

# Lecture 14

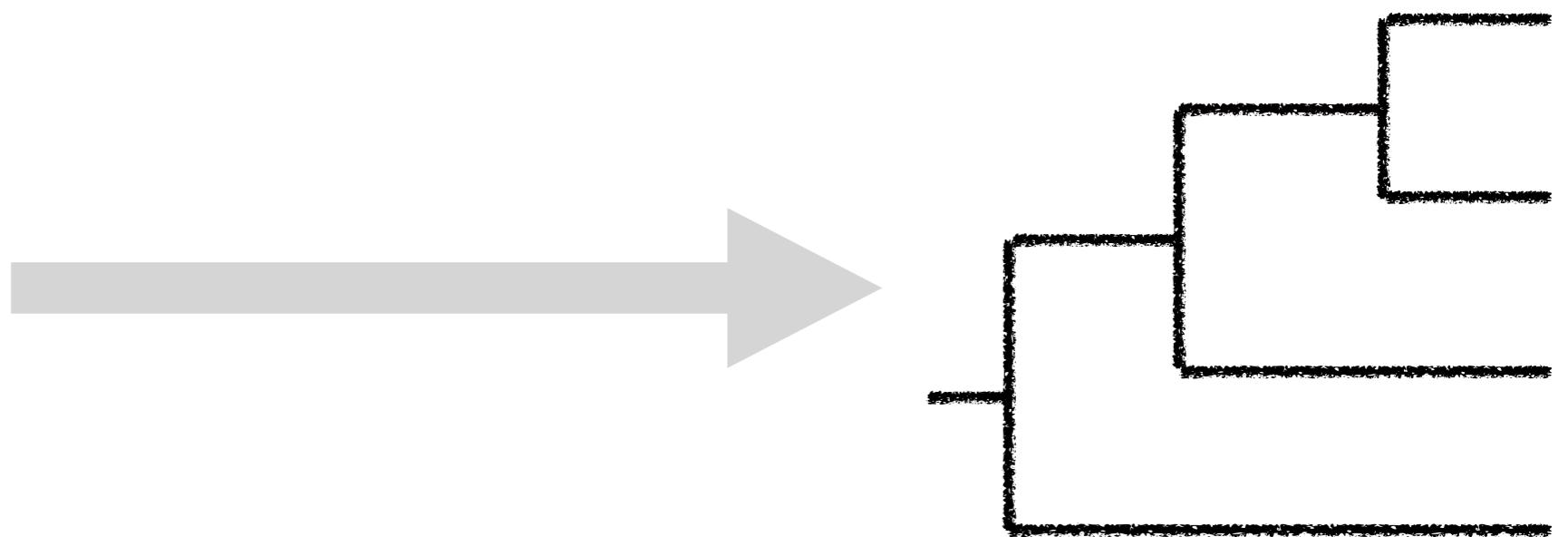
Coalescent-based methods  
Botany/Plant Path 563

- **Previous class check-up:**
  - We studied Bayesian phylogenetic inference
  - We practiced on MrBayes
- **Learning Objectives:** At the end of today's session, you will be able to
  - Explain the coalescent model on a species tree
  - Explain the steps in coalescent-based methods and the comparison with concatenation approaches
- **Pre-class work**
  - Read HAL 3.1 and 3.3
- **Important: start thinking about final project!**

As mentioned in slack, I added some comments to your GitHub repos.

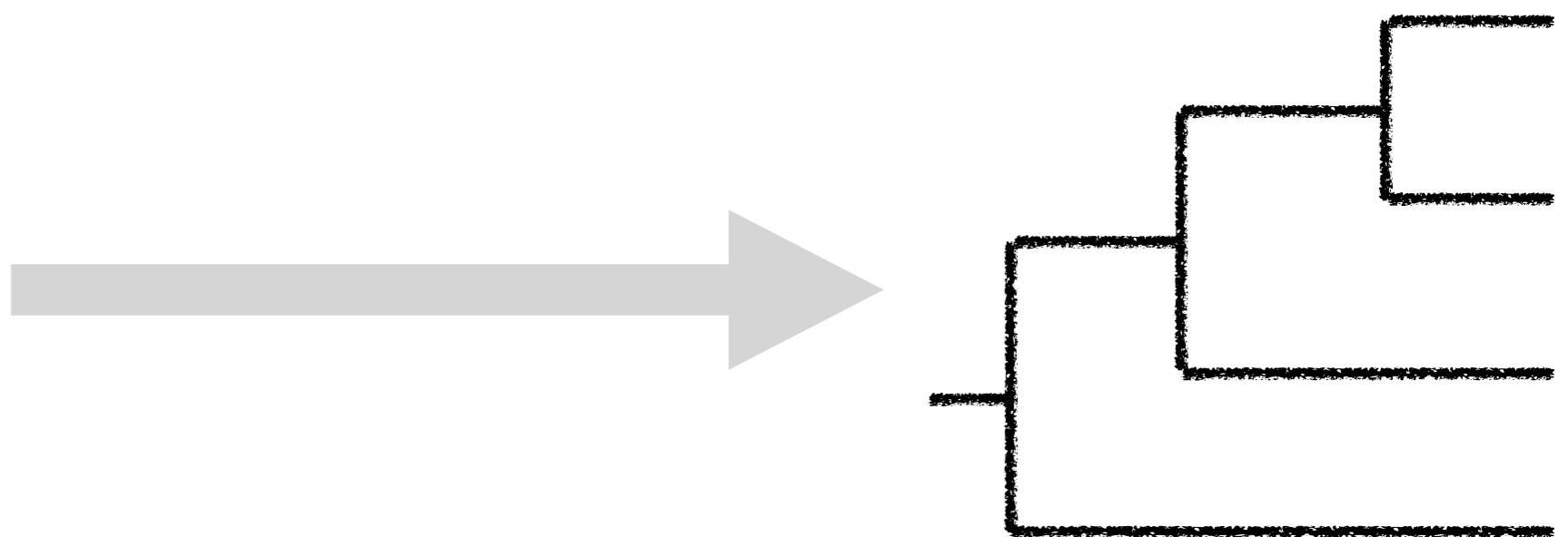
# Phylogenetic inference

AAGTCTAG  
AAGTCTAG  
AACTCTAG  
AATTCTAG



# Phylogenetic inference

AAGTCTAG  
AAGTCTAG  
AACTCTAG  
AATTCTAG



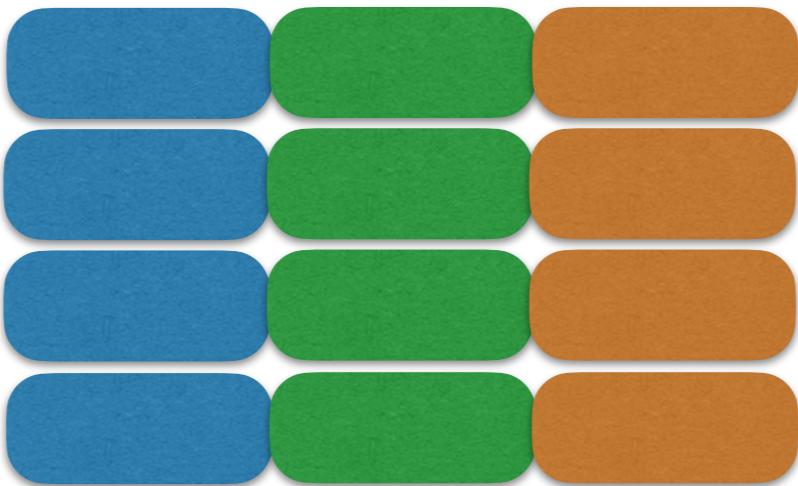
Gene  
Locus  
Region  
Whole genome





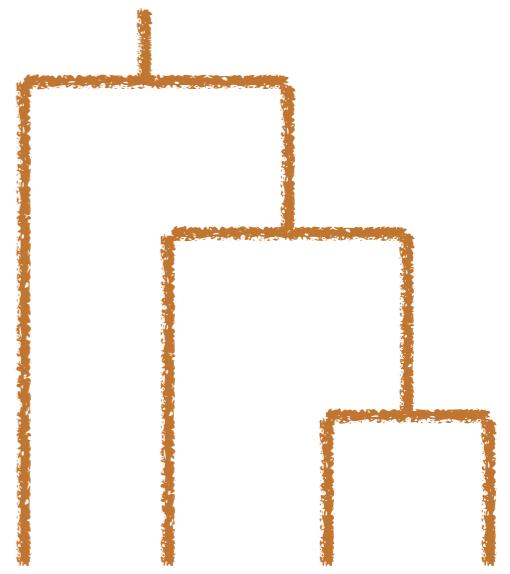
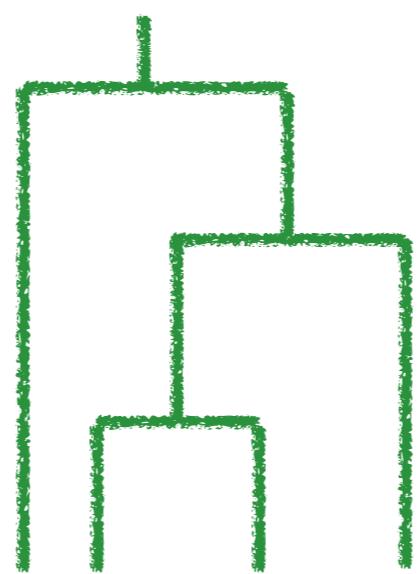
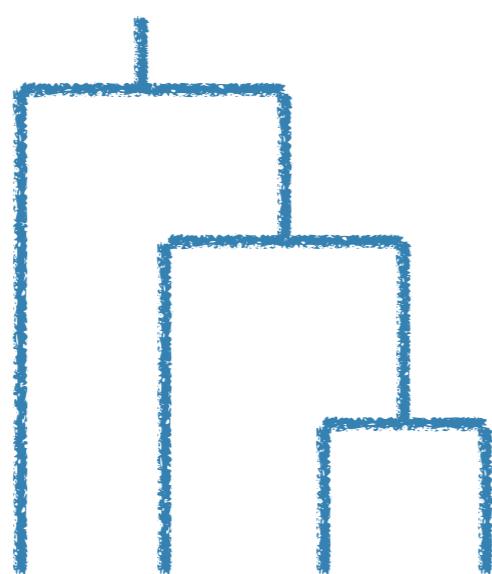
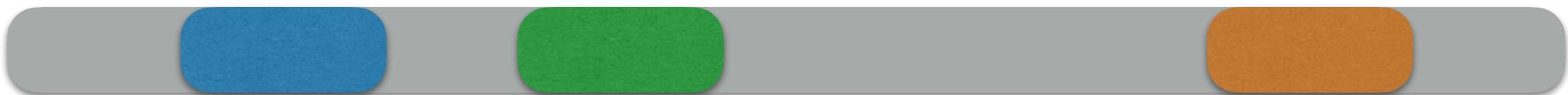
Statistically  
inconsistent

(Kubatko, Degnan, 2007)  
(Roch, Steel, 2015)

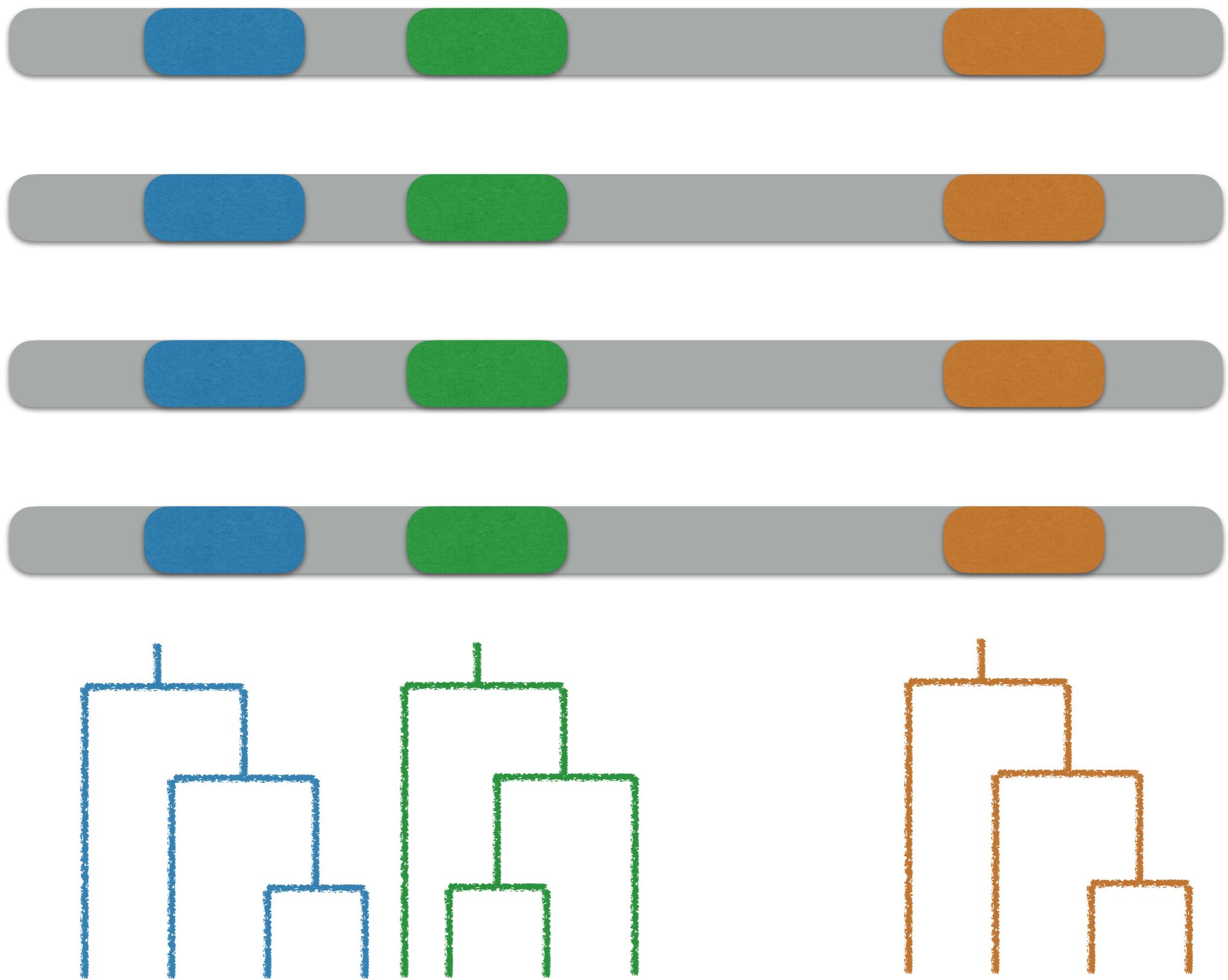


Concatenation or supermatrix





Ortholog  
Recombination-free  
MDL (Ané, 2011, GBE)



Ortholog  
Recombination-free  
MDL (Ané, 2011, GBE)

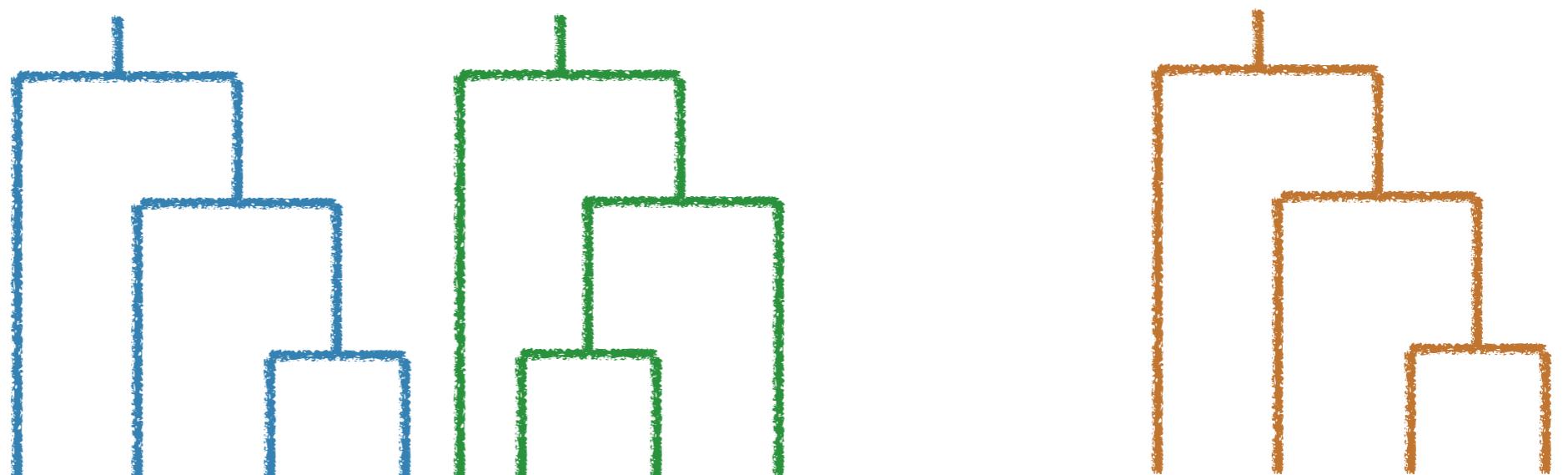


Estimate  
gene trees

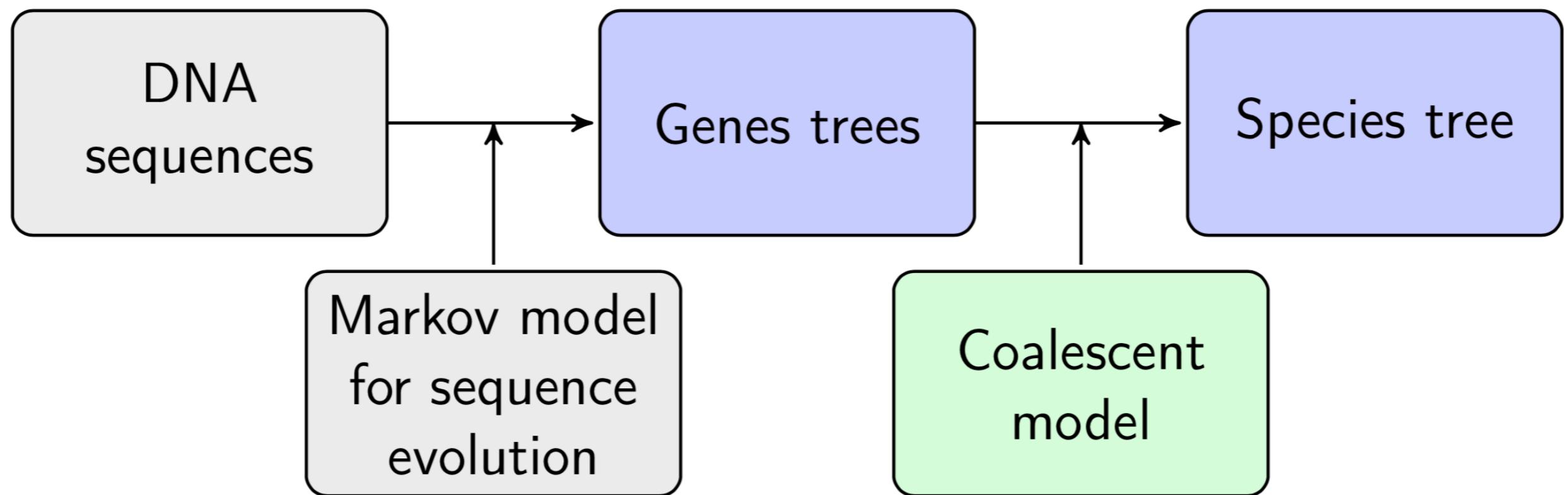
MrBayes  
(Huelsenbeck, Ronquist, 2001)

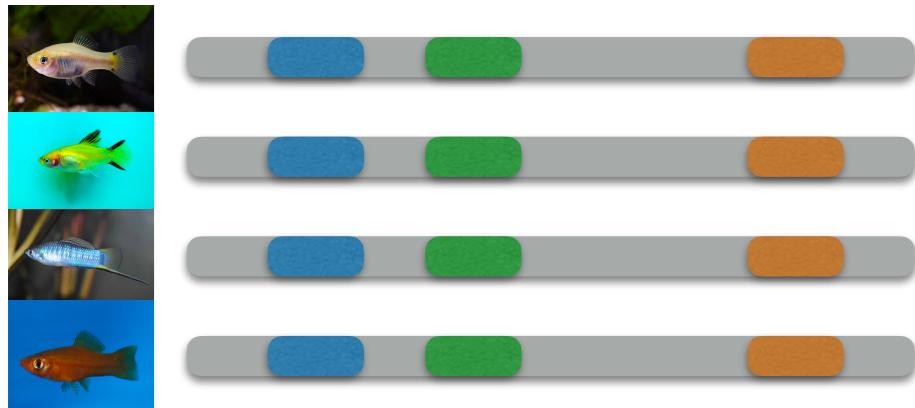
RAxML  
(Stamatakis, 2014)

IQ-tree 2  
(Minh et al, 2020)

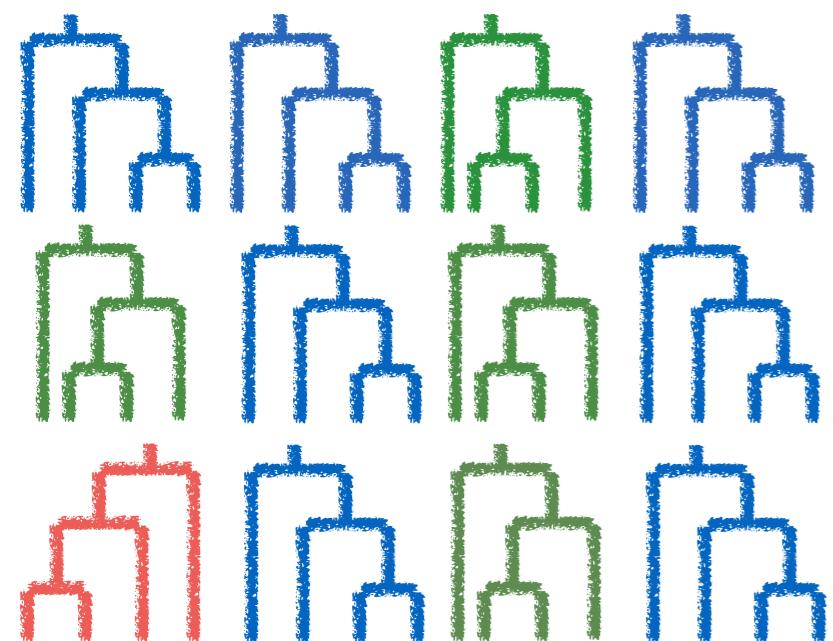


# Phylogenetic inference



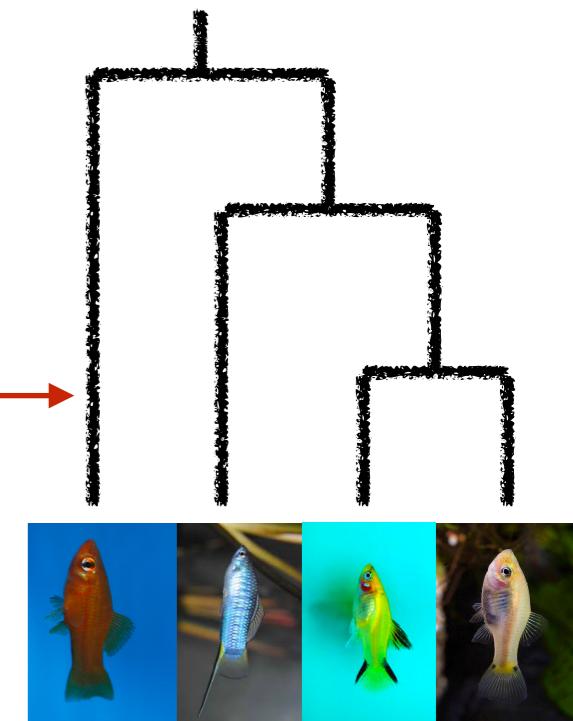


Distances  
Parsimony  
Likelihood  
(Bayesian)

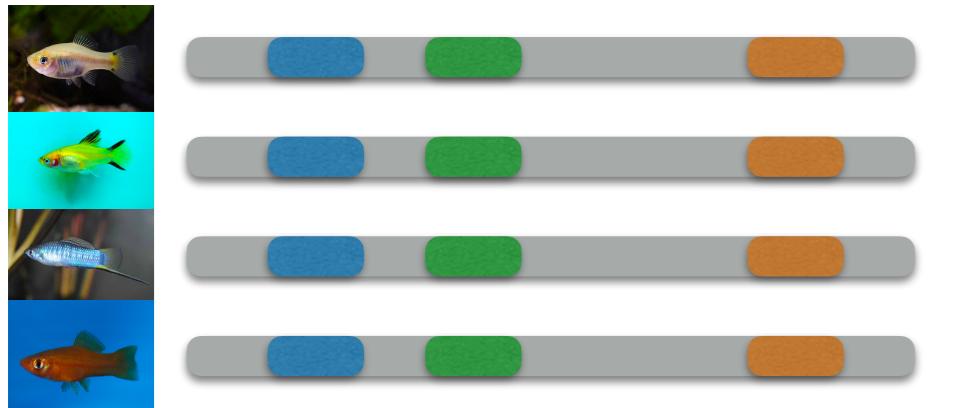


$$L(T, \theta) = \prod_{i=1}^L P(G_i | T, \theta)$$

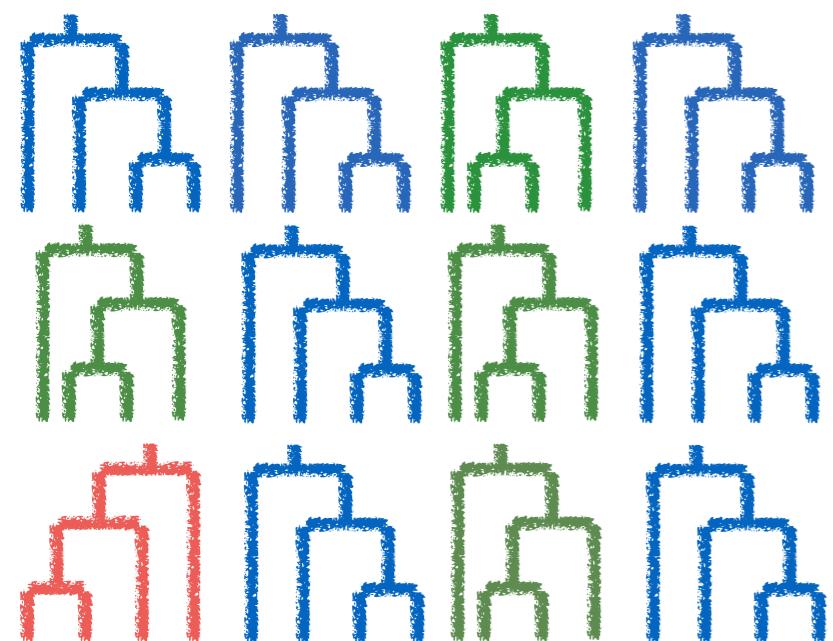
Max. Lik.



**Data**



Distances  
Parsimony  
Likelihood  
(Bayesian)



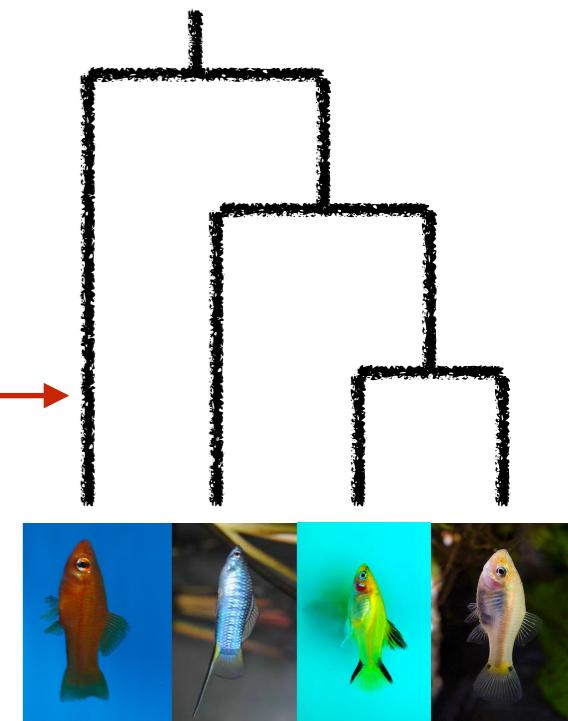
**Data**



$$L(T, \theta) = \prod_{i=1}^L P(G_i | T, \theta)$$

**Multispecies  
Coalescent  
Model**

Max. Lik.



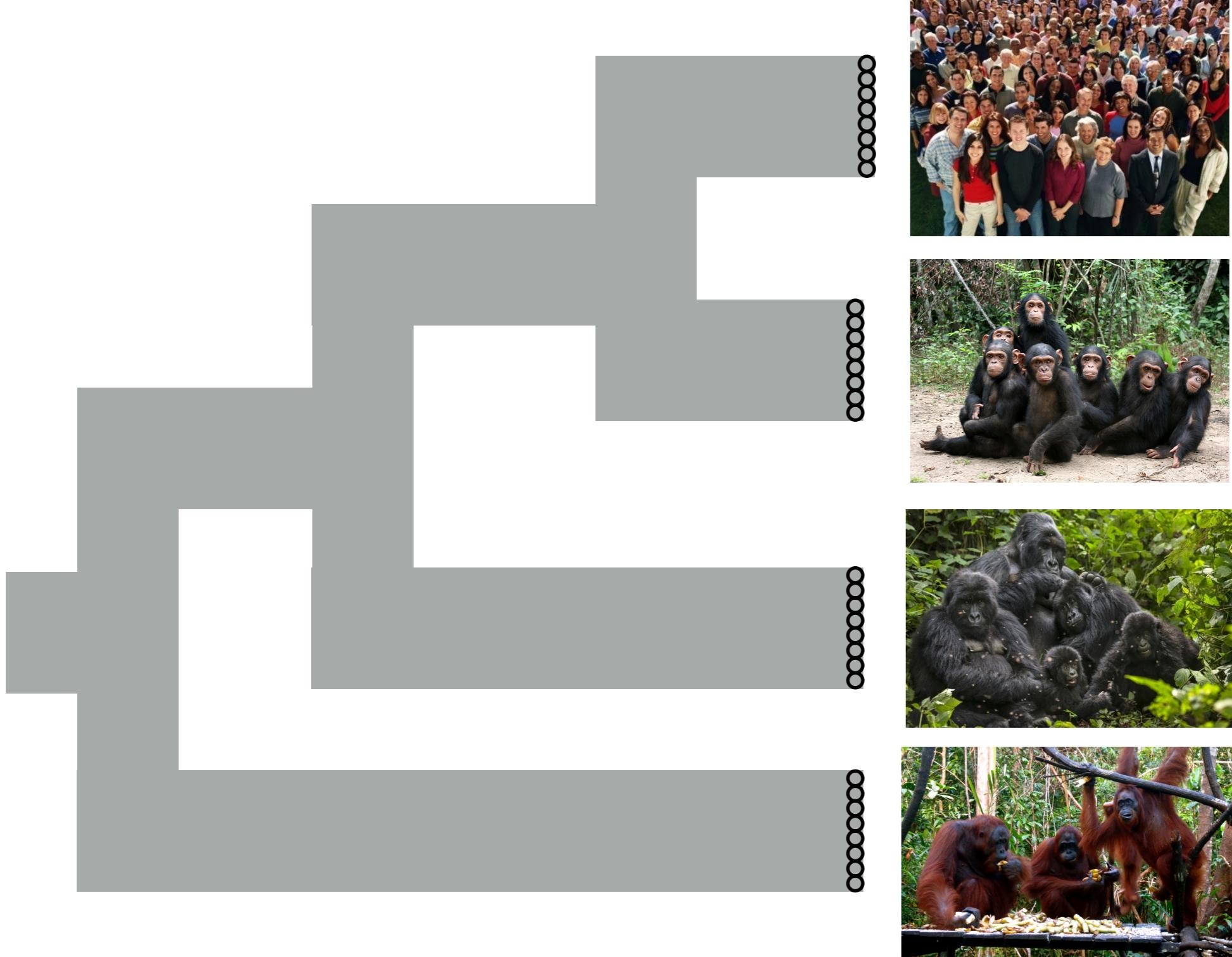
# Coalescent model



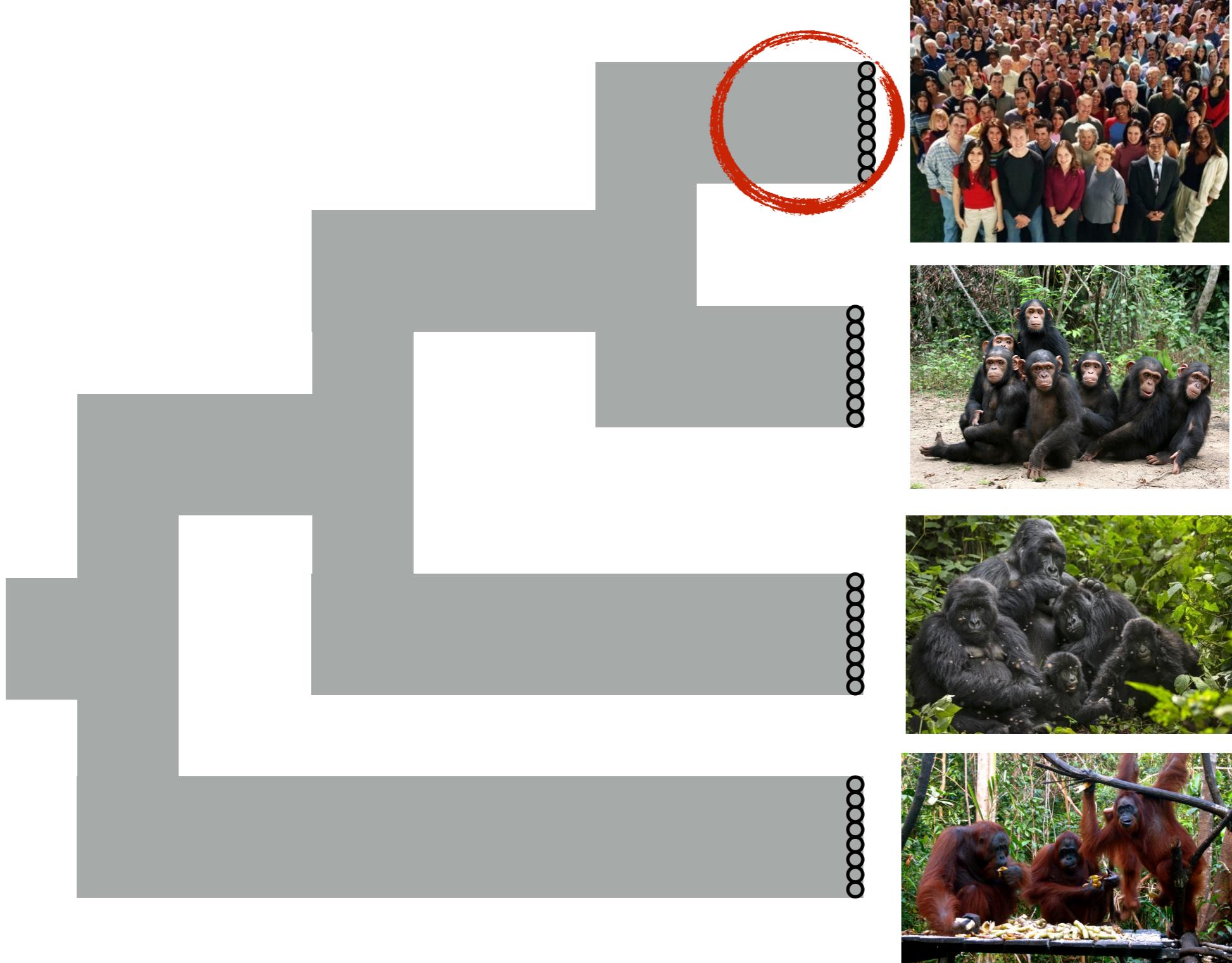
# Coalescent model



# Coalescent model

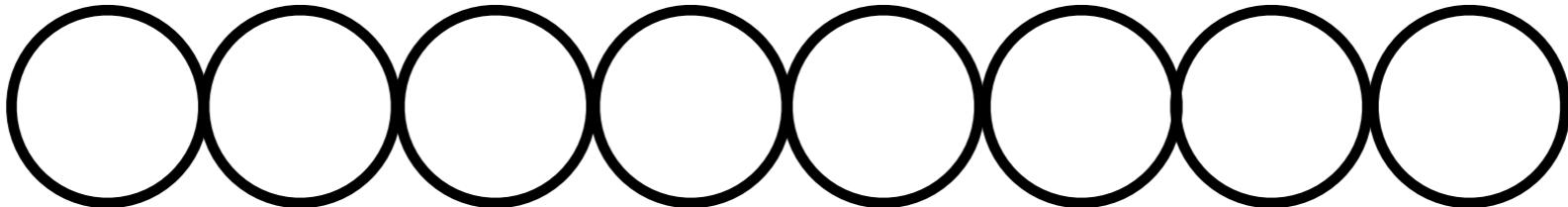


# Coalescent model



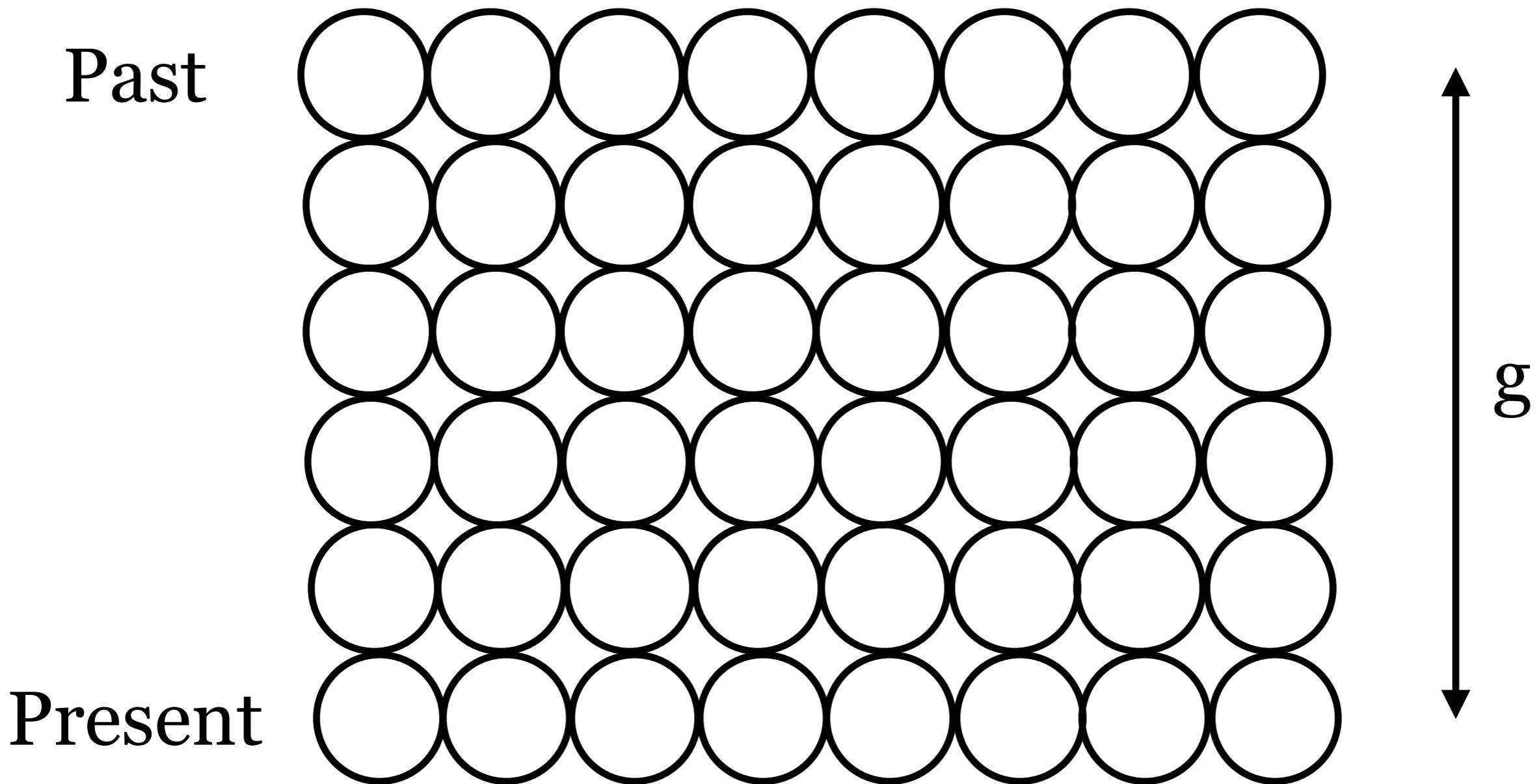
# Coalescent model within I population

Present



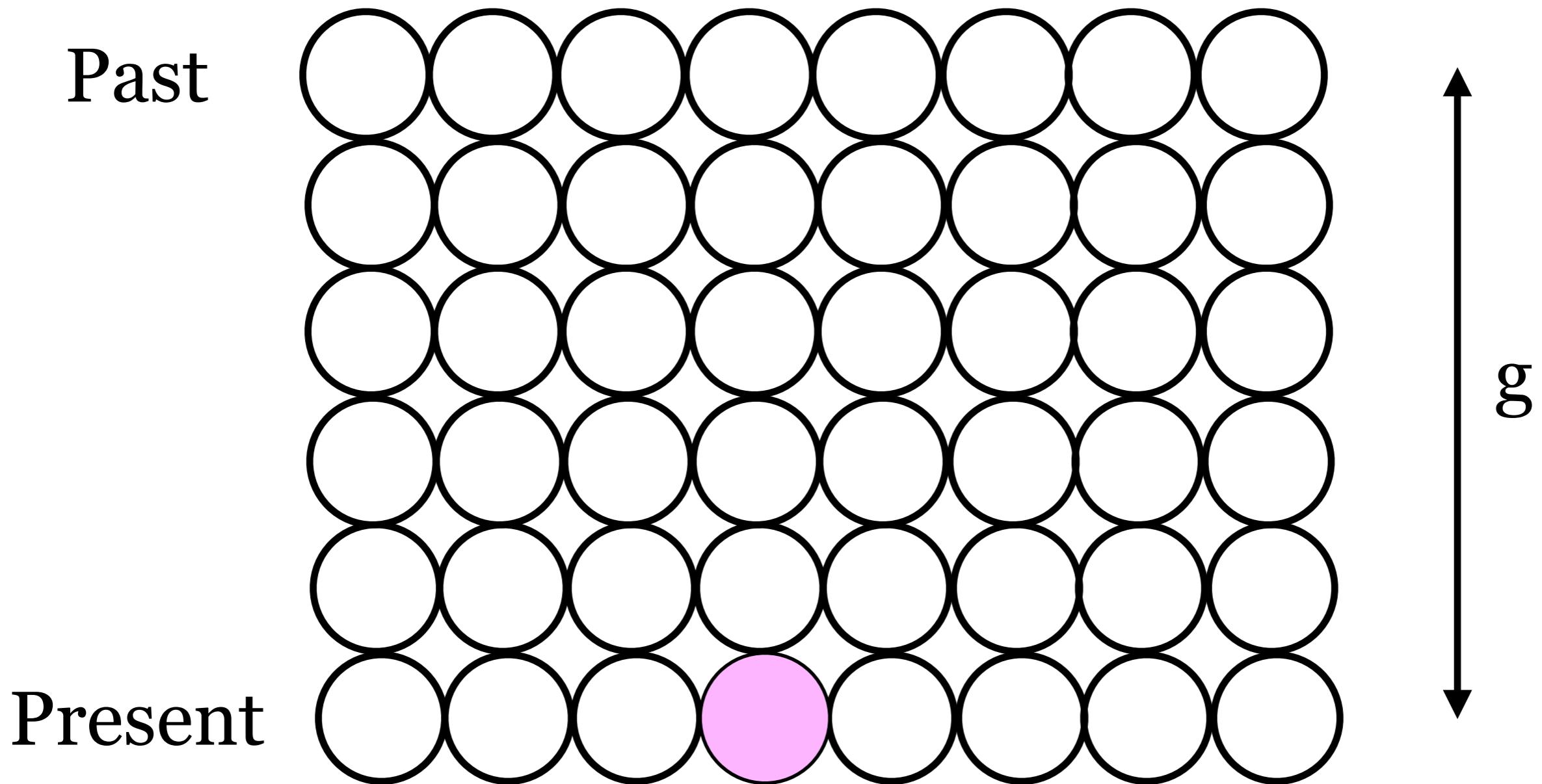
# Coalescent model within I

population



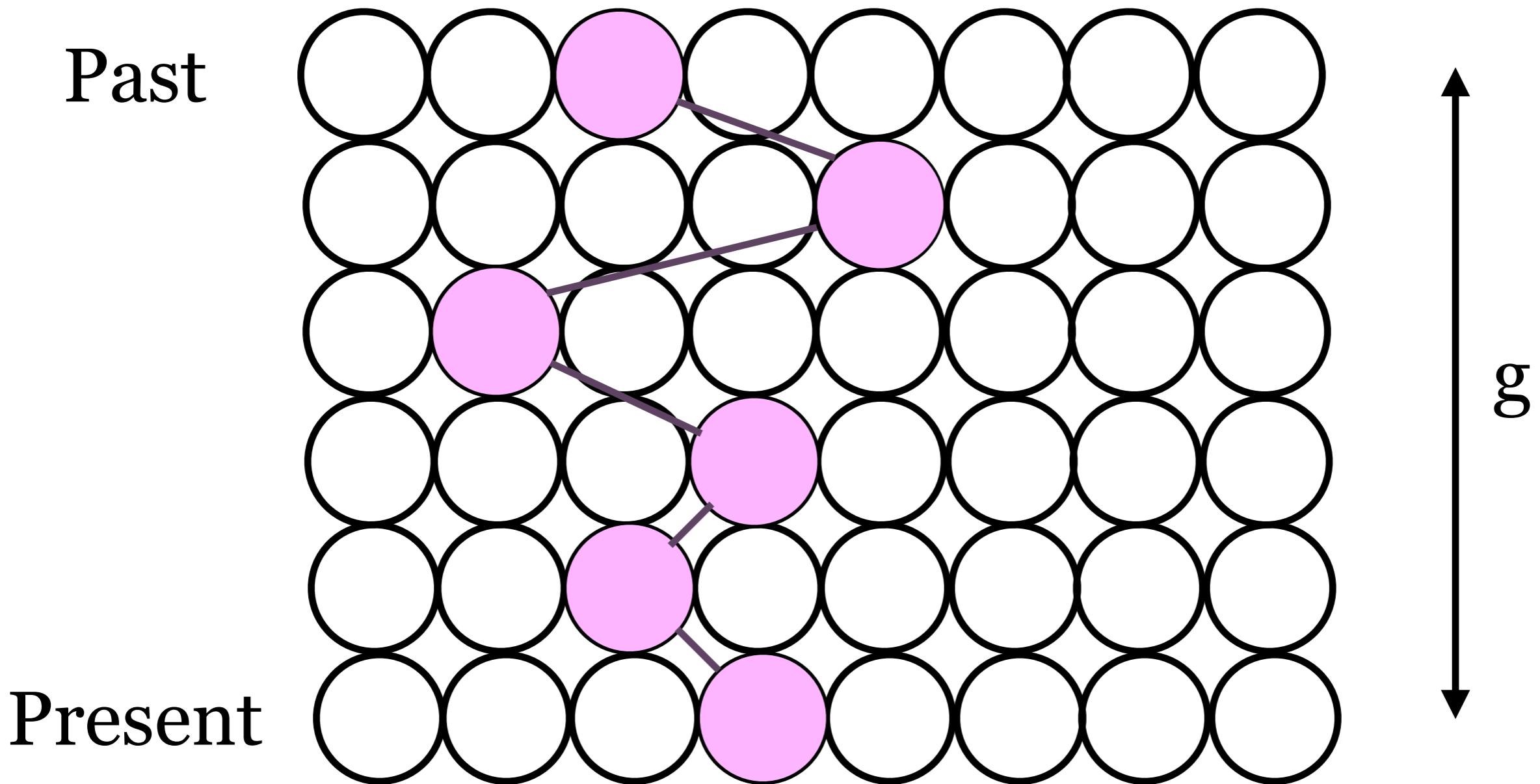
# Coalescent model within I

population



# Coalescent model within I

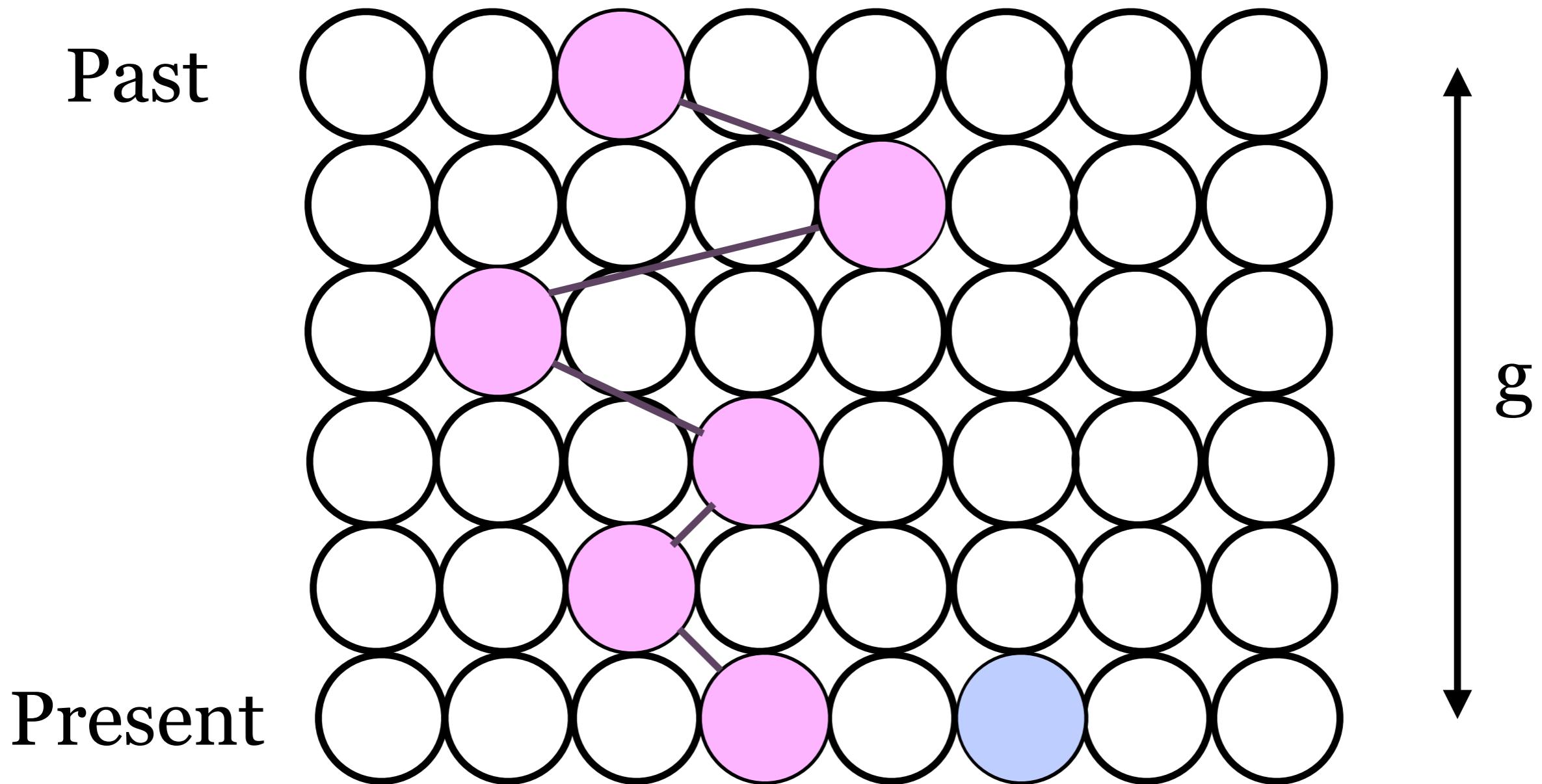
population



No selection: uniform probability  $1/N$

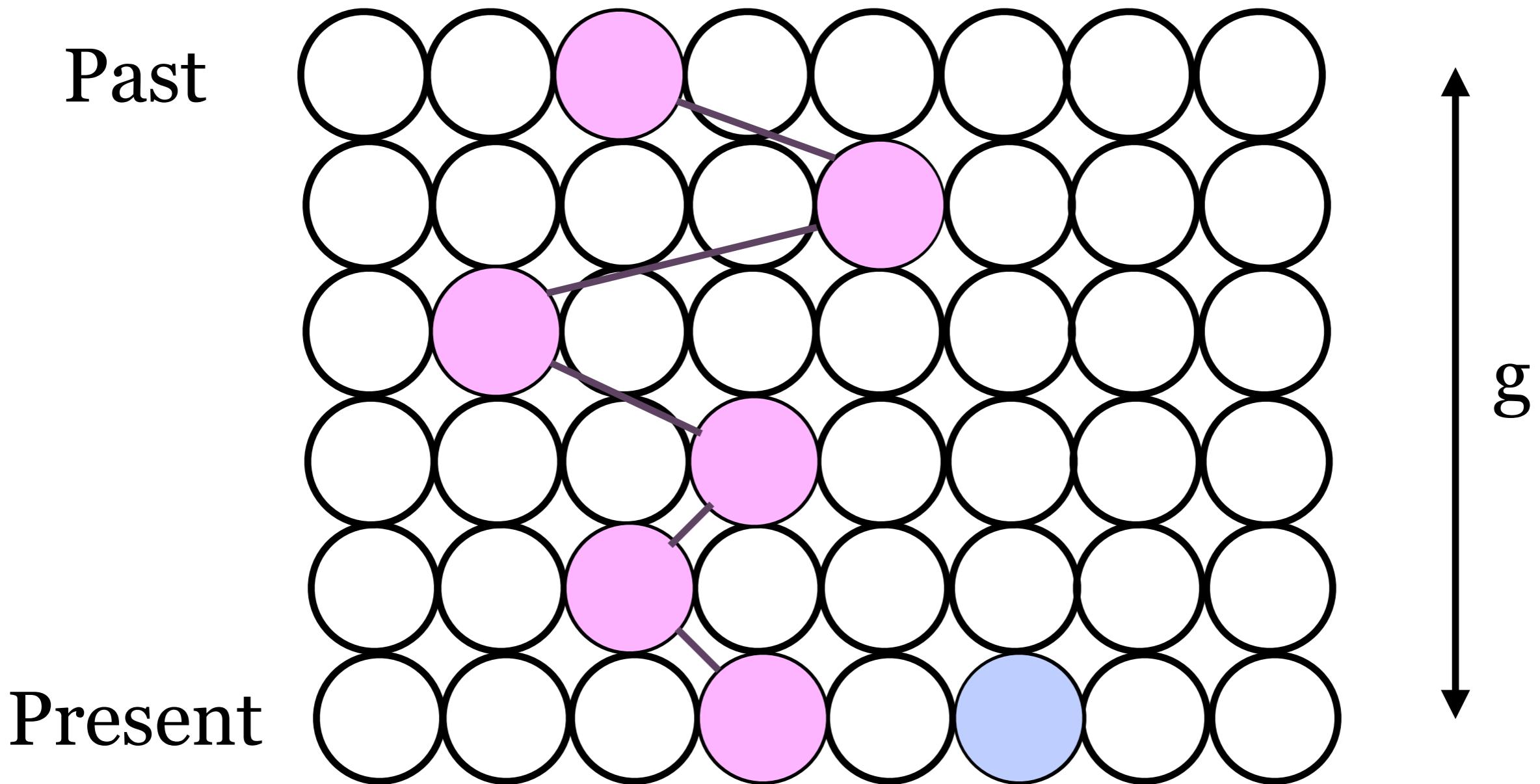
# Coalescent model within I

population



# Coalescent model within I

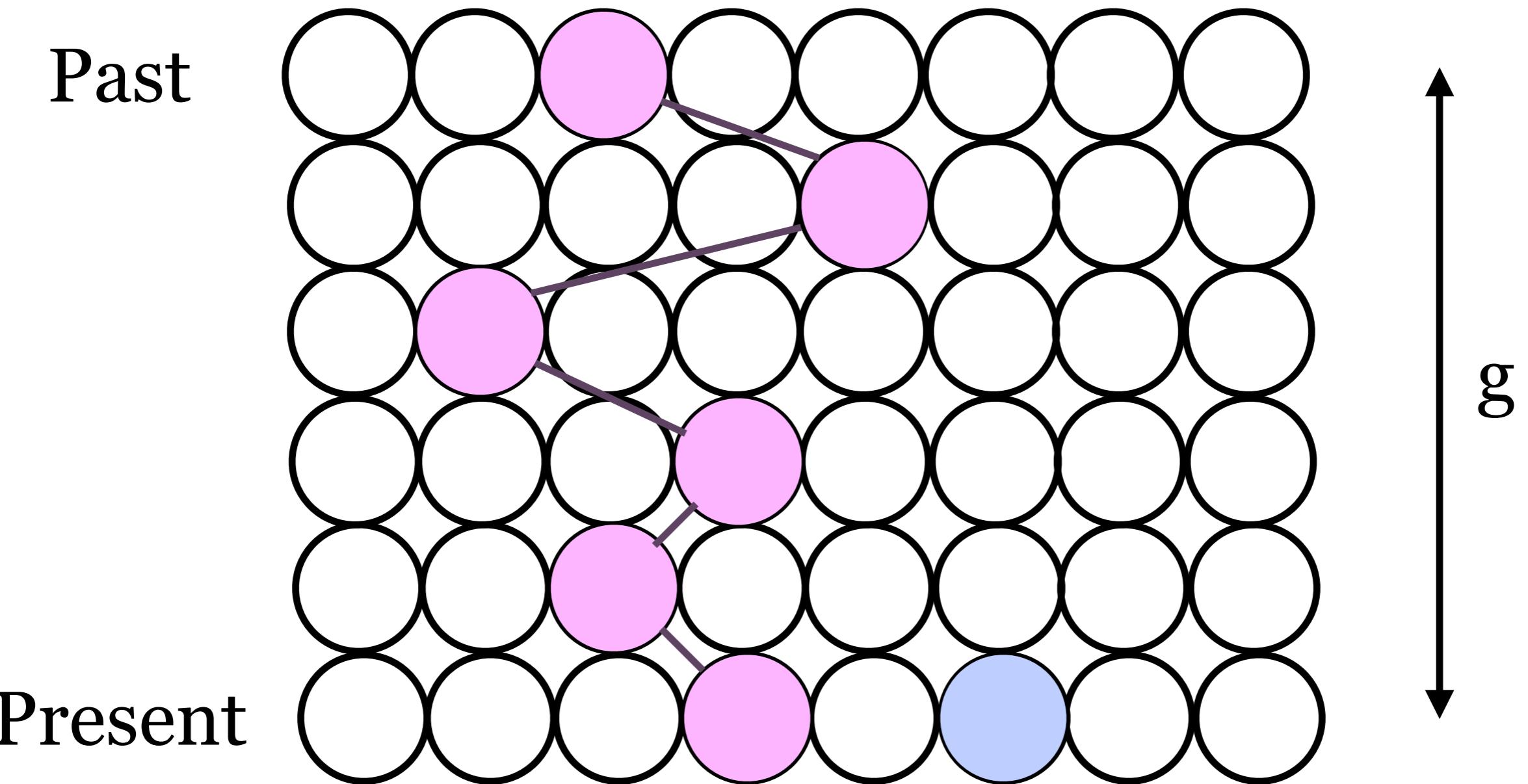
## population



How many generations do we have to wait for these  
two individuals to reach a common ancestor?

# Coalescent model within I

## population



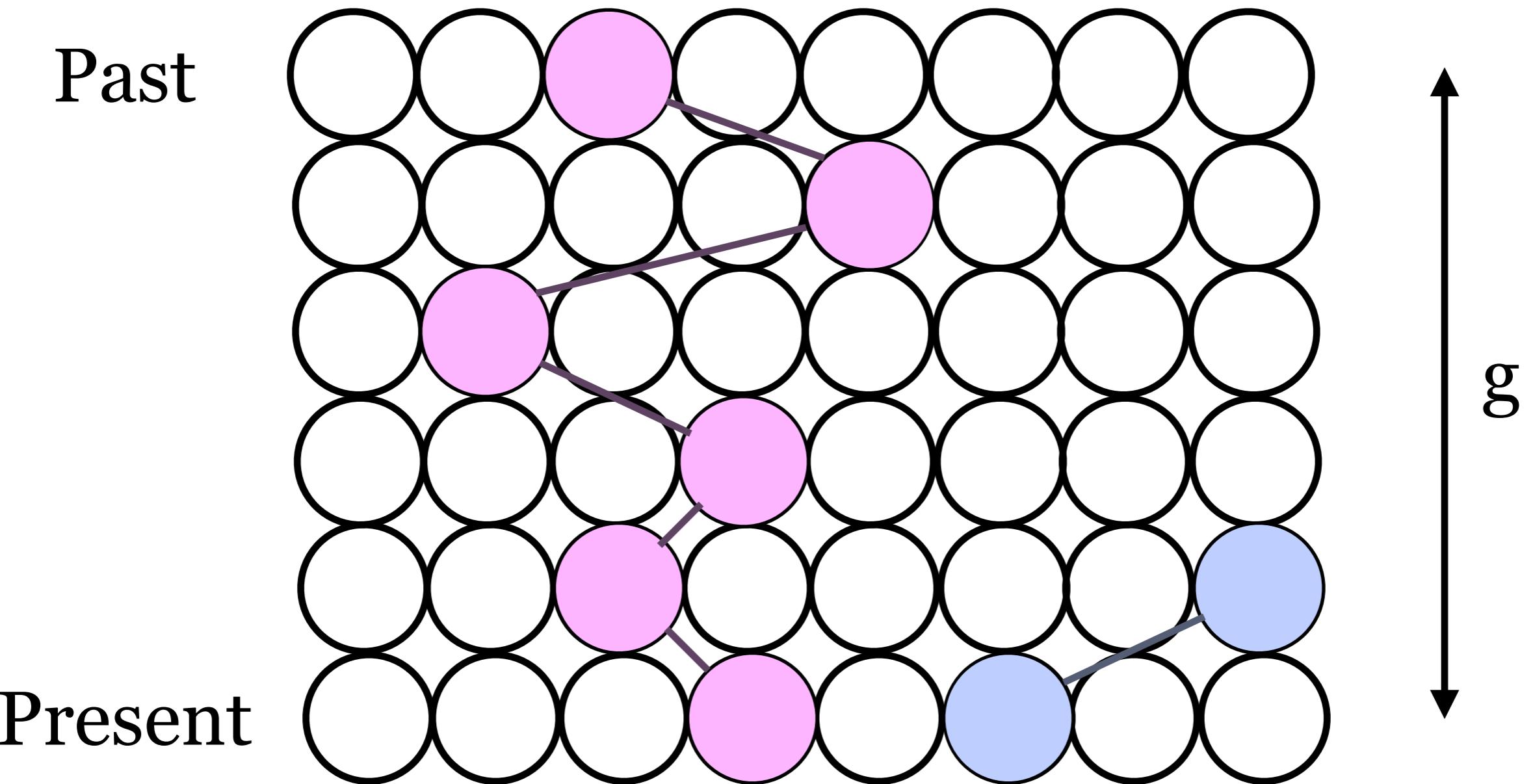
How many generations do we have  
to wait for these two individuals to  
reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

# Coalescent model within I

## population



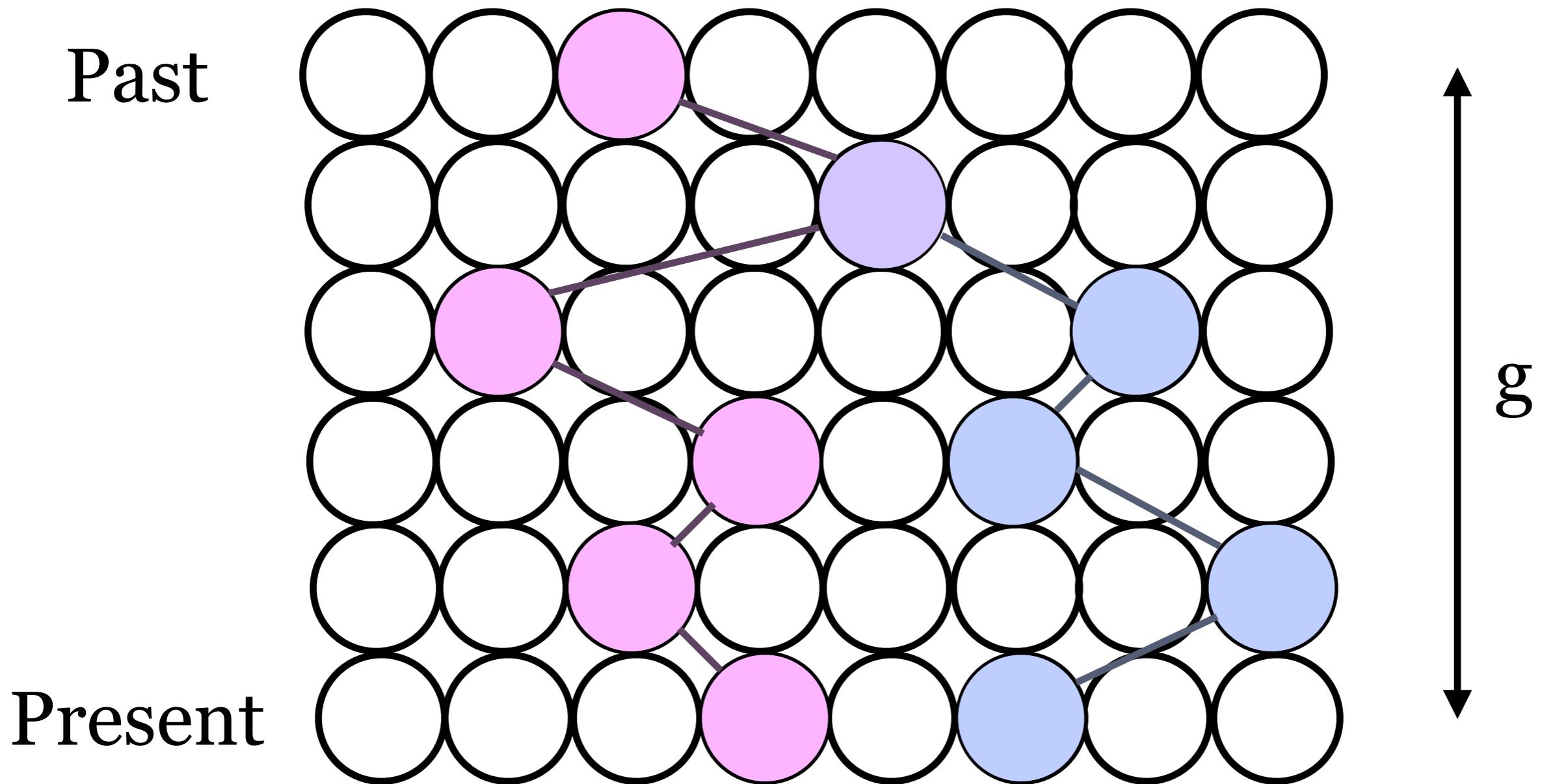
How many generations do we have  
to wait for these two individuals to  
reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

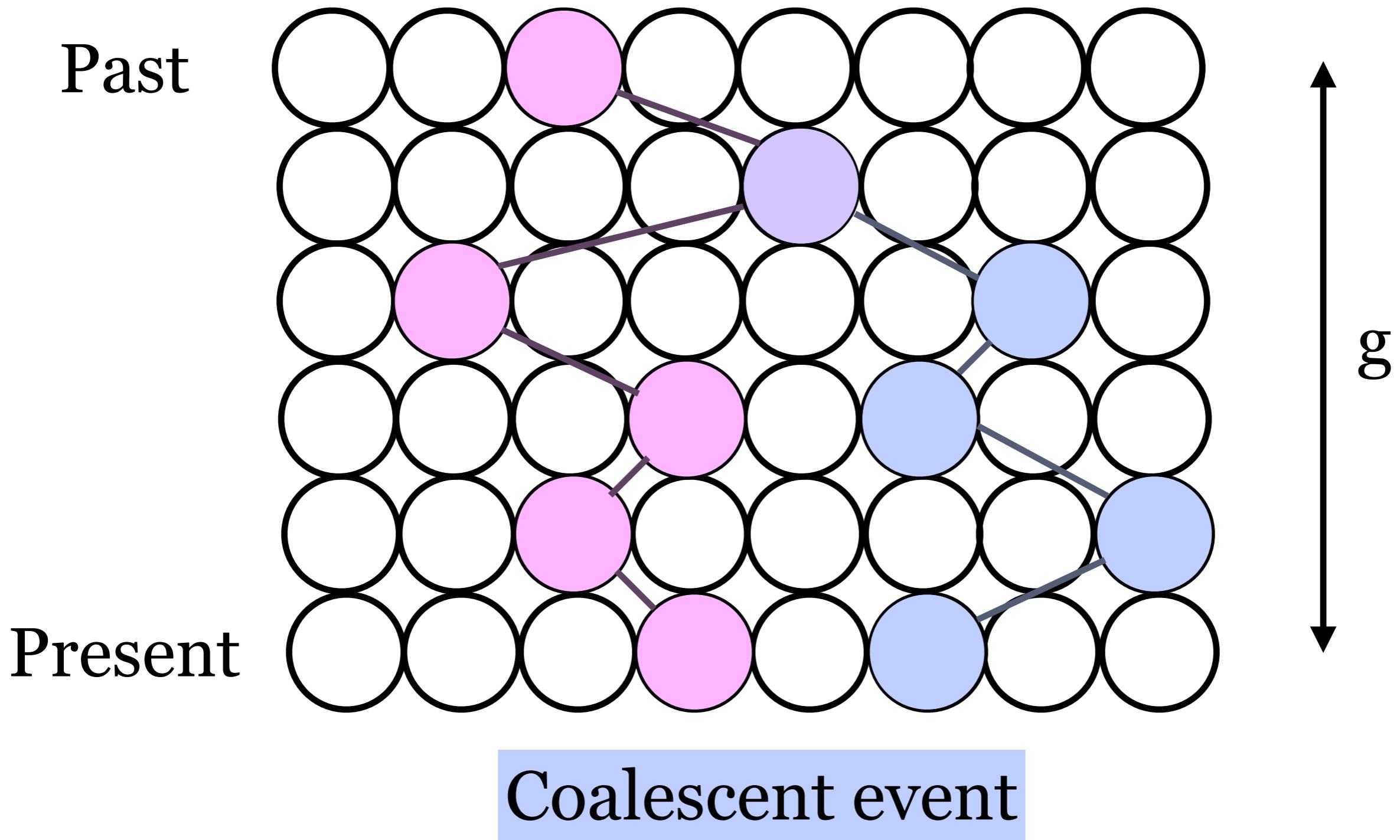
# Coalescent model within I

population

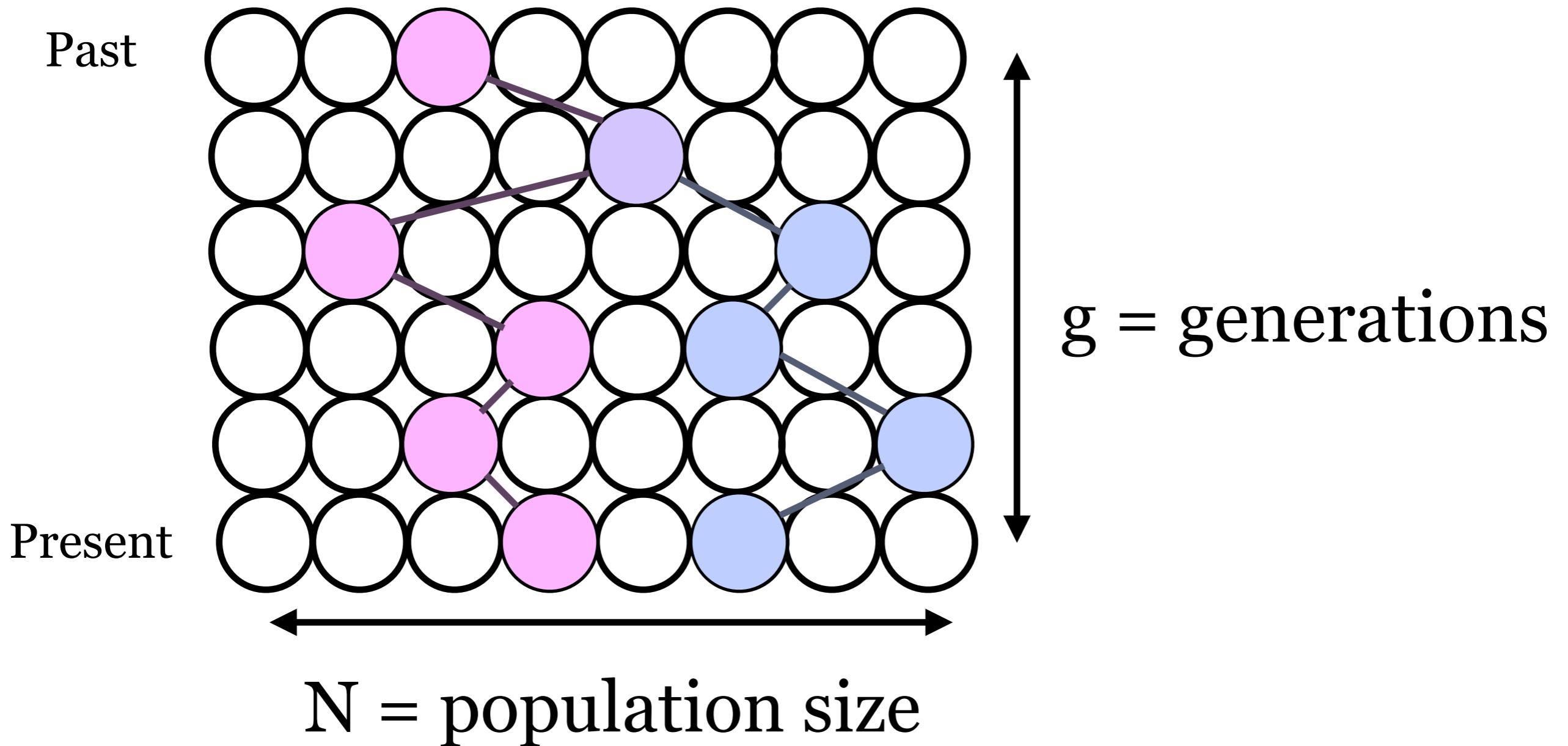


# Coalescent model within I

population



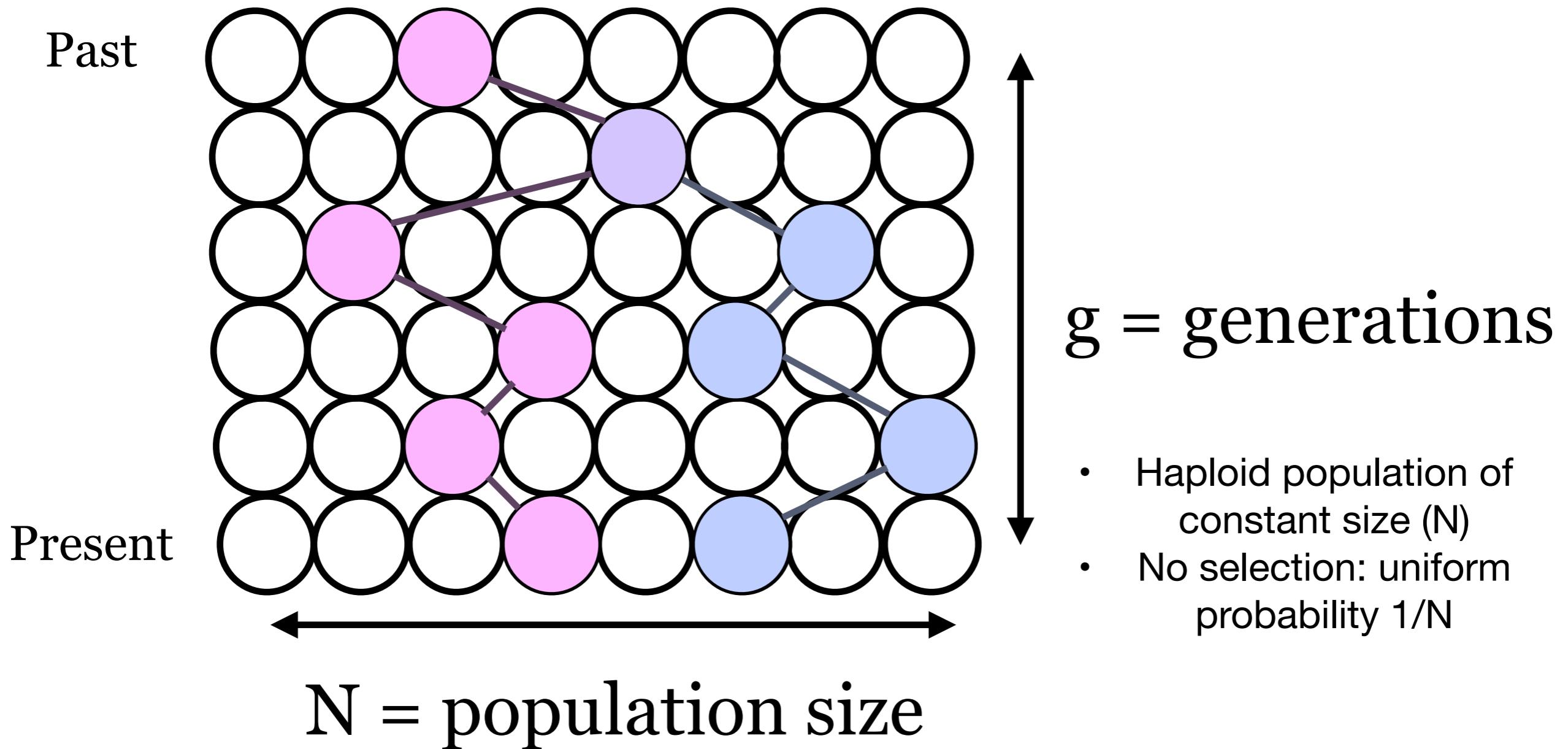
# Coalescent model within 1 population



Probability of no coalescence in  $g$  generations:

$$\left(1 - \frac{1}{N}\right)^g$$
$$t = g/N \Rightarrow \left(1 - \frac{t}{Nt}\right)^{Nt} \xrightarrow[N \rightarrow \infty]{} e^{-t}$$

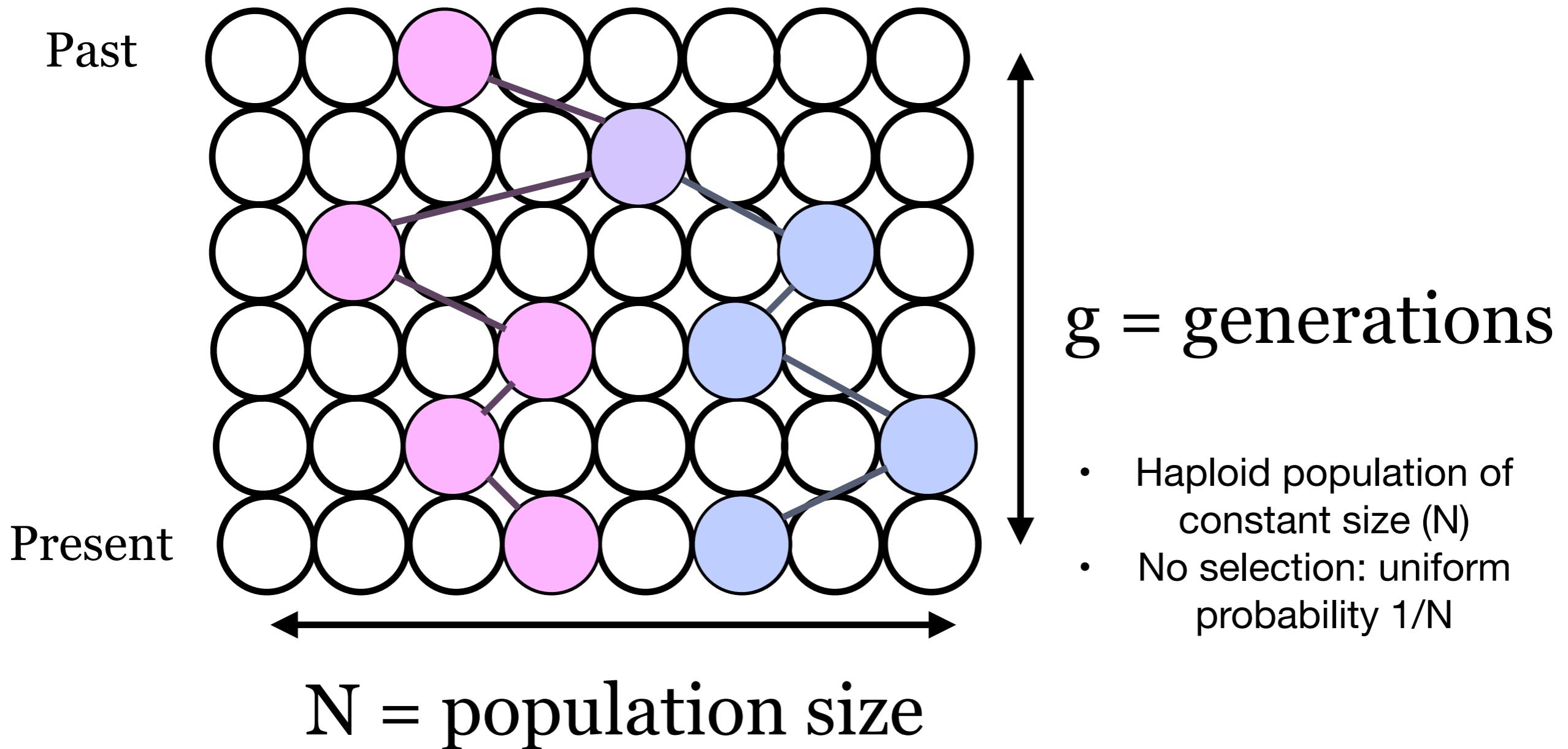
# Coalescent model within 1 population



Probability of no coalescence in  $g$  generations:

$$\left(1 - \frac{1}{N}\right)^g$$
$$t = g/N \Rightarrow \left(1 - \frac{t}{Nt}\right)^{Nt} \xrightarrow[N \rightarrow \infty]{} e^{-t}$$

# Coalescent model within 1 population



Probability of no coalescence in  $g$  generations:

$$\left(1 - \frac{1}{N}\right)^g$$
$$t = g/N \Rightarrow \left(1 - \frac{t}{Nt}\right)^{Nt} \xrightarrow[N \rightarrow \infty]{} e^{-t}$$

# Multispecies coalescent on a tree

Probability of no coalescence in  $g$  generations:  $\left(1 - \frac{1}{N}\right)^g$

$$t = g/N \Rightarrow \left(1 - \frac{t}{Nt}\right)^{Nt} \xrightarrow[N \rightarrow \infty]{} e^{-t}$$

$$T = \frac{g}{N} \text{ coalescent units} \sim \text{Exp}(1)$$

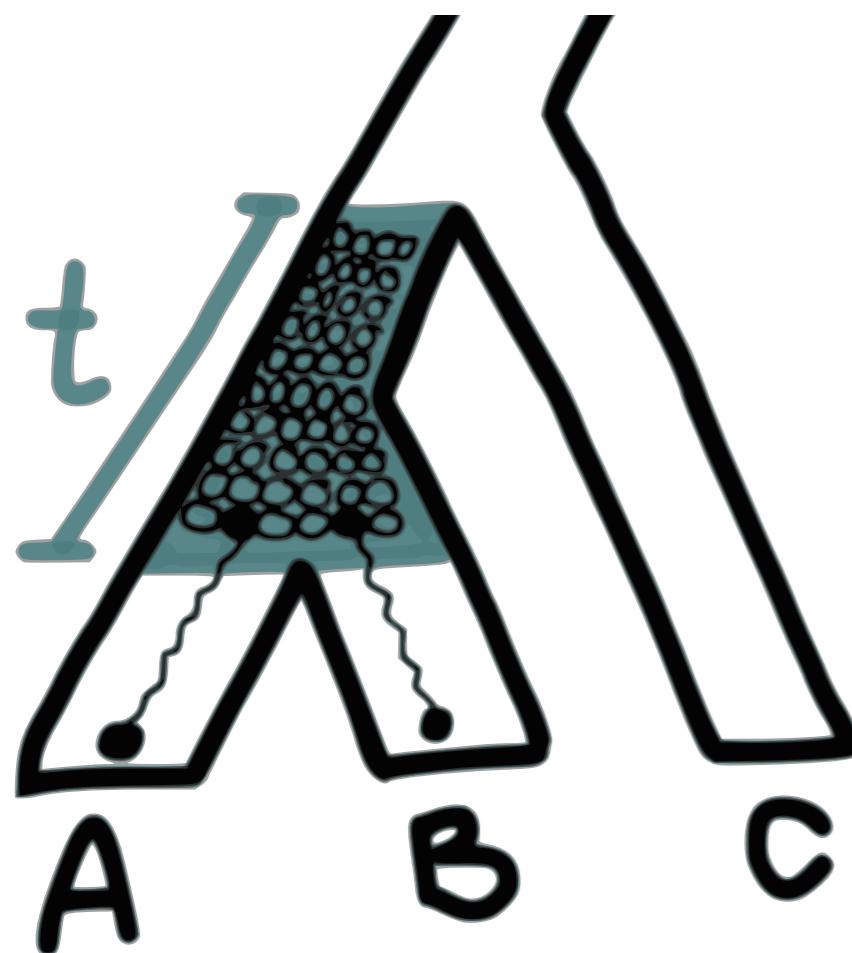
↑  
Time to coalesce

$$P(T > t) = e^{-t}$$

Probability of no  
coalescence in time  $t$

# Multispecies coalescent on a tree

Probability of no coalescence in  $g$  generations:  $\left(1 - \frac{1}{N}\right)^g$

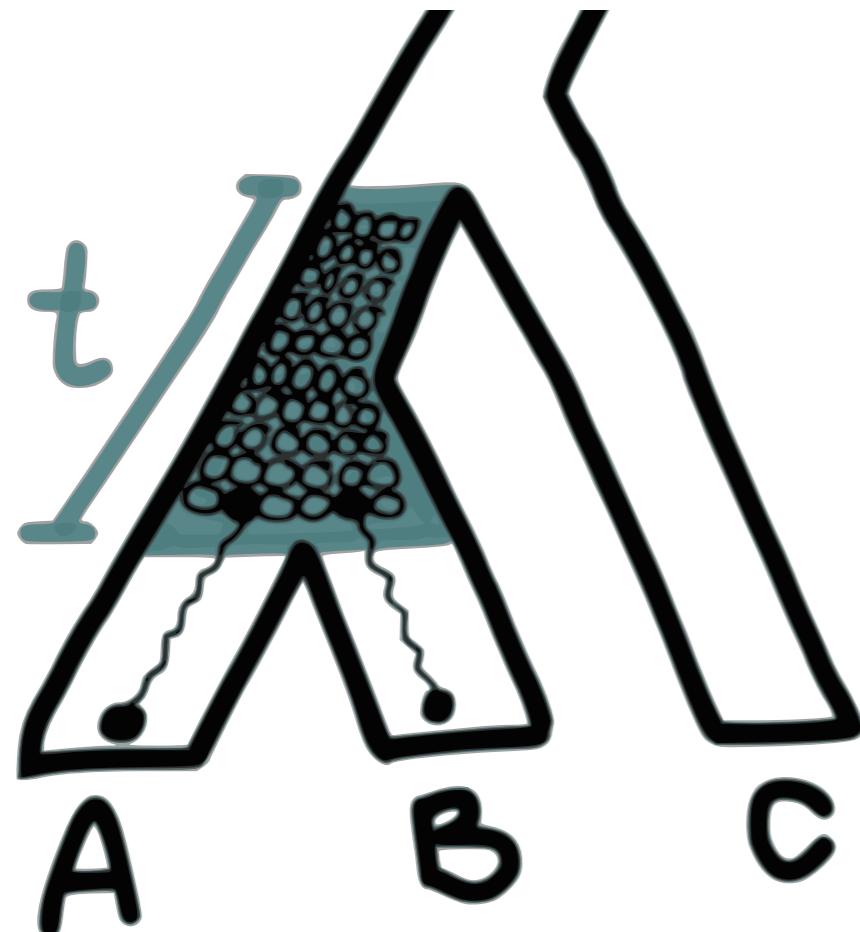
$$t = g/N \Rightarrow \left(1 - \frac{t}{Nt}\right)^{Nt} \xrightarrow[N \rightarrow \infty]{} e^{-t}$$


$$T = \frac{g}{N} \text{ coalescent units} \sim \text{Exp}(1)$$

Time to coalesce

$$P(T > t) = e^{-t} \quad \begin{matrix} \text{Probability of no} \\ \text{coalescence in time } t \end{matrix}$$

# Multispecies coalescent on a tree



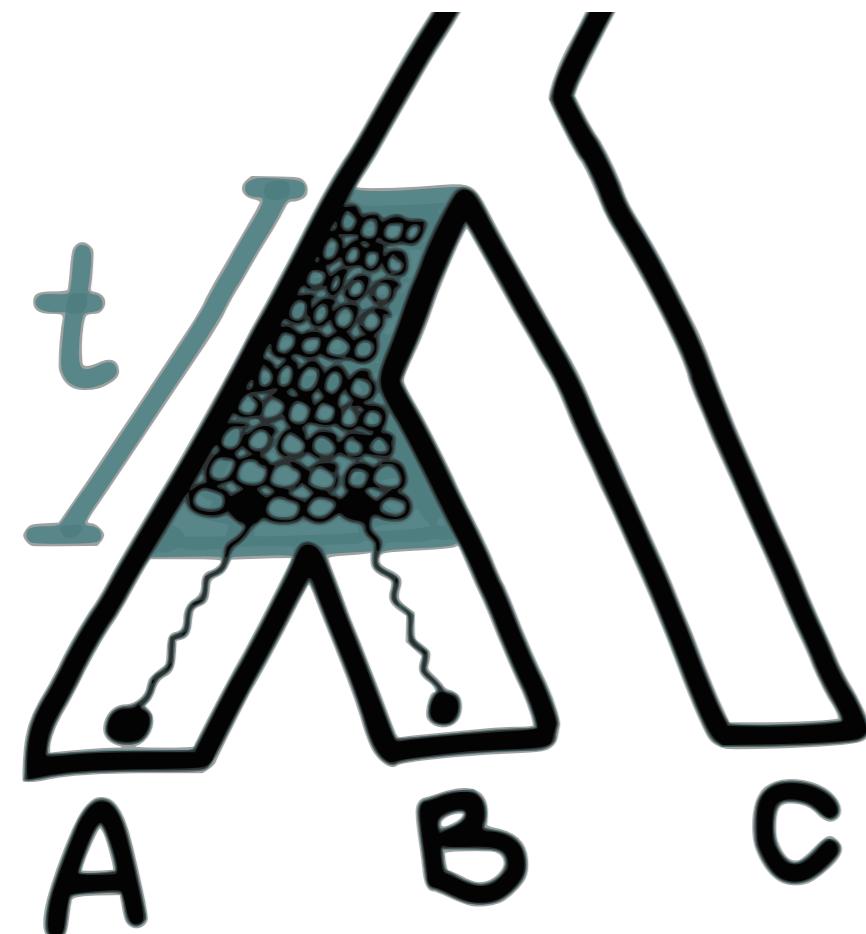
$$P(\text{ } \text{ } \text{ } \text{ } \text{ } \text{ } ) =$$

A large black letter P followed by a large black parenthesis containing a phylogenetic tree with three tips labeled A, B, and C. The entire expression is followed by an equals sign.

$$P(T > t) = e^{-t}$$

Probability of no  
coalescence in time t

# Multispecies coalescent on a tree

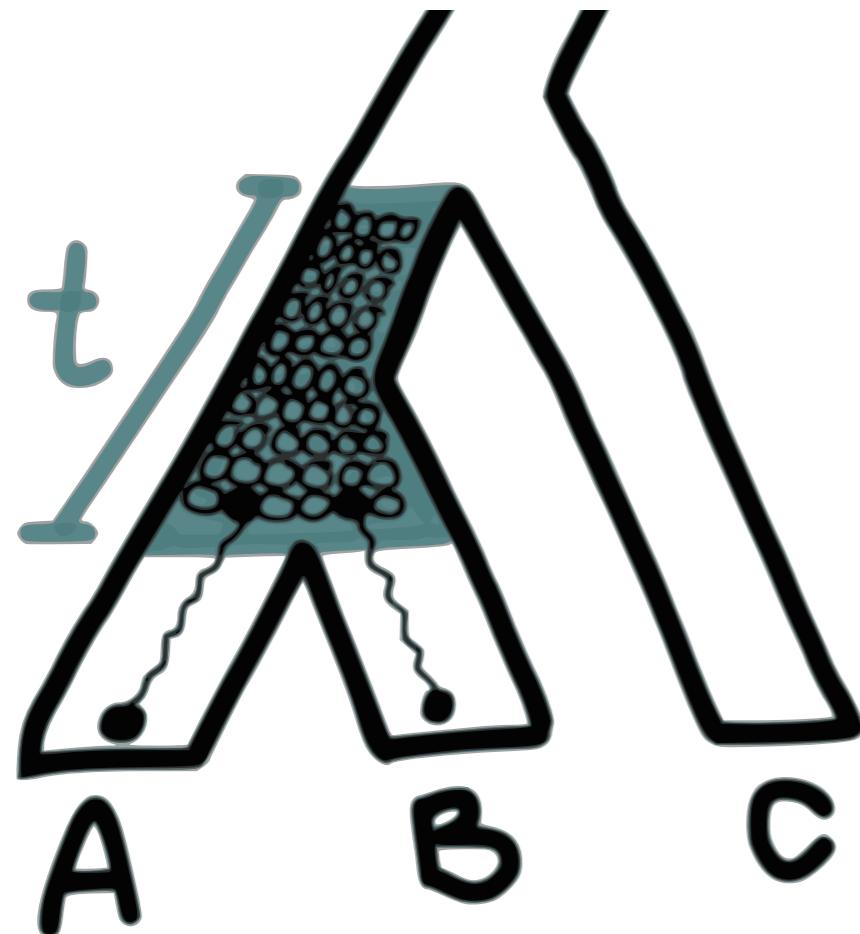


$$P(\text{ } \text{ } \text{ } \text{ } \text{ } \text{ } ) = \\ 1 - e^{-t}$$

$$P(T > t) = e^{-t}$$

Probability of no  
coalescence in time t

# Multispecies coalescent on a tree

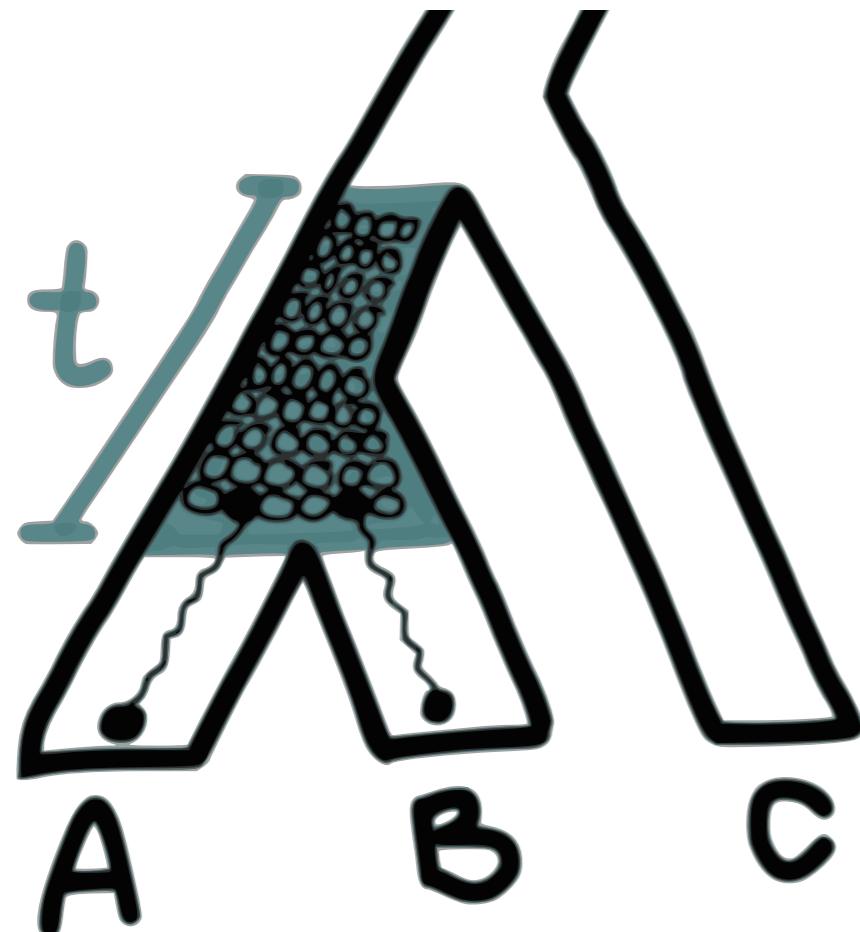


$$P(\text{ } \text{ } \text{ } \text{ } \text{ } \text{ } ) = \\ 1 - e^{-t} \\ +$$

$$P(T > t) = e^{-t}$$

Probability of no  
coalescence in time t

# Multispecies coalescent on a tree

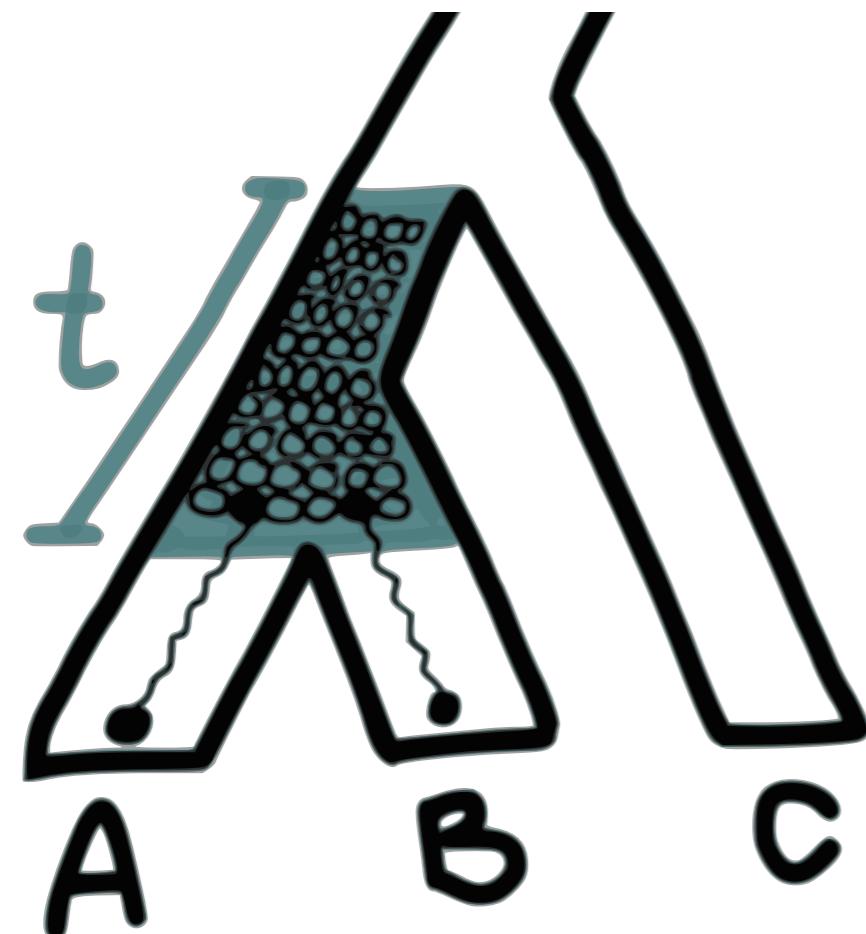


$$P(\text{No coalescence by time } t) = 1 - e^{-t} + e^{-t} \times 1/3$$

$$P(T > t) = e^{-t}$$

Probability of no  
coalescence in time t

# Multispecies coalescent on a tree



$$P(T > t) = e^{-t}$$

Probability of no  
coalescence in time t

$$P(\text{ } \text{ } \text{ } \text{ } \text{ } \text{ } ) =$$

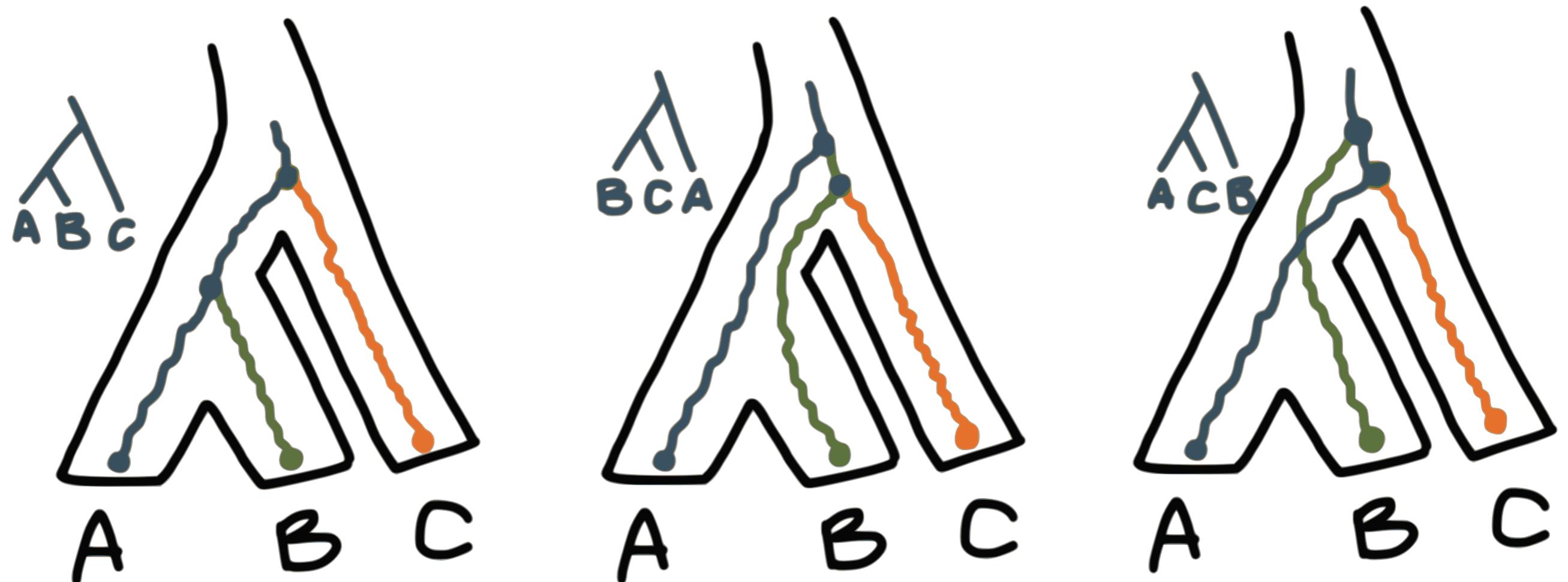
$$1 - e^{-t}$$

+

$$e^{-t} \times 1/3$$

$$= 1 - \frac{2}{3}e^{-t}$$

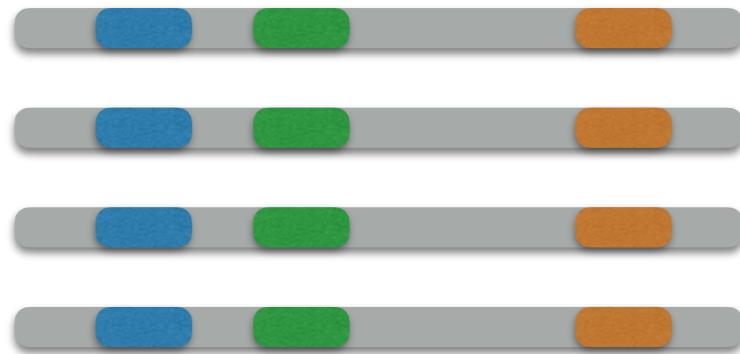
# Multispecies coalescent on a tree



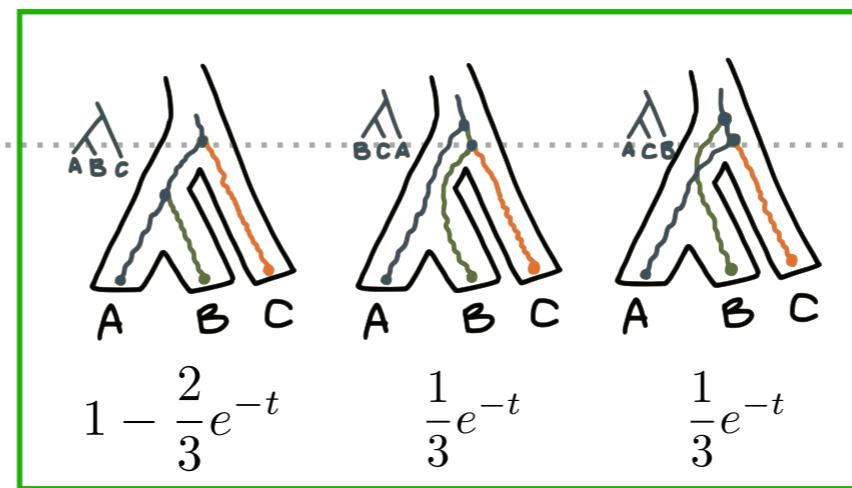
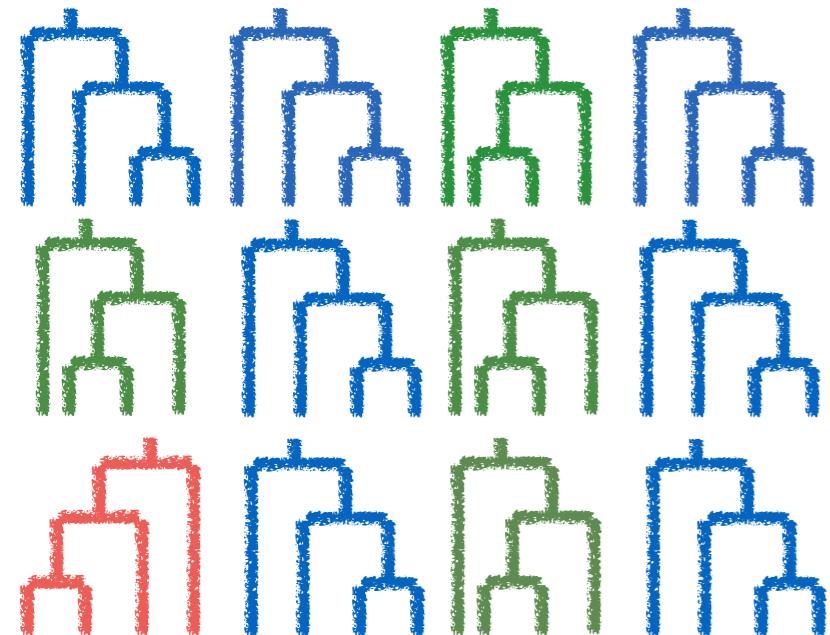
$$1 - \frac{2}{3}e^{-t}$$

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

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

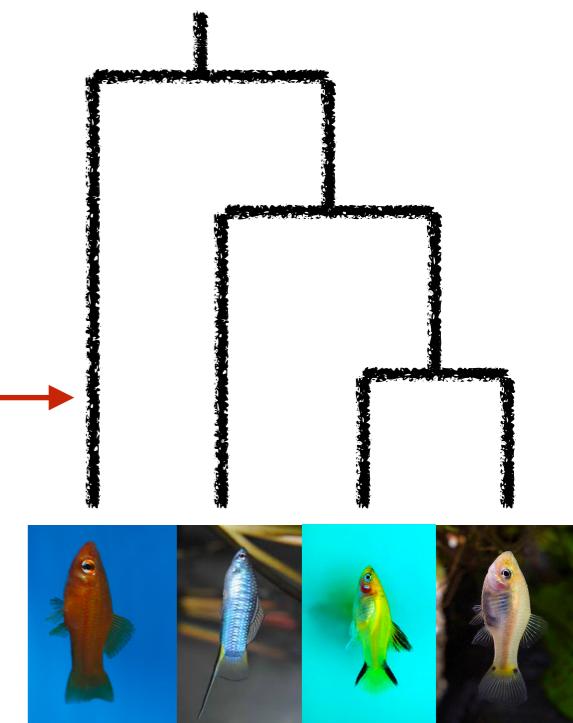


Distances  
Parsimony  
Likelihood  
(Bayesian)

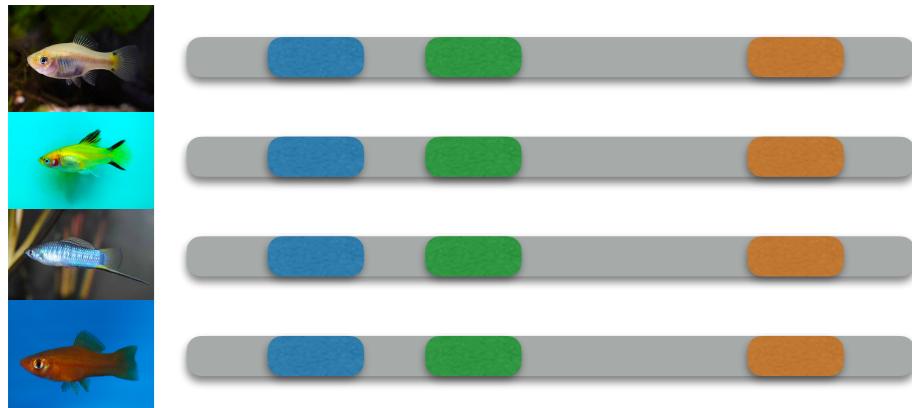


$$L(T, \theta) = \prod_{i=1}^L P(G_i | T, \theta)$$

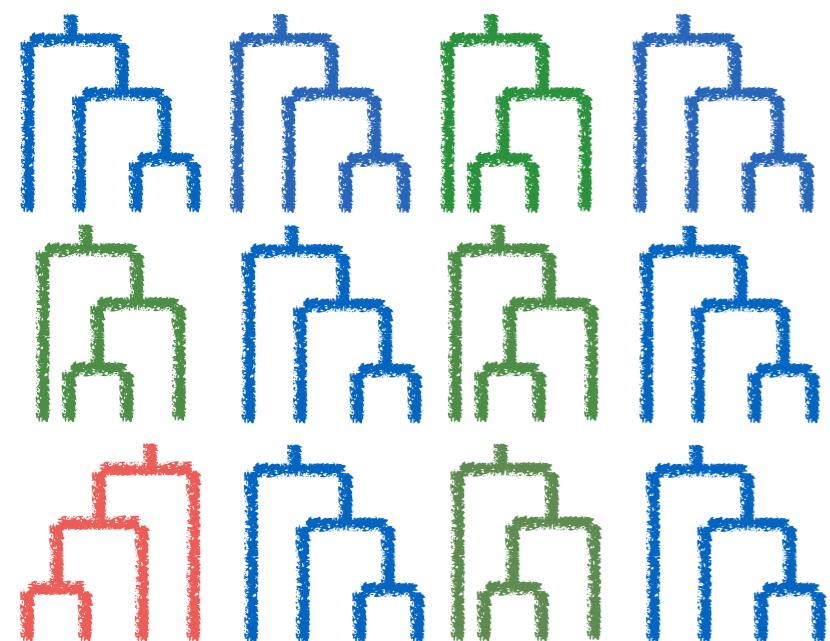
Max. Lik.



Data



Distances  
Parsimony  
Likelihood  
(Bayesian)

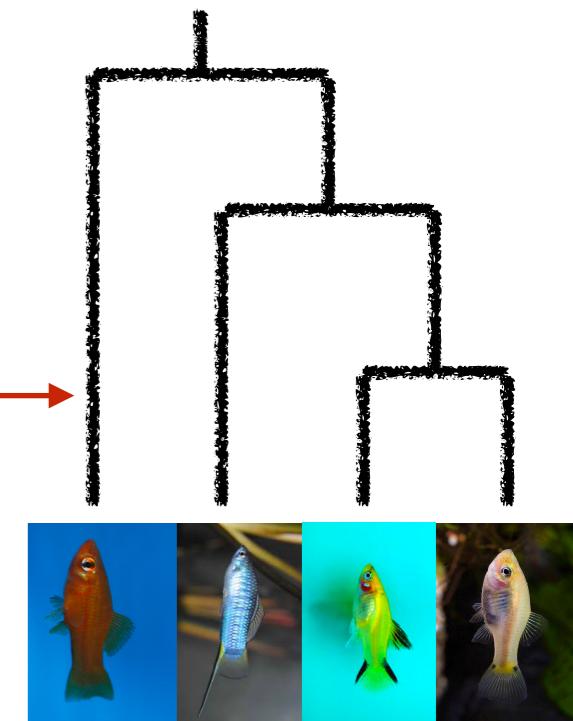


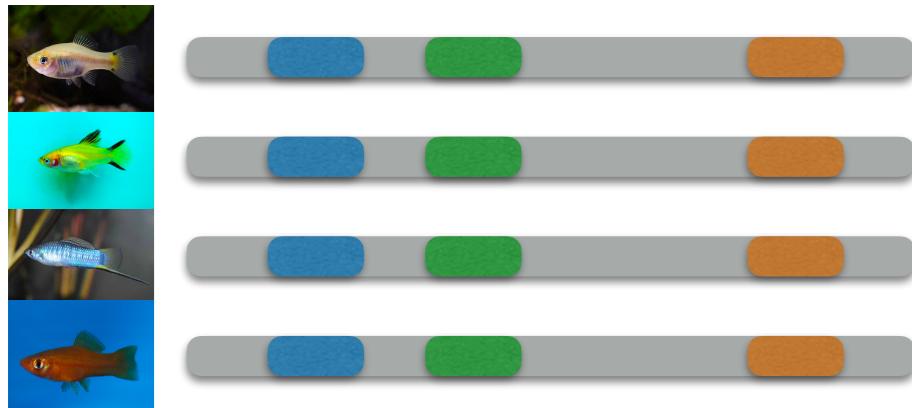
**Data**

$$L(T, \theta) = \prod_{i=1}^L P(G_i | T, \theta)$$

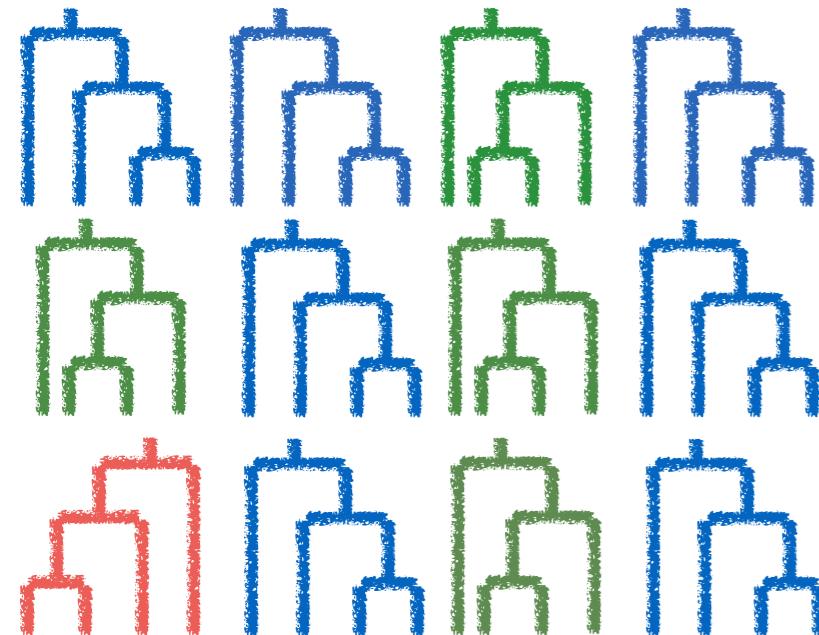
Max. Lik.

1. Guess a species tree
2. Evaluate likelihood of data (gene trees) given species tree
3. Search space of trees for species tree that maximizes likelihood





Distances  
Parsimony  
Likelihood  
(Bayesian)



$$P(T, \theta | G) \propto \pi(T) \pi(\theta) \prod_{i=1}^L P(G_i | T, \theta)$$

Prior Tree      Multispecies Coalescent

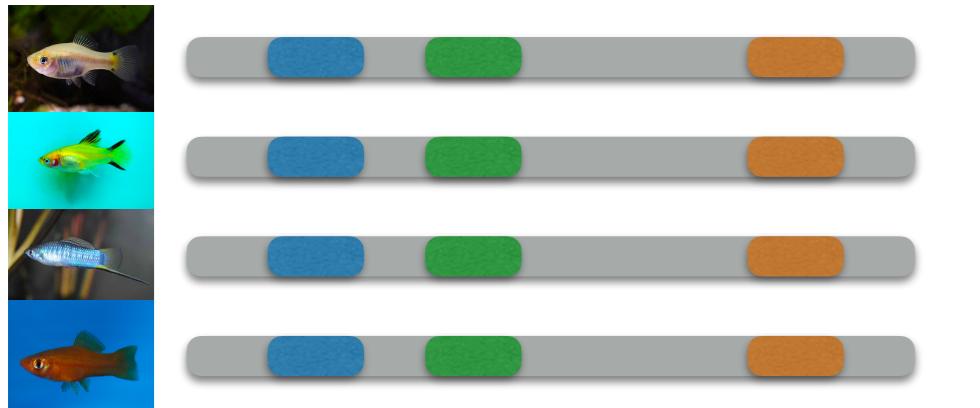


$$L(T, \theta) = \prod_{i=1}^L P(G_i | T, \theta)$$

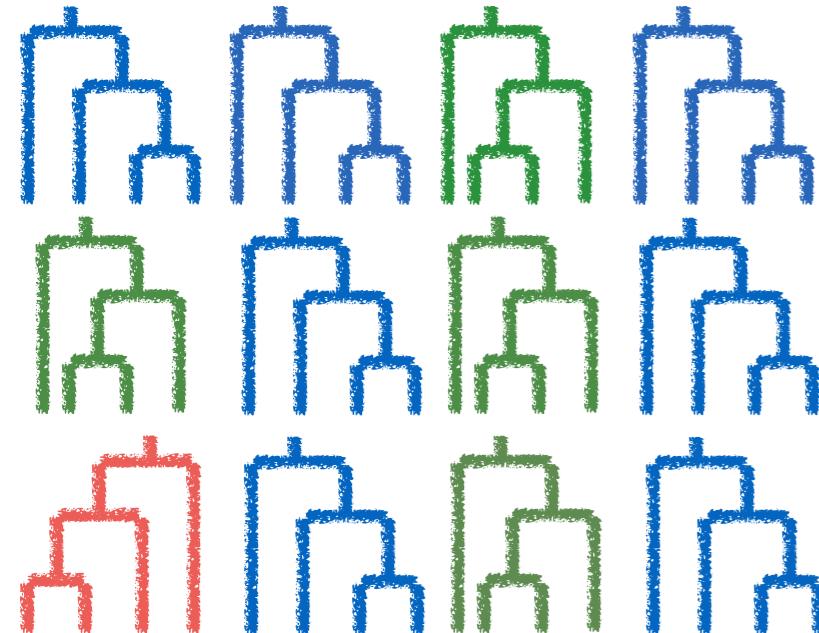
Bayesian

Max. Lik.

Summary methods: ASTRAL, BUCKy, MP-EST  
 (Zhang et al, 2018) (Larget et al, 2010) (Liu et al, 2010)



Distances  
Parsimony  
Likelihood  
(Bayesian)



$$P(T, \theta | G) \propto \pi(T) \pi(\theta) \prod_{i=1}^L P(G_i | T, \theta)$$

Prior Tree      Multispecies Coalescent



$$L(T, \theta) = \prod_{i=1}^L P(G_i | T, \theta)$$

Bayesian

Max. Lik.

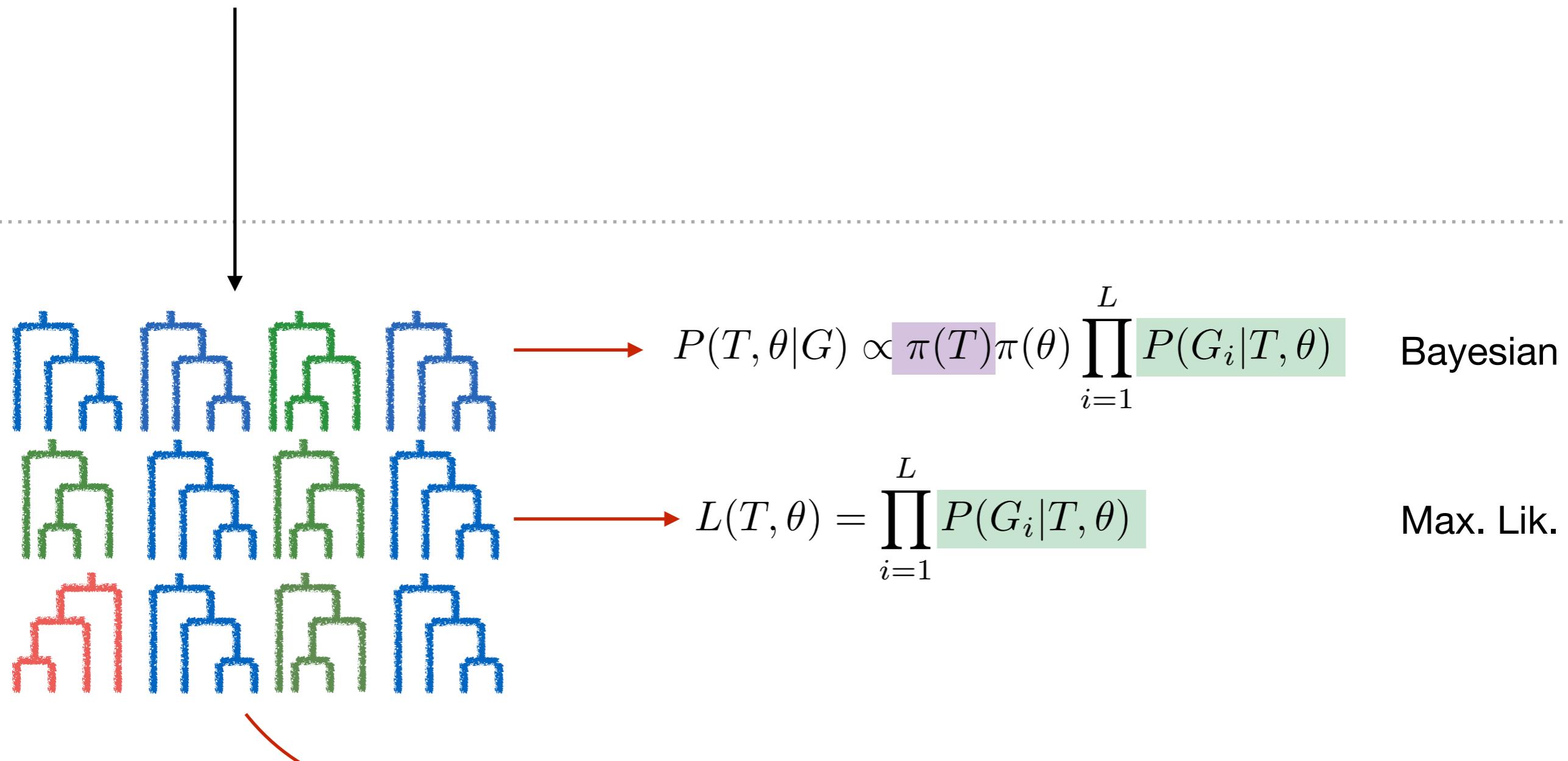
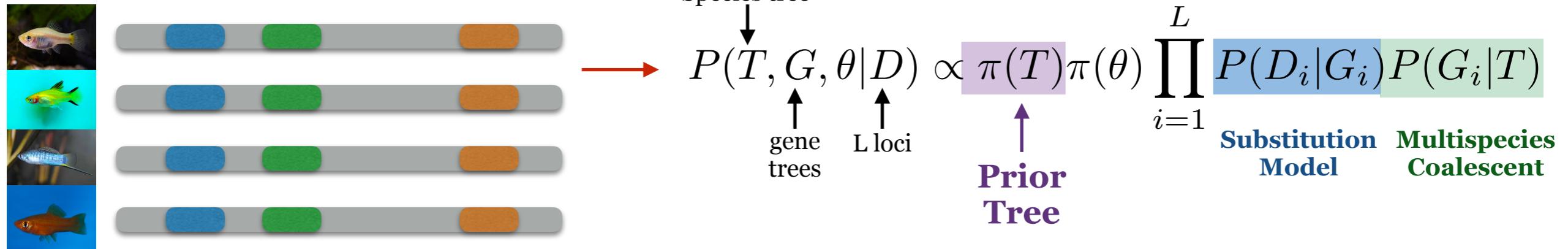
Summary methods: ASTRAL, BUCKy,  
(Zhang et al, 2018) (Larget et al, 2010)

Do not search tree space

Approx lik

MP-EST  
(Liu et al, 2010)

## Co-estimation (lecture 15)



Summary methods: ASTRAL, BUCKy, MP-EST  
 (Larget et al, 2010) (Liu et al, 2010)

# Anomaly zone

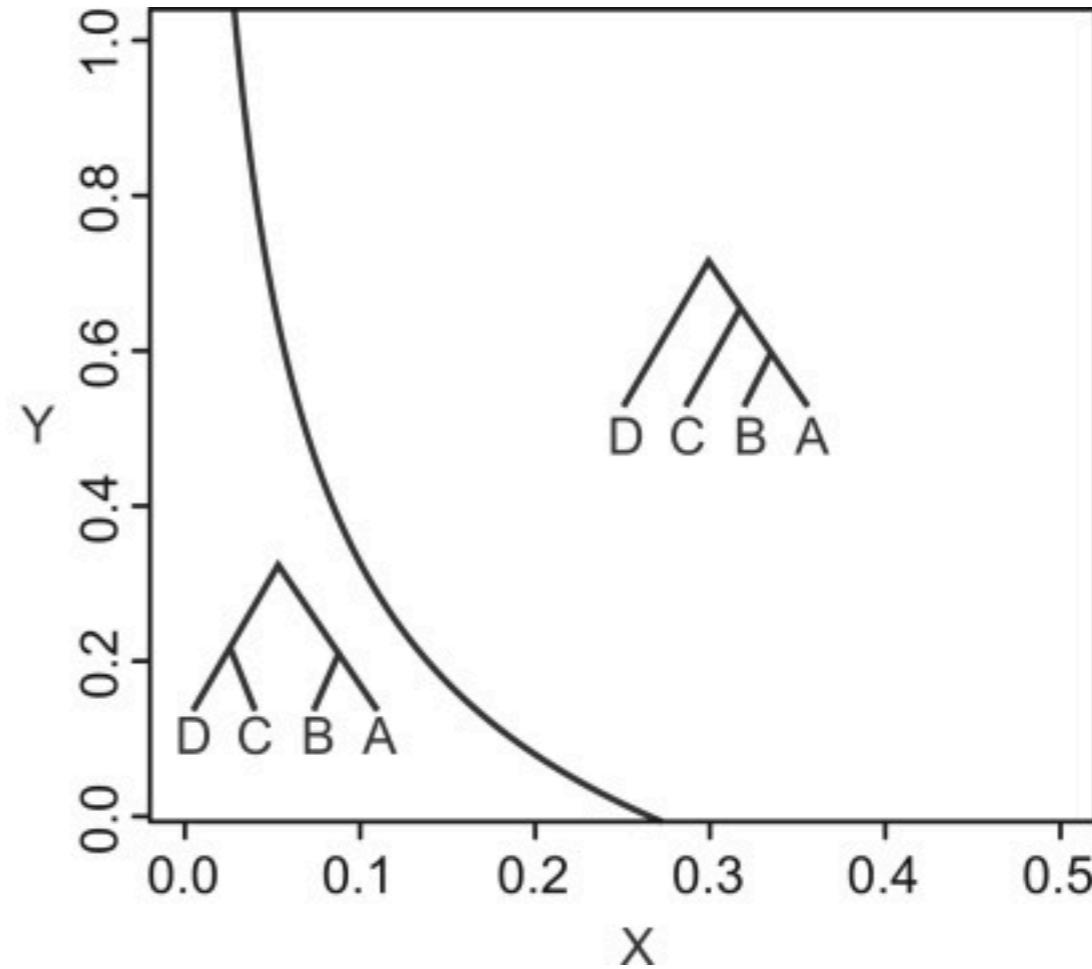
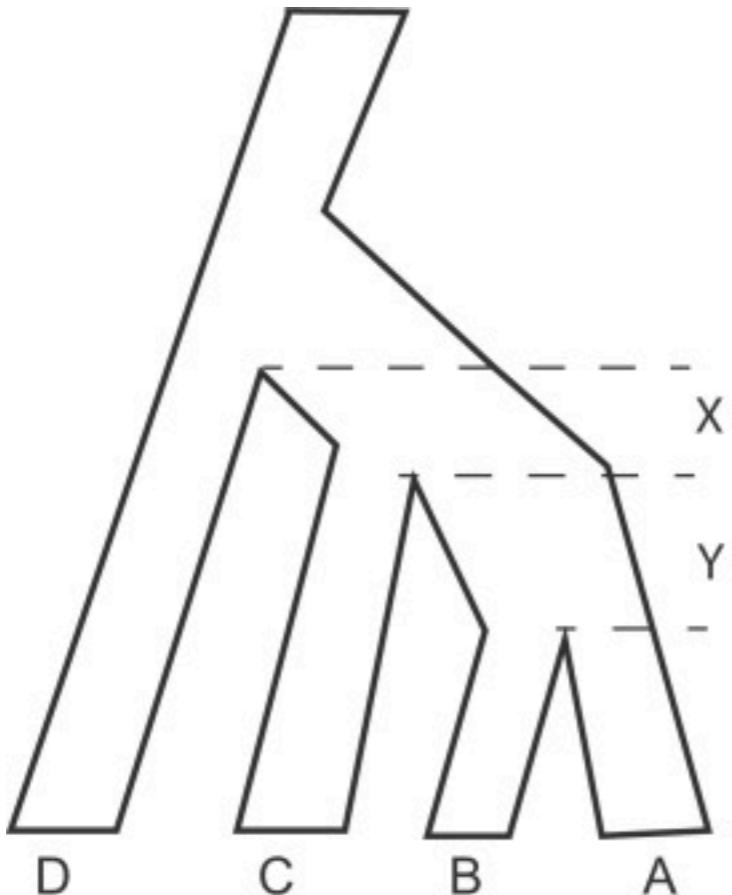
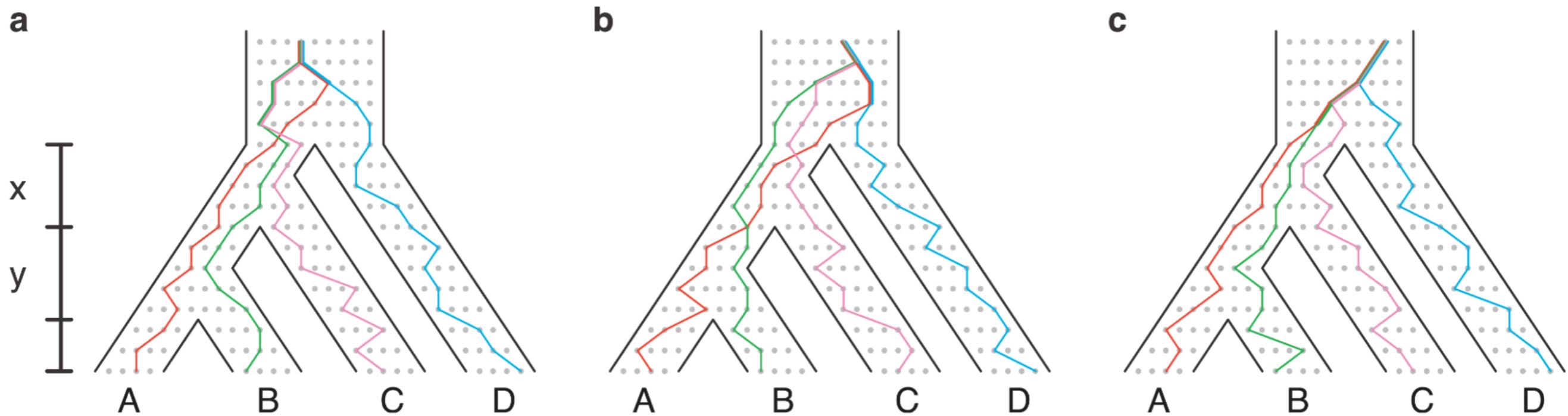
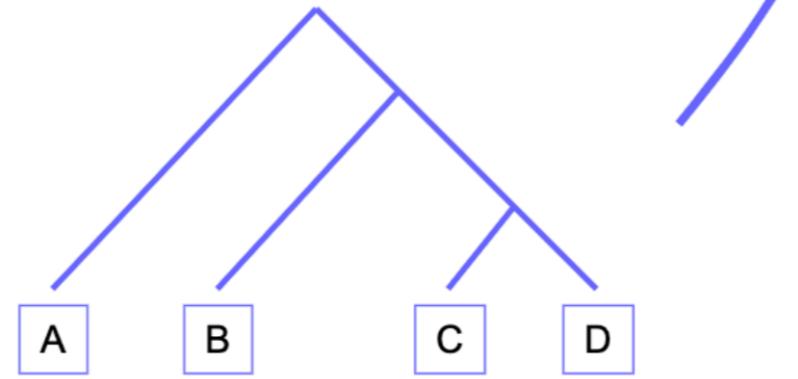
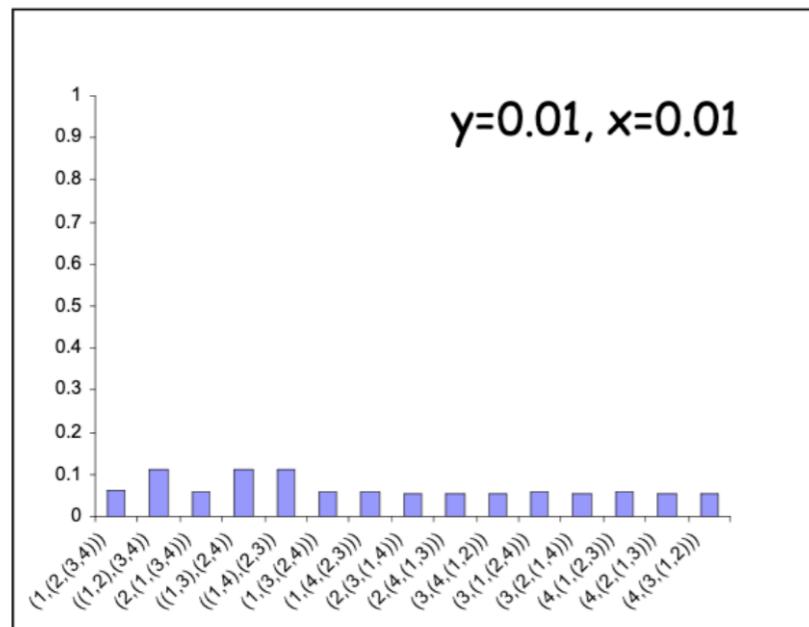
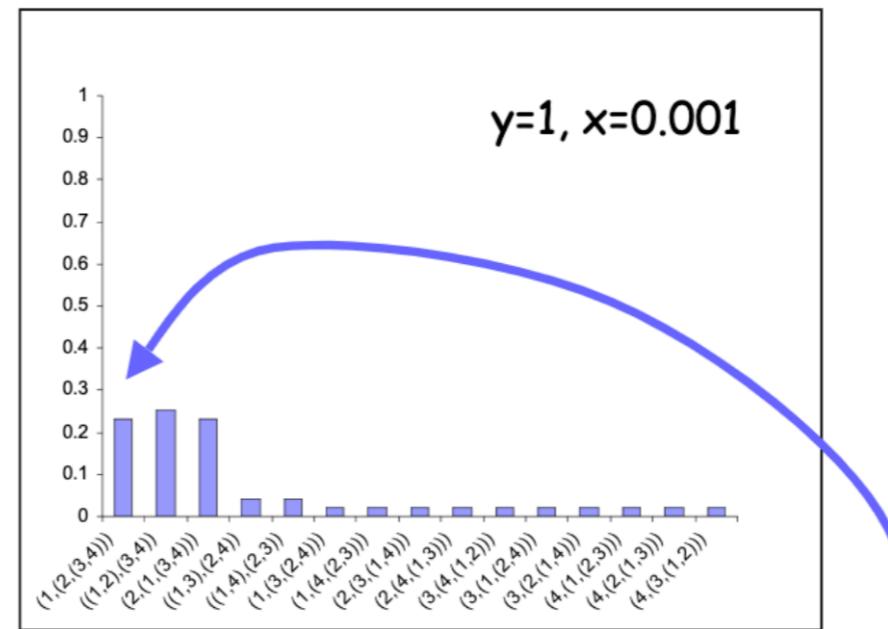
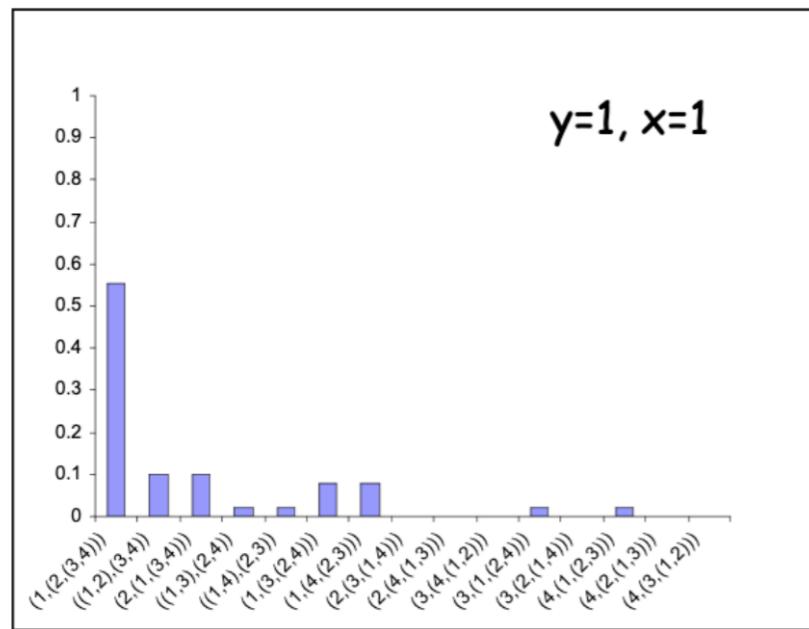
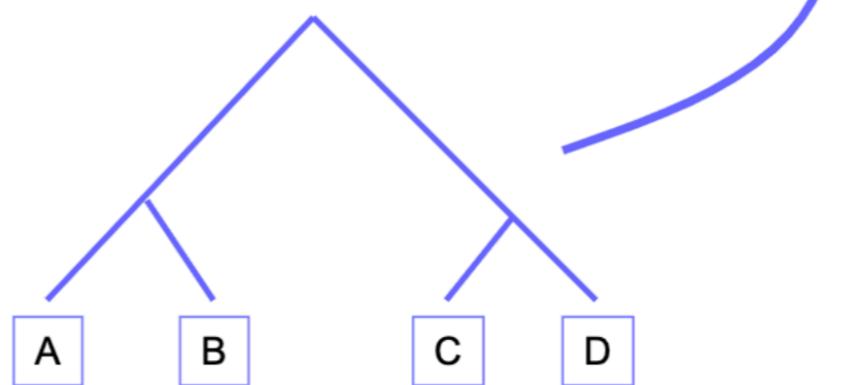
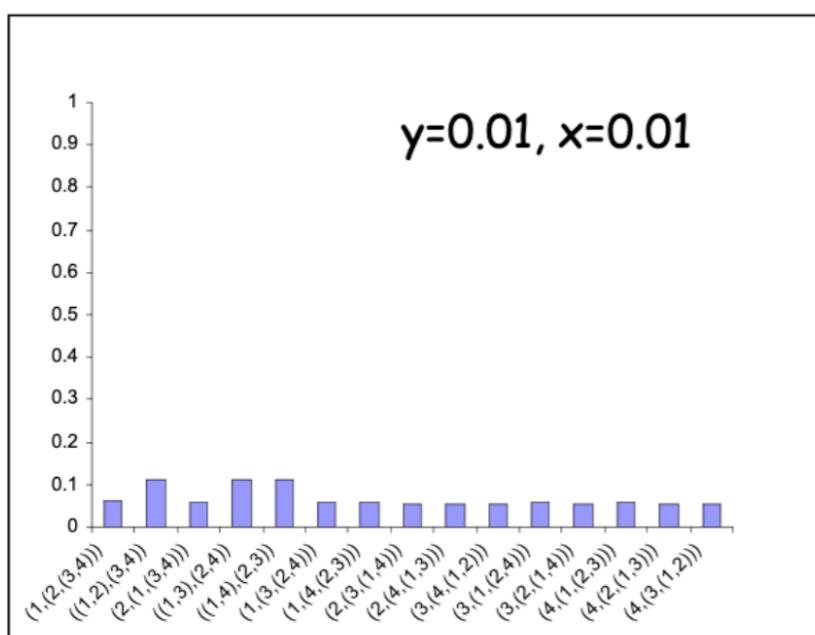
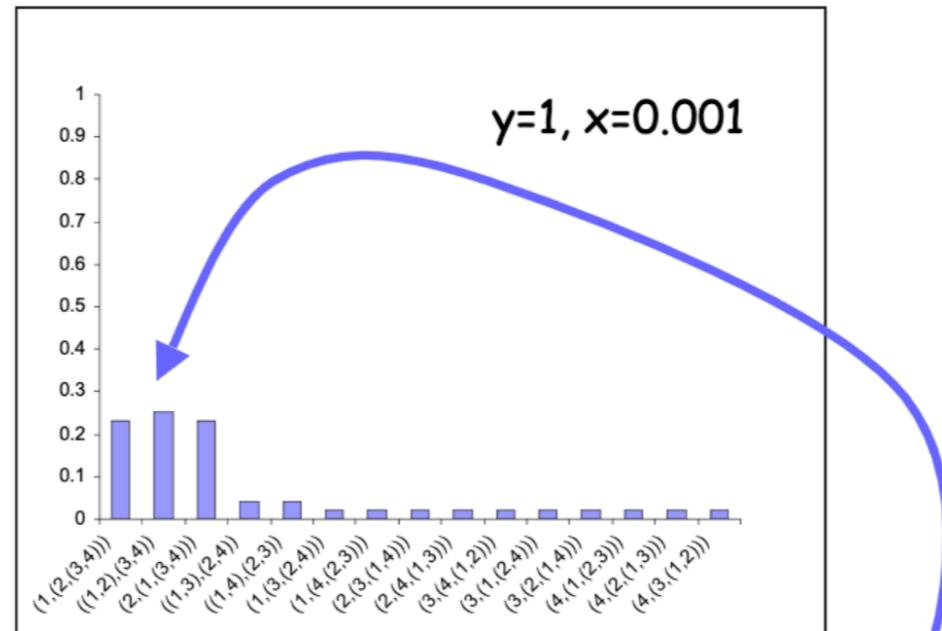
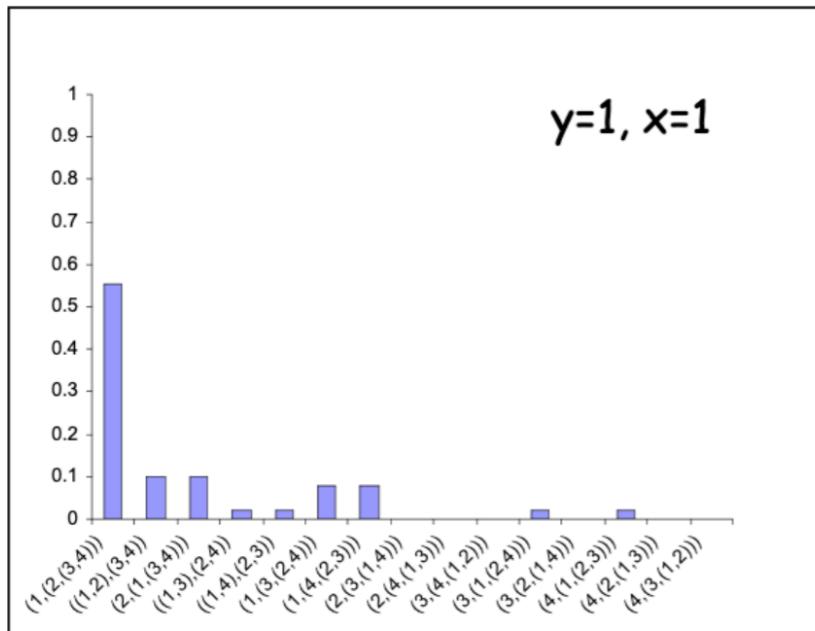


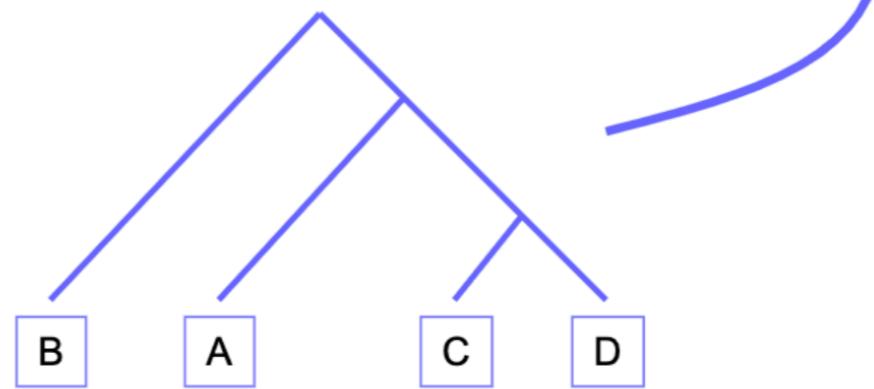
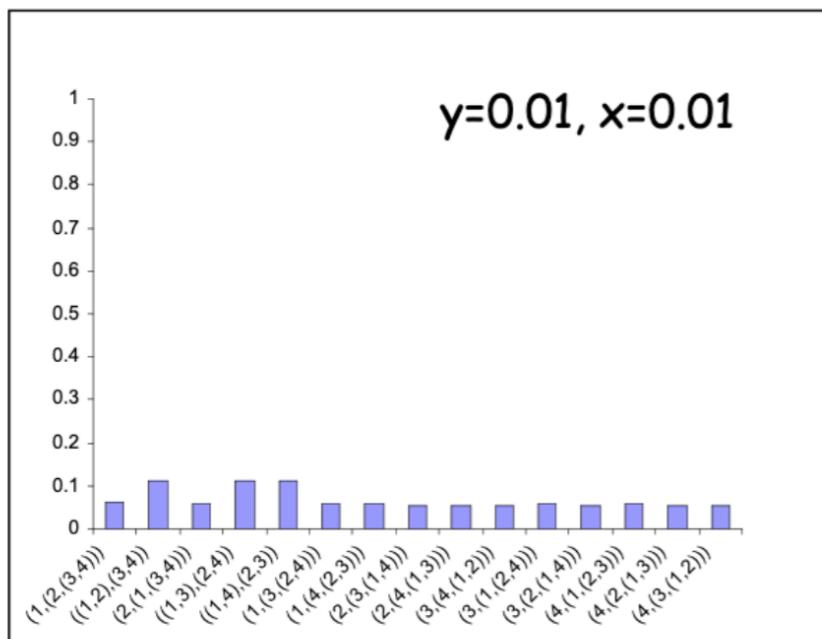
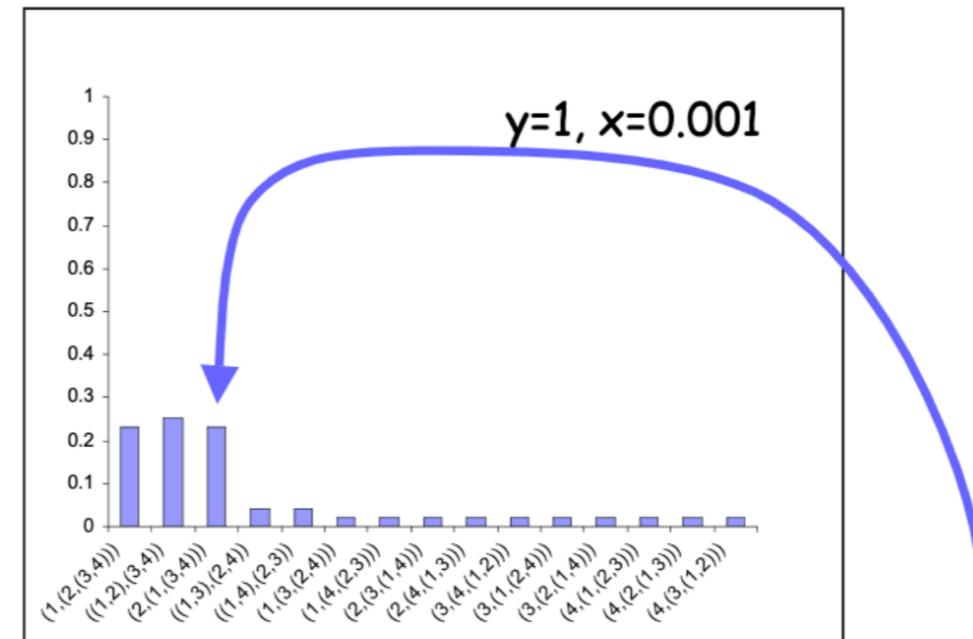
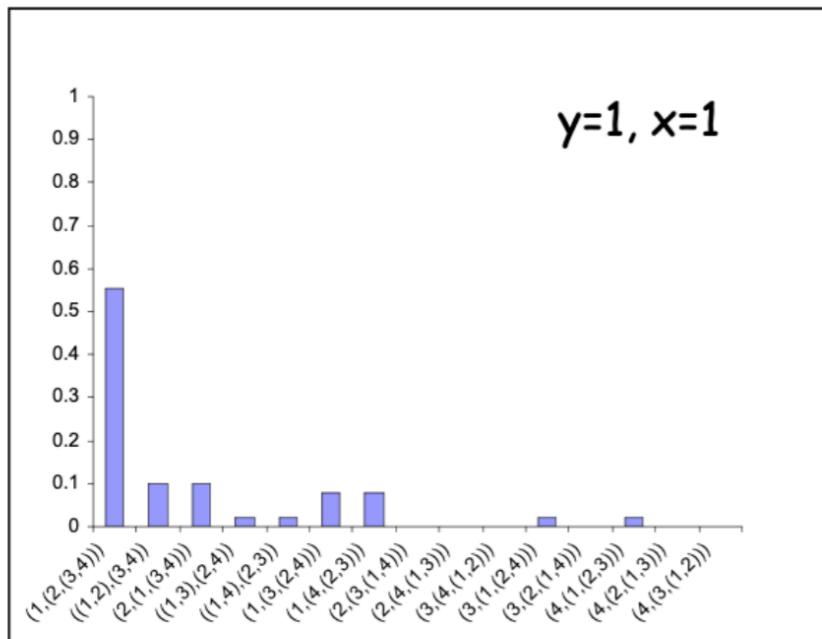
Figure 1 in [Linkem et al \(2016\)](#)



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)







- Concatenation: assumes all genes follow the same tree-like history
- Coalescent-based tree methods: accounts for ILS (and sometimes gene tree estimation error), but does not allow for other sources of gene tree discordance (like GDL or gene flow)
- Extensions to coalescent-based tree methods:
  - Coalescent-based network methods: accounts for ILS, gene tree estimation error and gene flow
  - Coalescent+GDL methods (Li et al, 2020)

# For next class:

- We will go over ASTRAL and BUCKy
- Each student is assigned to one software and has to read the paper for that software
- We will have a class discussion followed by installing and using the software