

Introduction to Coalescent theory

Matthieu Foll

22.11.2011

Population Genomics course, Helsinki

Introduction to Coalescent Theory

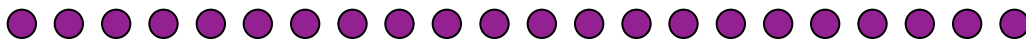
Classical population genetics theory tries to predict what will happen in the future of a given population. It is a **prospective** approach.

Coalescent theory is a **retrospective approach** to population genetics based on the genealogy of gene copies.

It uses mathematics for describing the characteristics of the joining of lineages **back in time** to a **common ancestor**.

This lineage joining is referred to as **coalescence**.

Present

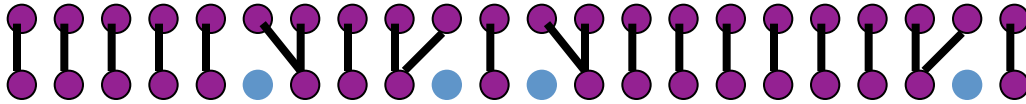


22 individuals

Time



Present



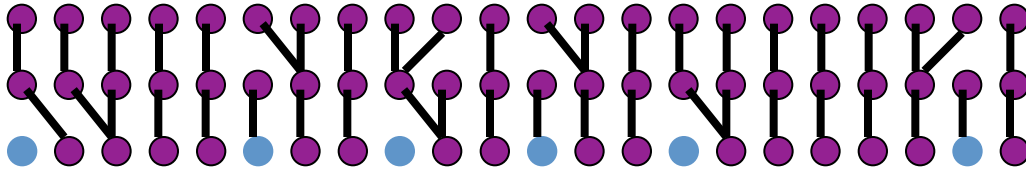
22 individuals

18 ancestors

Time



Present



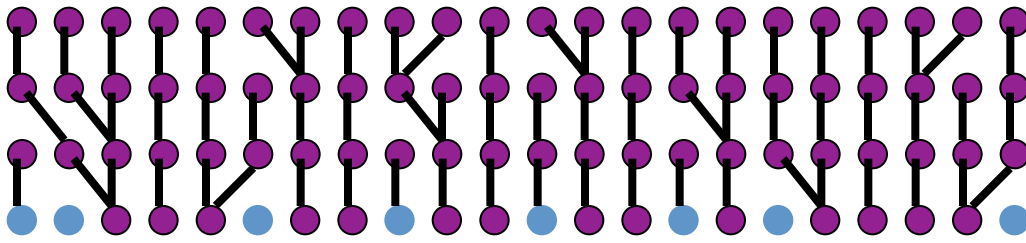
22 individuals

18 ancestors

16 ancestors

Time

Present



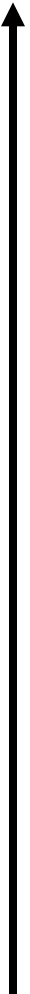
22 individuals

18 ancestors

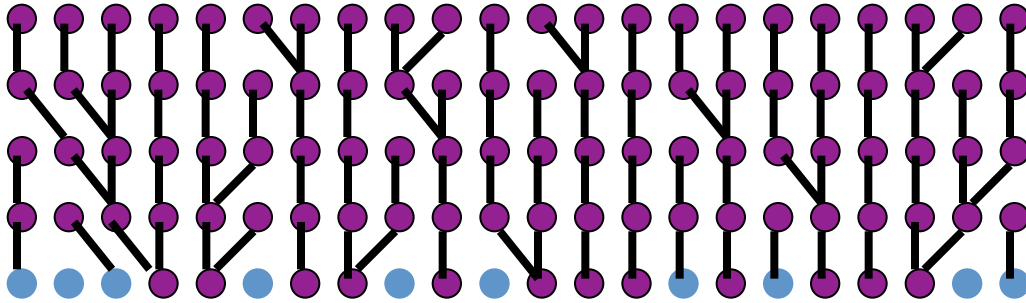
16 ancestors

14 ancestors

Time



Present



22 individuals

18 ancestors

16 ancestors

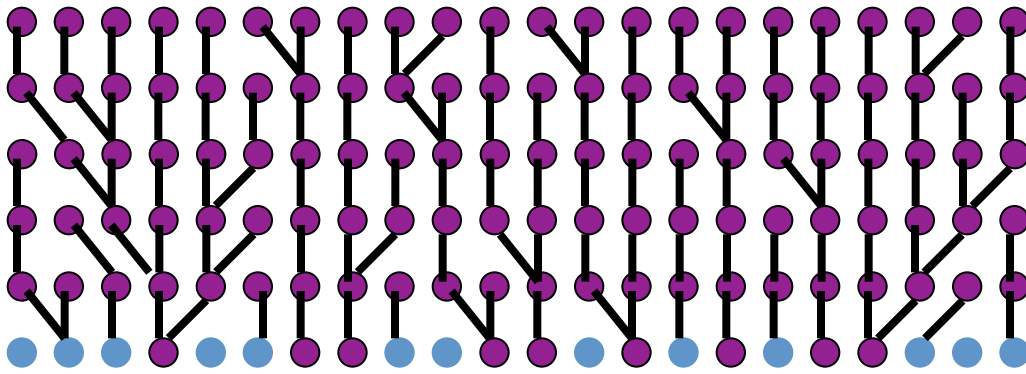
14 ancestors

12 ancestors

Time



Present



22 individuals

18 ancestors

16 ancestors

14 ancestors

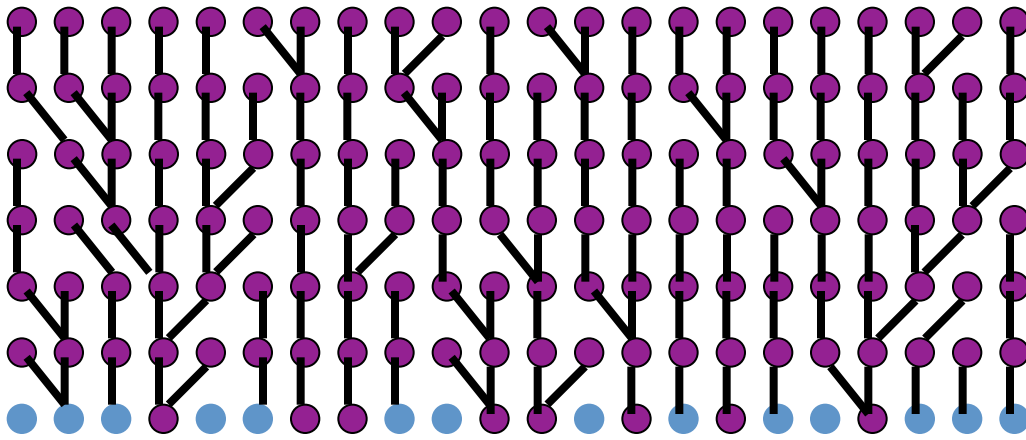
12 ancestors

9 ancestors

Time

Present

Time



22 individuals

18 ancestors

16 ancestors

14 ancestors

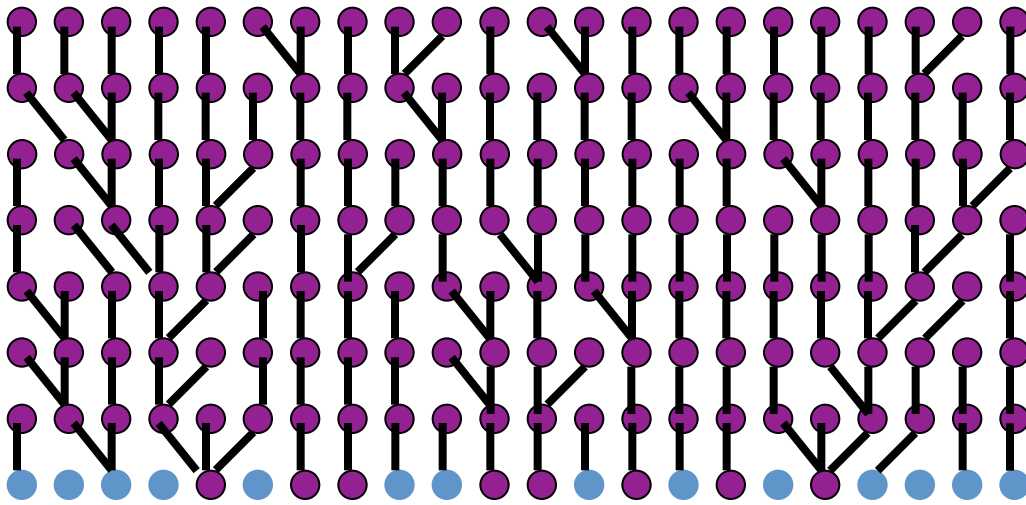
12 ancestors

9 ancestors

8 ancestors

Present

Time



22 individuals

18 ancestors

16 ancestors

14 ancestors

12 ancestors

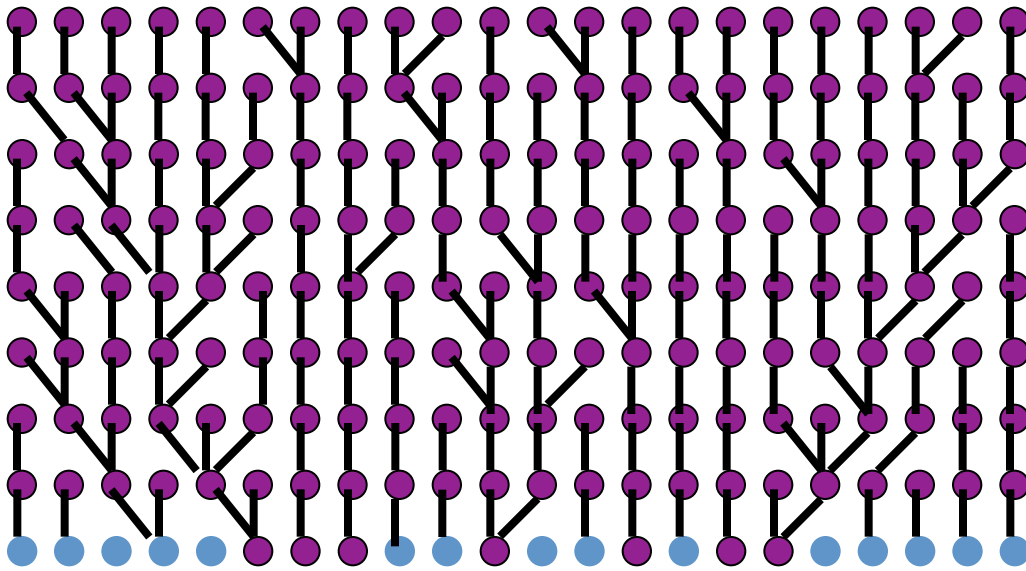
9 ancestors

8 ancestors

8 ancestors

Present

Time



22 individuals

18 ancestors

16 ancestors

14 ancestors

12 ancestors

9 ancestors

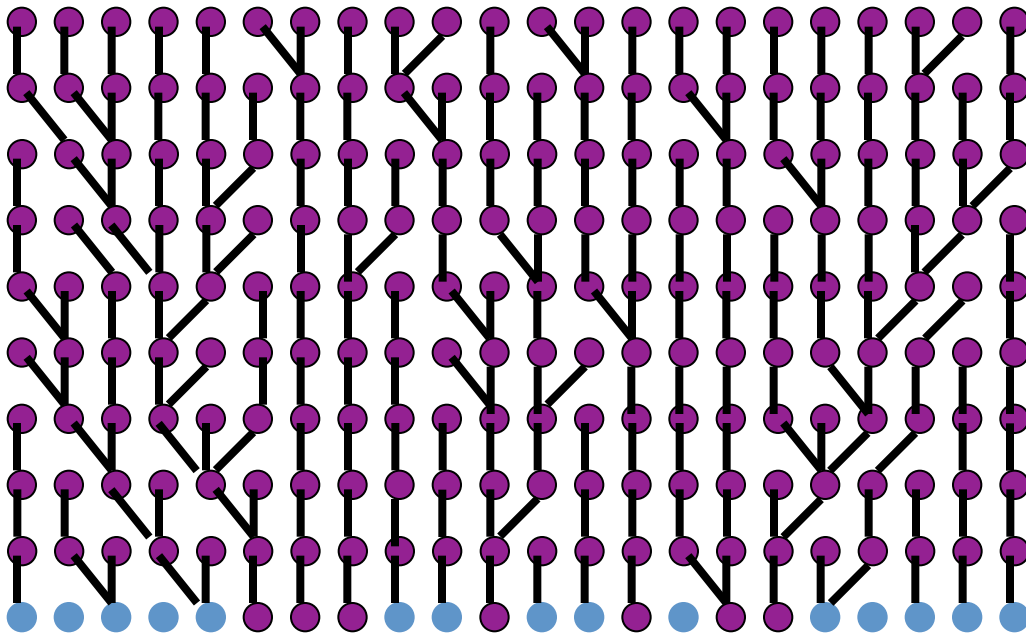
8 ancestors

8 ancestors

7 ancestors

Present

Time



22 individuals

18 ancestors

16 ancestors

14 ancestors

12 ancestors

9 ancestors

8 ancestors

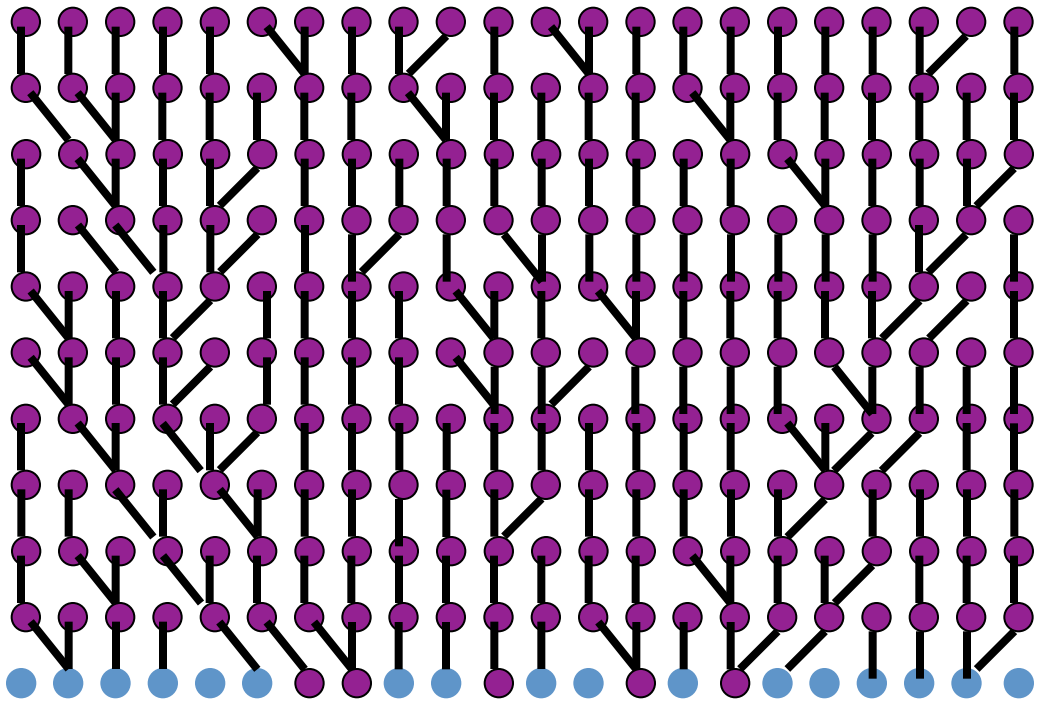
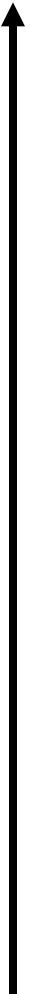
8 ancestors

7 ancestors

7 ancestors

Present

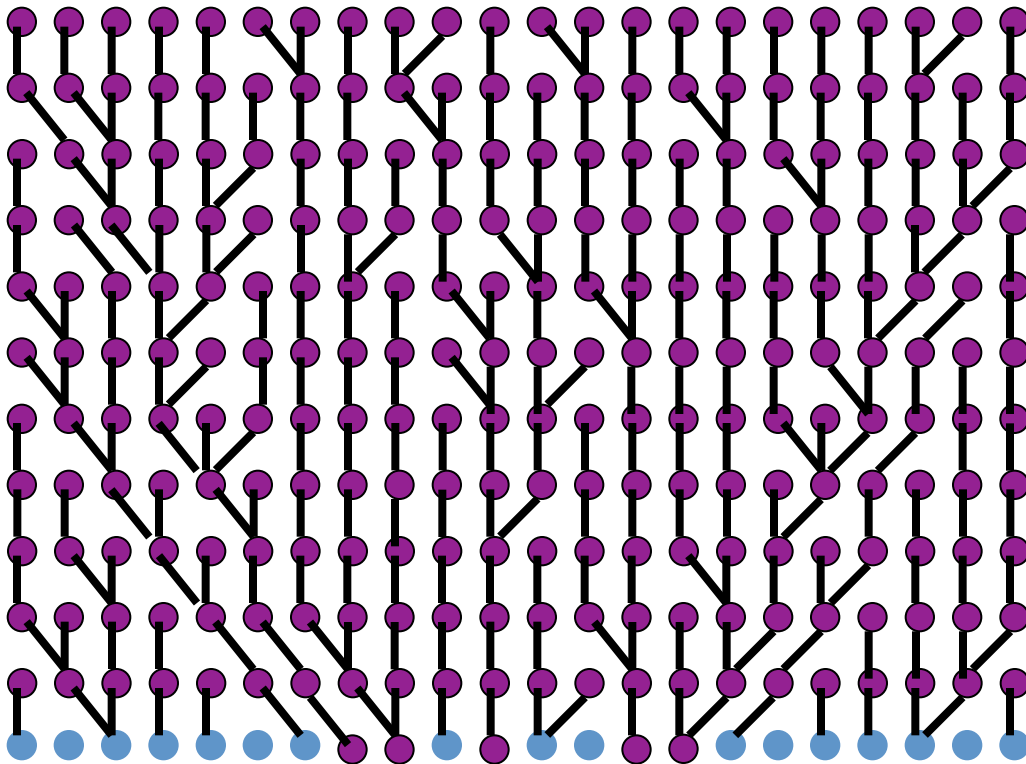
Time



- 22 individuals**
- 18 ancestors**
- 16 ancestors**
- 14 ancestors**
- 12 ancestors**
- 9 ancestors**
- 8 ancestors**
- 8 ancestors**
- 7 ancestors**
- 7 ancestors**
- 5 ancestors**

Present

Time



22 individuals

18 ancestors

16 ancestors

14 ancestors

12 ancestors

9 ancestors

8 ancestors

8 ancestors

7 ancestors

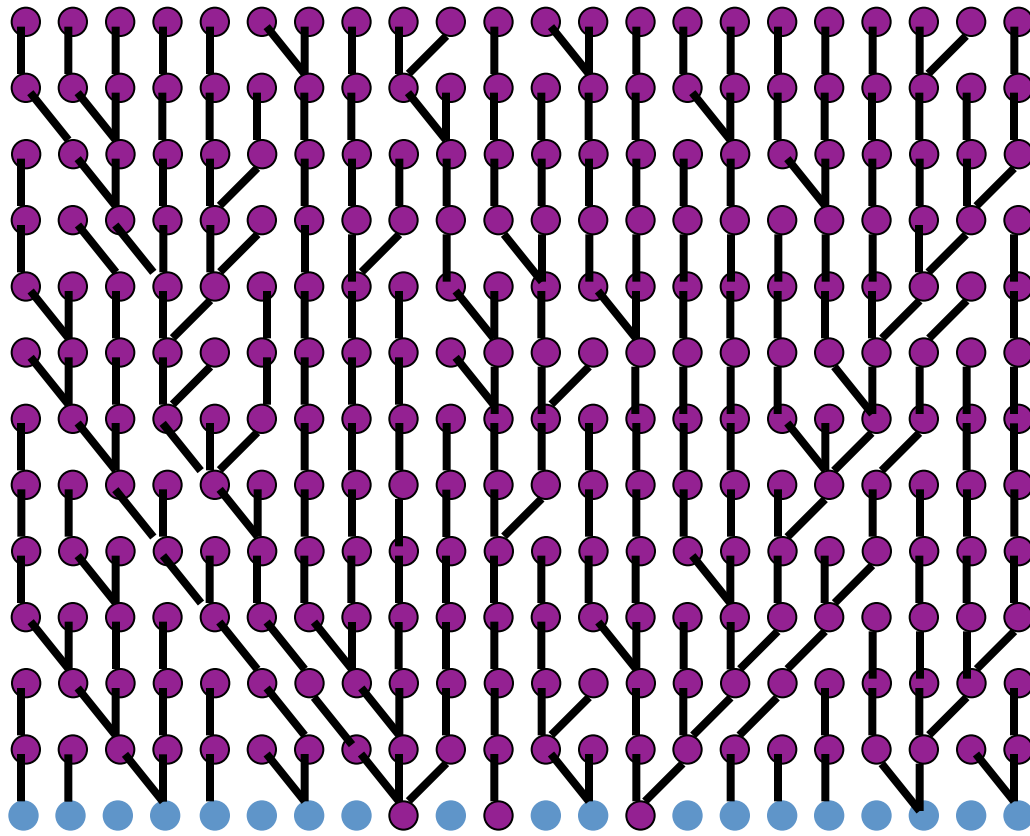
7 ancestors

5 ancestors

5 ancestors

Present

Time



22 individuals

18 ancestors

16 ancestors

14 ancestors

12 ancestors

9 ancestors

8 ancestors

8 ancestors

7 ancestors

7 ancestors

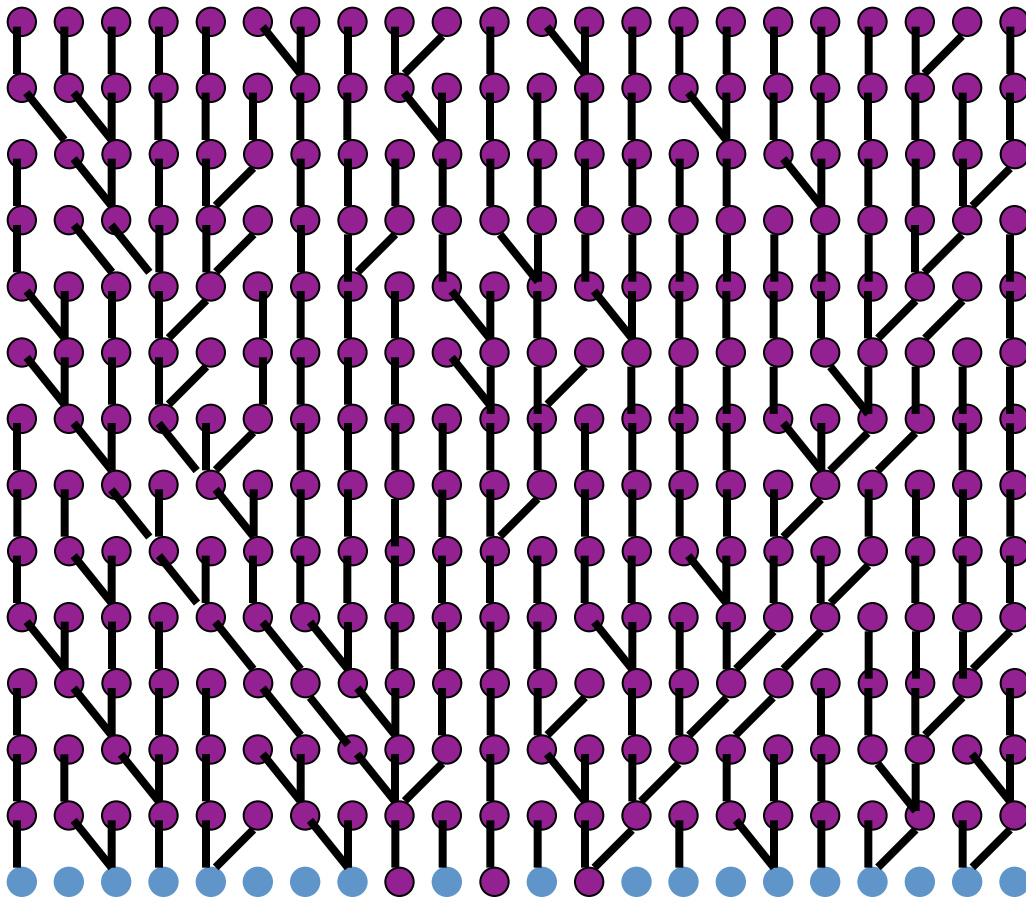
5 ancestors

5 ancestors

3 ancestors

Present

Time



22 individuals

18 ancestors

16 ancestors

14 ancestors

12 ancestors

9 ancestors

8 ancestors

8 ancestors

7 ancestors

7 ancestors

5 ancestors

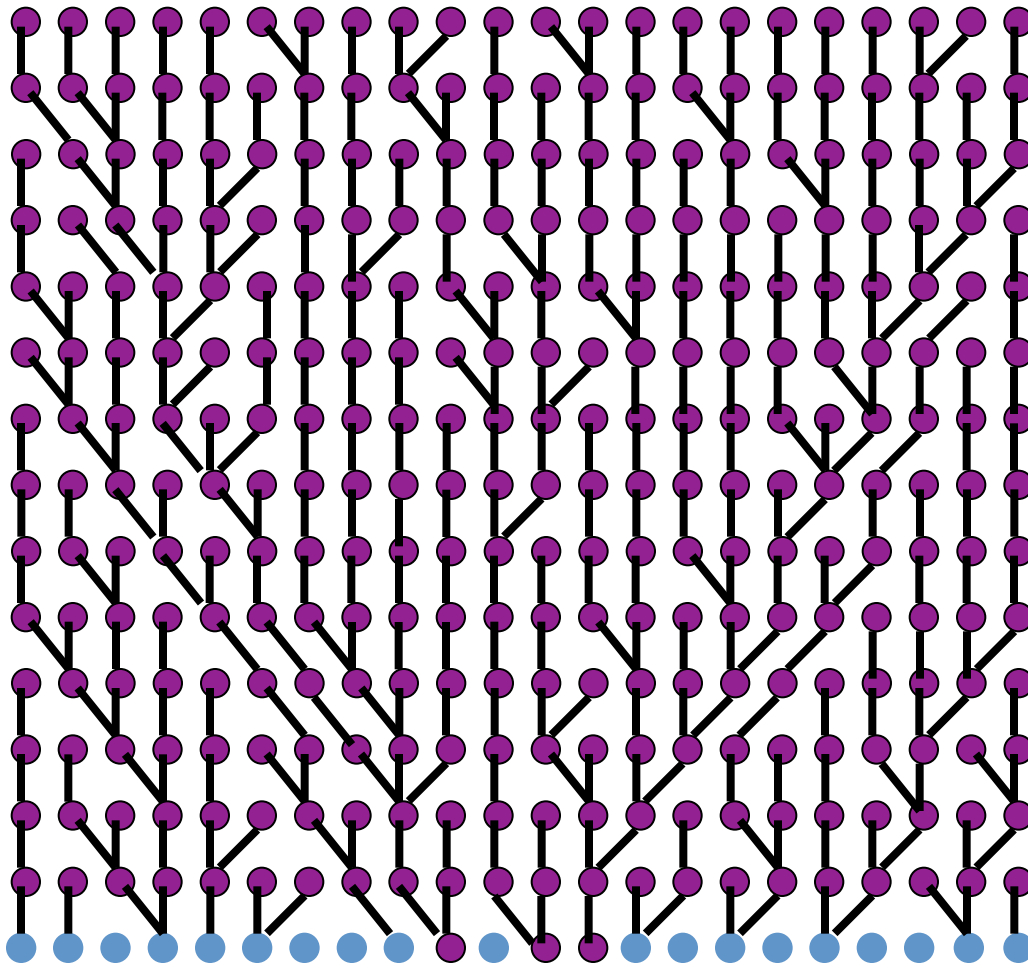
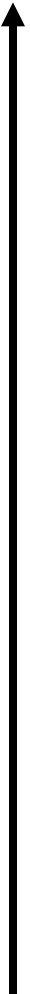
5 ancestors

3 ancestors

3 ancestors

Present

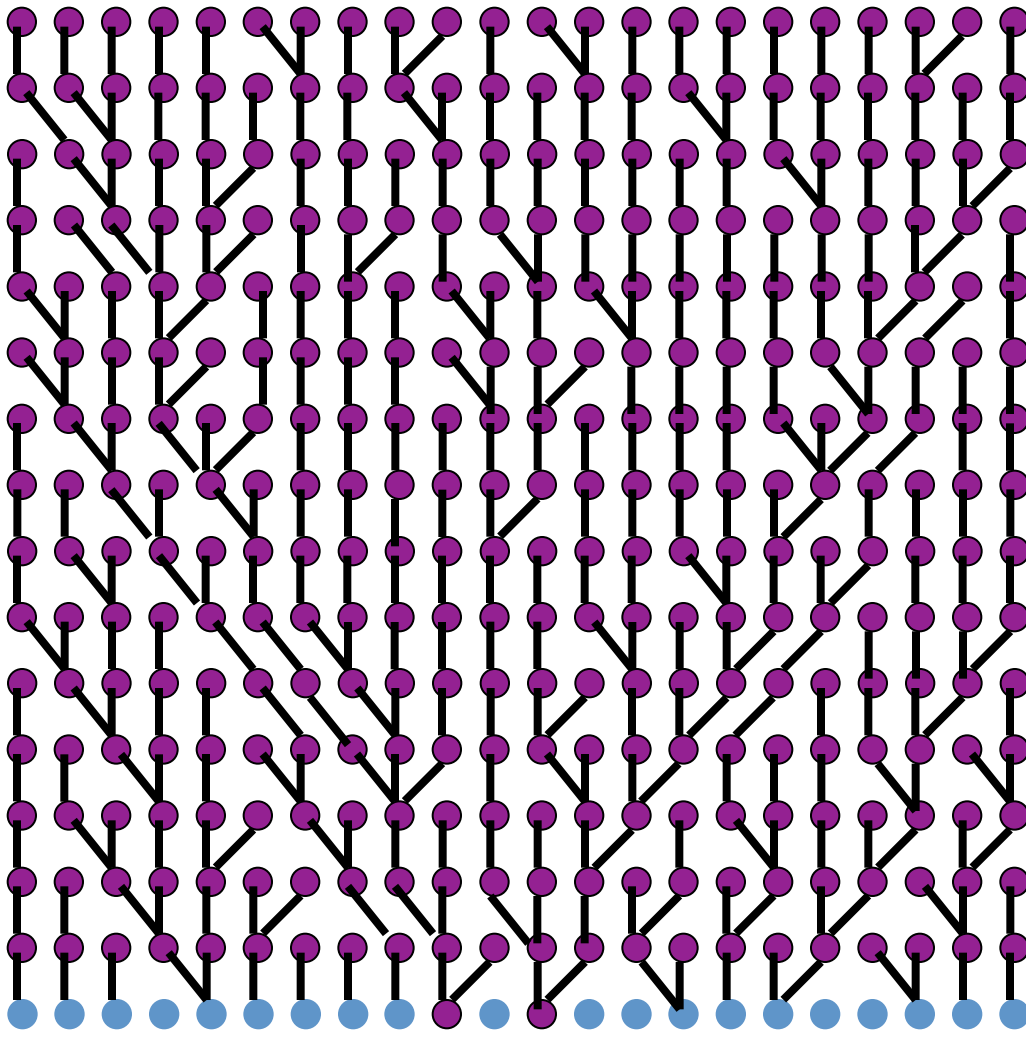
Time



- 22 individuals**
- 18 ancestors**
- 16 ancestors**
- 14 ancestors**
- 12 ancestors**
- 9 ancestors**
- 8 ancestors**
- 8 ancestors**
- 7 ancestors**
- 7 ancestors**
- 5 ancestors**
- 5 ancestors**
- 3 ancestors**
- 3 ancestors**
- 3 ancestors**

Present

Time



22 individuals

18 ancestors

16 ancestors

14 ancestors

12 ancestors

9 ancestors

8 ancestors

8 ancestors

7 ancestors

7 ancestors

5 ancestors

5 ancestors

3 ancestors

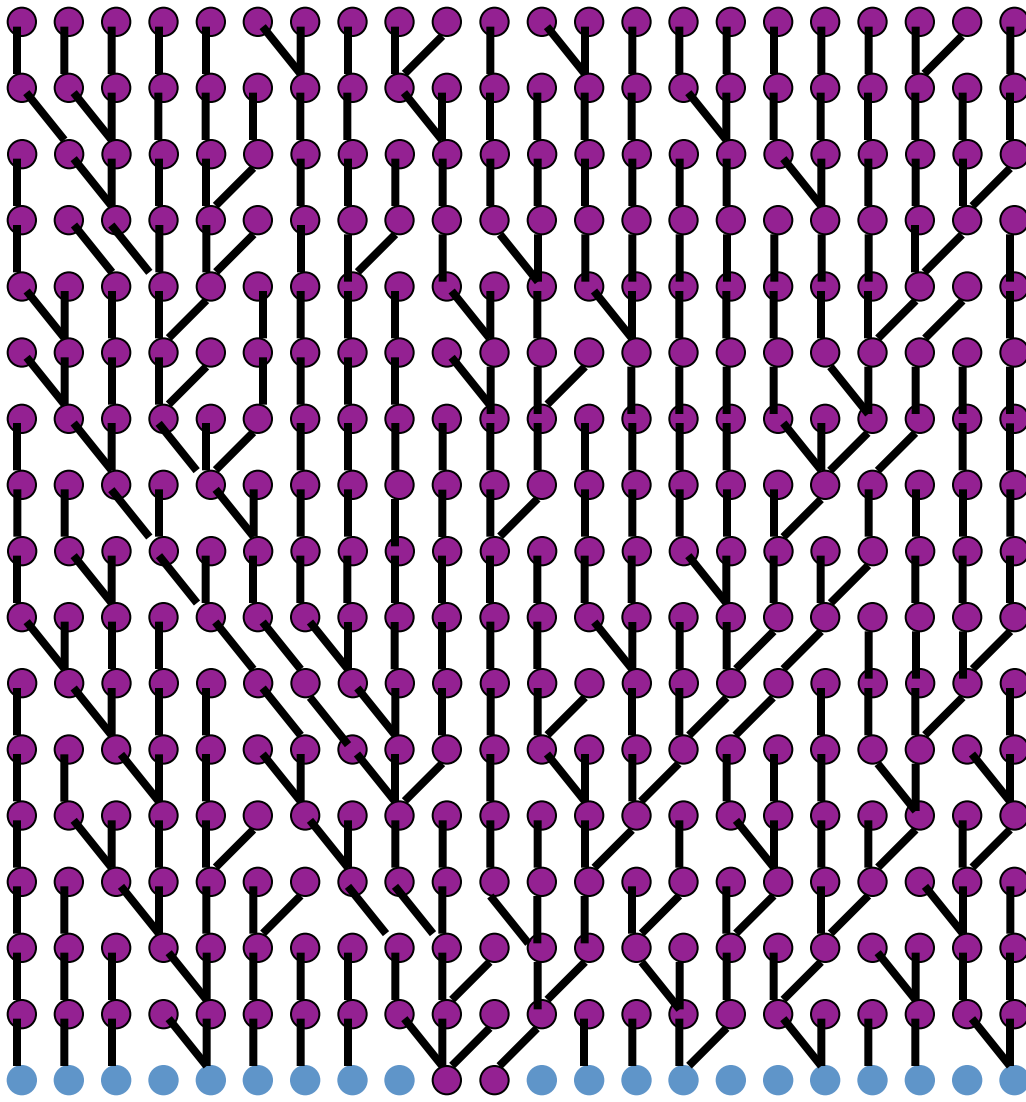
3 ancestors

3 ancestors

2 ancestors

Present

Time



22 individuals

18 ancestors

16 ancestors

14 ancestors

12 ancestors

9 ancestors

8 ancestors

8 ancestors

7 ancestors

7 ancestors

5 ancestors

5 ancestors

3 ancestors

3 ancestors

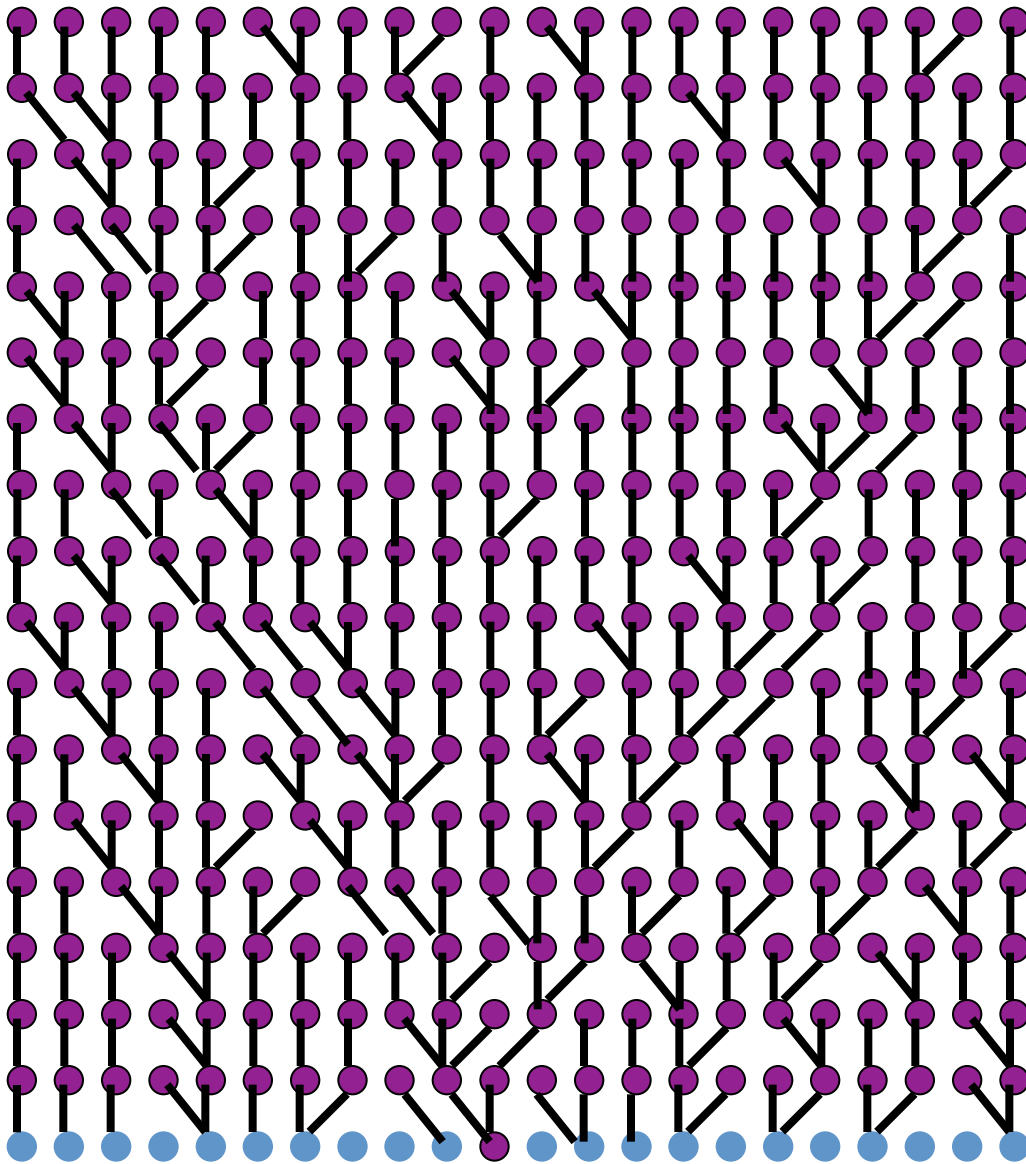
3 ancestors

2 ancestors

2 ancestors

Present

Time



22 individuals

18 ancestors

16 ancestors

14 ancestors

12 ancestors

9 ancestors

8 ancestors

8 ancestors

7 ancestors

7 ancestors

5 ancestors

5 ancestors

3 ancestors

3 ancestors

3 ancestors

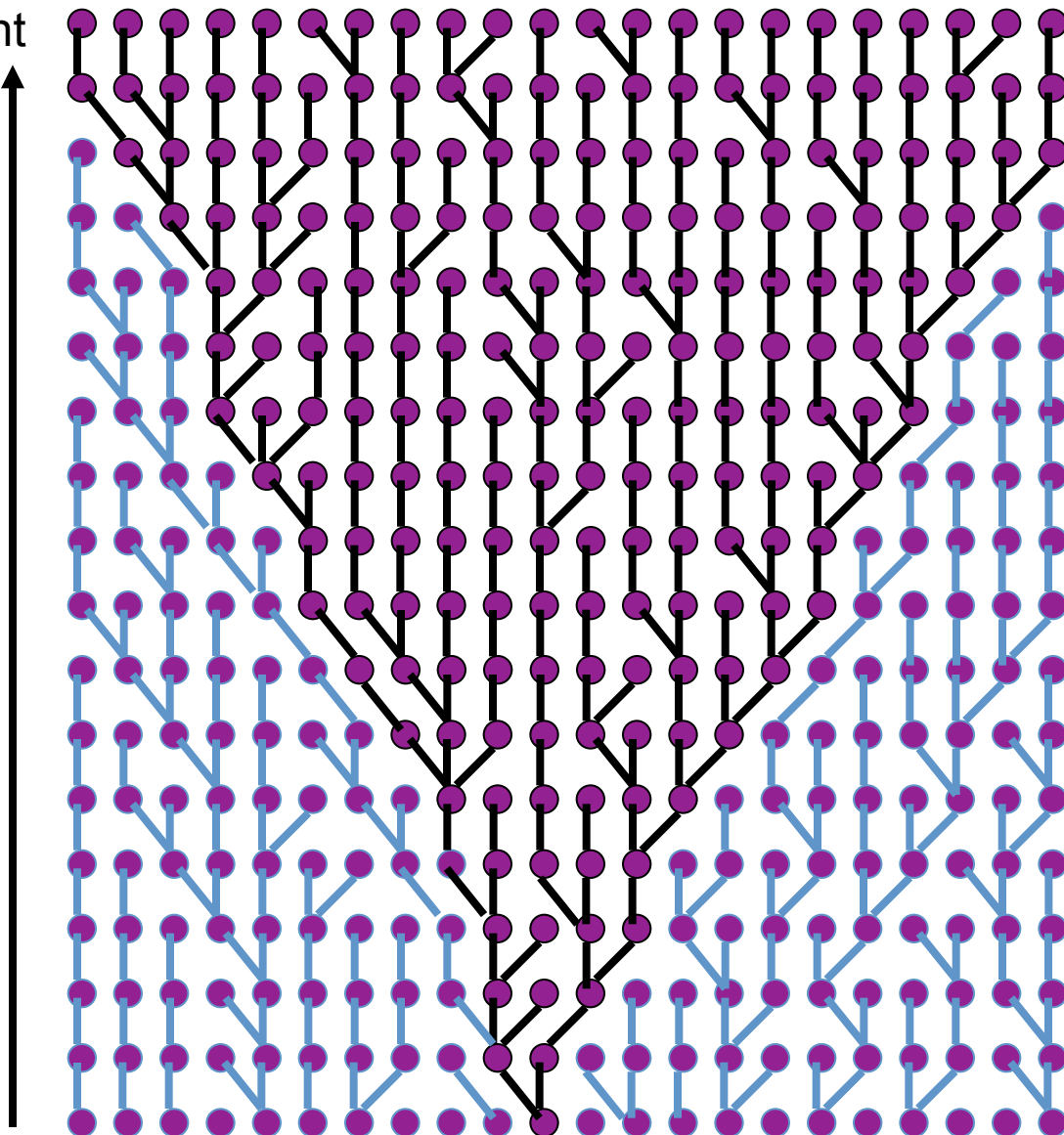
2 ancestors

2 ancestors

1 ancestor

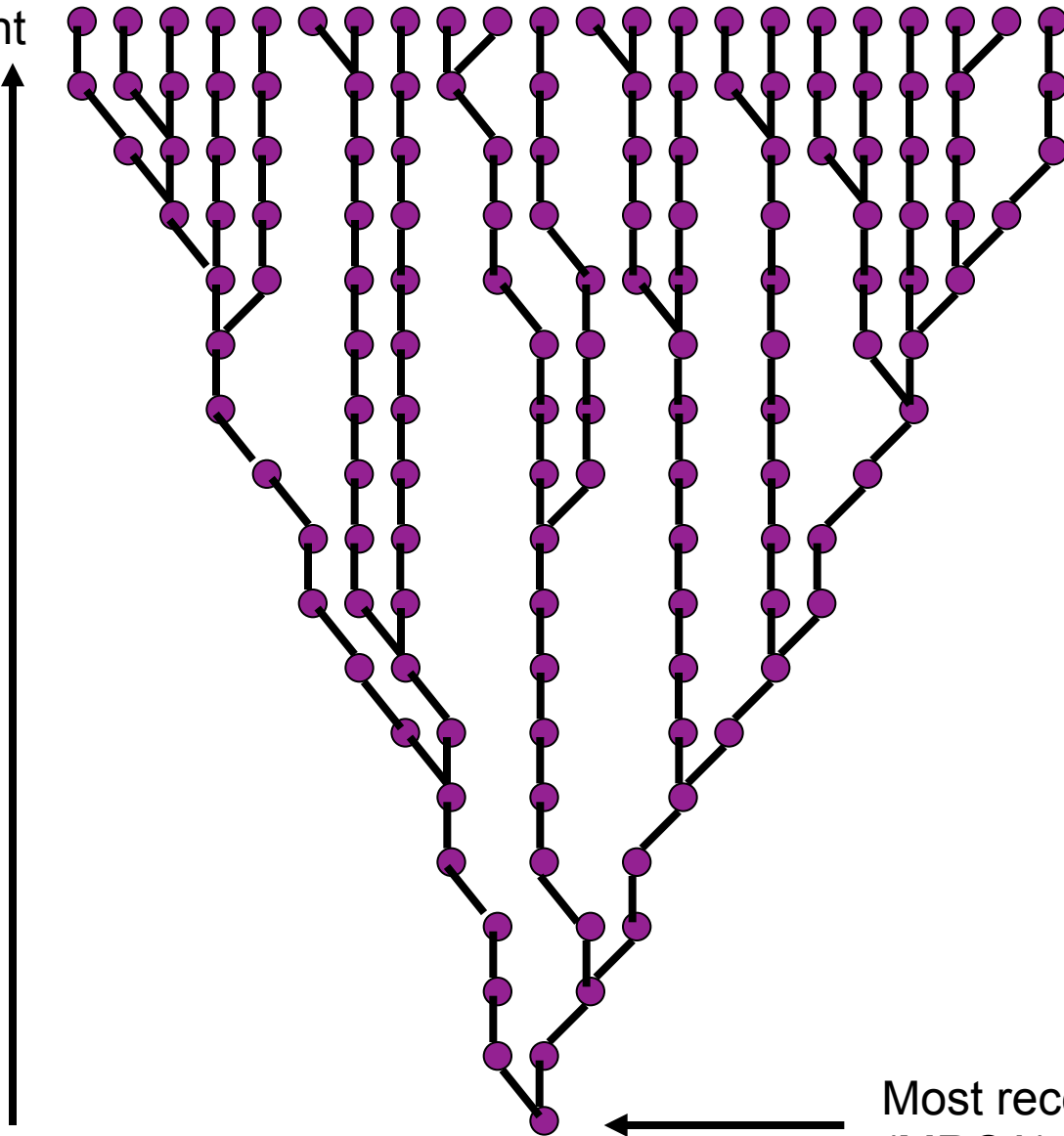
Present

Time



Present

Time

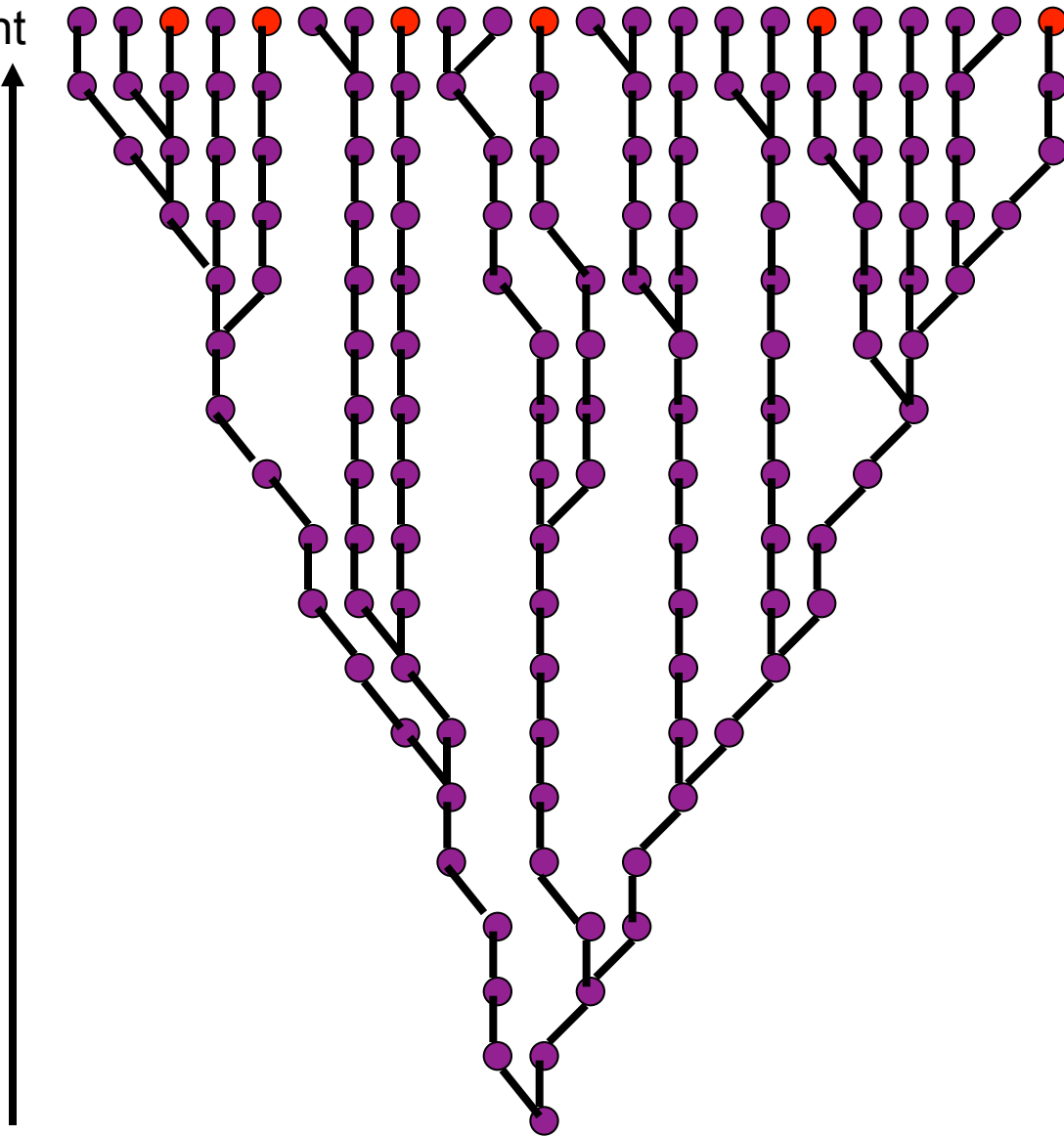


← Most recent common ancestor (MRCA)

- Most of the time, we are interested in the genealogy of a sample taken from the whole population.

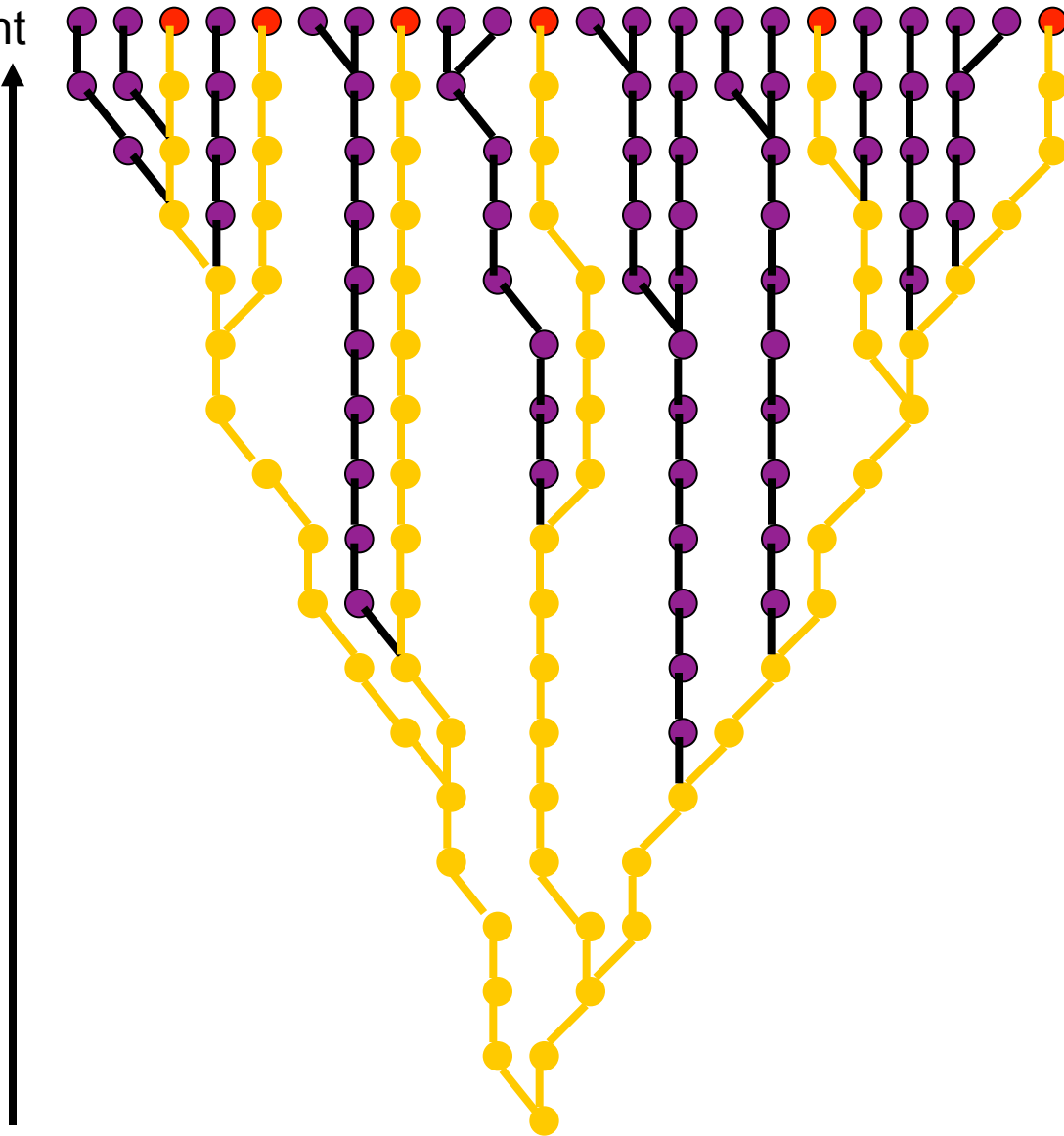
Present

Time



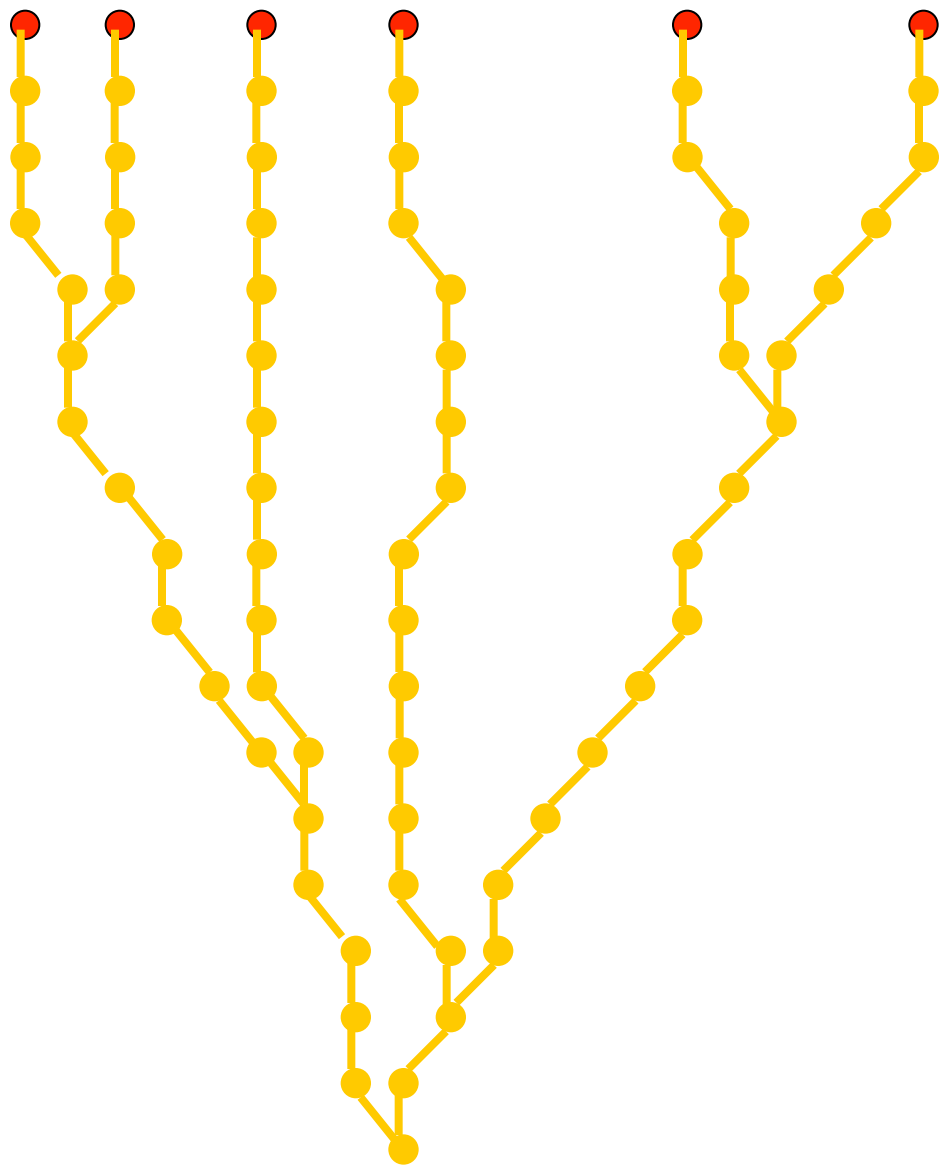
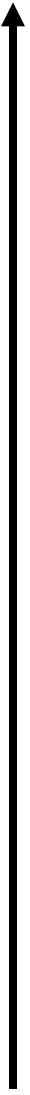
Present

Time



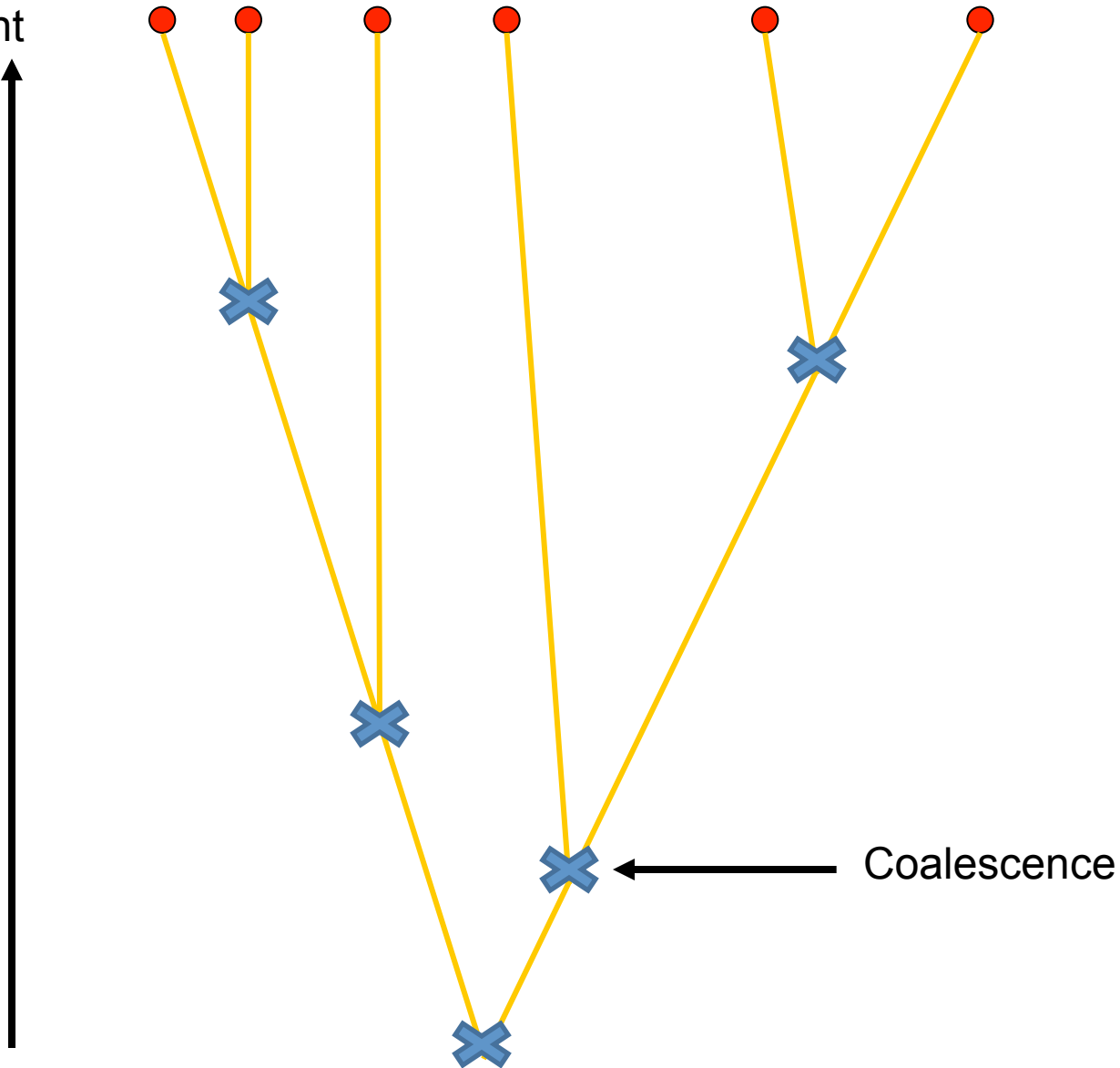
Present

Time



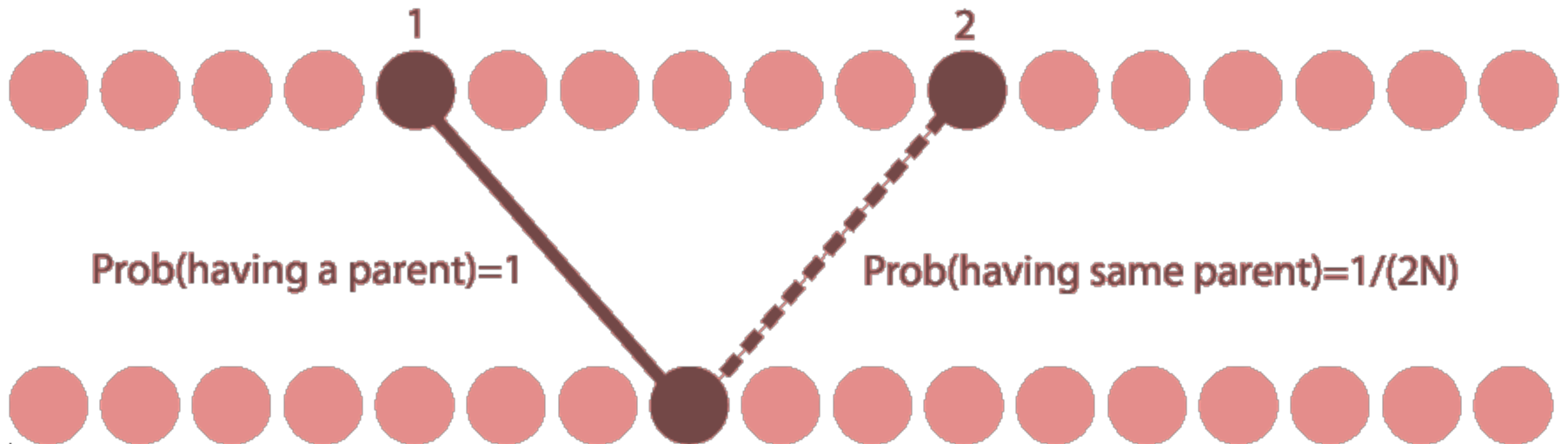
Present

Time



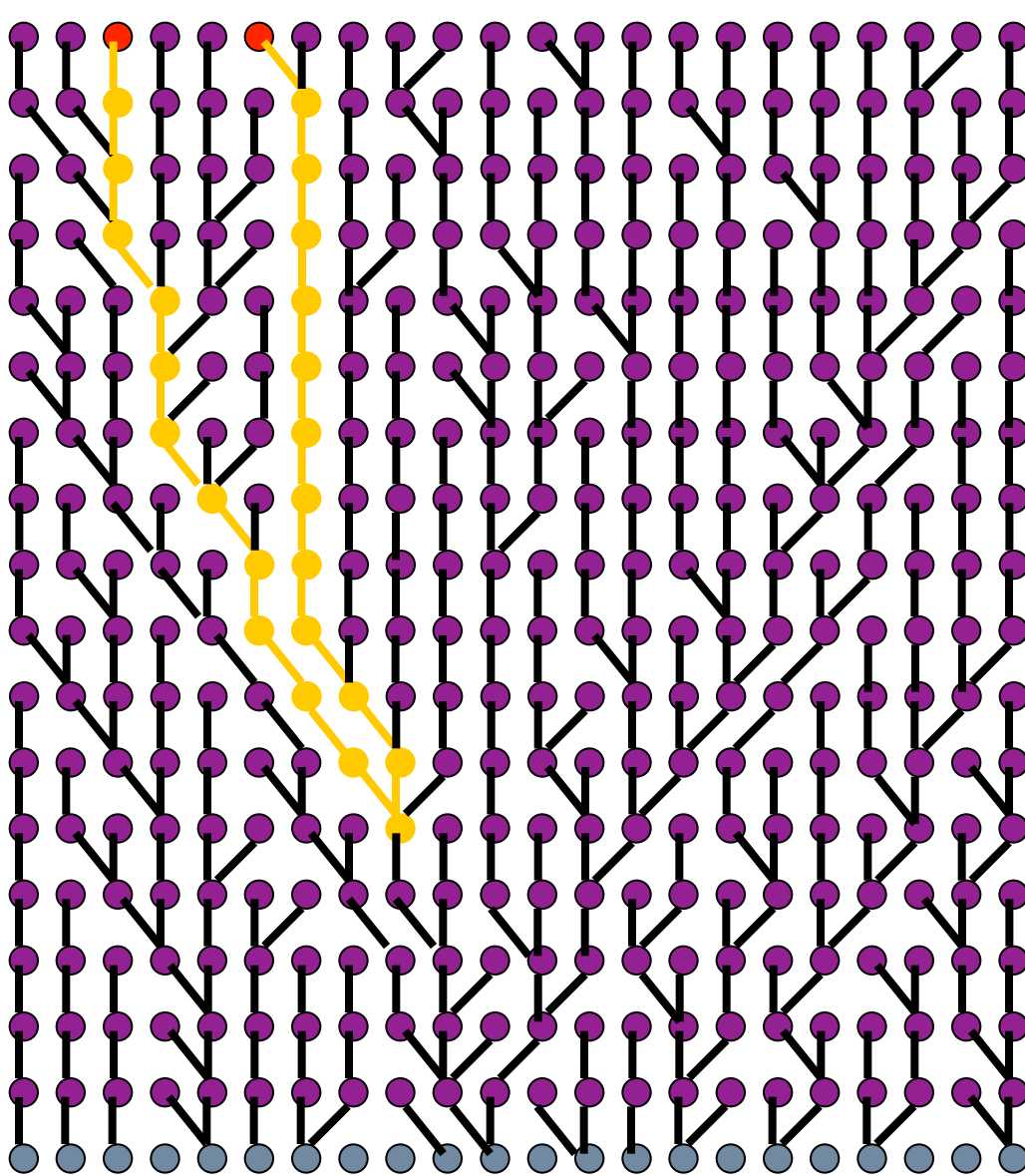
Wright-Fisher demographic model

- Forwards-in-time model of a population in a **constant-size, random-mating**, evolving in **discrete** generations (non-overlapping)



Present

Time



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



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

$$P = \left(1 - \frac{1}{2N}\right)^{11} * \frac{1}{2N}$$

Wright-Fisher model

- The time to coalesce for two genes follows a **geometric distribution** with parameter $1/2N$
- The probability that two genes coalesce t generations ago is given by:

$$\left(1 - \frac{1}{2N}\right)^{t-1} * \frac{1}{2N}$$

- The **expected time** to coalesce is **$2N$**
- But the **variance is big**: $2N*(2N-1) \sim N^2$

Kingman's "n-coalescent"

- We now consider k genes
- There is more chance to observe a coalescent event: $1/2N$ for each possible pair among the k
- Number of possible pairs: $\binom{k}{2} = \frac{k * (k - 1)}{2}$
- The total probability of any one pair to coalesce in the former generation is then

$$P = \frac{k * (k - 1)}{4N}$$

Kingman's "n-coalescent"

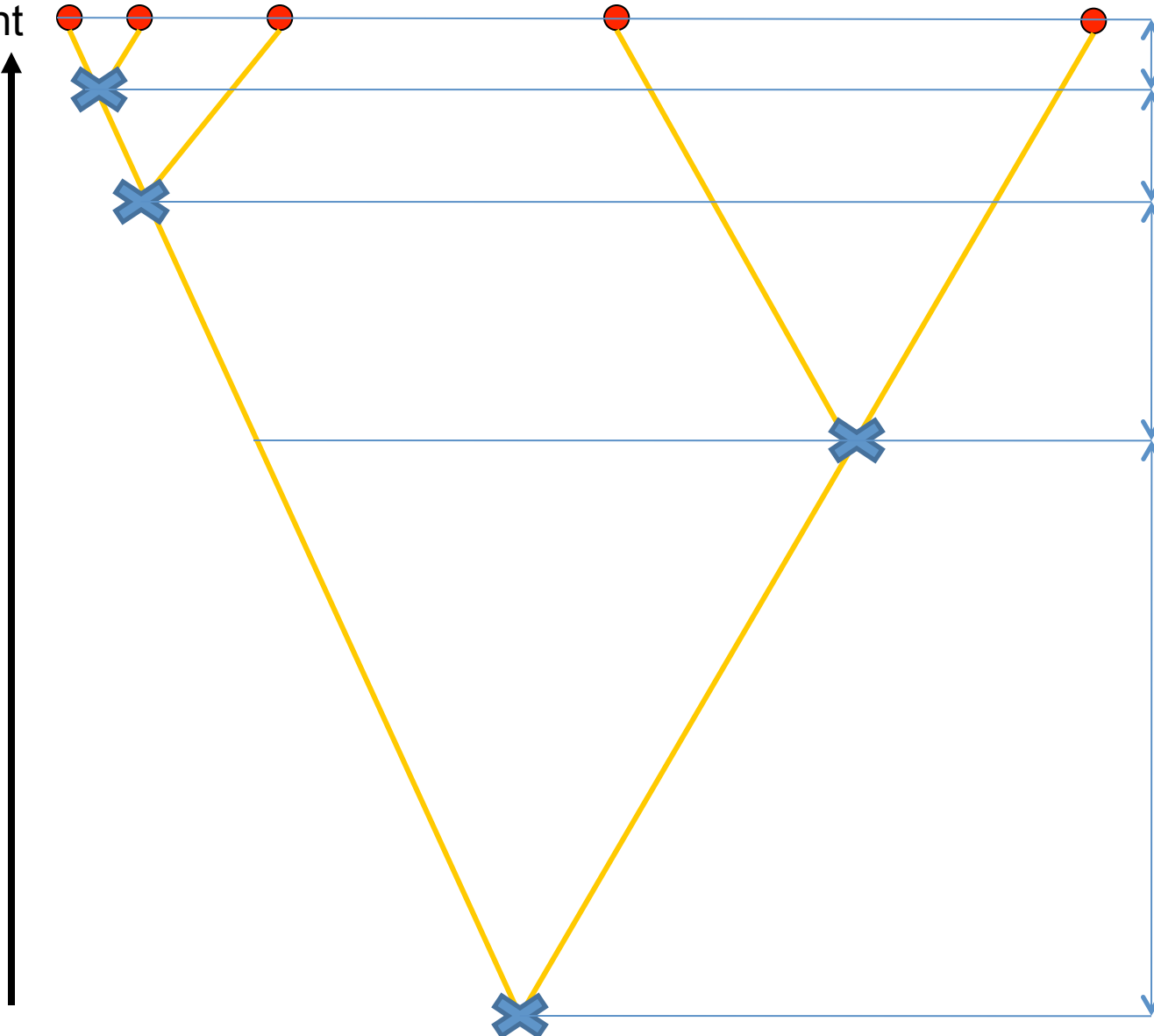
- now depends on k
- The time to coalesce follows a geometric distribution, but with parameter $P = \frac{k * (k - 1)}{4N}$
- It now depends on k and we have:

$$T(k) = \frac{4N}{k * (k - 1)}$$

- We still have $T(2) = 2N$

Present

Time



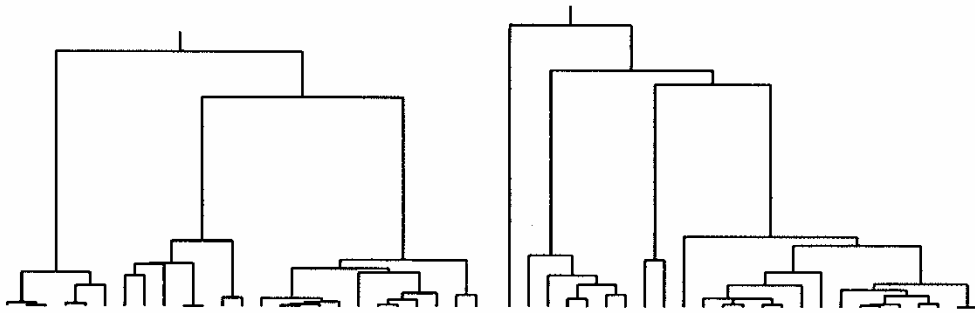
$$T(5) = \frac{2N}{10}$$

$$T(4) = \frac{2N}{6}$$

$$T(3) = \frac{2N}{3}$$

$$T(2) = 2N$$

Coalescent in a stationary population



Gene genealogies are extremely variable in stationary populations, both for the topology and the branch length



Generally we'll have long internal branch length and small external branch length.

$$T(k) = \frac{4N}{k * (k-1)}$$

Time to coalesce:

- The total time for the k genes to coalesce is:

$$T_{MRCA}(k) = T(2) + T(3) + \dots + T(k-1) + T(k)$$

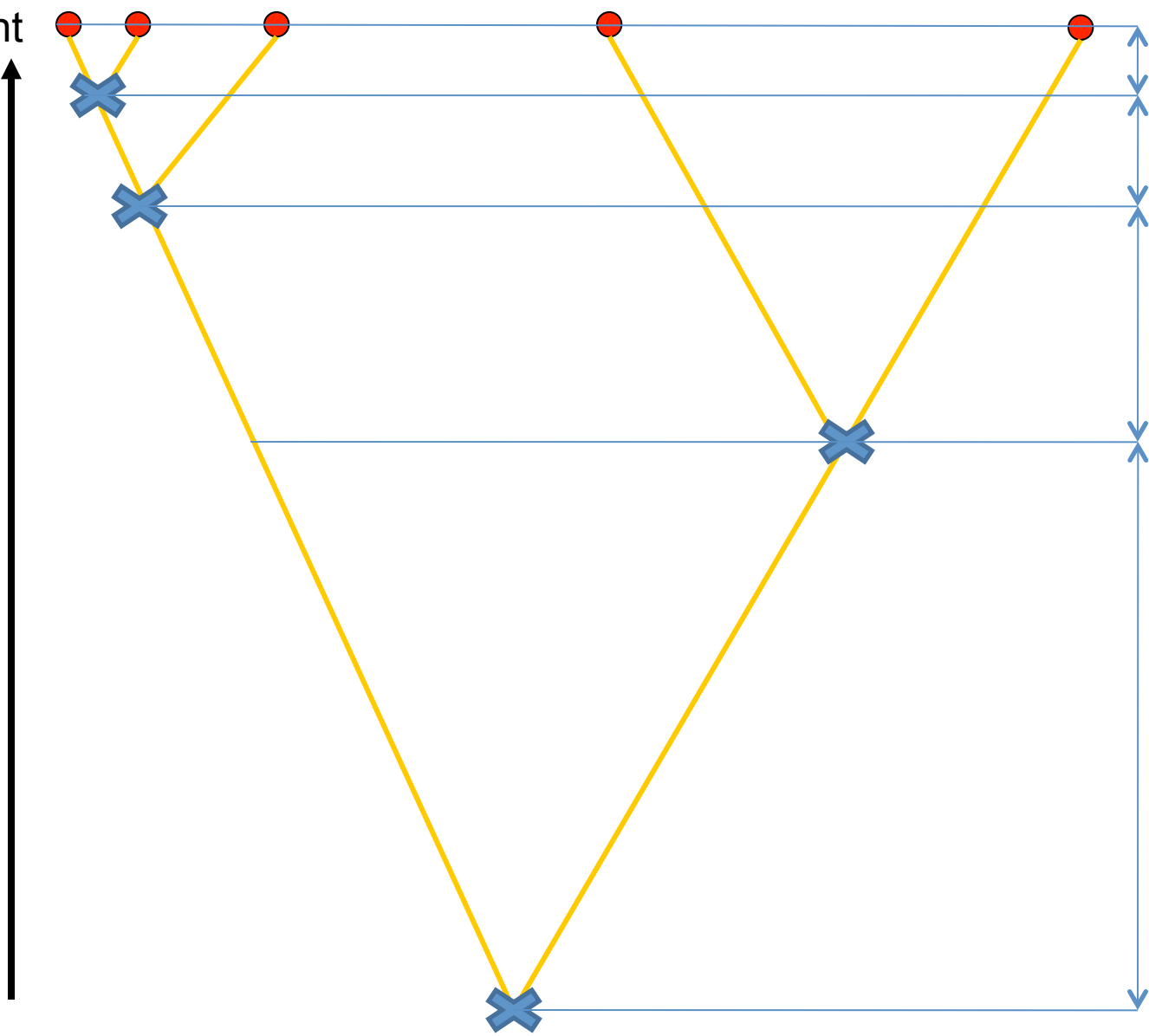
$$= 4N \left(\frac{1}{\underbrace{2 * 1}_{=\frac{1}{1} \frac{1}{2}}} + \frac{1}{\underbrace{3 * 2}_{=\frac{1}{2} \frac{1}{3}}} + \dots + \frac{1}{\underbrace{(k-1) * (k-2)}_{=\frac{1}{k-2} \frac{1}{k-1}}} + \frac{1}{\underbrace{k * (k-1)}_{=\frac{1}{k-1} \frac{1}{k}}} \right)$$

$$= 4N \left(1 - \frac{1}{k} \right)$$

- The expected time during which there are only two branches ($2N$) is greater than half the expected total tree height

Present

Time



$$T(5) = \frac{2N}{10}$$

$$T(4) = \frac{2N}{6}$$

$$T(3) = \frac{2N}{3}$$

$$T(2) = 2N$$

Total length of the tree:

- The total length is simply given by:

$$T_{Total}(k) = 2T(2) + 3T(3) + \dots + kT(k)$$

$$= 4N \sum_{i=2}^k \frac{i}{i(i-1)}$$

$$= 4N \left(\frac{1}{1} + \frac{1}{2} + \dots + \frac{1}{k-1} \right)$$

$$= 4N \sum_{i=1}^{k-1} \frac{1}{i}$$

$$T(k) = \frac{4N}{k * (k-1)}$$

Sampling effect

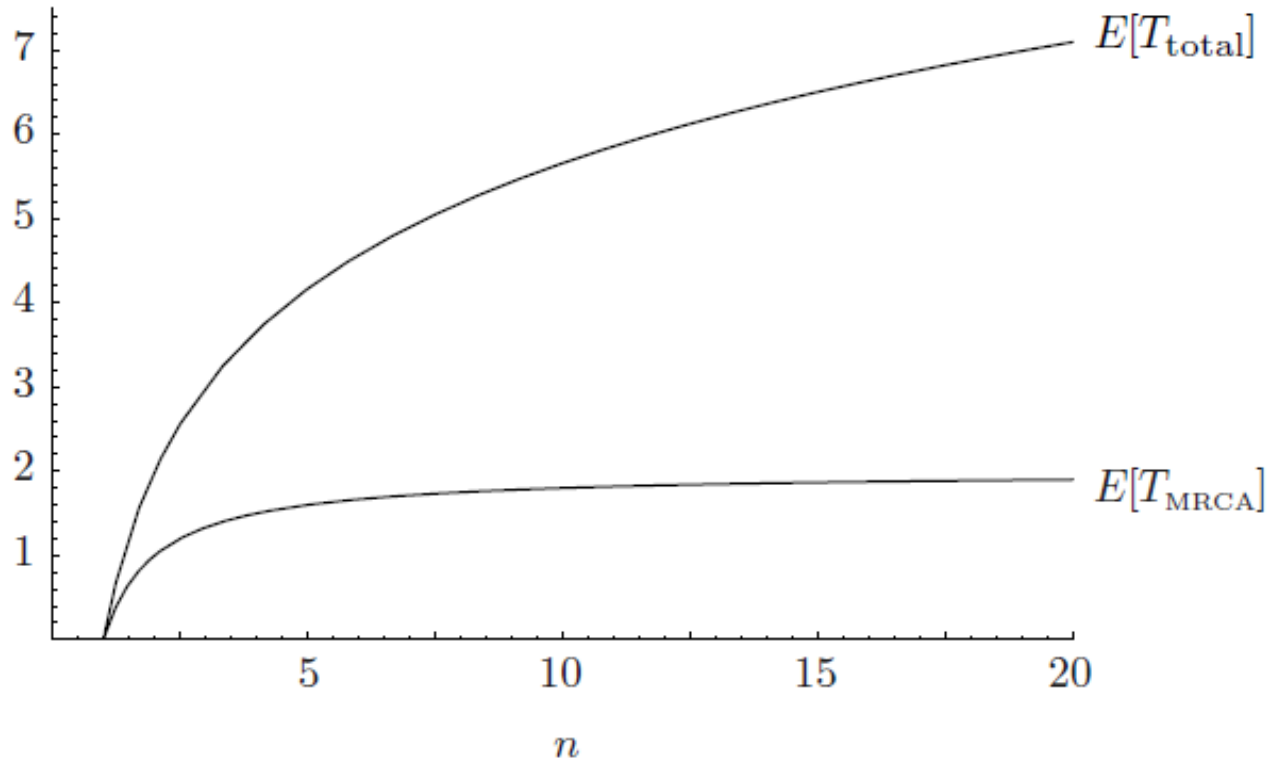


Figure 3.3: The relationship between sample size and the expected values of T_{MRCA} and T_{total} .

Consequences:

- The larger the sample size the greater the rate of coalescence
- the larger the population size the slower the rate of coalescence
- Time to coalescence gets longer as the process moves toward the most recent common ancestor
- No need to take a lot of genes

Continuous-time version

- The **geometric (discrete)** distribution can be approximated with an **exponential (continuous)** distribution as long as **N is big**:

$$P = \frac{1}{2N} \left(1 - \frac{1}{2N} \right)^{t-1} \approx \frac{1}{2N} e^{-\frac{t}{2N}}$$

- This is the exponential distribution with parameter $\lambda = \frac{1}{2N}$
- We often also rescale the time so that $T=1$ corresponds to $t=2N$: $P = e^{-T}$

Scaled continuous-time “n-coalescent”

- The probability (density) to have a coalescence event at time T is:

$$P_k(T) = \frac{k(k-1)}{2} e^{-\frac{k(k-1)}{2}T}$$

- This is the basic equation for **genealogies**
- But what can we do with this now?

A first simulation algorithm:

1. Start with k genes
2. Simulate the time $T(k)$ from the exponential distribution with parameter
$$P = \frac{k * (k - 1)}{2}$$
3. Choose a random pair of genes and merge them into one
4. Decrease the sample size $k \rightarrow k - 1$
5. If $k > 1$, go to 2, otherwise stop

Demo

- ms, fastsimcoal and Figtree

```
./ms 5 1 -T
```

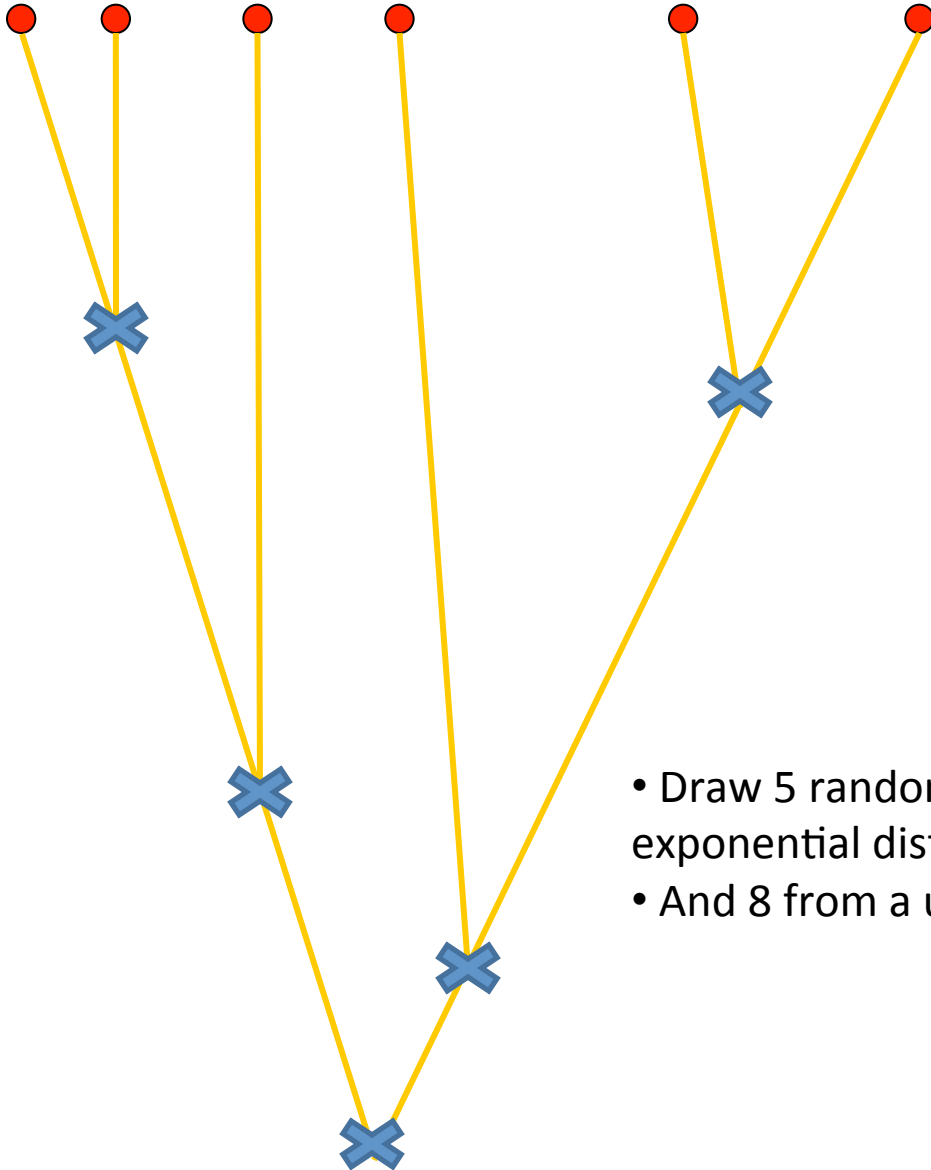
```
./fastsimcoal -i 1PopDNA_sta.par -n 10 -T
```


So what?

- This simulation algorithm is extremely **efficient** compared to a forward simulation of the Wright-Fisher model
- You only simulate **what you need**
- The complexity increases **linearly** with the number of genes

Present

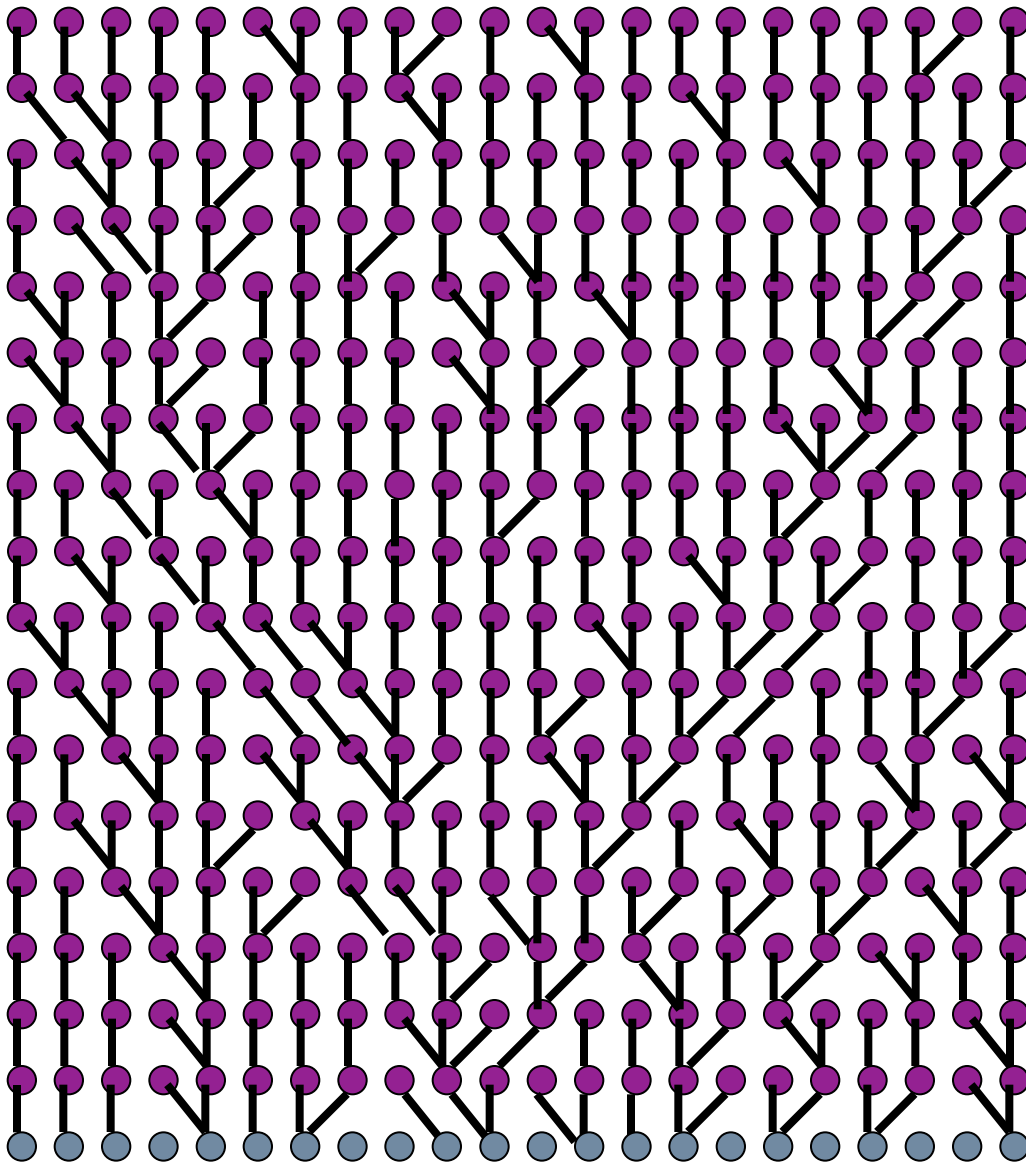
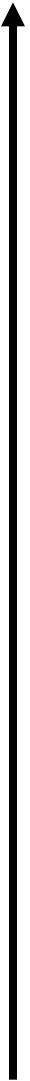
Time



- Draw 5 random numbers from an exponential distribution 
- And 8 from a uniform distribution

Present

Time



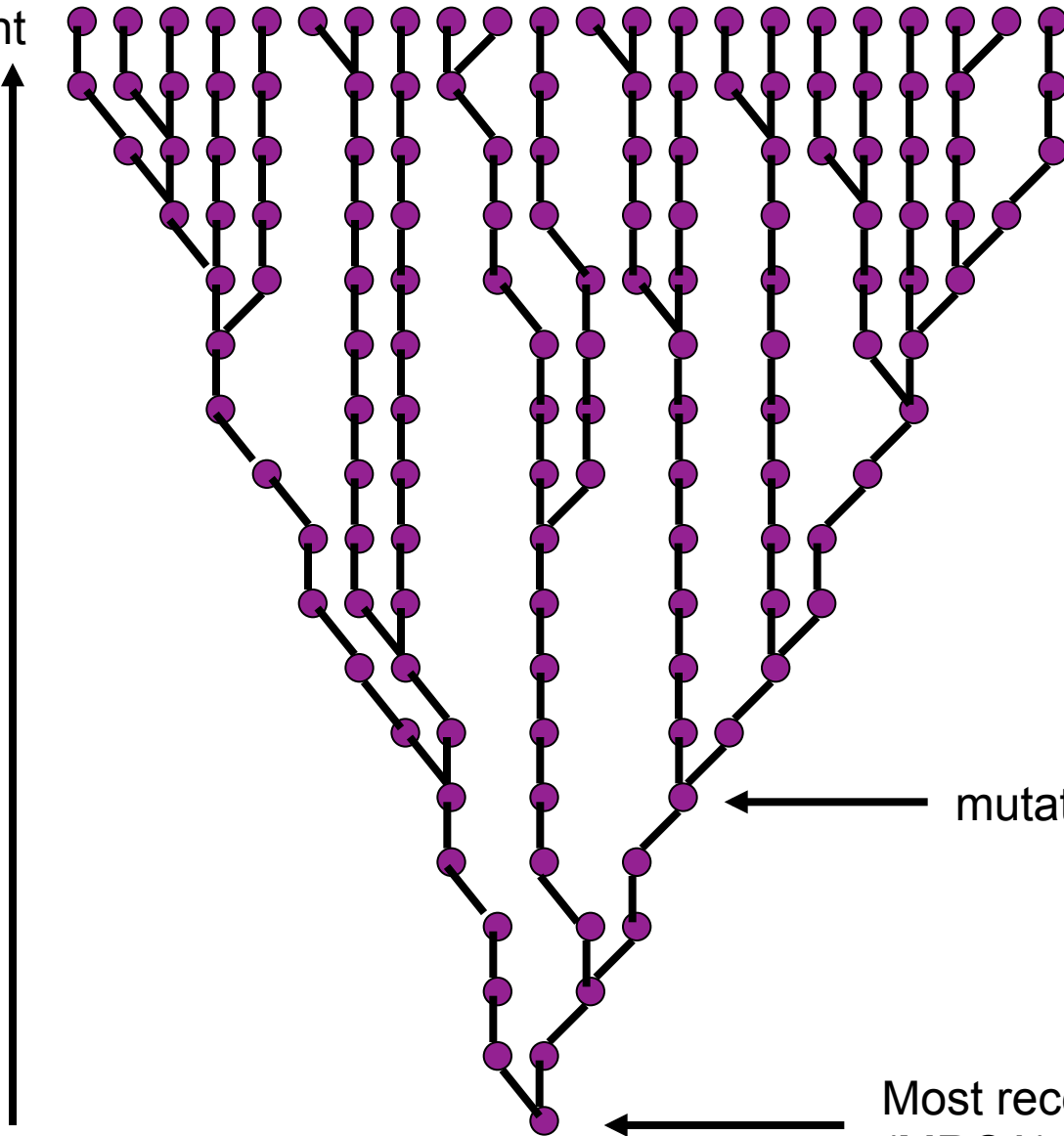
And here???

Adding neutral mutations

- The shape of neutral coalescent trees only depend on the population demography, and not on the mutational process. The mutational process can be modeled as an independent process superimposed on a realized coalescent tree.

Present

Time



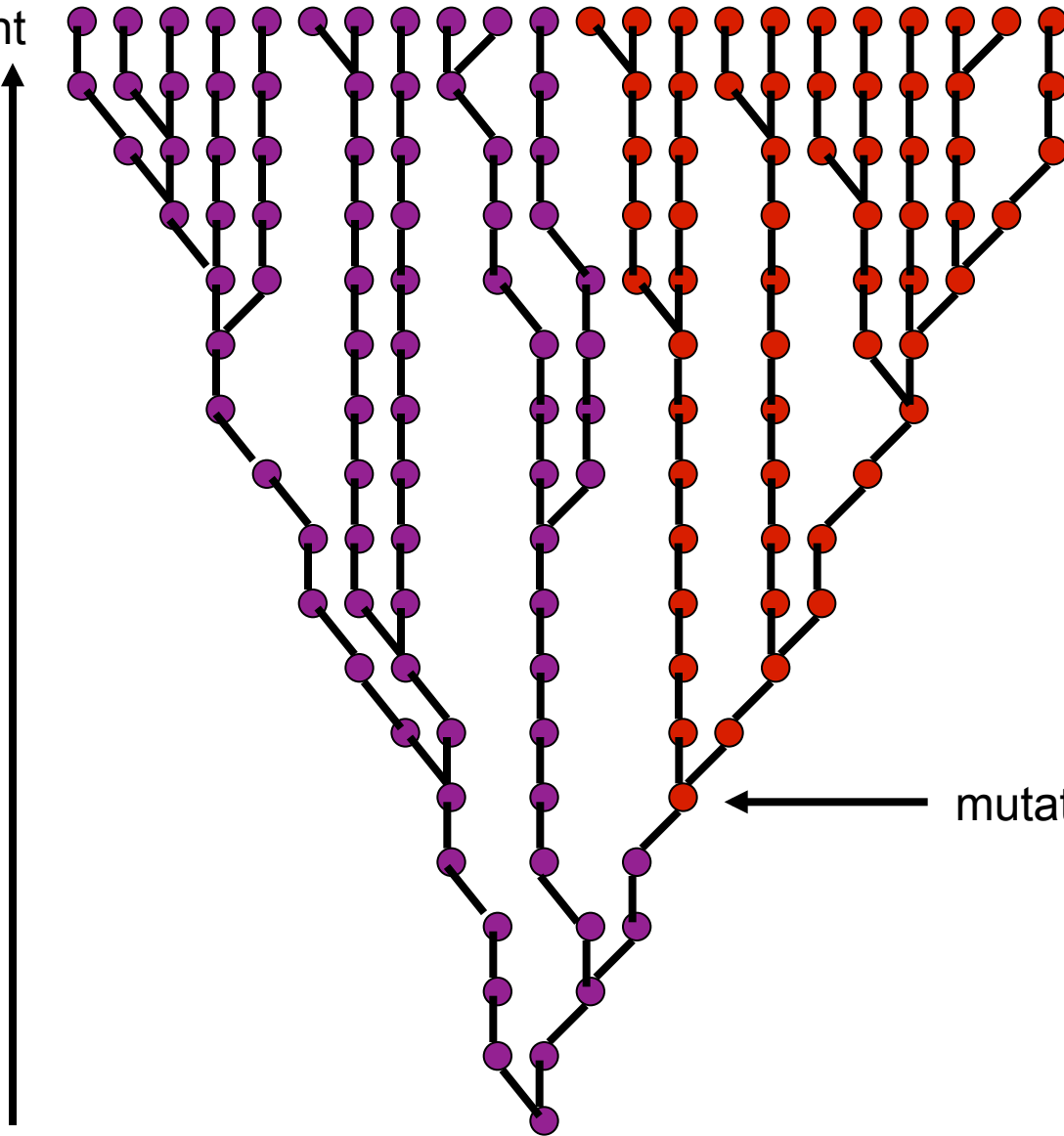
TCGAGGTATTAAC
T

← mutation

← Most recent common ancestor (MRCA)

Present

Time



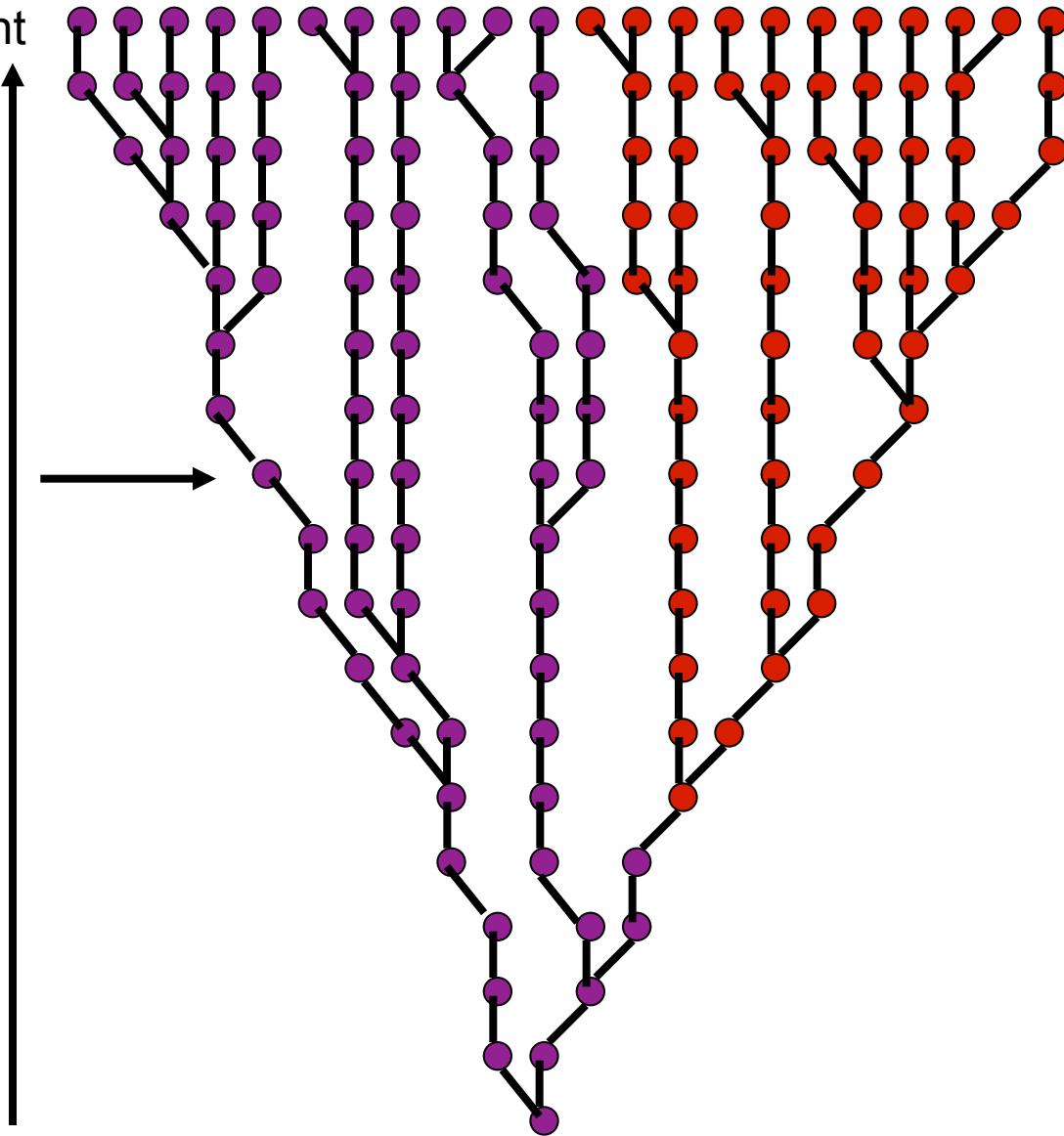
TCGAGGTATTAAC

TCTAGGTATTAAC

mutation

Present

Time



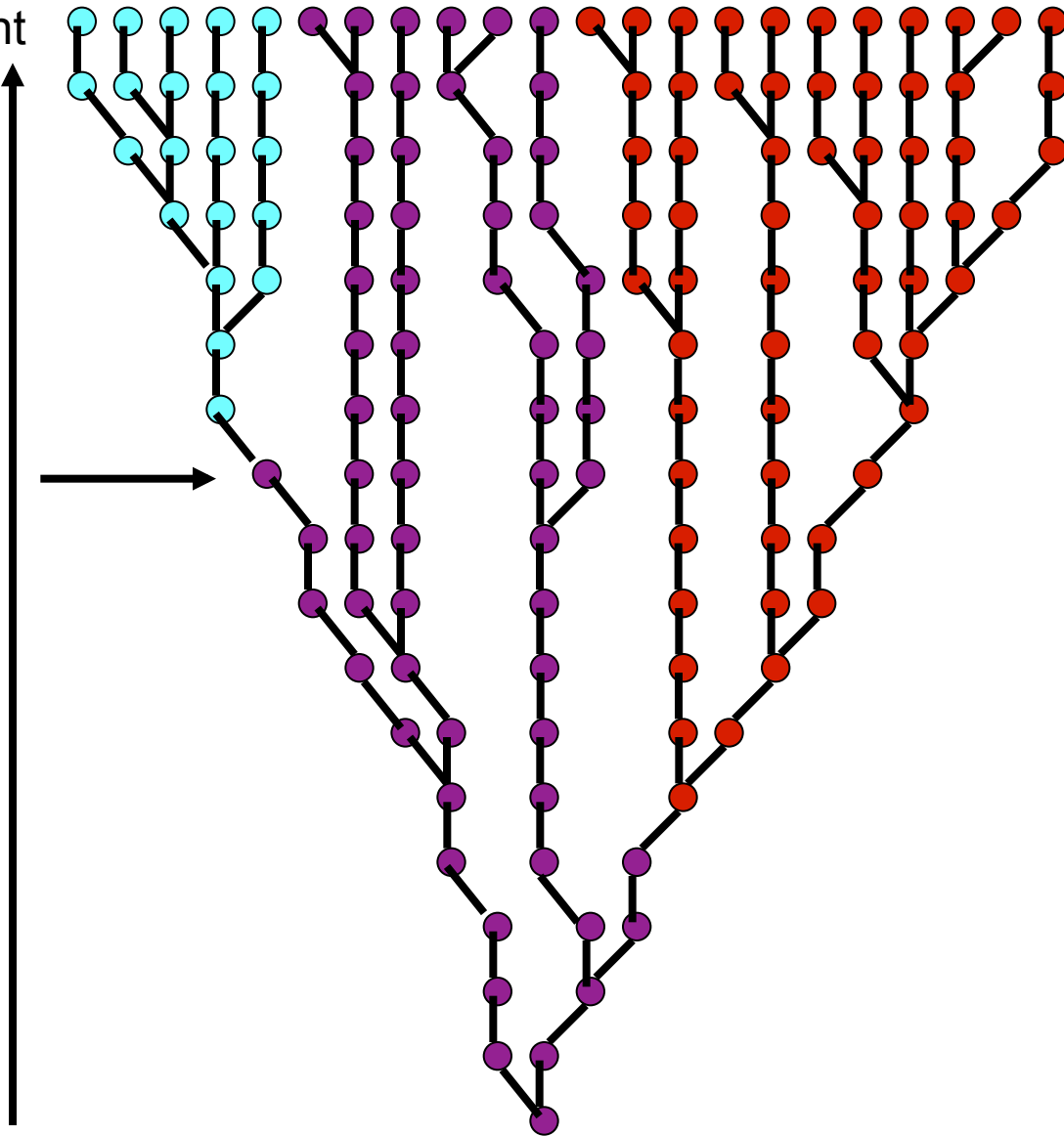
TCGAGGTATTAAC

TCTAGGTATTAAC

C

Present

Time



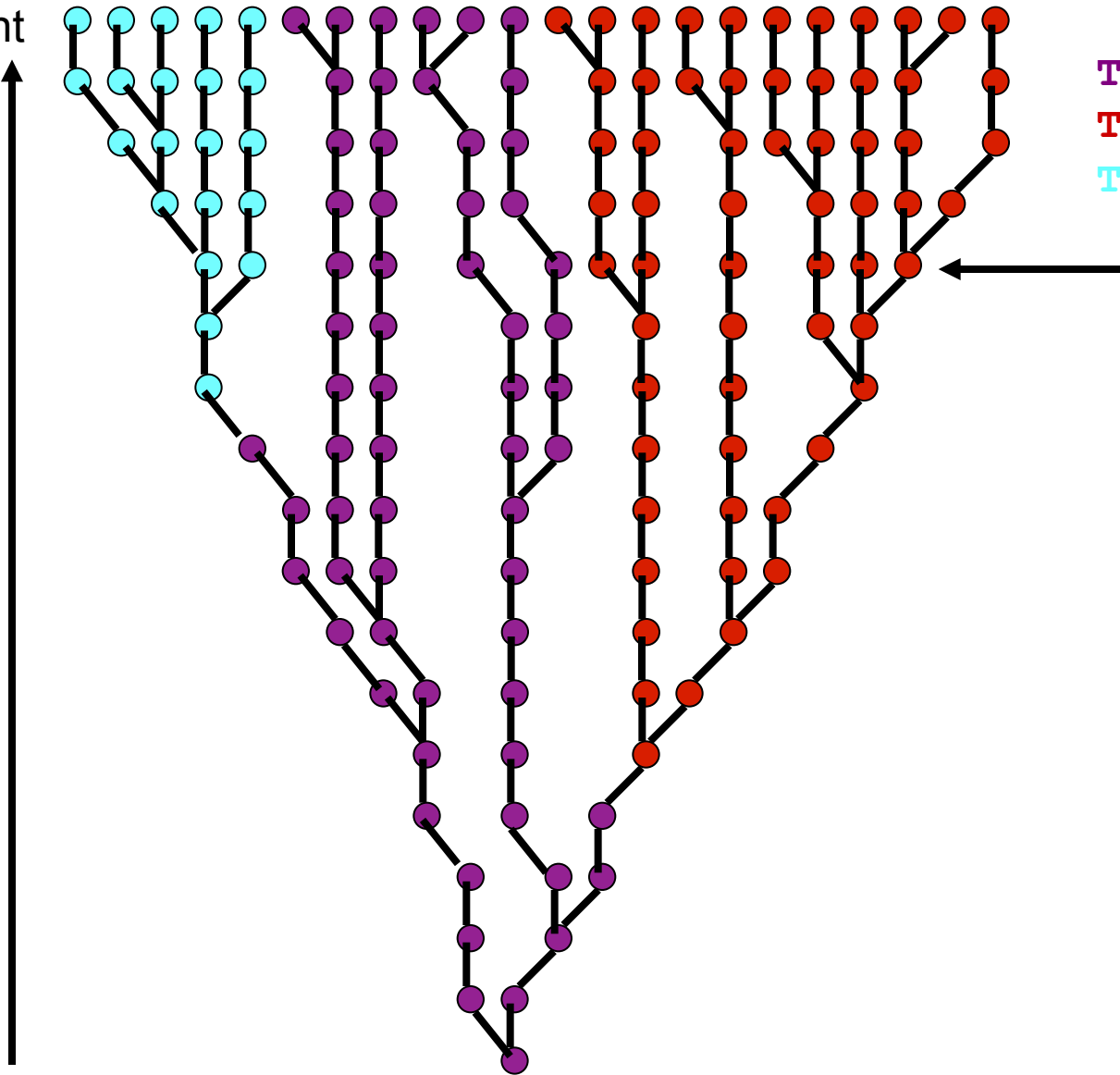
TCGAGGTATTAAC

TCTAGGTATTAAC

TCGAGGCATTAAC

Present

Time



TCGAGGTATTAAC

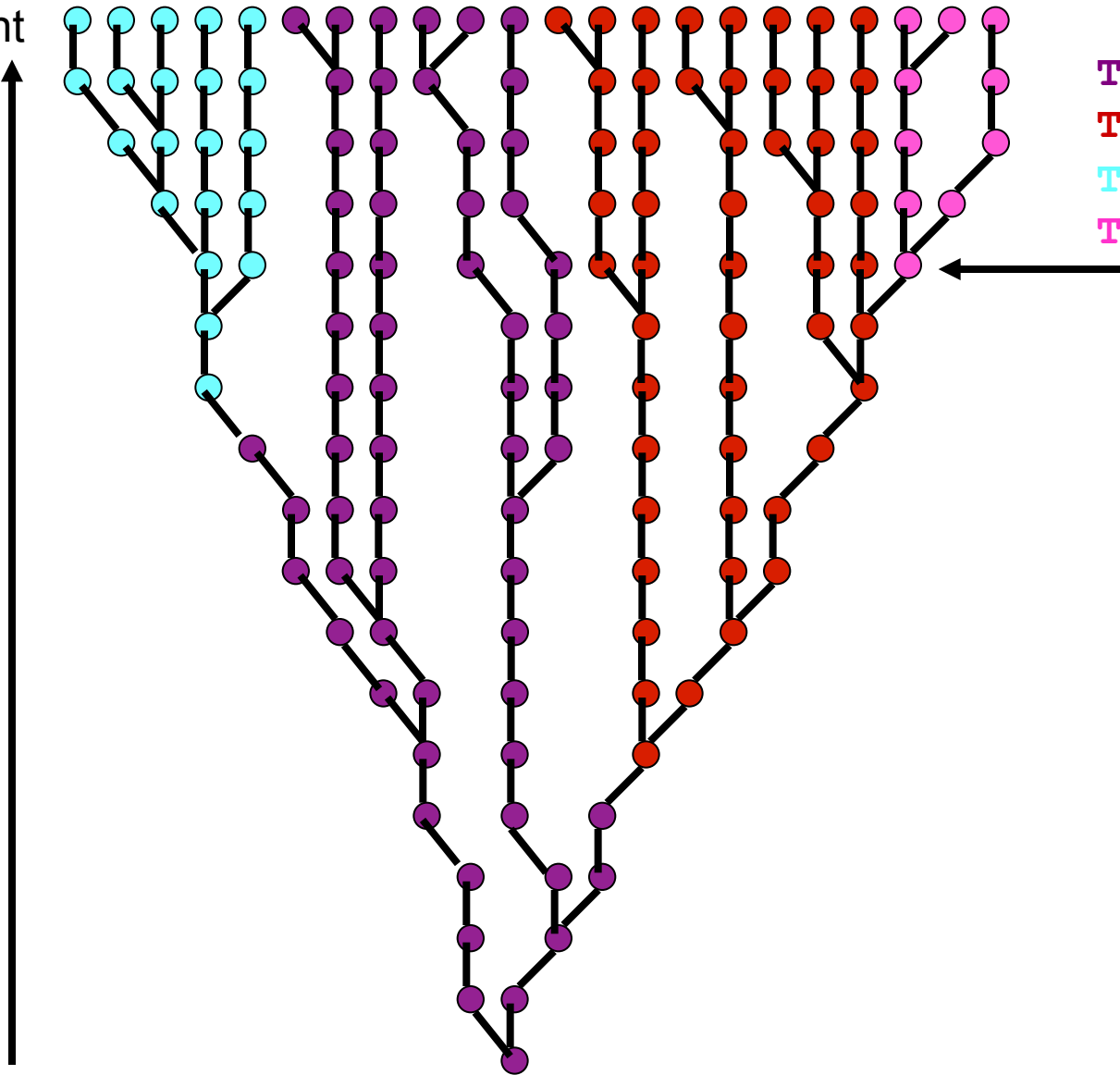
TCTAGGTATTAAC

TCGAGGCATTAAC

G

Present

Time



TCGAGGTATTAAC

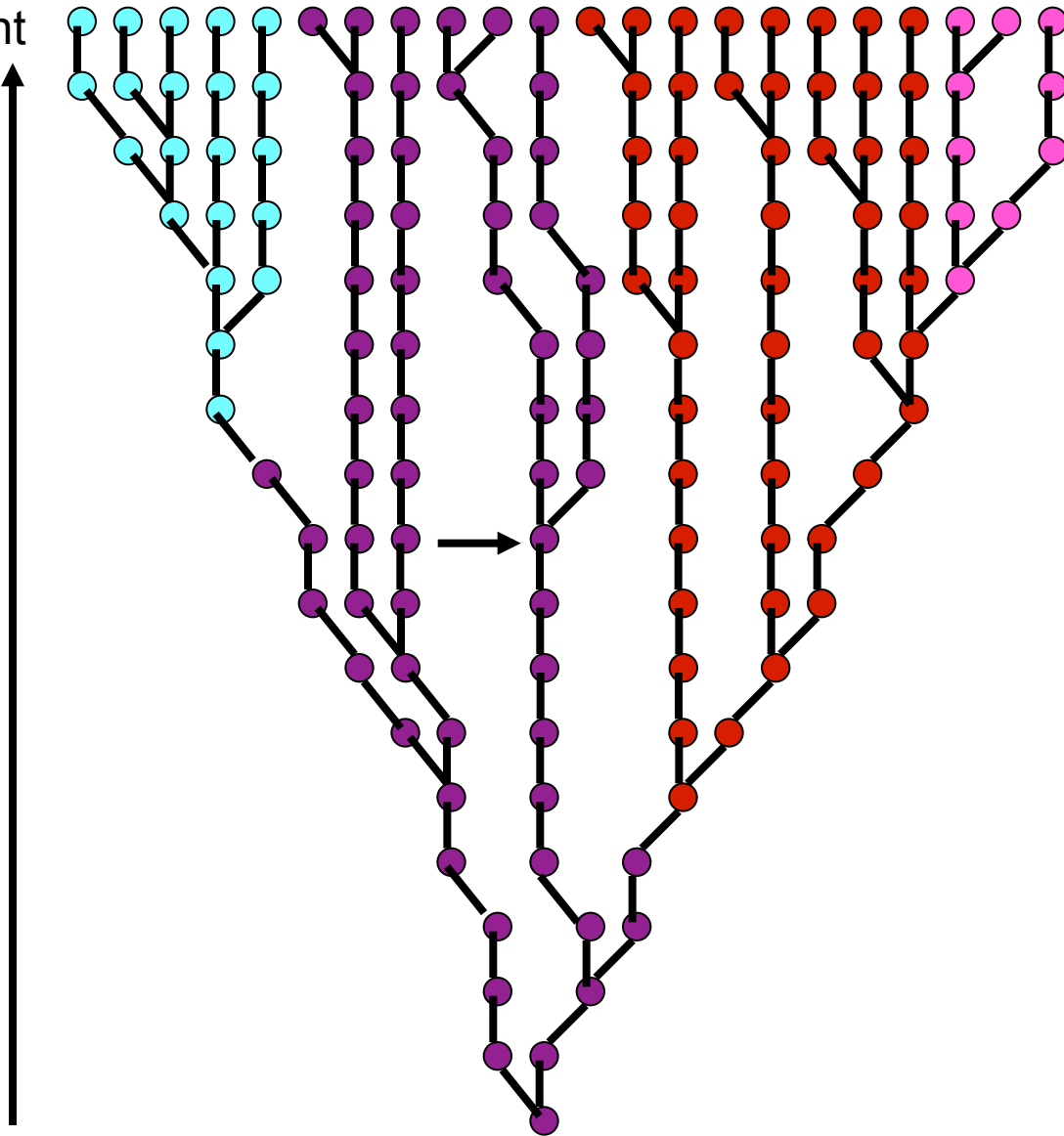
TCTAGGTATTAAC

TCGAGGCATTAAC

TCTAGGTGTTAAC

Present

Time



TCGAGGTATTAAC

TCTAGGTATTAAC

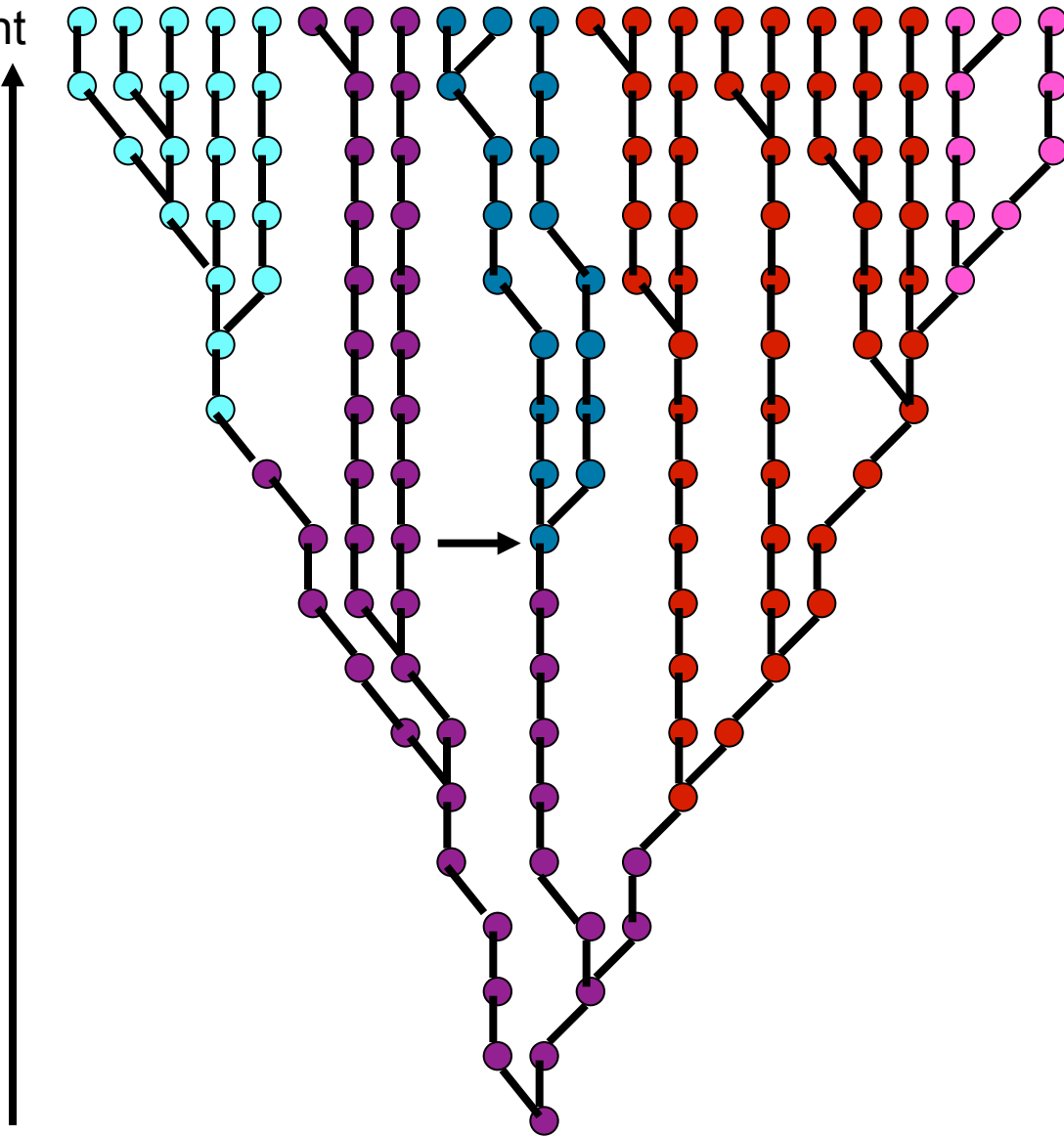
TCGAGGCATTAAC

TCTAGGTGTTAAC

G

Present

Time



TCGAGGTATTAAC

TCTAGGTATTAAC

TCGAGGCATTAAC

TCTAGGTGTTAAC

TCGAGGTATTAGC

Present

Time



TCGAGGTATTAAC

TCTAGGTATTAAC

TCGAGGCATTAAC

TCTAGGTGTTAAC

TCGAGGTATTAGC

C

Present

Time



TCGAGGTATTAAC

TCTAGGTATTAAC

TCGAGGCATTAAC

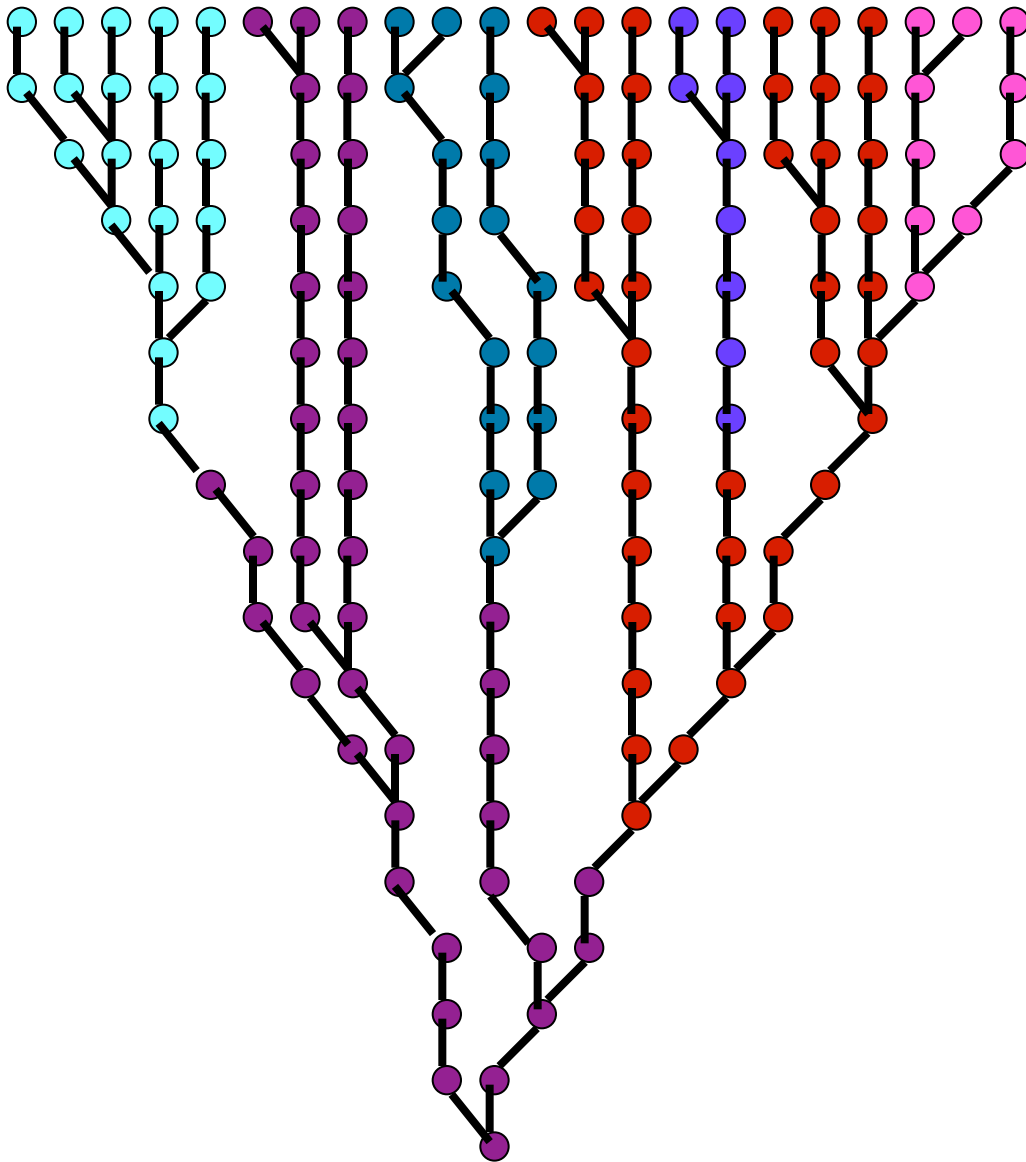
TCTAGGTGTTAAC

TCGAGGTATTAGC

TCTAGGTATCAAC

Present

Time



TCGAGGTATTAAC

TCTAGGTATTAAC

TCGAGGCATTAAC

TCTAGGTGTTAAC

TCGAGGTATTAGC

TCTAGGTATCAAC

*

**

*

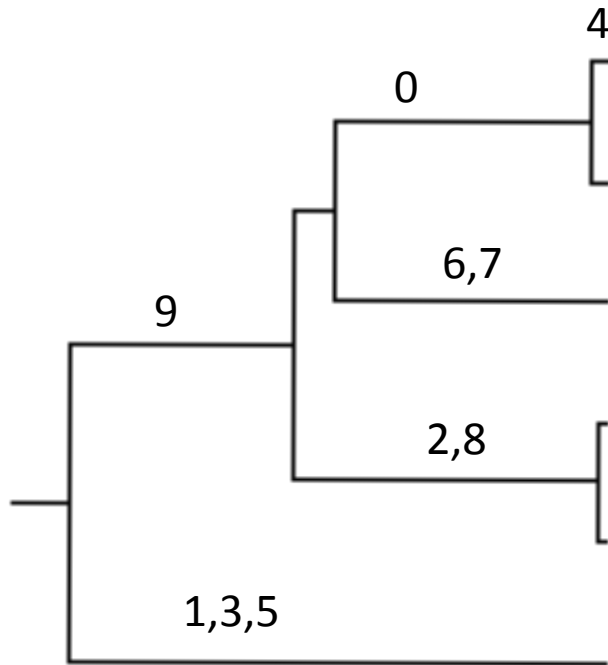
*

Simulating population data

- Mutations just accumulate along the branches of the tree according to a Poisson process with rate $\lambda = \mu t$ for a branch of length t .
- The Poisson process is stochastic but it should be immediately obvious that long branches will carry more mutations than short branches

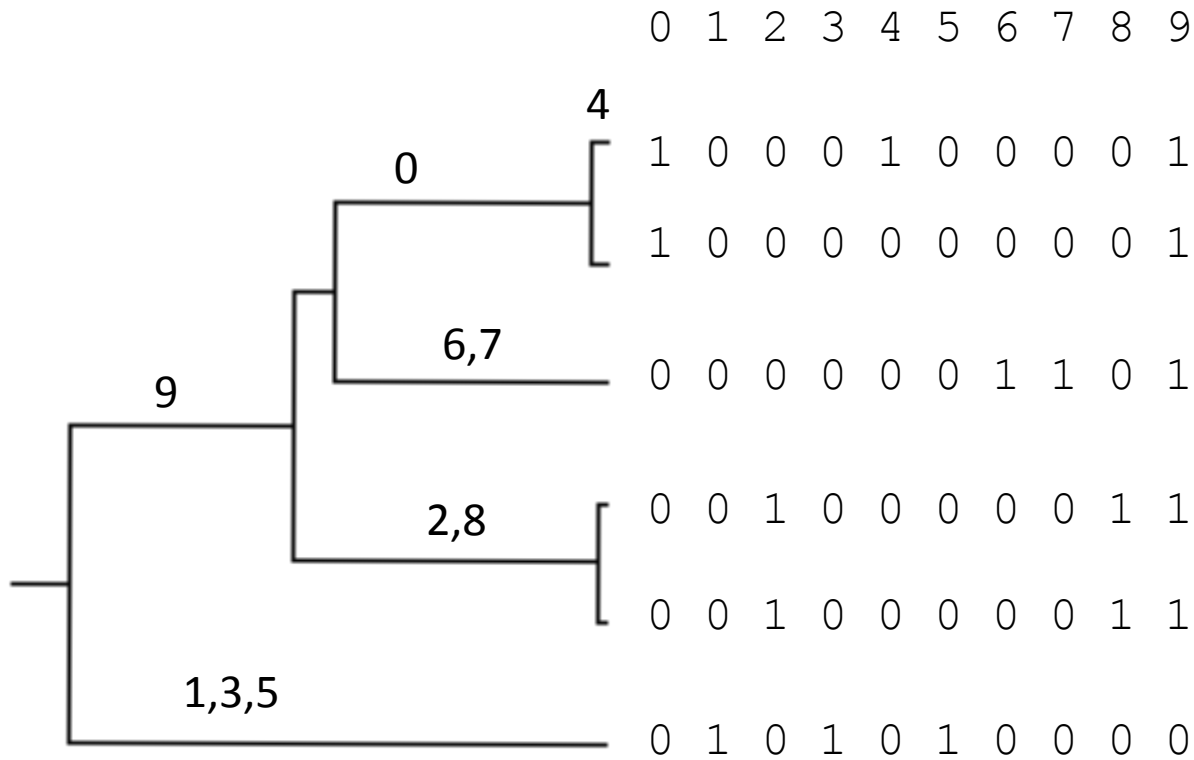
Simulating population data

- Generate a coalescent (Topology + Branch lengths)
- For each branch length t , drop mutations with rate μt
- Based on infinite sites, each mutation is at a unique location



Simulating population data

- Generate Sequences



Demo

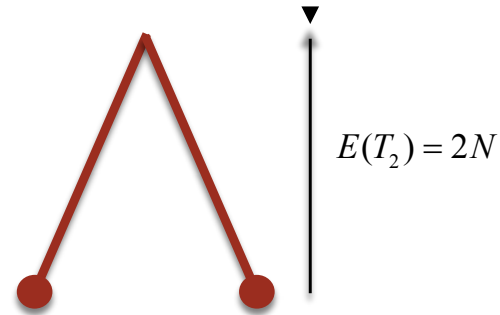
- ms, fastsimcoal, Figtree, Arlequin input file

```
./ms 6 1 -s 10 -T
```

```
./fastsimcoal -i 1PopDNA_sta.par -n 10 -T
```

Average number of pairwise differences (π)

- Since the expected coalescent time between a pair of gene is $2N$ generations, the average number of mutations expected between a pair of genes (also called the average number of pairwise differences under the infinite site model) is:



$$E(\pi | N, \mu) = 2 \times E(T_2 \times \mu) = 2\mu \times 2N = 4N\mu = \theta$$

- This shows that coalescent theory provides a very powerful way to obtain classical population genetics results.

Number of segregating sites (S)

- It is very simple to derive the expected number of segregating (polymorphic) sites S in a sample of size n under the infinite site model as:

$$E(S | n, N, \mu) = E(T_{total} \times \mu) = \mu E(T_{total}) = 4N\mu \sum_{i=1}^{n-1} \frac{1}{i} = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

- A result that was originally obtained by Ewens (1974) and Watterson (1975) using much more complex approaches based on classical forward population genetics.

Demo

- Fastsimcoal, arlsumstat, R

```
./fastsimcoal -i 1PopDNA_sta.par -n 100  
./LaunchArlsumstatDirMac.sh 1PopDNA_sta SettingsDNASStats.ars stats.txt
```

```
stats=read.table("1PopDNA_sta/stats.txt",header=T)
```

```
hist(stats$Pi_1)  
theta=2*20000*0.00000002*100000  
mean(stats$Pi_1)
```

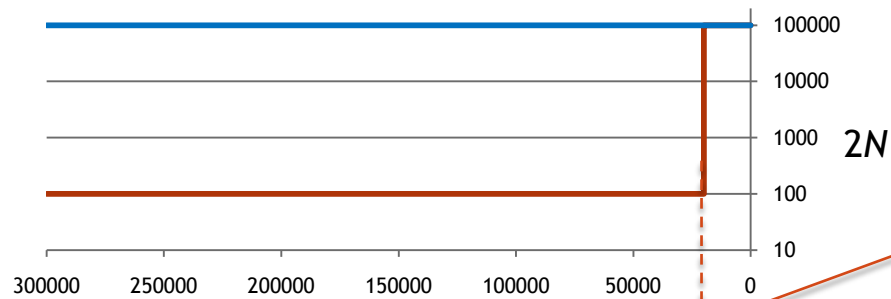
```
hist(stats$S_1)  
theta*sum(1/(1:9))  
mean(stats$S_1,br=20)
```

Variable population size

- Intuitively, coalescent events will tend to be rare when the population size is large and frequent when it is small.
- Actually population size changes only require a rescaling of branch lengths and have no effect on the topology of the tree.
- Assuming that current population size is N_0 and that t generations ago it was $N(t) = N_0\lambda(t)$, then a branch generated under a coalescent process occurring at rate N_0 between times t_1 and t_2 should just be rescaled by a factor:

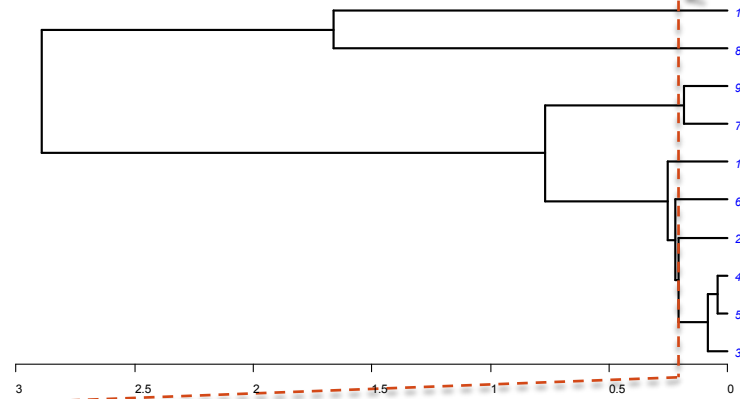
$$\Lambda = \int_{t_1}^{t_2} \frac{1}{\lambda(s)} ds$$

Past demographic expansion

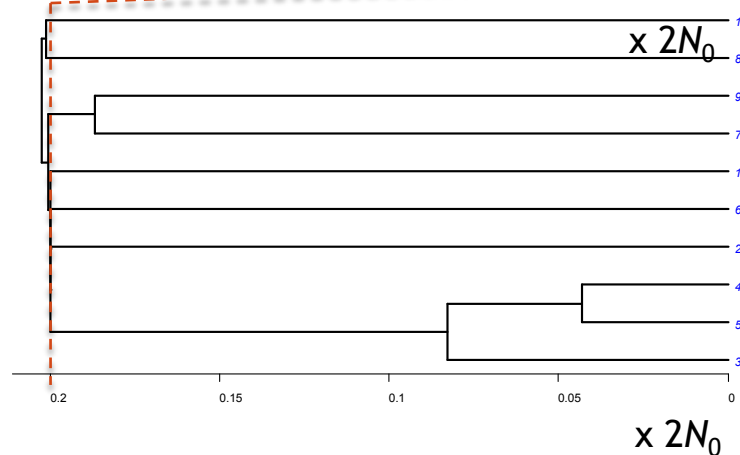


Population size
change
 $0.2 \times 2N$
generations ago

Coalescent in a
stationary
population

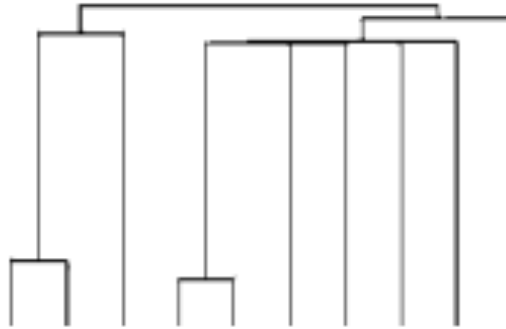
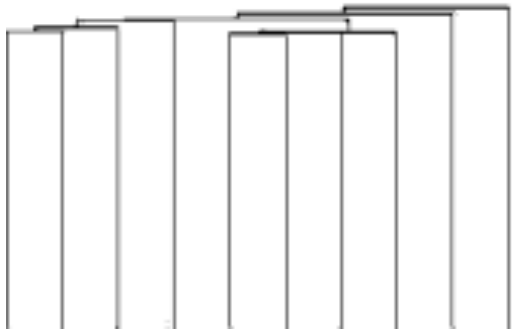
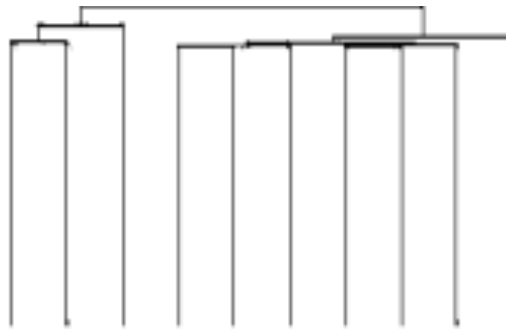
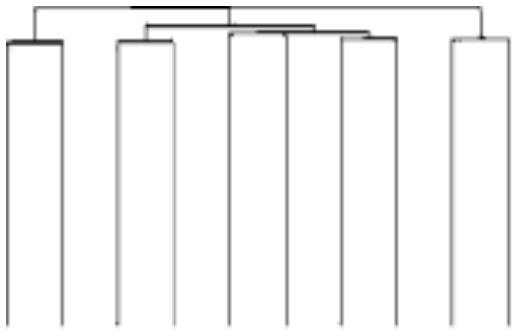


Rescaled coalescent



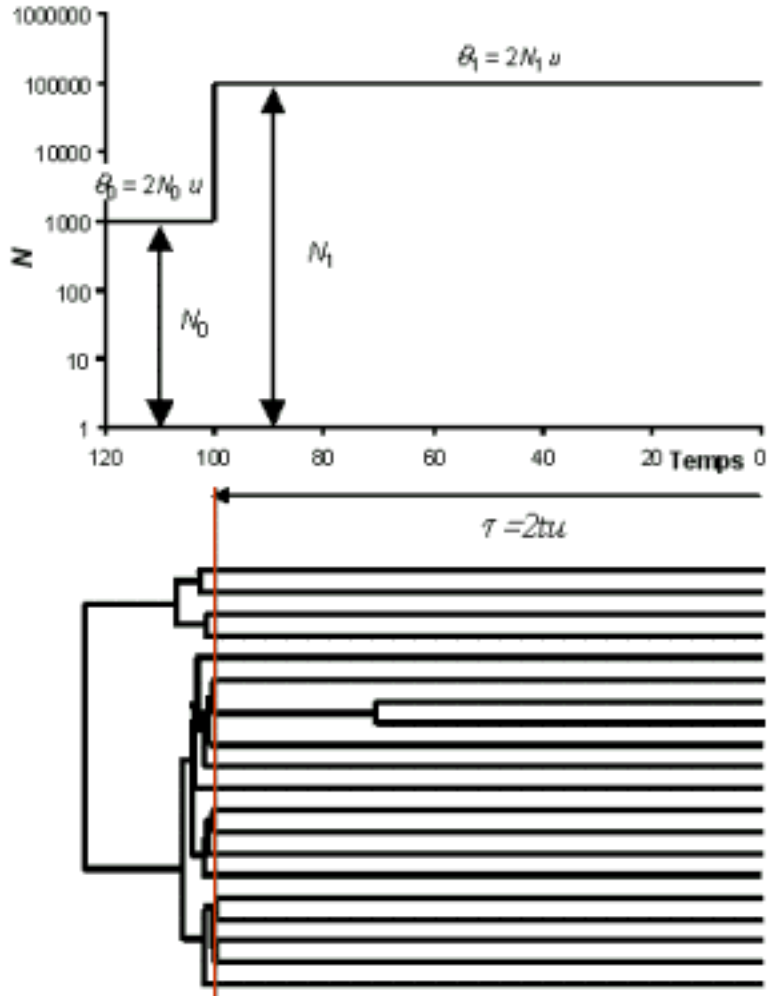
Note the long
terminal branch
lengths after a
population
expansion

Past demographic expansion

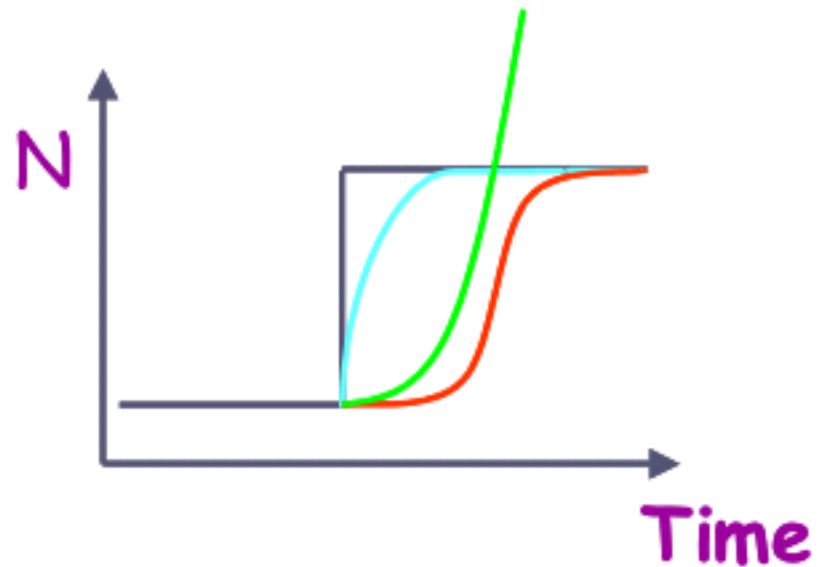


Genealogies in expanding populations have usually short internal branch lengths and long external branch lengths. Comb-like or star-like genealogies

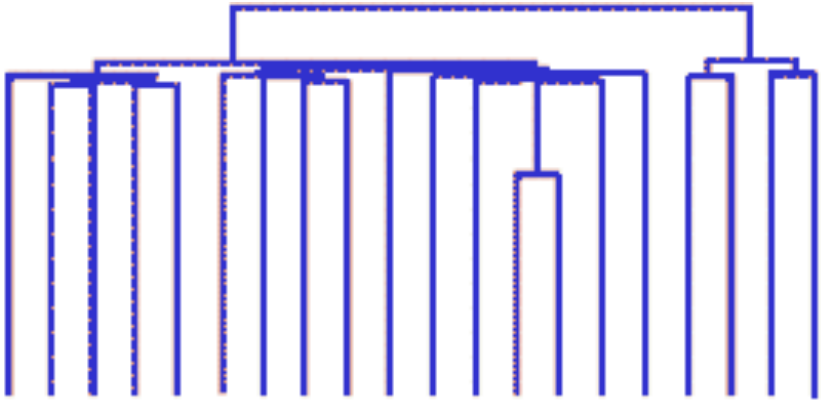
The effect of a sudden expansion



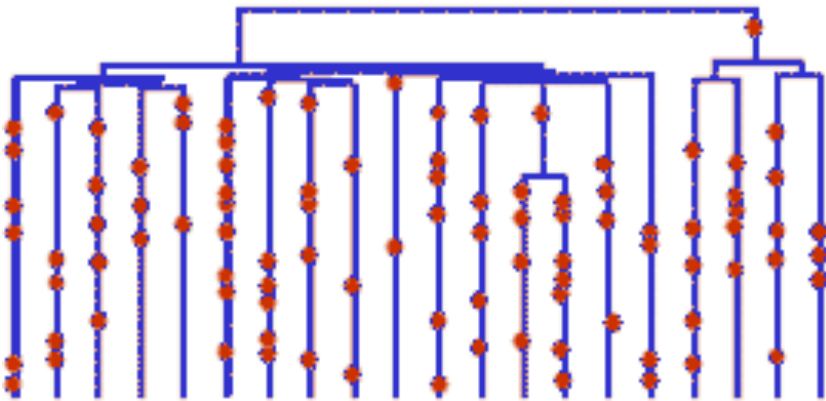
Coalescent events are very unlikely in large populations, but much more likely in small populations



Mutations in expanding populations

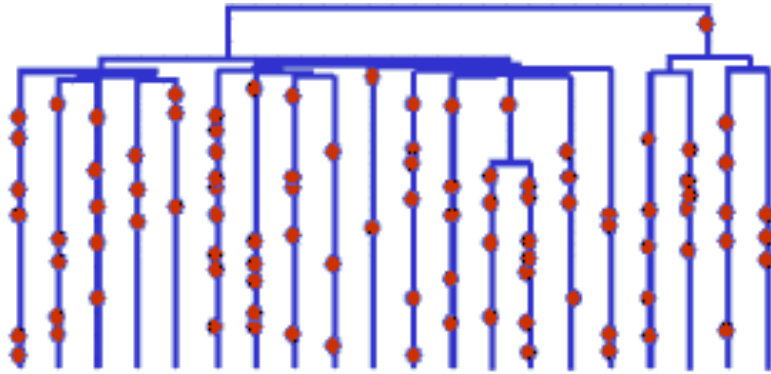


↓ *Mutations*

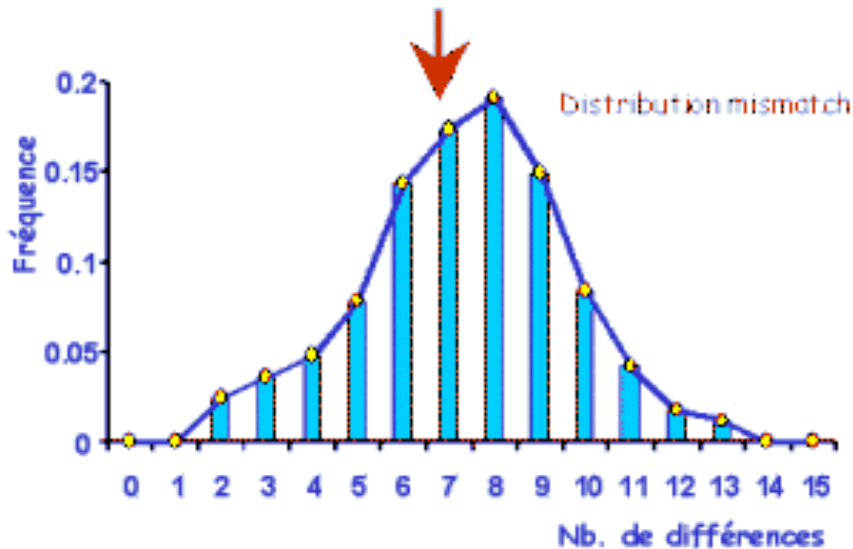


Mutations will accumulate preferentially after the expansion

Mismatch distribution

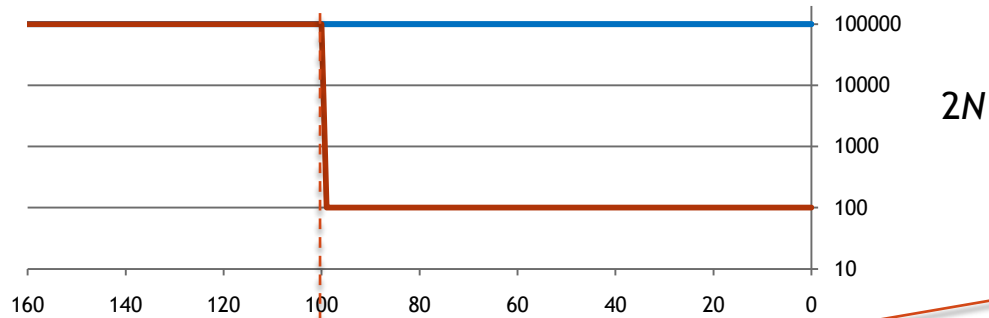


The molecular diversity of a sample may be summarized by plotting the distribution of the number of pairwise differences between genes



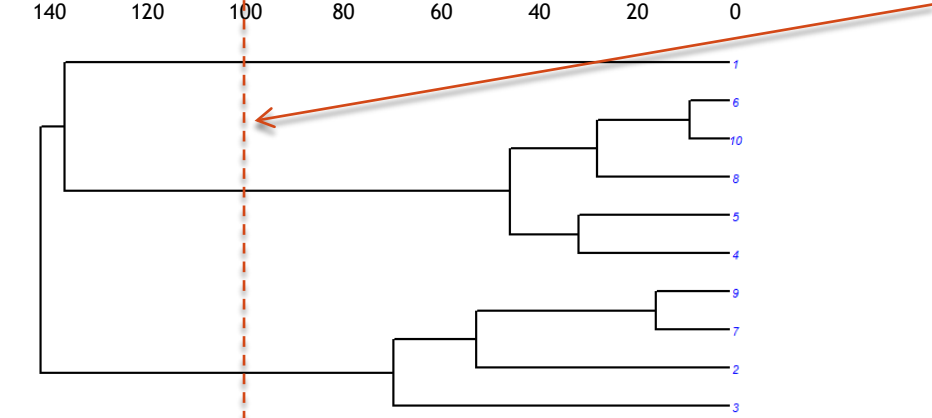
This distribution is often called the **mismatch distribution**

Past demographic contraction

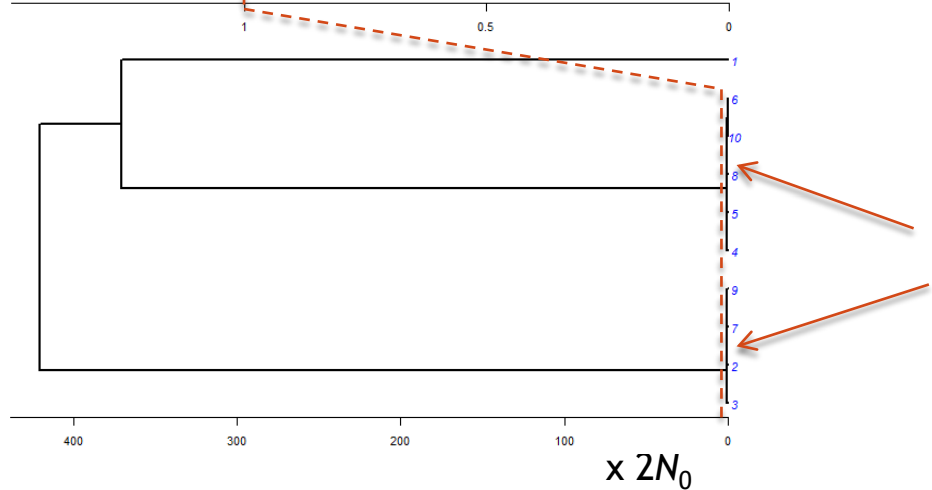


Population size change
 $2N$ (100)
generations ago

Coalescent in a stationary population

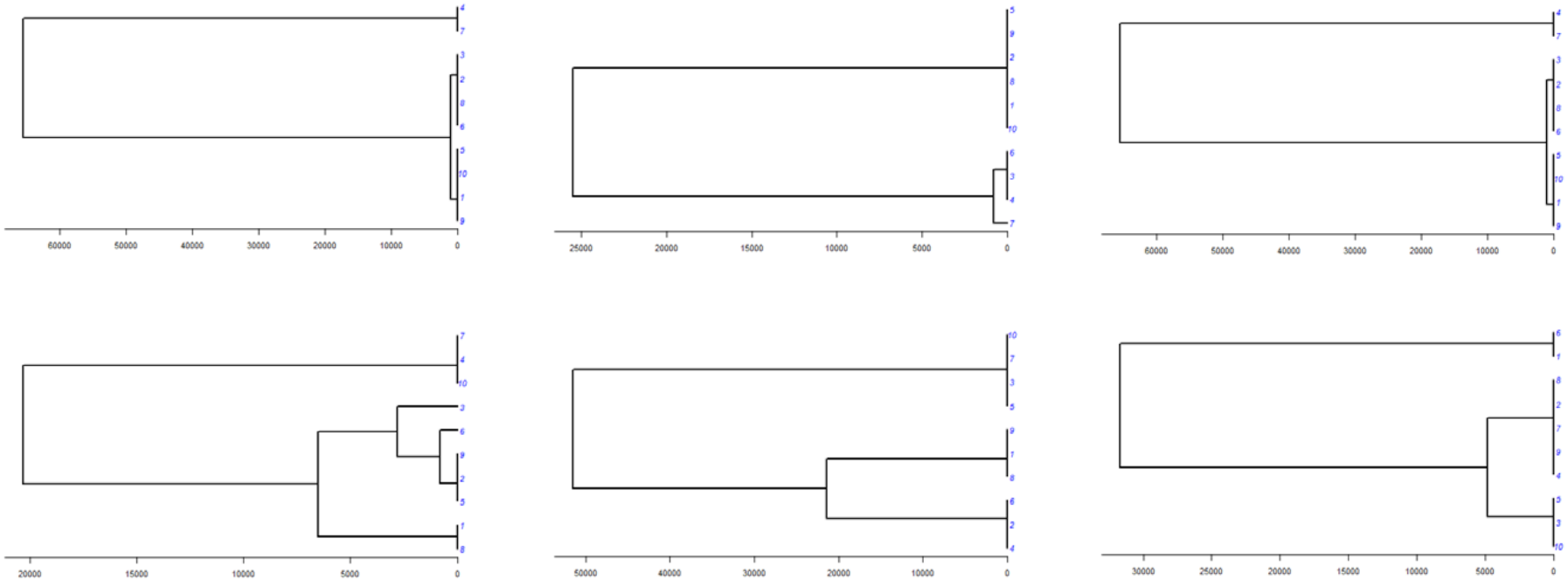


Rescaled coalescent



Note the long internal branches and tiny external branches after a population contraction

Past demographic contraction



Contractions often leads to the observation of deep lineages and two main clades

Demo

- Fastsimcoal, Figtree, arlsumstat, R

```
./fastsimcoal -i 1PopDNA_bot.par -n 100 -T
```

```
./fastsimcoal -i 1PopDNA_exp.par -n 100 -T
```

```
./LaunchArlsumstatDirMac.sh 1PopDNA_exp SettingsDNASTats.ars stats.txt
```

```
stats=read.table("1PopDNA_exp/stats.txt",header=T)
```

```
hist(stats$Pi_1)
```

Simulating the coalescent

- A big advantage of coalescent approaches is that they lead themselves to very efficient simulations, as compared to forward approaches.
- Advantages:
 - Speed
 - Small memory footprint
 - Direct simulation of the sample, no sub-sampling
 - No need to specify initial conditions (initial allele frequencies)
 - Easy to integrate into estimation procedures (ABC, likelihood-based...)
- Disadvantages:
 - Approximation (multiple coalescent events not allowed)
 - Difficult to simulate non-neutral diversity
 - Difficult to include realistic factors linked to life-history traits
 - Simulations involving recombination become tedious to program