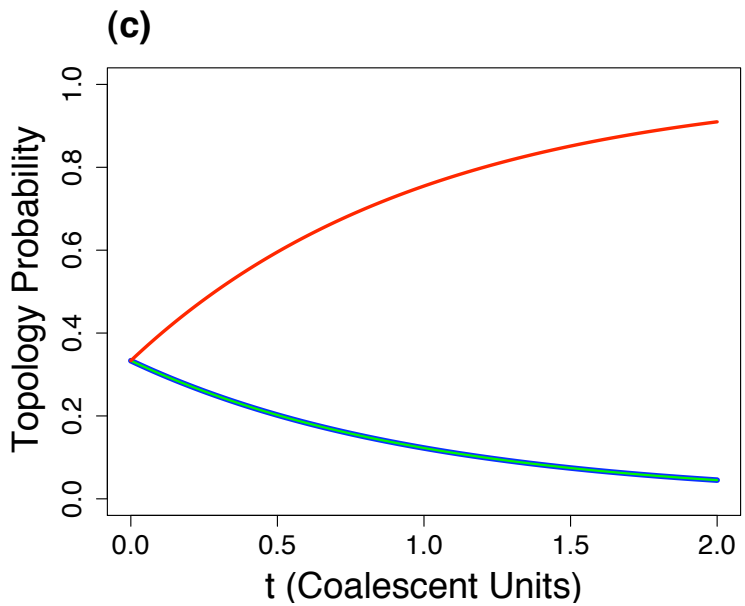


$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
0.40	0.20	0.20	0.20
0.85	0.05	0.05	0.05

Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

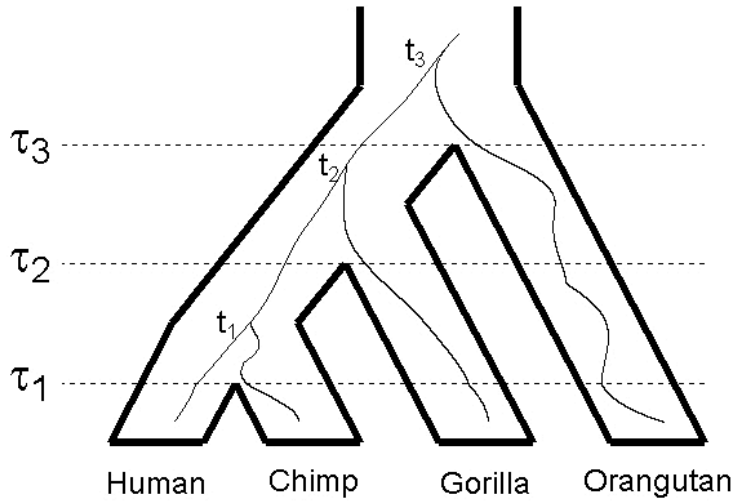
- What are these probabilities like as a function of t, the length of time between speciation events?





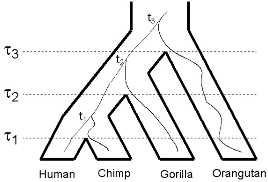
Example: a slightly larger case

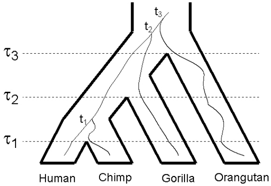
- Consider 4 taxa – the human-chimp-gorilla problem

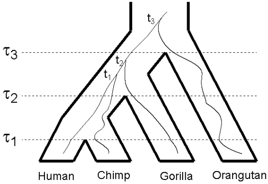


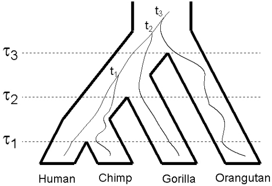
Coalescent histories for the 4-taxon example

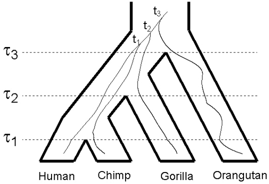
- There are 5 possible histories for this example:



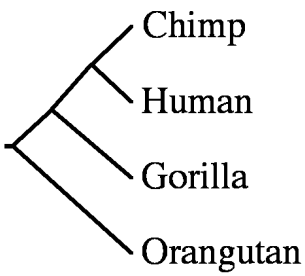


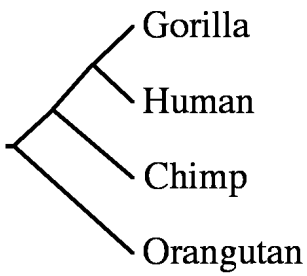


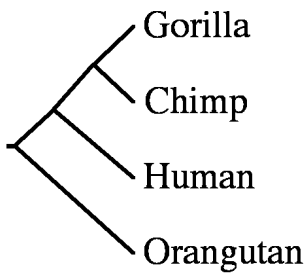




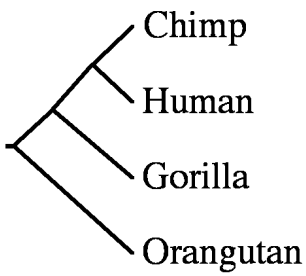
Applications of the topology distribution - example 1



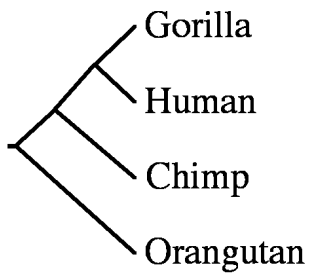




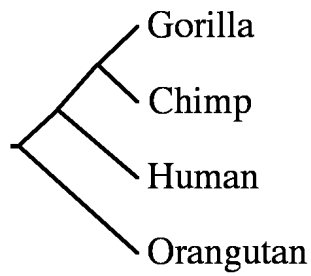
Applications of the topology distribution - example 1



76.6%



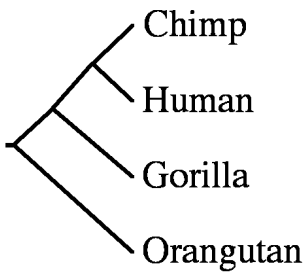
11.4%

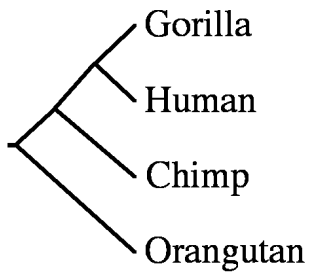


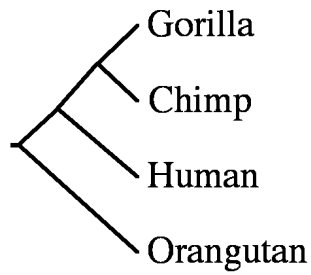
11.5%

Observed proportions of each gene tree among ML phylogenies

Applications of the topology distribution - example 1







76.6% 11.4% 11.5%

79.1% 9.9% 9.9%

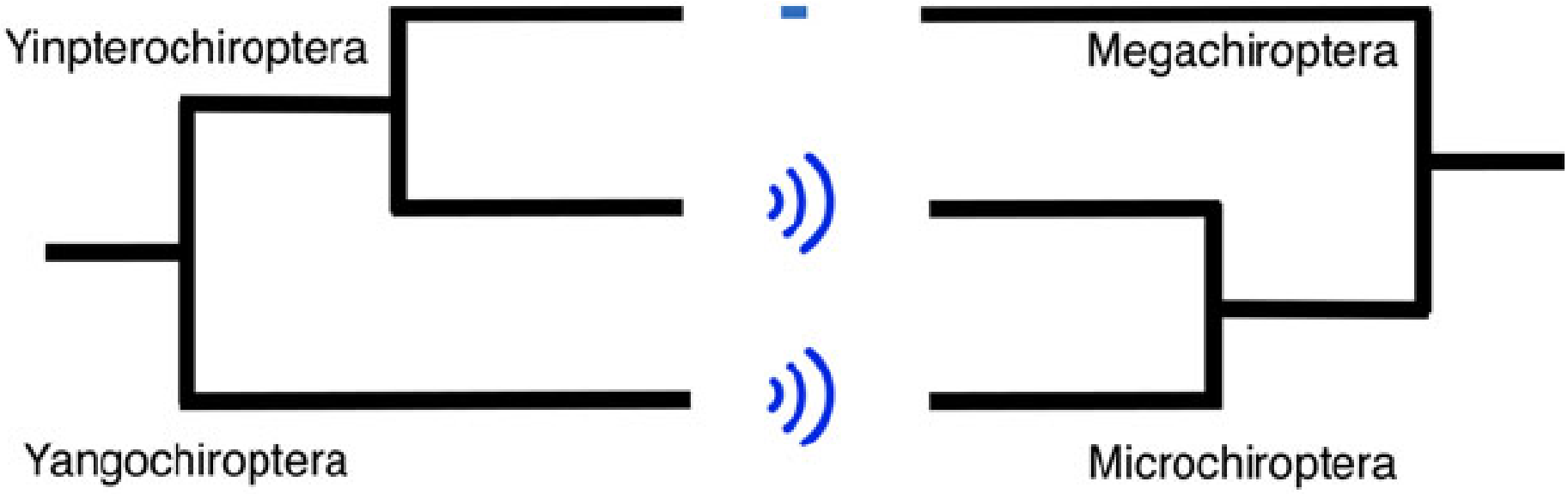
Observed proportions of each gene tree among ML phylogenies

Predicted proportions using parameters from Rannala & Yang, 2003.

Applications of the topology distribution - example 2

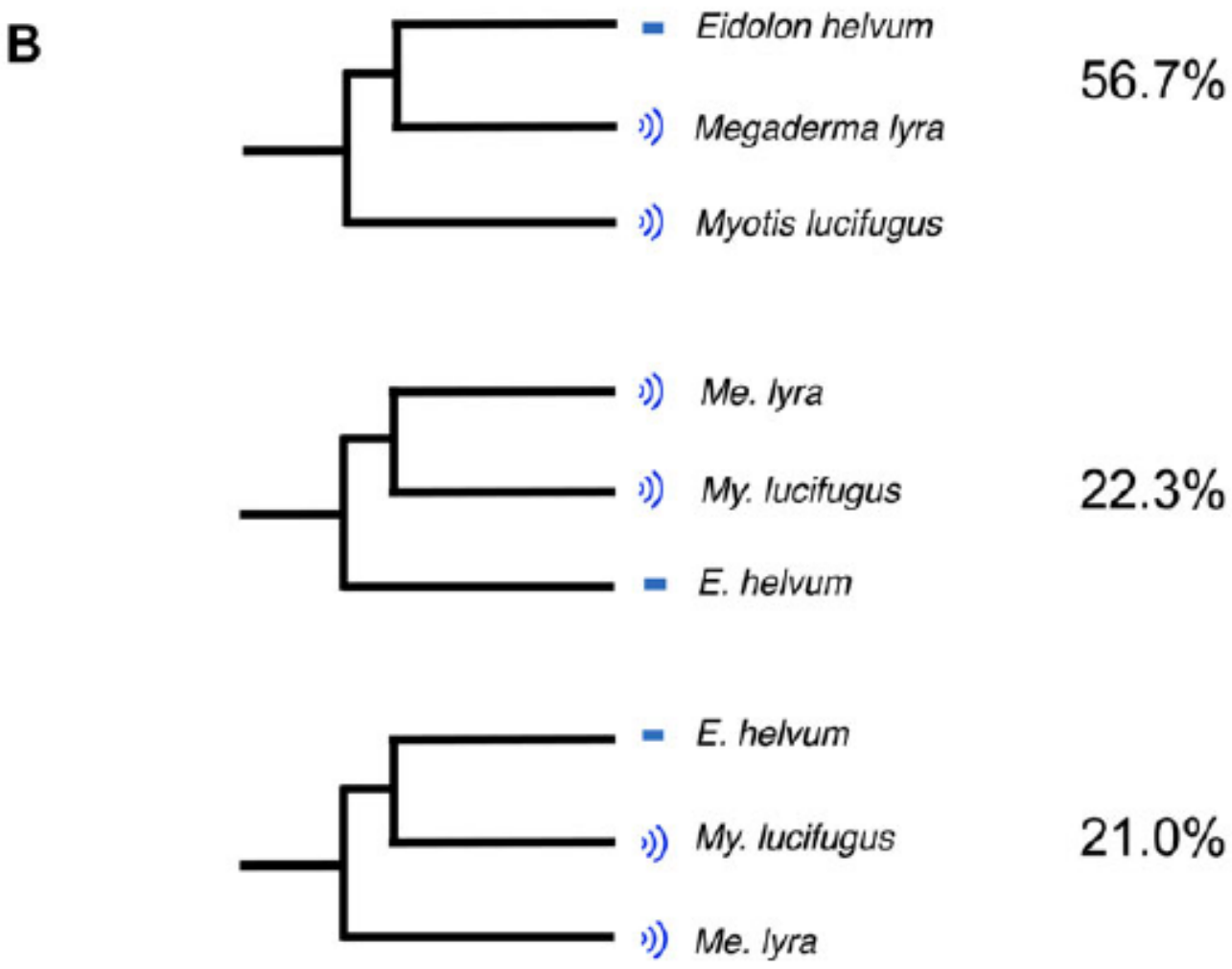
- In the previous example, one topology is clear preferred
- Must the distribution always look this way?
- Examine the entire distribution when the number of taxa is small

Different gene trees can drive different conclusions



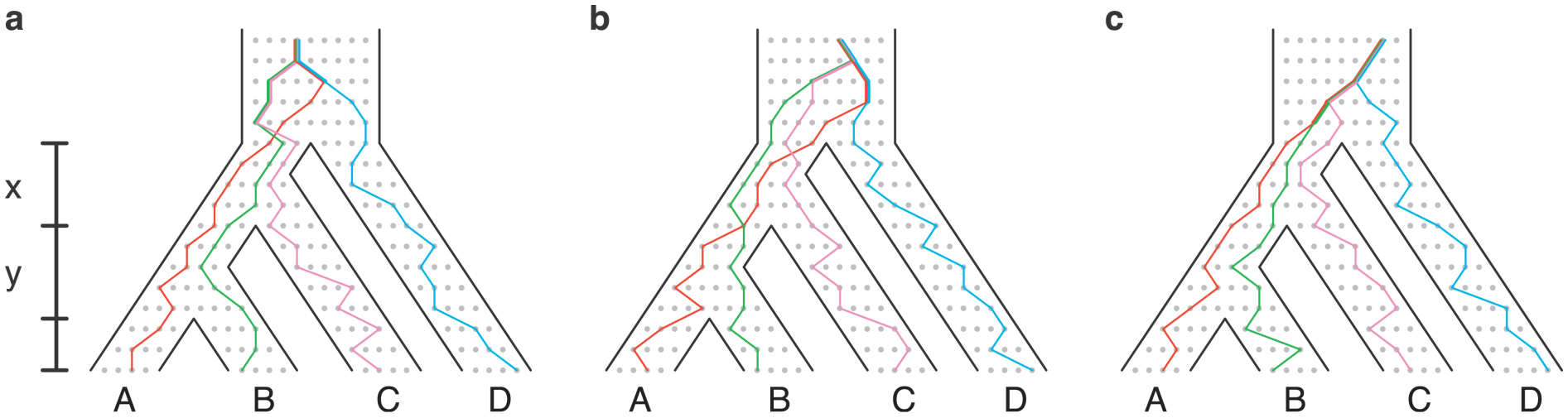
Species relationships between echolocating and nonecholocating bats (after Teeling 2009). Left: inferences from DNA sequence data.
Right: traditional species relationships inferred from morphological characters (and limited sequence data). (Hahn and Nakhleh, 2016)

Do you even need the species tree?



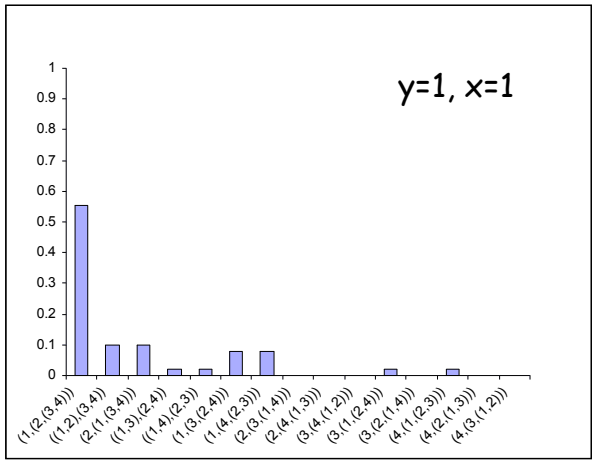
(Hahn and Nakhleh, 2016)

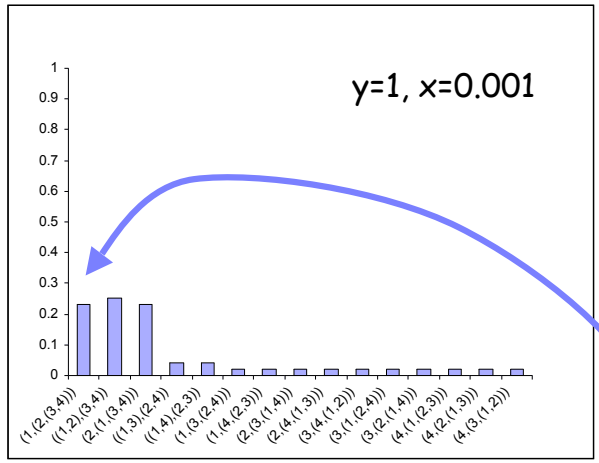
Is the gene tree that matches the species tree always the most common gene tree?

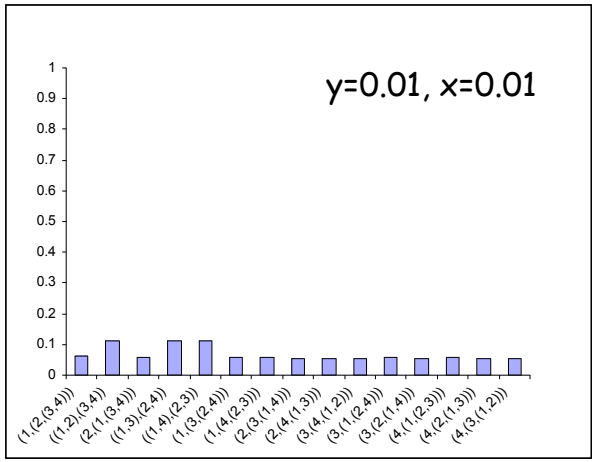


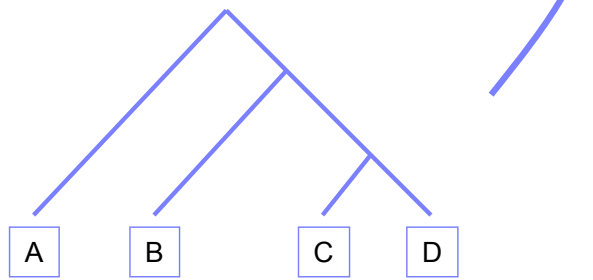
If the internal branches of the species tree—x and y—are short so that coalescences occur deep in the tree, the two sequences of coalescences that produce a given symmetric gene tree topology together have higher probability than the single sequence that produces the topology that matches the species tree. (a) and (b) Two coalescence sequences leading to gene tree topology ((AD)(BC)). In (a), the lineages from B and C coalesce more recently than those from A and D, and in (b), the reverse is true. (c) The single sequence of coalescences leading to gene tree topology (((AB)C)D). (Degnan and Rosenberg, 2006)

Applications of the topology distribution - example 2

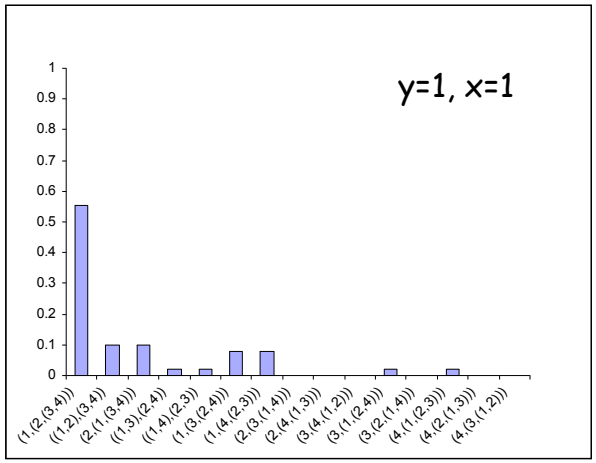


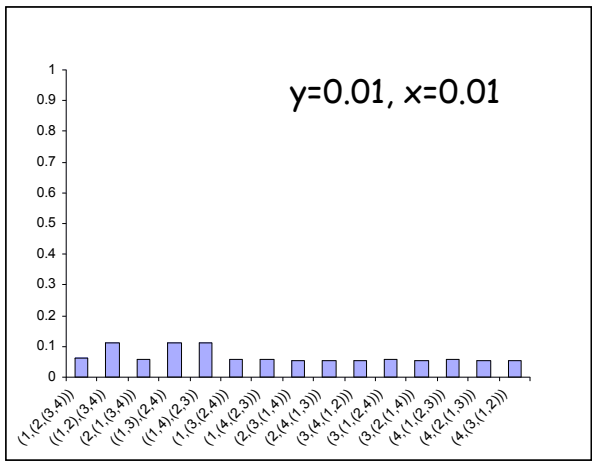


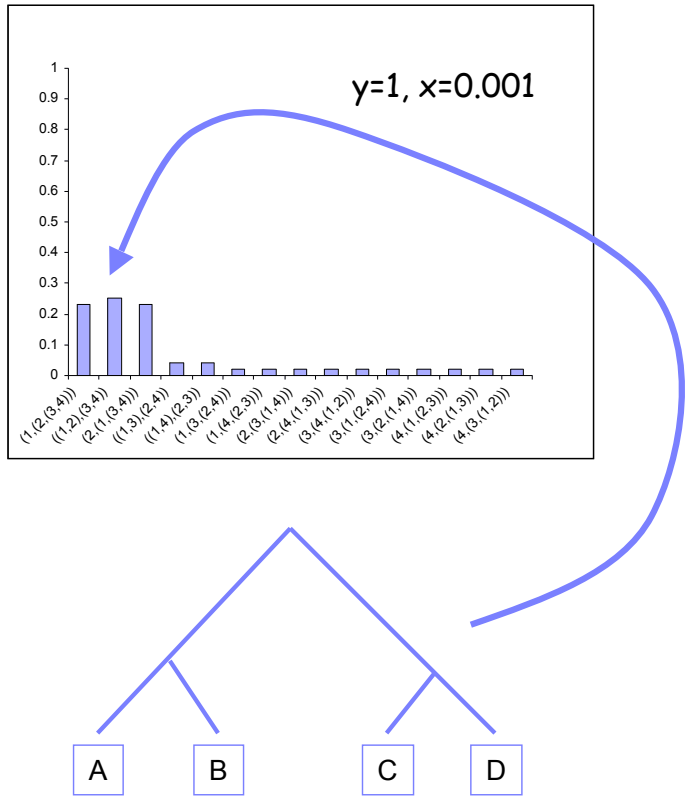




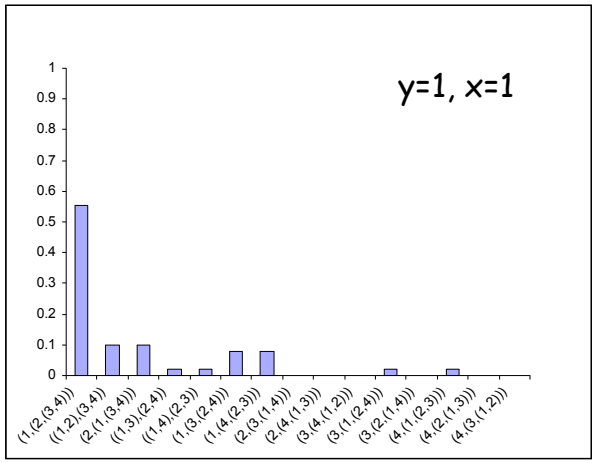
Applications of the topology distribution - example 2

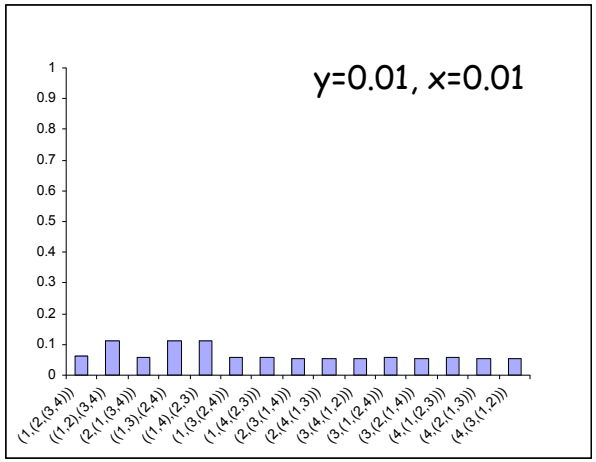


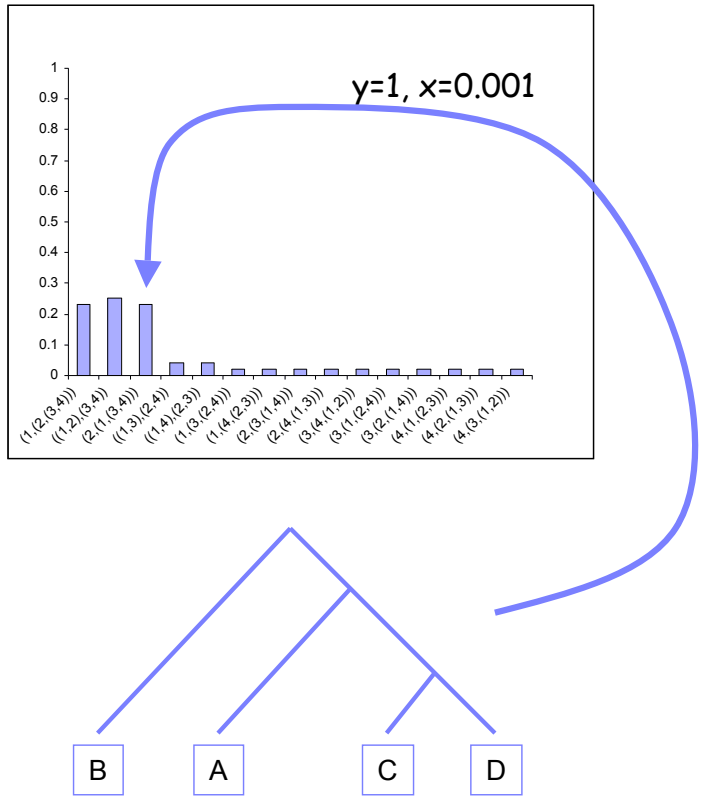




Applications of the topology distribution - example 2







Summary:

- Genome scale data gives us a lot of information about relationships, but have contrasting signal from different loci
- Failure to incorporate coalescent expectations into models can result in incorrect inferences
- Lots of ongoing development in new approaches to analyzing multi locus data.

Next week:

- Walk through some different methods for estimating species trees from multi locus data
- Clarification of any confusing concepts
- Paper discussion!

Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70(1):7–17.