

# Molecular Evolution and Population Genetics

A few notes on population genetics  
of interest in phylogenetics

*Thomas Mailund*  
<mailund@stats.ox.ac.uk>

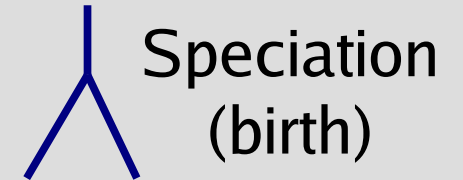
# Species Evolution

- A common view:
  - Death/birth process of species
    - Binary tree with inner nodes being speciation events

|

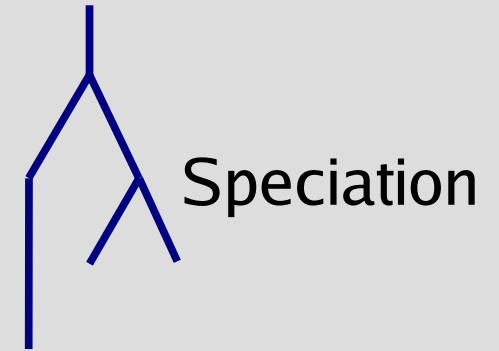
# Species Evolution

- A common view:
  - Death/birth process of species
    - Binary tree with inner nodes being speciation events



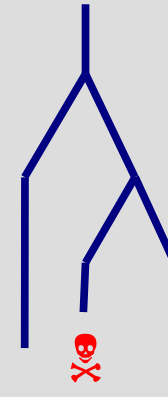
# Species Evolution

- A common view:
  - Death/birth process of species
    - Binary tree with inner nodes being speciation events



# Species Evolution

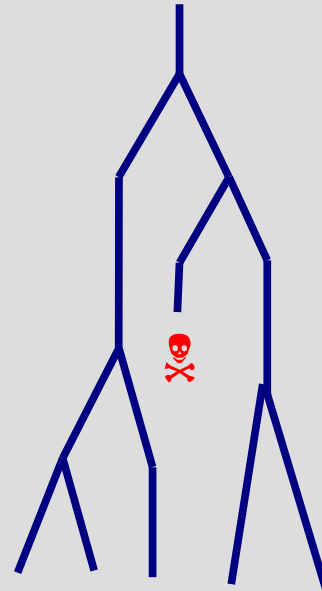
- A common view:
  - Death/birth process of species
    - Binary tree with inner nodes being speciation events



Extinction  
(death)

# Species Evolution

- A common view:
  - Death/birth process of species
    - Binary tree with inner nodes being speciation events



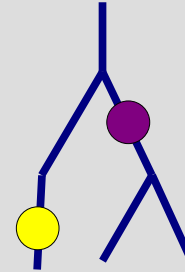
# Species Evolution

- A common view:
  - Death/birth process of species
    - Binary tree with inner nodes being speciation events
  - Mutations accumulate on edges



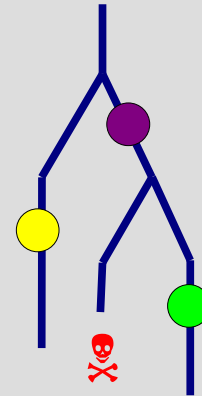
# Species Evolution

- A common view:
  - Death/birth process of species
    - Binary tree with inner nodes being speciation events
  - Mutations accumulate on edges



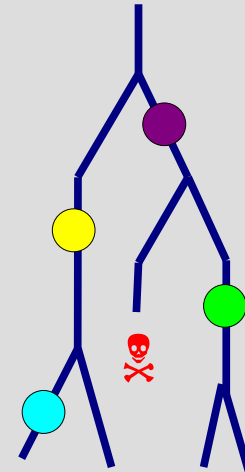
# Species Evolution

- A common view:
  - Death/birth process of species
    - Binary tree with inner nodes being speciation events
  - Mutations accumulate on edges



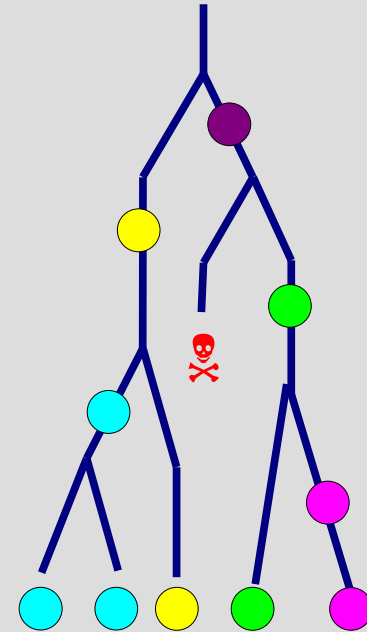
# Species Evolution

- A common view:
  - Death/birth process of species
    - Binary tree with inner nodes being speciation events
  - Mutations accumulate on edges



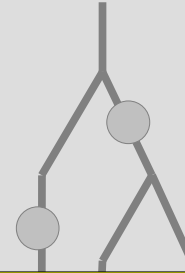
# Species Evolution

- A common view:
  - Death/birth process of species
    - Binary tree with inner nodes being speciation events
  - Mutations accumulate on edges



# Species Evolution

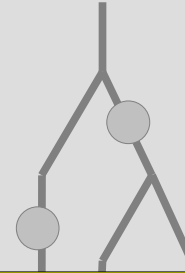
- A common view:
  - Death/birth process of species
  - Binary tree with inner



**There is no such process!**

# Species Evolution

- A common view:
  - Death/birth process of species
  - Binary tree with inner

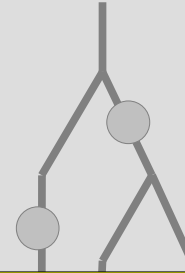


**There is no such process!**

I dare you to observe a 'species'...

# Species Evolution

- A common view:
  - Death/birth process of species
  - Binary tree with inner

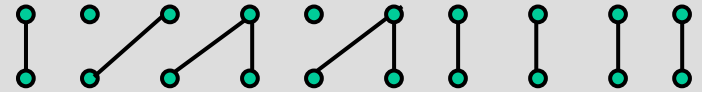


**There is no such process!**

An abstraction over a more complicated process...

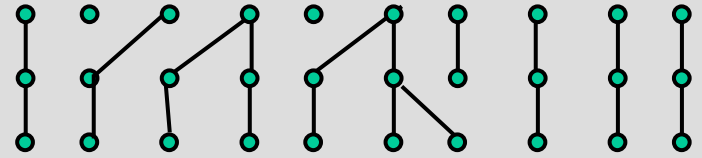
# Population Evolution

- Wright-Fisher model
  - Discrete, non-overlapping generations
  - Constant population size
  - Each individual in one generation is a random copy of an individual from the previous generation



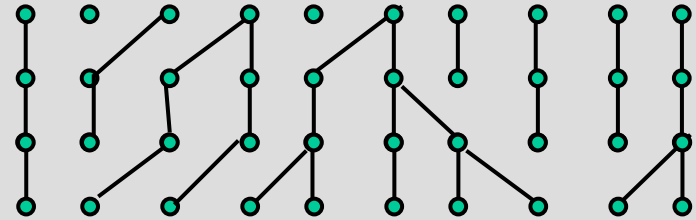
# Population Evolution

- Wright-Fisher model
  - Discrete, non-overlapping generations
  - Constant population size
  - Each individual in one generation is a random copy of an individual from the previous generation



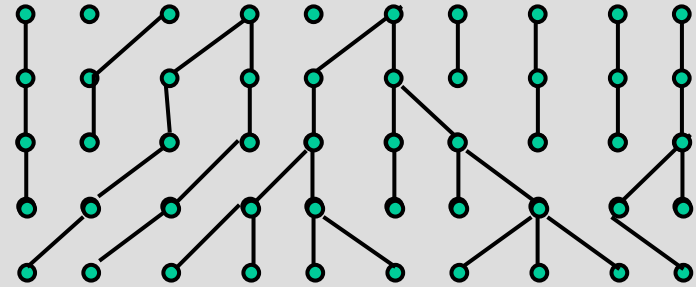
# Population Evolution

- Wright-Fisher model
  - Discrete, non-overlapping generations
  - Constant population size
  - Each individual in one generation is a random copy of an individual from the previous generation



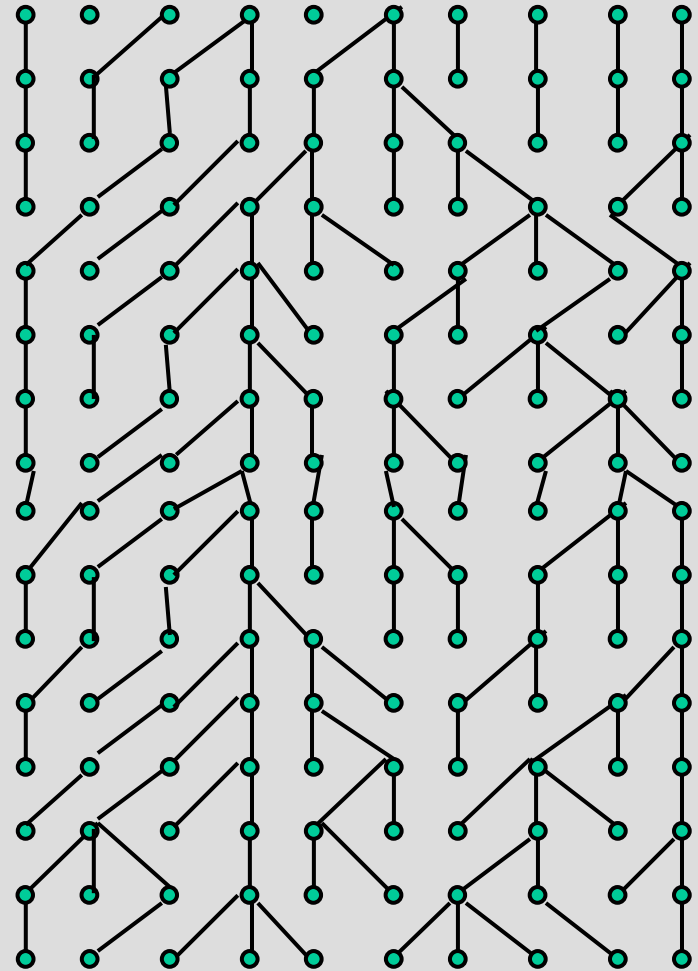
# Population Evolution

- Wright-Fisher model
  - Discrete, non-overlapping generations
  - Constant population size
  - Each individual in one generation is a random copy of an individual from the previous generation

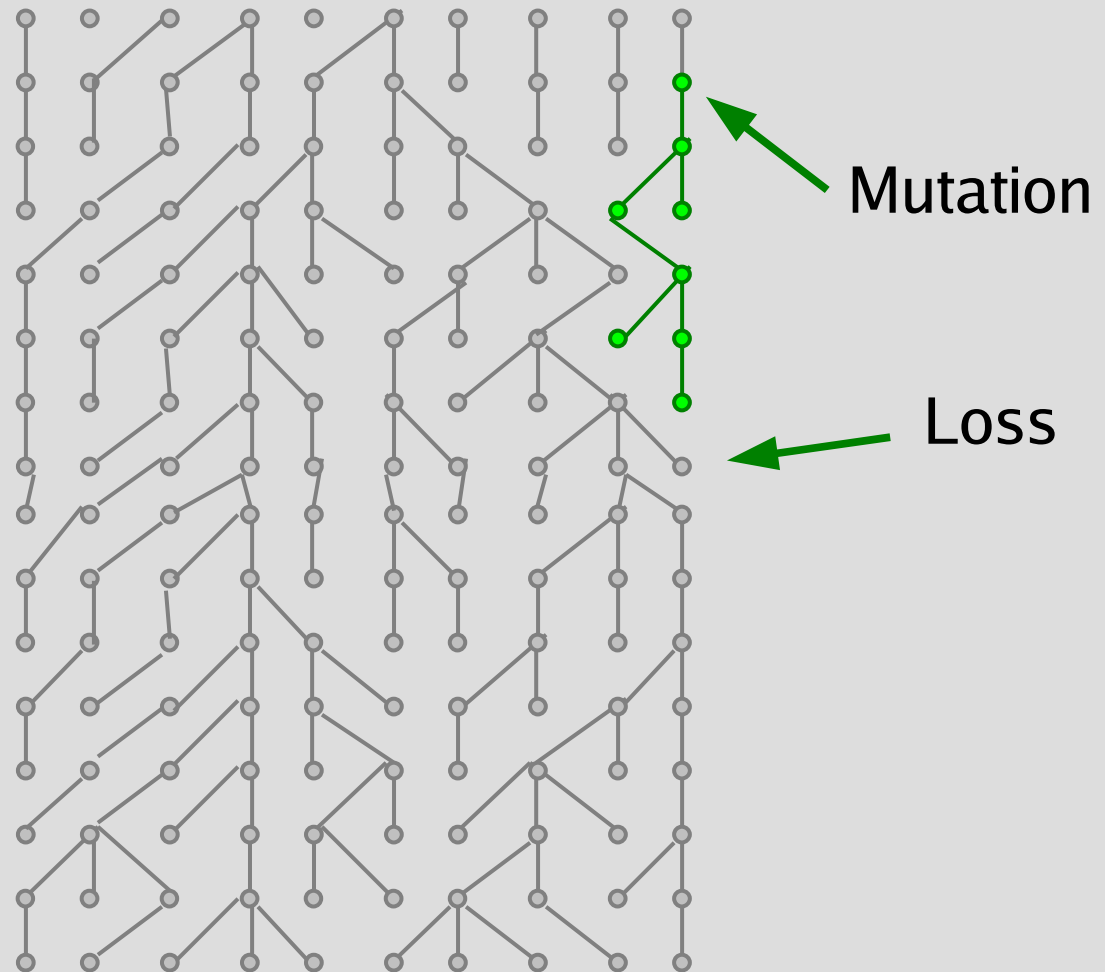


# Population Evolution

- Wright-Fisher model
  - Discrete, non-overlapping generations
  - Constant population size
  - Each individual in one generation is a random copy of an individual from the previous generation

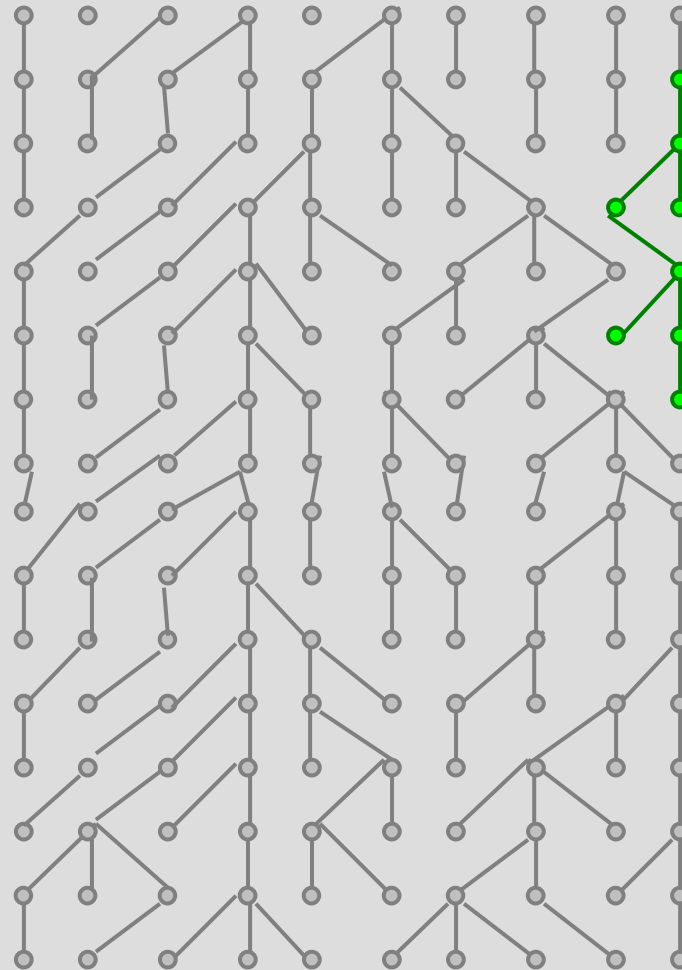


# Population Evolution

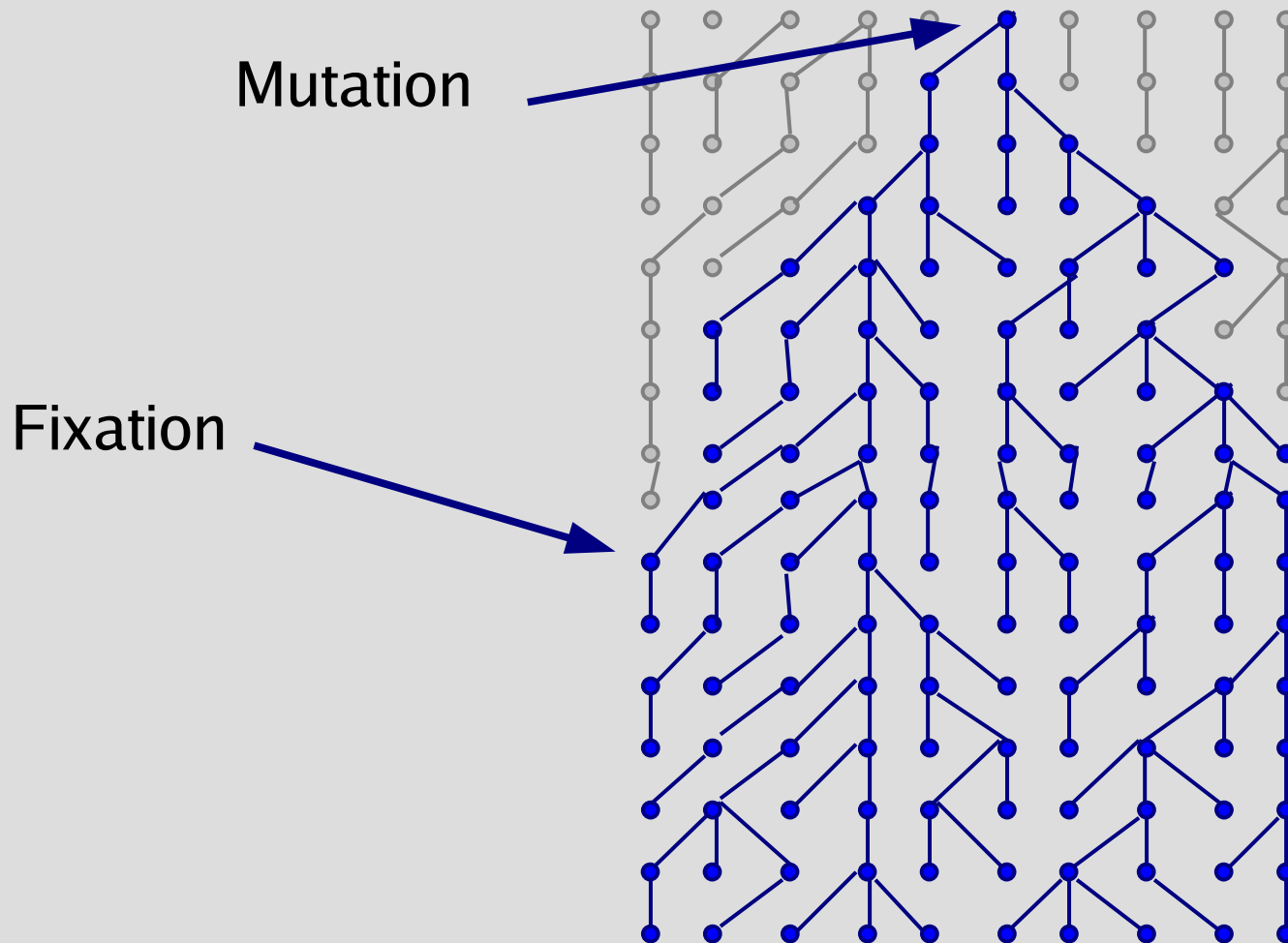


# Population Evolution

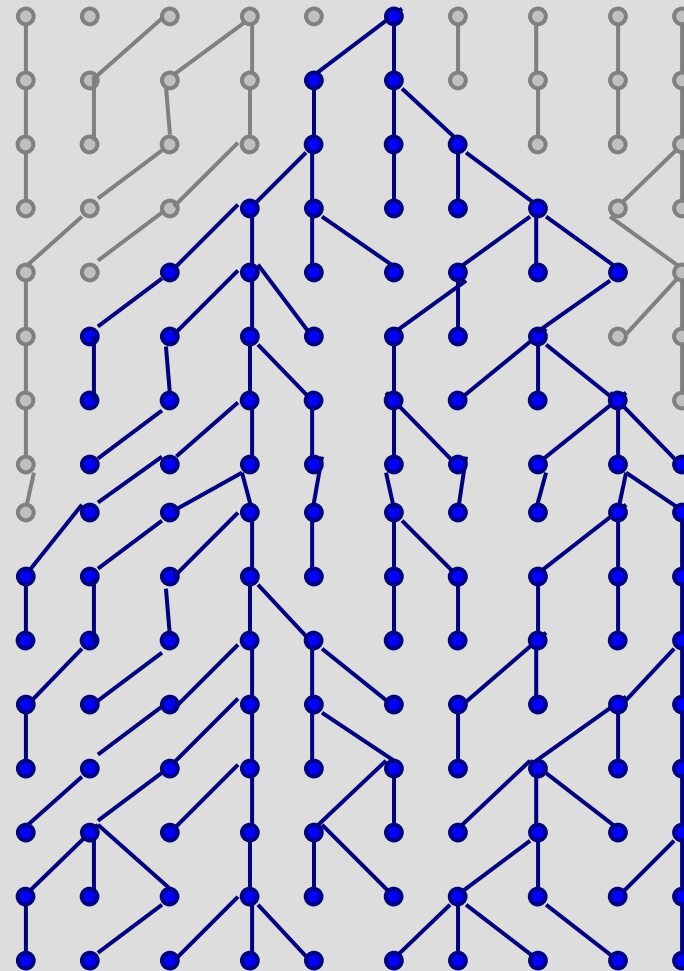
Species  
view



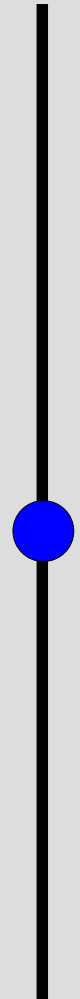
# Population Evolution



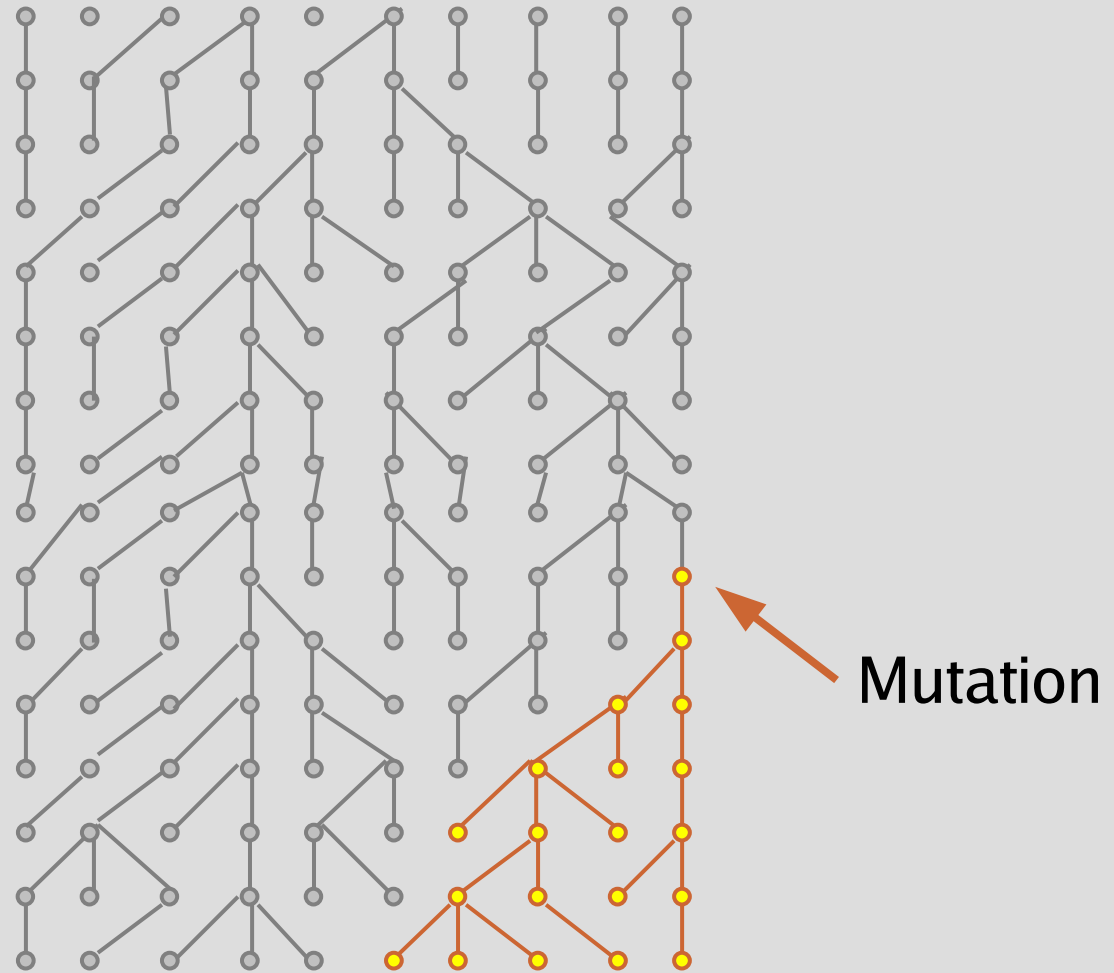
# Population Evolution



Species  
view

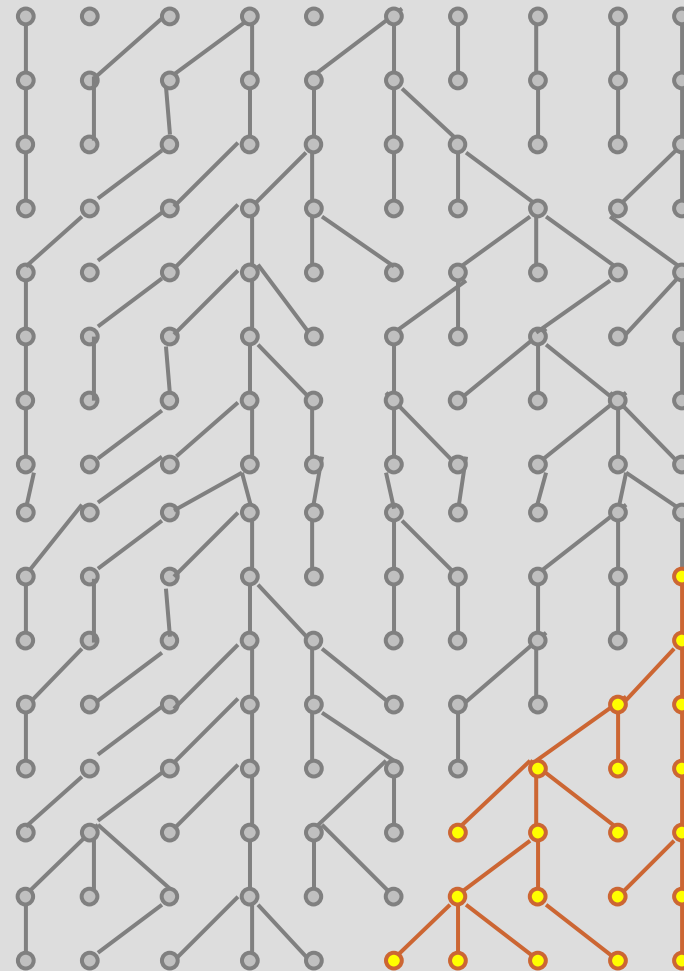


# Population Evolution

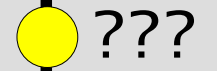


Polymorphism...

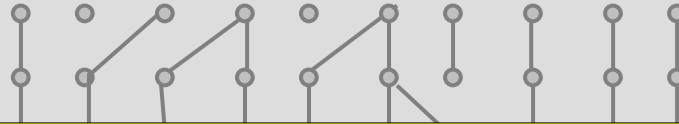
# Population Evolution



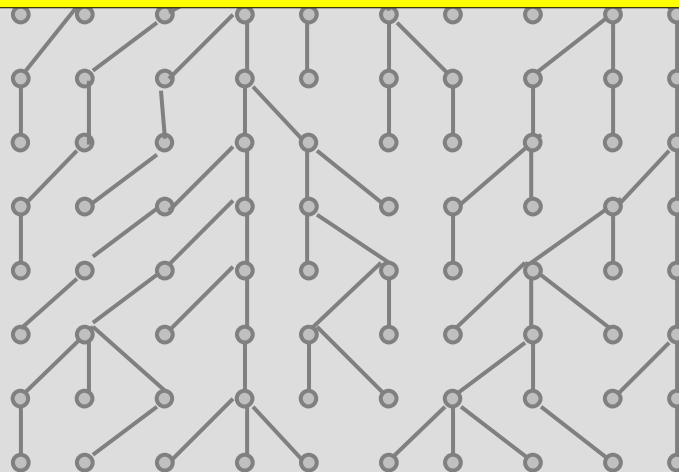
Species  
view



# Population Evolution



**There is no such process!**

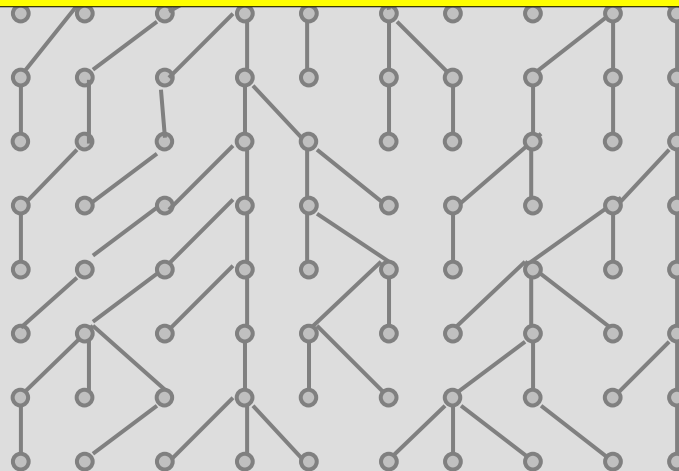


I dare you to observe a 'generation'...

# Population Evolution



**There is no such process!**



This abstraction is good enough for now...

# Unsupported Claims

- Species are populations on a larger scale
- Speciation events are population splits plus time

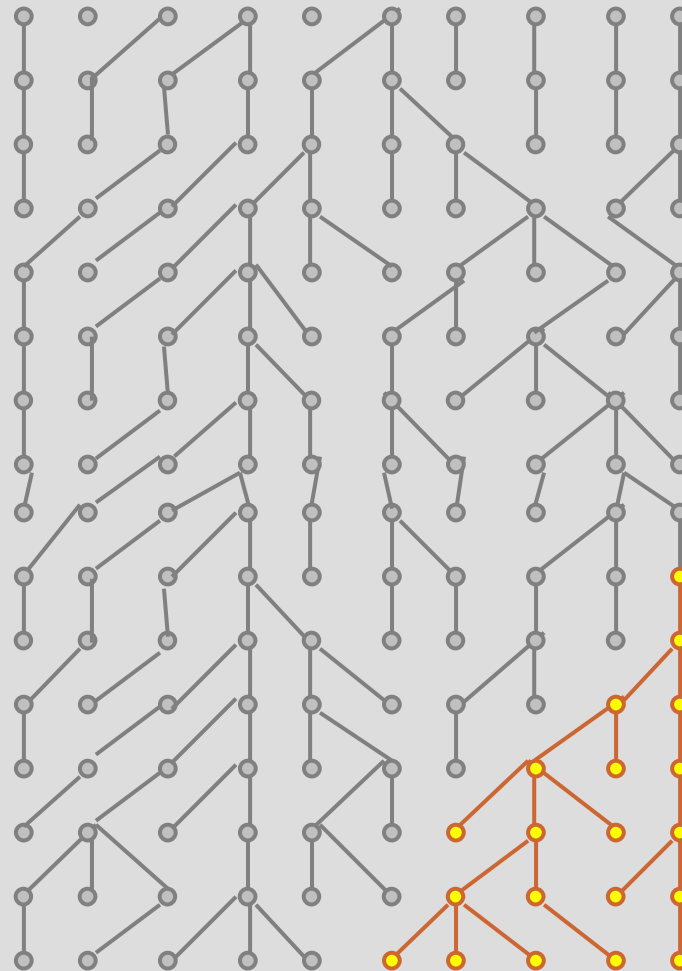
# Unsupported Claims

- Species are populations on a larger scale
- Speciation events are population-splits plus time

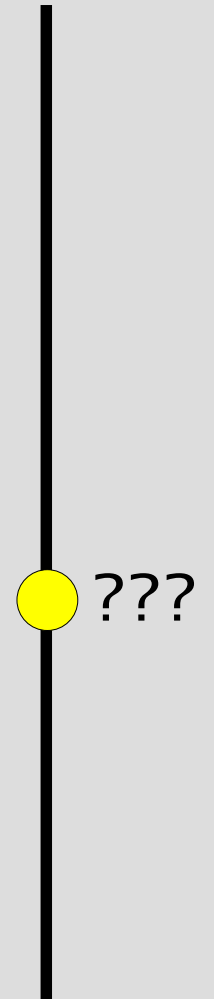
(no one really knows how speciation occurs...)

# Population Evolution

What 'type'  
is the species  
gene here?

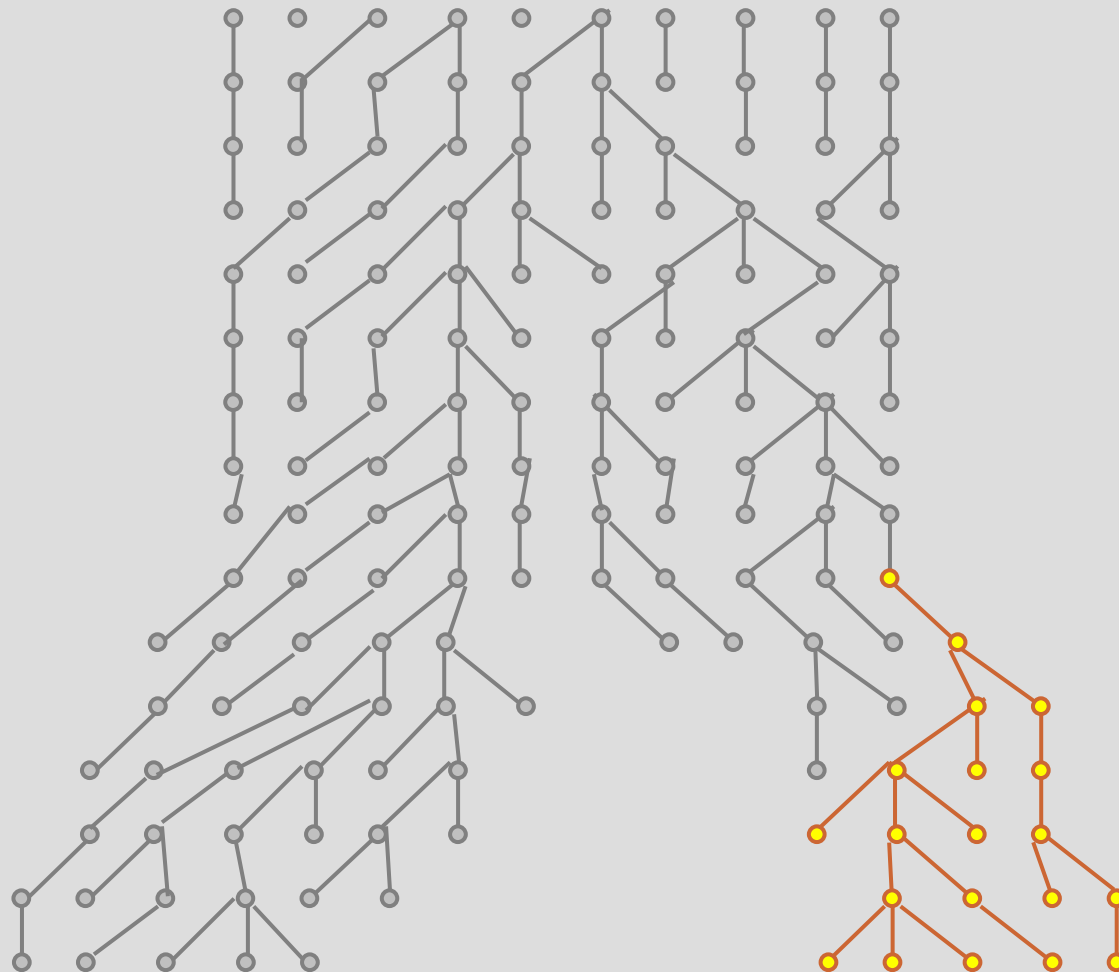


Species  
view

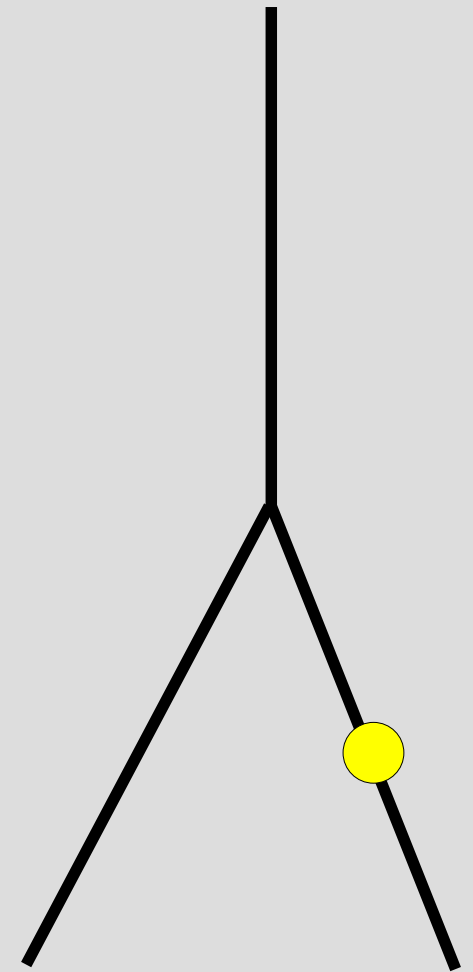


# Population Evolution

Individuals  
view

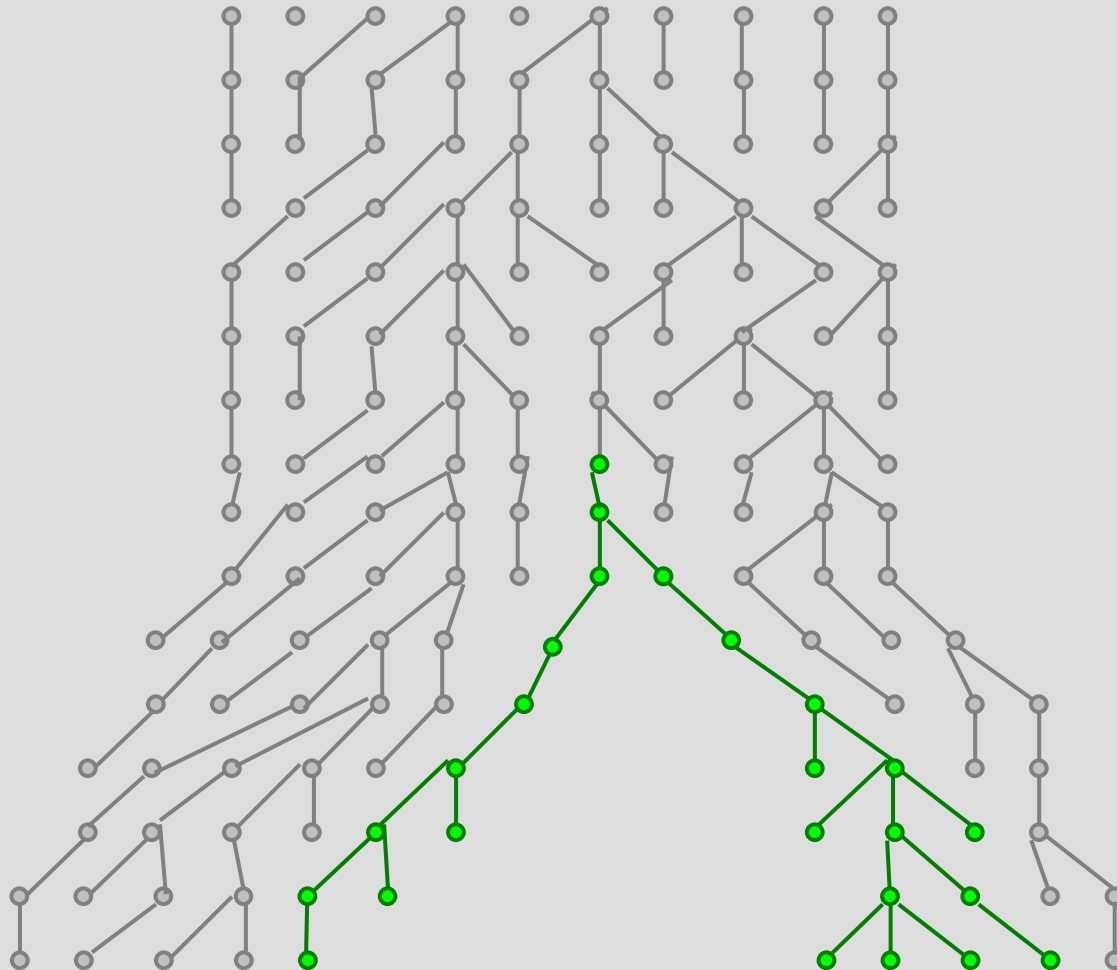


Species  
view

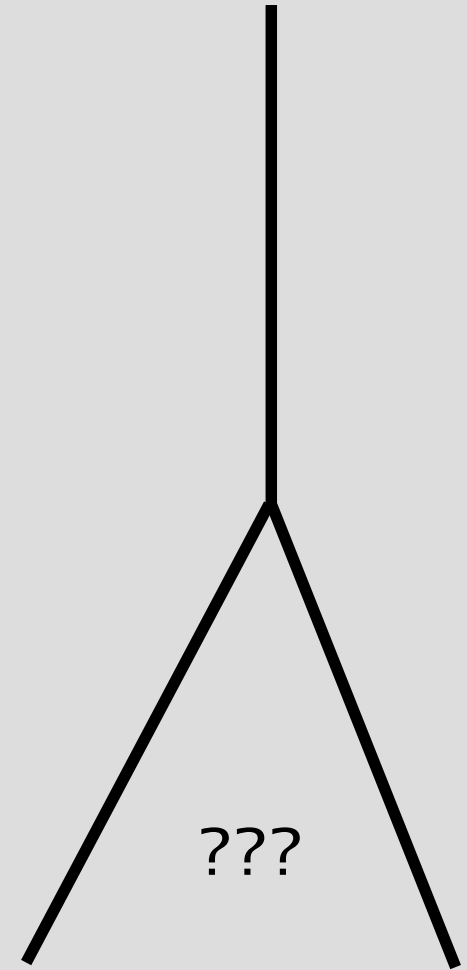


# Population Evolution

Individuals  
view

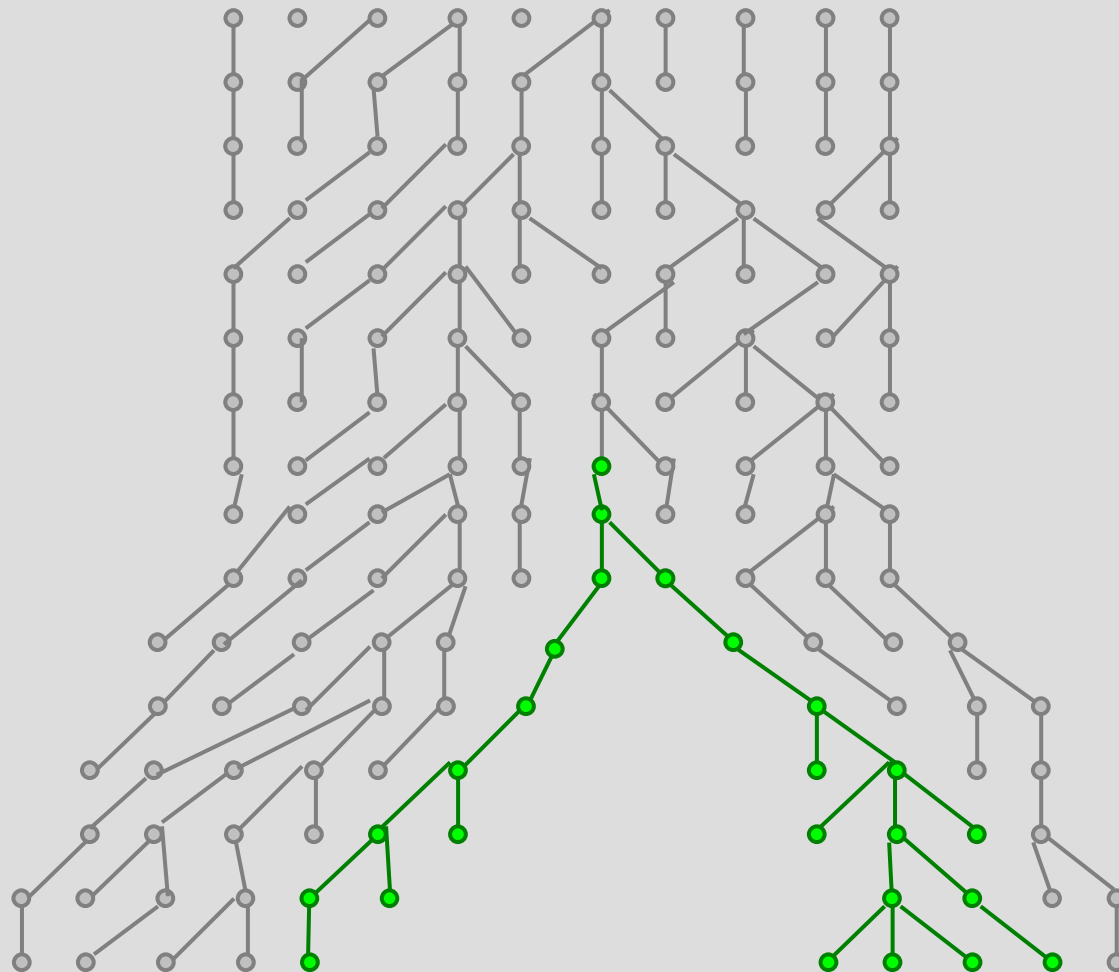


Species  
view

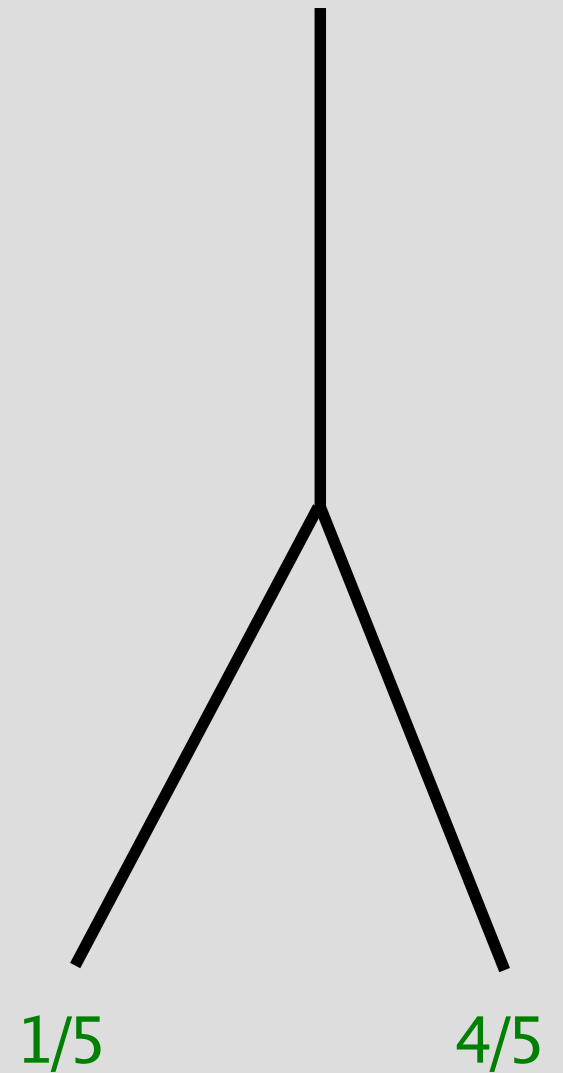


# Population Evolution

Individuals  
view

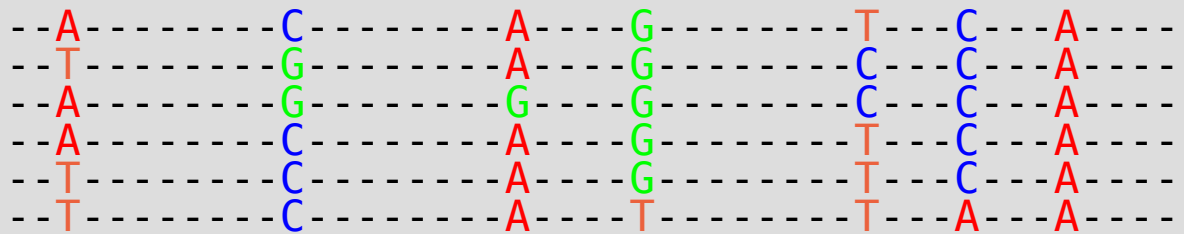


Species  
view

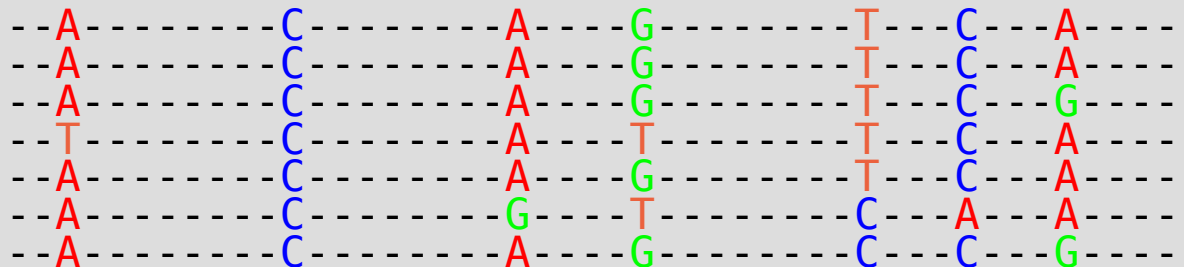


# Dealing with polymorphism

Two populations/species with polymorphic sites

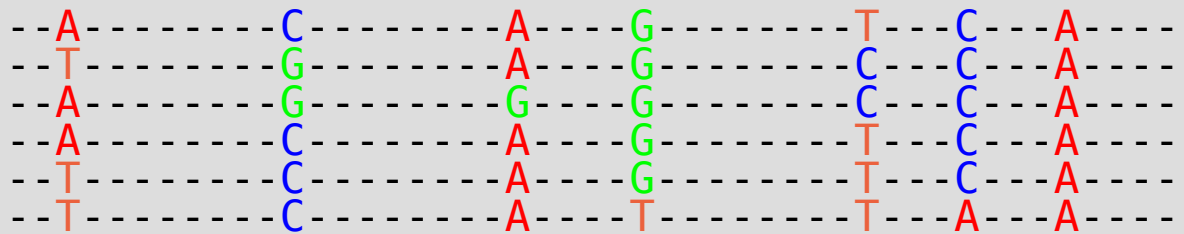


Alleles polymorphic in both population

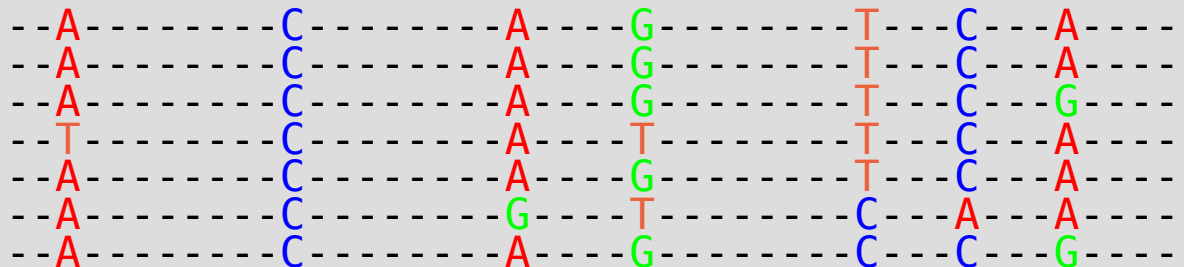


# Dealing with polymorphism

Two populations/species with polymorphic sites



Alleles fixed in one population



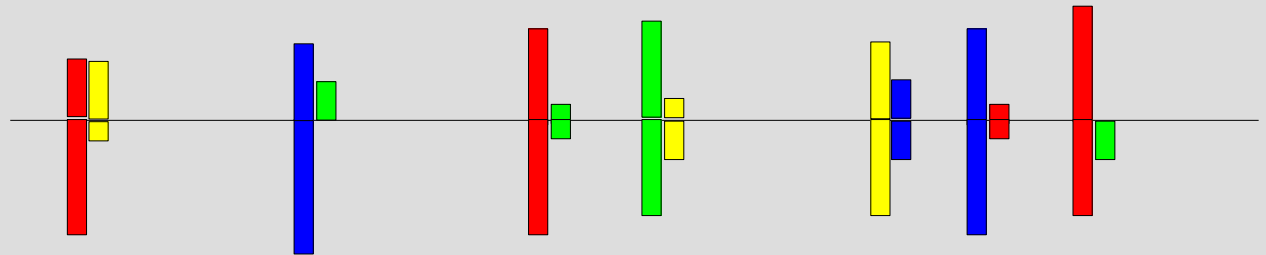
# Dealing with polymorphism

The difference is only in frequencies...



A	C	A	G	T	C	A
T	G	A	G	C	C	A
A	G	G	G	C	C	A
A	C	A	G	T	C	A
T	C	A	G	T	C	A
T	C	A	T	T	A	A

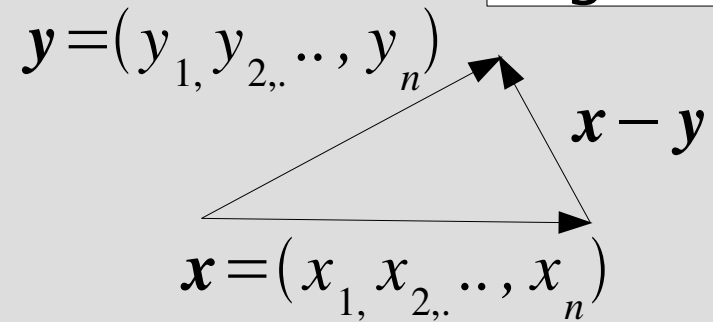
Different allele-frequencies  
for different genes



A	C	A	G	T	C	A
A	C	A	G	T	C	A
A	C	A	G	T	C	G
T	C	A	T	T	C	A
A	C	A	G	T	C	A
A	C	G	T	C	A	A
A	C	A	G	C	C	G

# Distances Based on Frequencies

- Consider vectors of allele-frequencies
  - Here n alleles at a single gene



$$\text{dist}(x, y) = |x - y| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Take values in range  $[0, \sqrt{n}]$

and can be normalized as

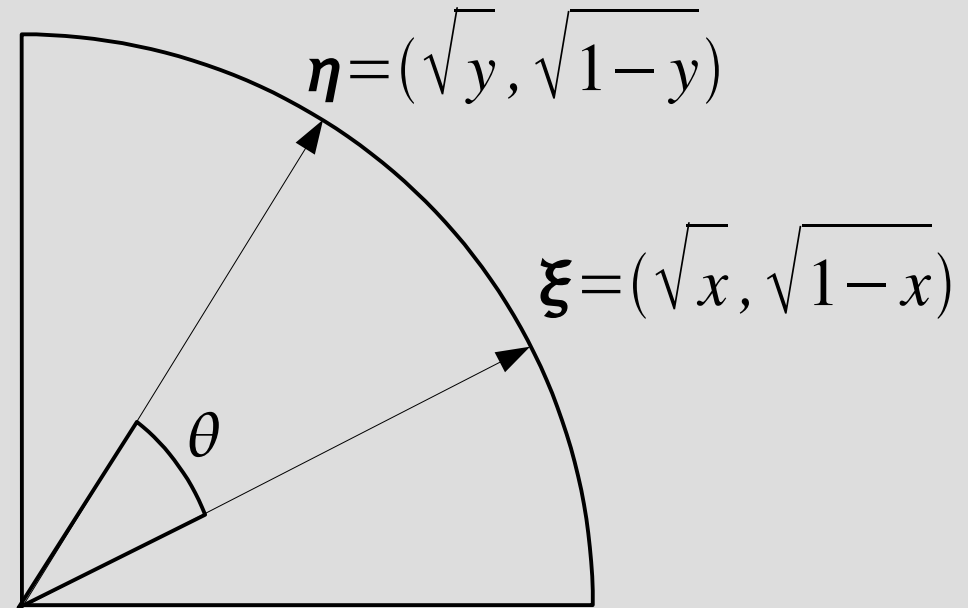
$$\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2}$$

- Population distance just average over genes

# Distances Based on Frequencies

## Bhattacharyya's Distance

- Consider single (two-allele) gene
  - Generalizes to hyperspheres for more alleles
- Distance as angle



$$\theta = \sqrt{\arccos(\eta \cdot \xi)} = \sqrt{\arccos(\sqrt{xy} + \sqrt{(1-x)(1-y)})}$$

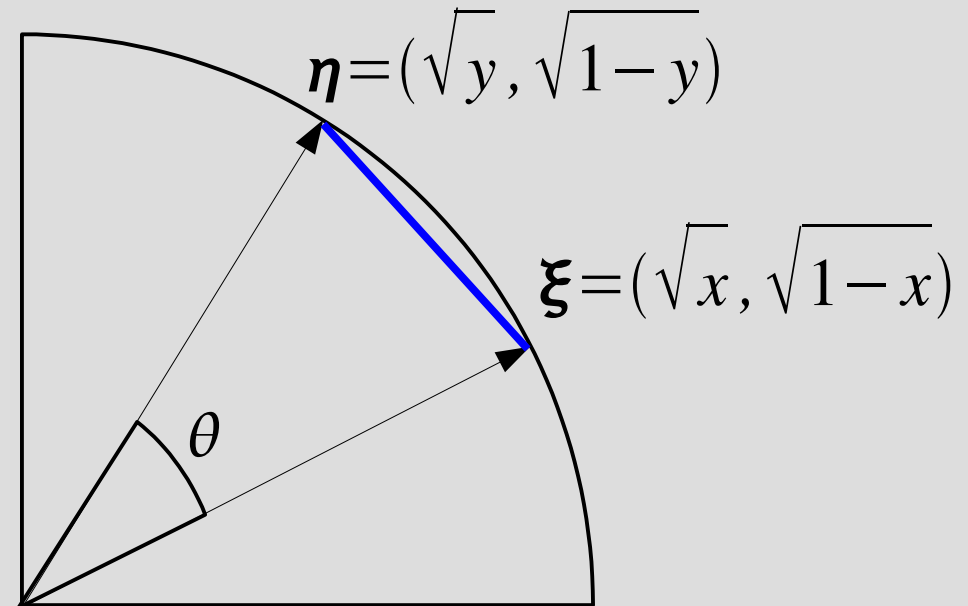
Take values in range  $\left[0, \frac{\pi}{2}\right]$  and can be normalized by  $\frac{2}{\pi}$

- Population distance just average over genes

# Distances Based on Frequencies

- Consider single (two-allele) gene
  - Generalizes to hyperspheres for more alleles
- Distance as cord-length

Cavalli-Sforza & Edwards' Distance



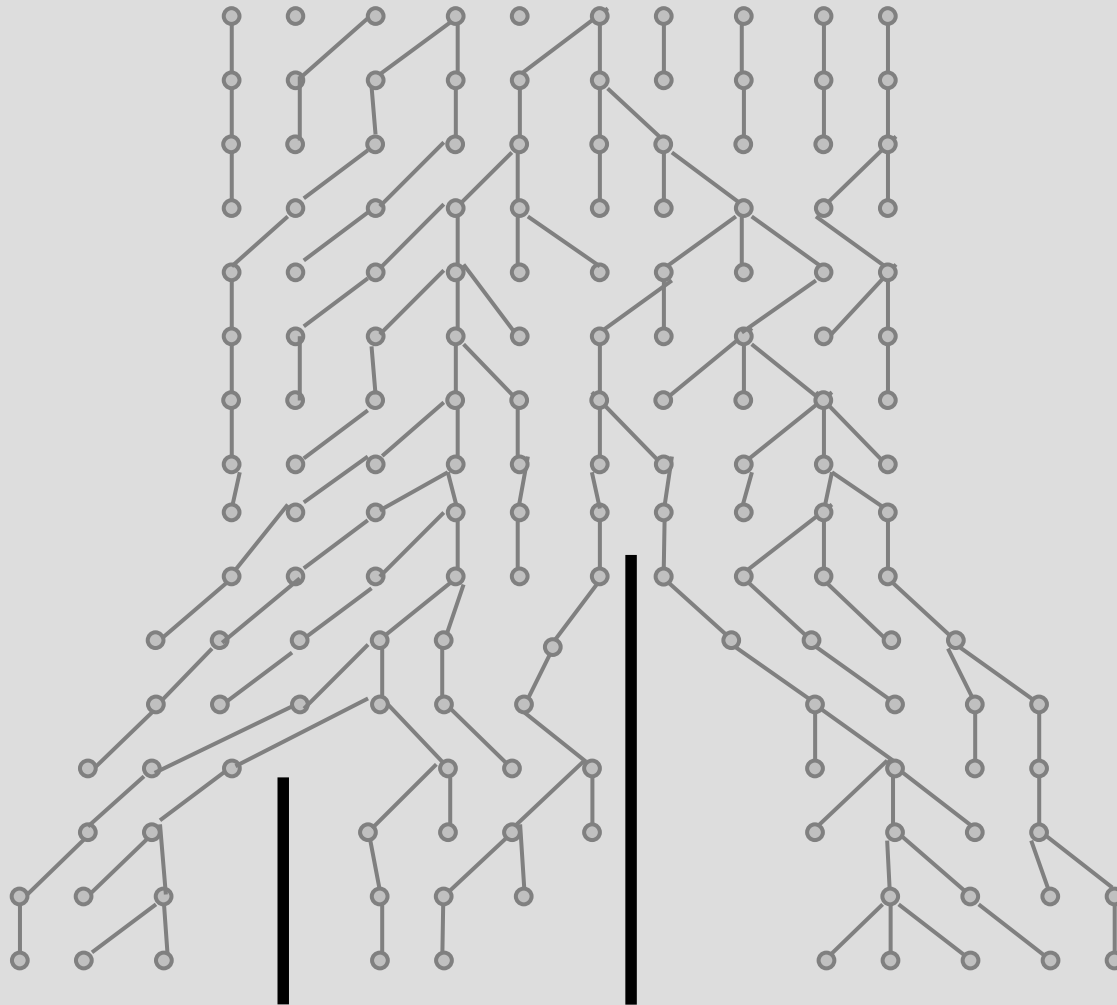
$$l = |\eta - \xi| = \sqrt{2(1 - \cos \theta)}$$

Take values in range  $[0, \sqrt{2}]$  and can be normalized by  $\frac{1}{\sqrt{2}}$

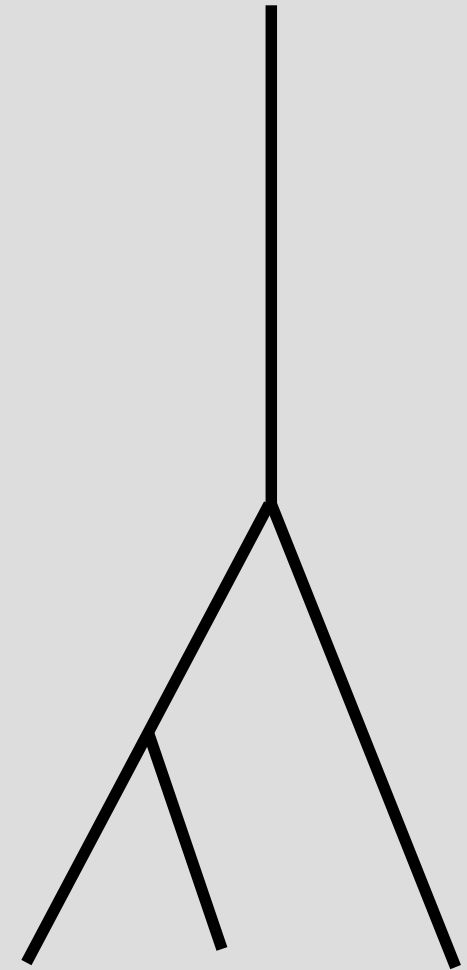
- Population distance just average over genes

# Population vs Individual Trees

Individuals  
view

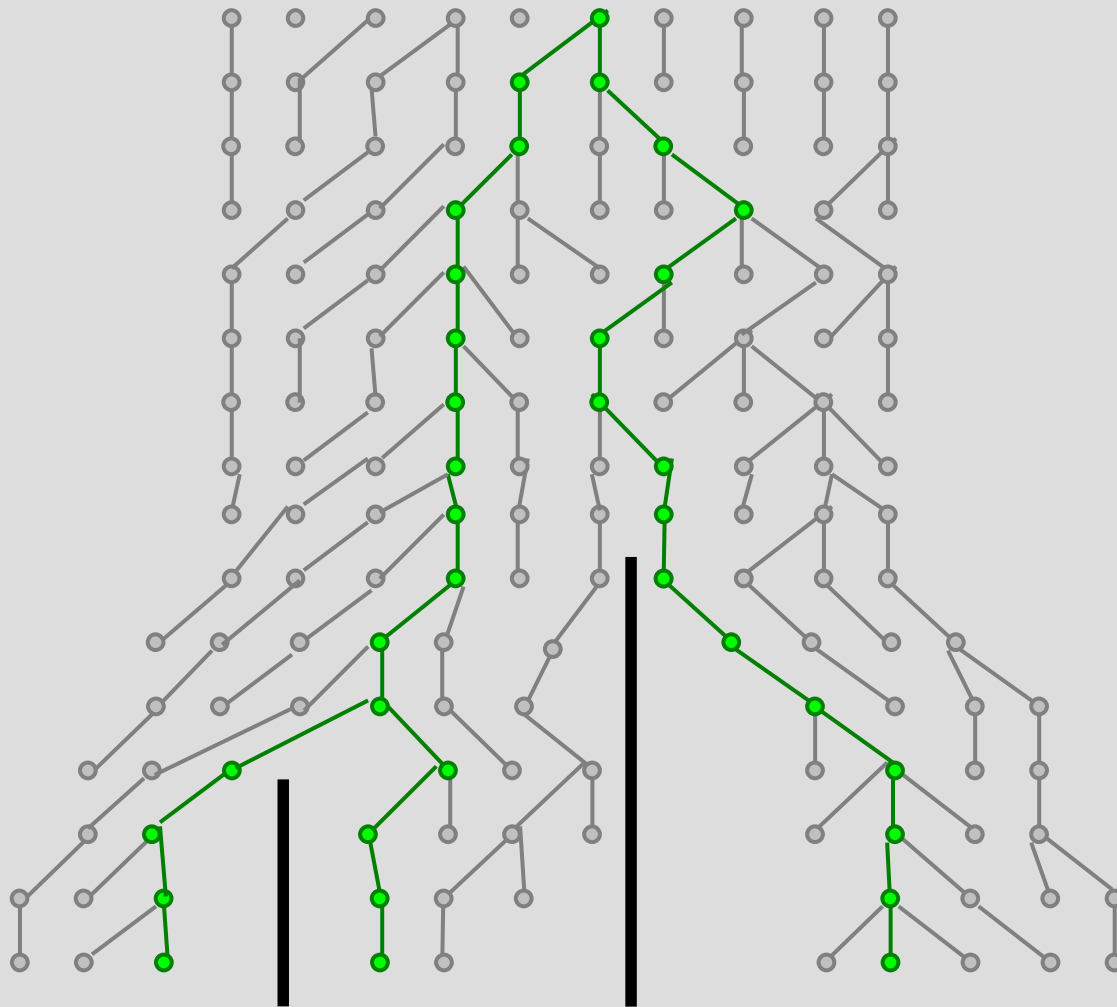


Population  
view

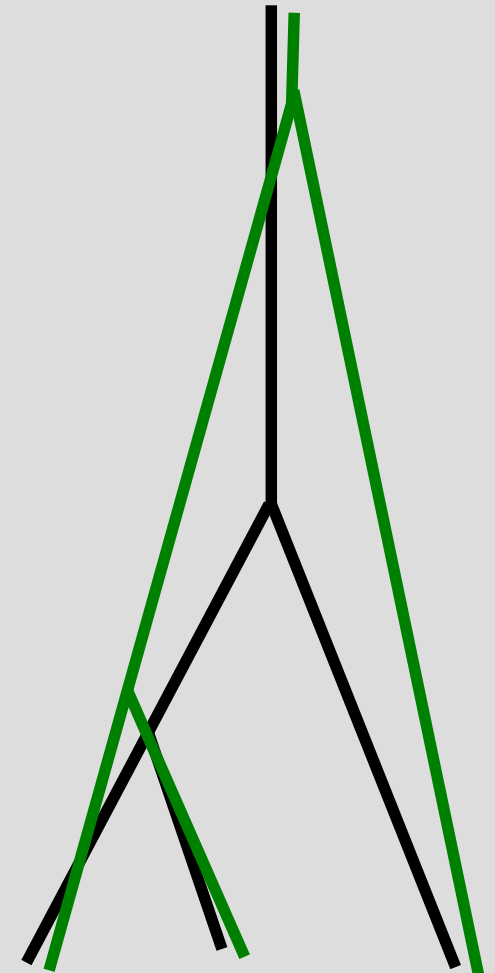


# Population vs Individual Trees

Individuals  
view

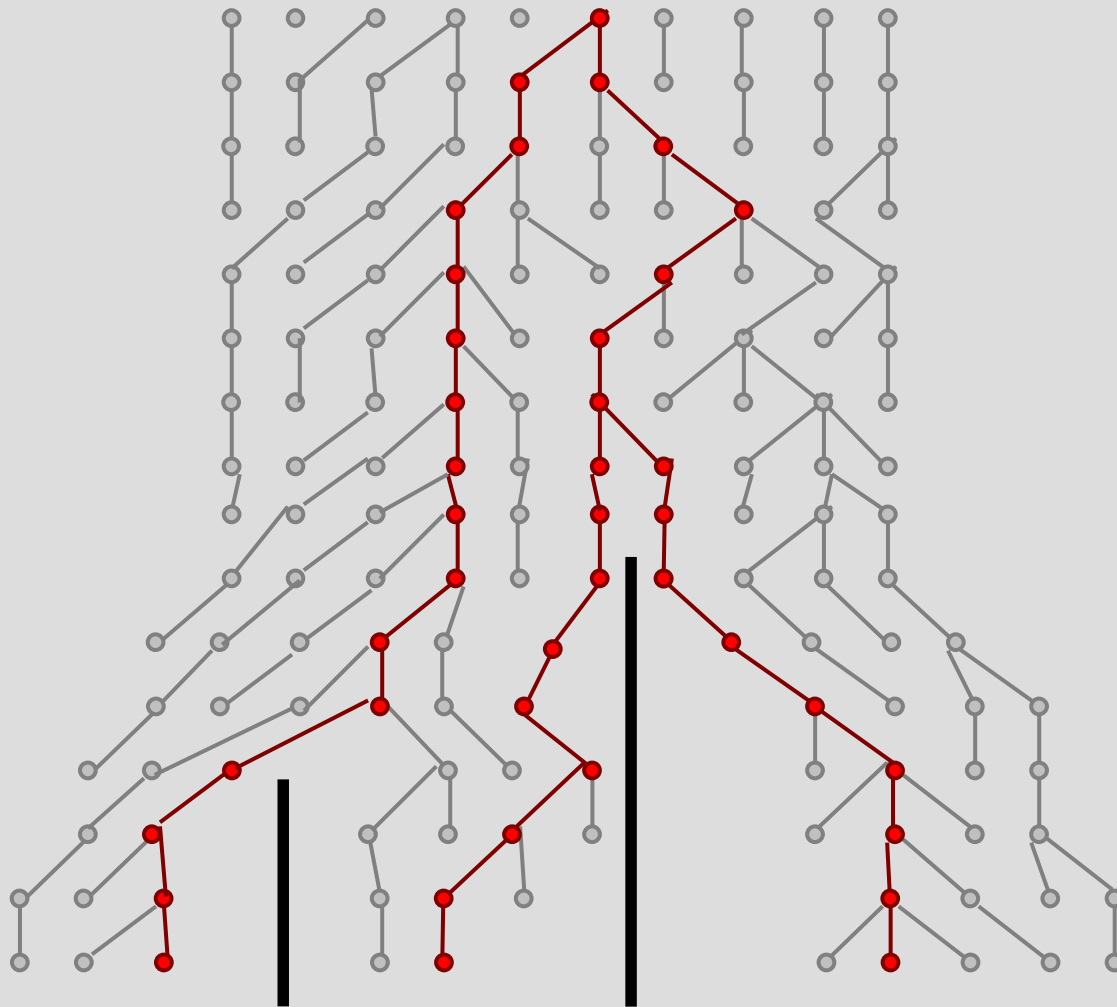


Population  
view

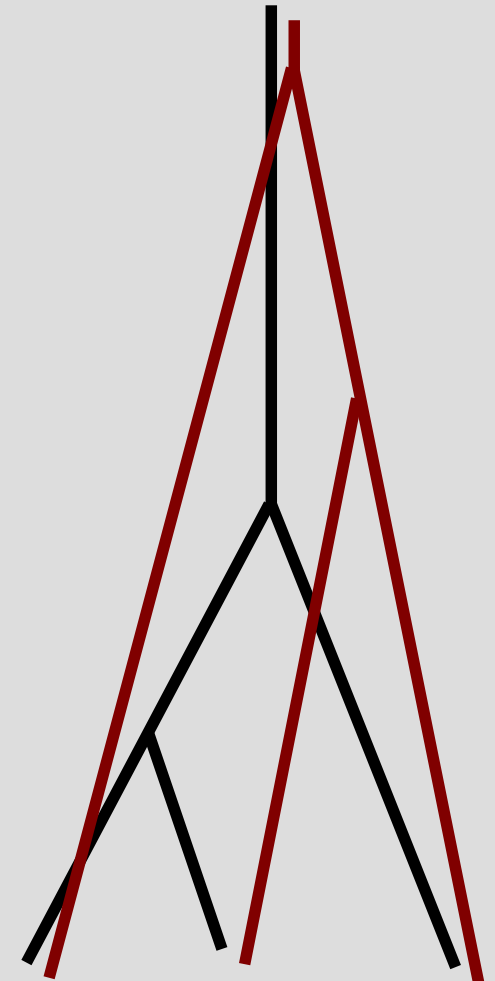


# Population vs Individual Trees

Individuals  
view



Population  
view



# Population vs Individual Trees

- Only a problem if
  - Splitting time is very fast
  - Or population sizes are very large

# Species vs Individual Trees

- Only a problem if
  - Splitting time is very fast
  - Or population sizes are very large
- Not really happening at the species level
  - At least hard to imagine *how*...

# Species vs *Gene* Trees

- Only a problem if
  - Splitting time is very fast
  - Or population sizes are very large
- Not really happening at the species level
  - At least hard to imagine *how...*
- But it *does* happen between species when looking at *some* parts of the genome!
  - Recombination splits the genome in segments
  - Some segments can have different (gene) trees than the species tree

# Recombination Recap

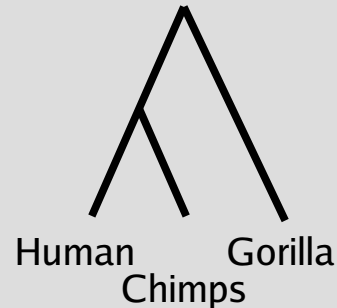
- Recombination events split the individual's genomes in segments
- Some coalesce sooner than others

# New Observation

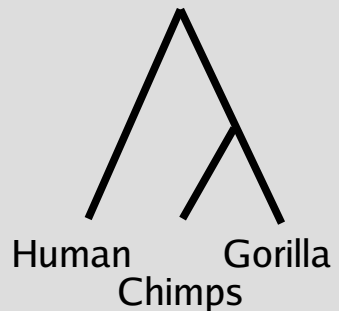
- Recombination events split the individual's genomes in segments
- Some coalesce sooner than others
- **Some might not coalesce until *after* (back in time) the species have merged**

# Example: Great Apes

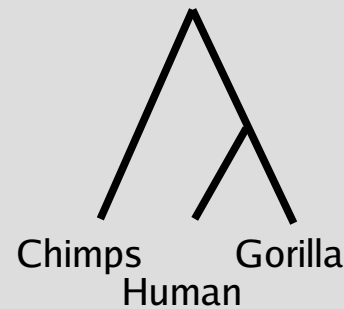
- From fossils and molecular data we know that the species tree is:



- But a large fraction of genes (10-20%) support:



or

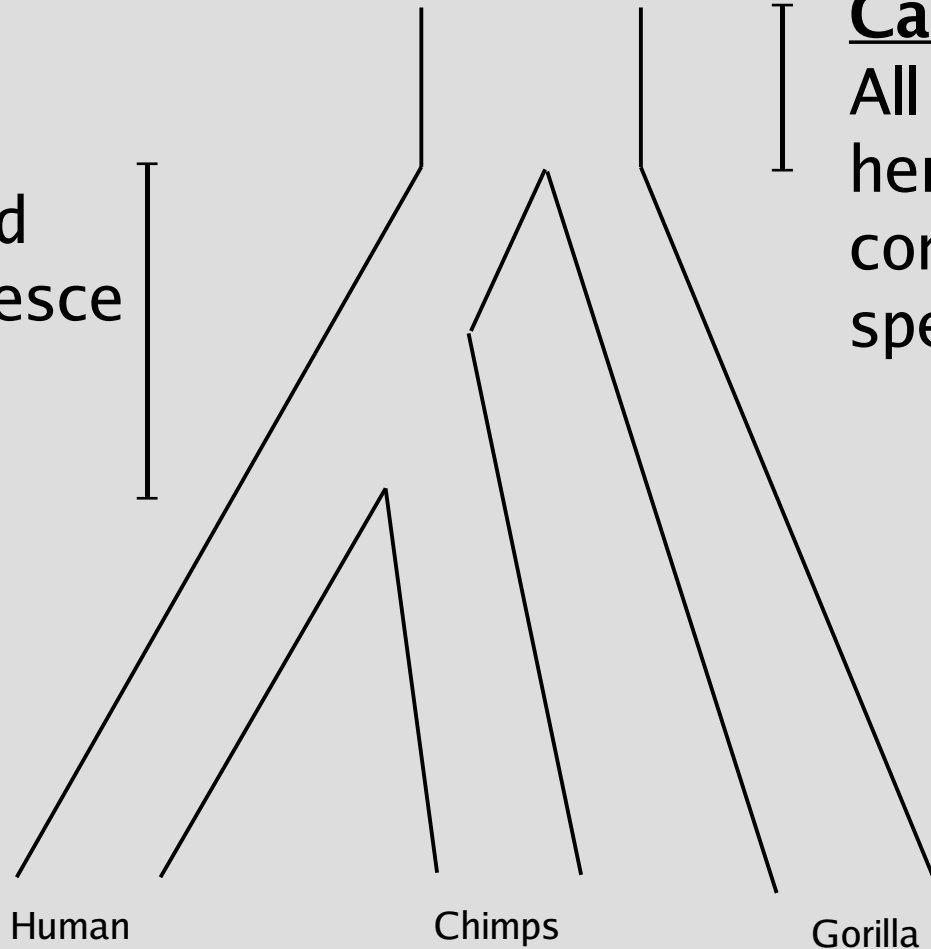




# Example: Great Apes

## Case A:

Only Human and Chimpanzee can coalesce here. Always consistent with species tree.

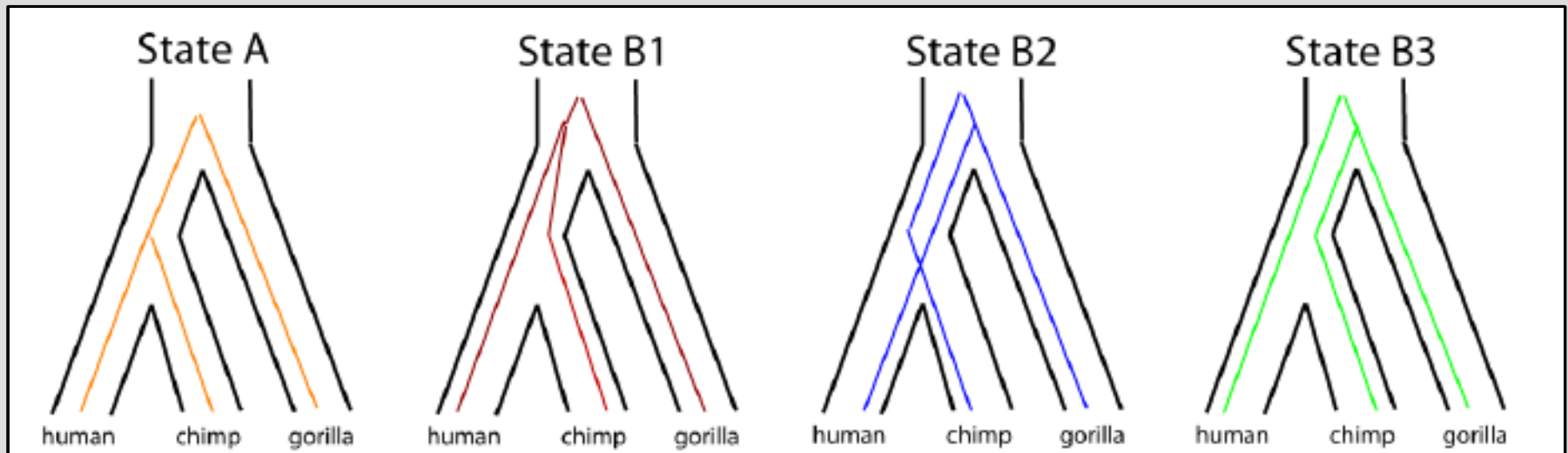


## Case B:

All species can coalesce here. Only a third consistent with species tree.

How often both H and C reach B depends on the H+C population size and the time between the H+C+B split and the H+C split.

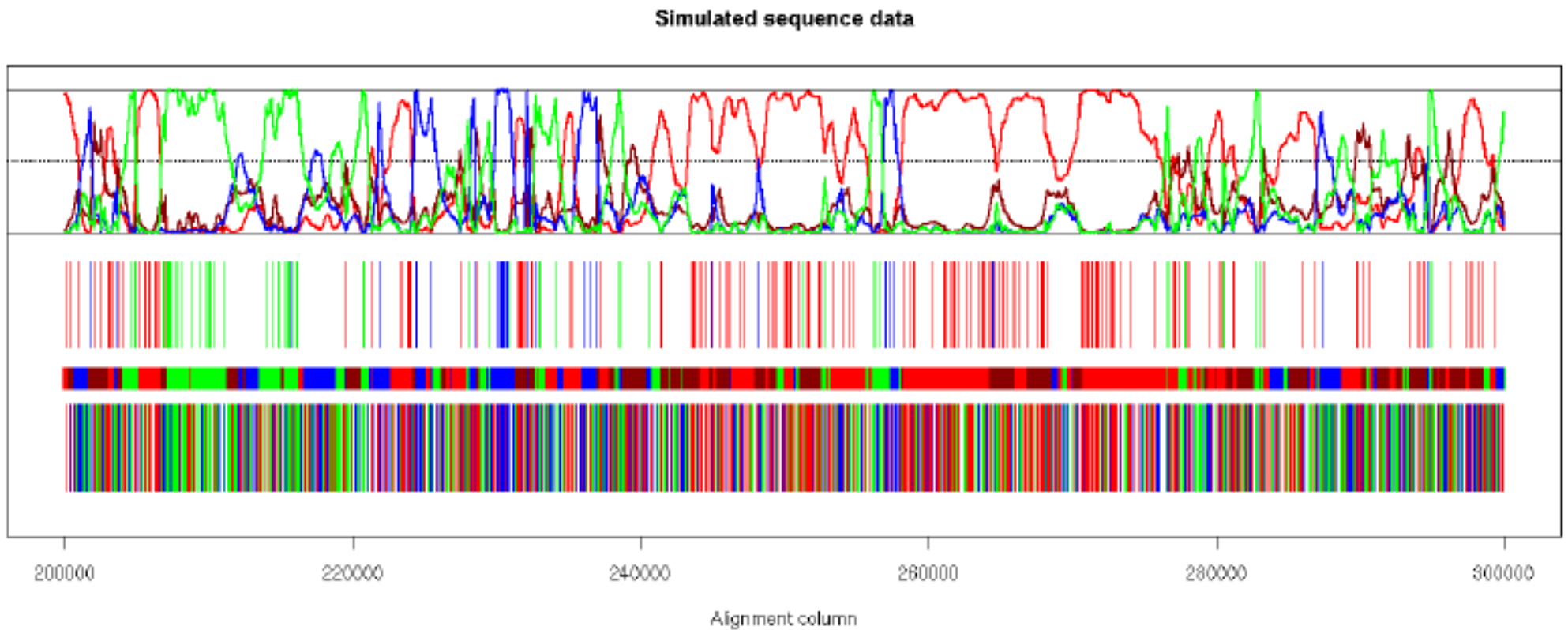
# Example: Great Apes





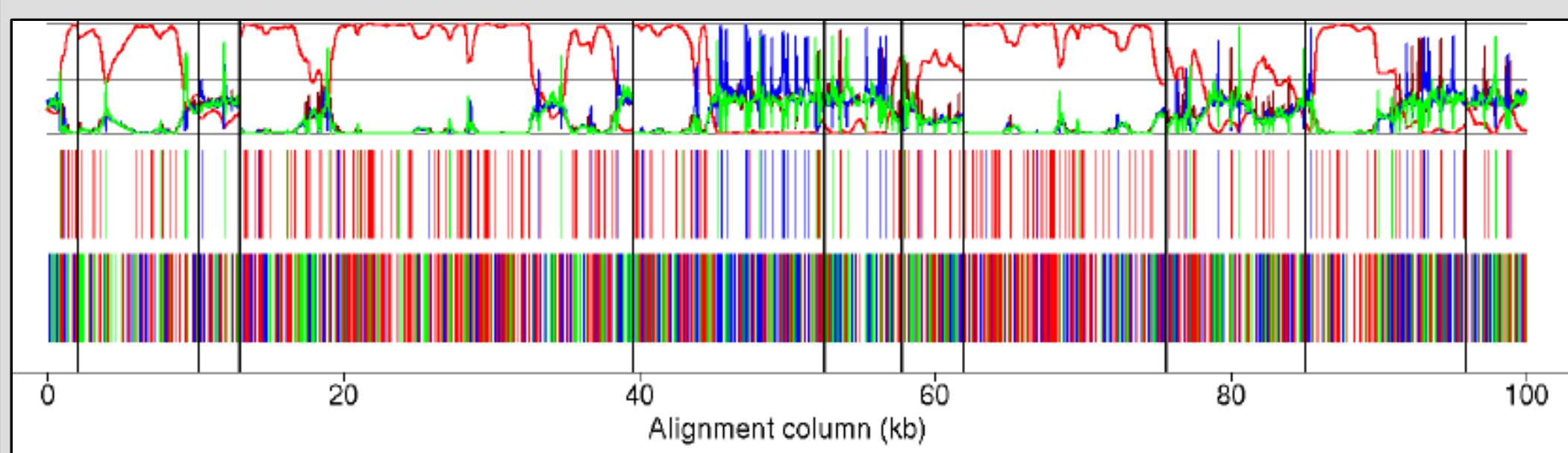
# coalHMM

- Simulated data:



# coalHMM

- 100 kbp real data:



# coalHMM

- From fraction of segments in state B and length of segments before state changes we can infer:
  - Species population sizes (both current species and ancestral species over the time span considered)
  - Speciation times

# Take-Home Message

- When dealing with populations, or closely related species:
  - Polymorphism must be taken into account
  - Recombination must be taken into account