# Introduction to population genetics:

# Basic terms and concepts

**Copenhagen PopGen Course, 2024**

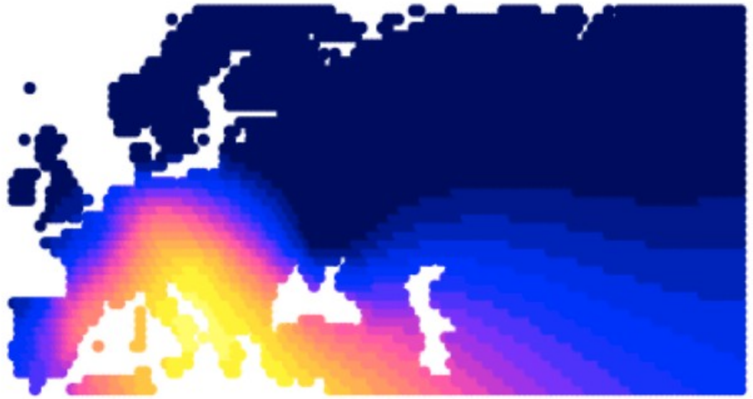
**Fernando Racimo**

# A bit about me and my research group

GLOBE Institute

## Paleogenomics and paleoproteomics

## Spatiotemporal modelling and paleoecology



## Academic activism

## Degrowth and socio-ecological transformation

# Beyond doing research and teaching, I am also a **scientist-activist**


If you’re interested in talking more about

**evolutionary genomics**

or **degrowth**

or **activism**

...feel free to come talk with me after lecture, or send an email to:
fracimo@sund.ku.dk

Always happy to chat about these topics!

# Introduction to population genetics:

# Basic terms and concepts

**Copenhagen PopGen Course, 2024**

**Fernando Racimo**

# Today

# Today

- Terminology

# Today

- Terminology
- Wright-Fisher Model and Genetic Drift

# Today

- Terminology
- Wright-Fisher Model and Genetic Drift
- Intro to coalescent theory

# Today

- Terminology
- Wright-Fisher Model and Genetic Drift
- Intro to coalescent theory
- Mutations and the coalescent

# Today

- Terminology
- Wright-Fisher Model and Genetic Drift
- Intro to coalescent theory
- Mutations and the coalescent

# Terminology

- **Locus**: a specific “spot” in the genome (could be a single base-pair or an entire gene or region, depending on context)

# Terminology

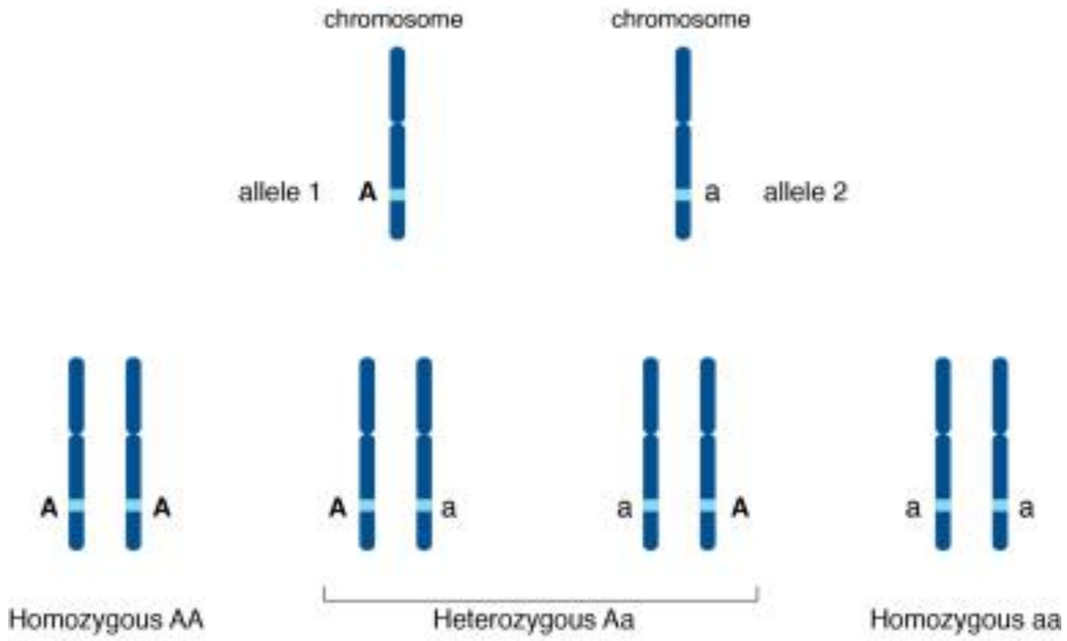
- **Locus**: a specific “spot” in the genome (could be a single base-pair or an entire gene or region, depending on context)
- **Allele**: 1 of the alternative forms of a locus that exist in a population

# Terminology

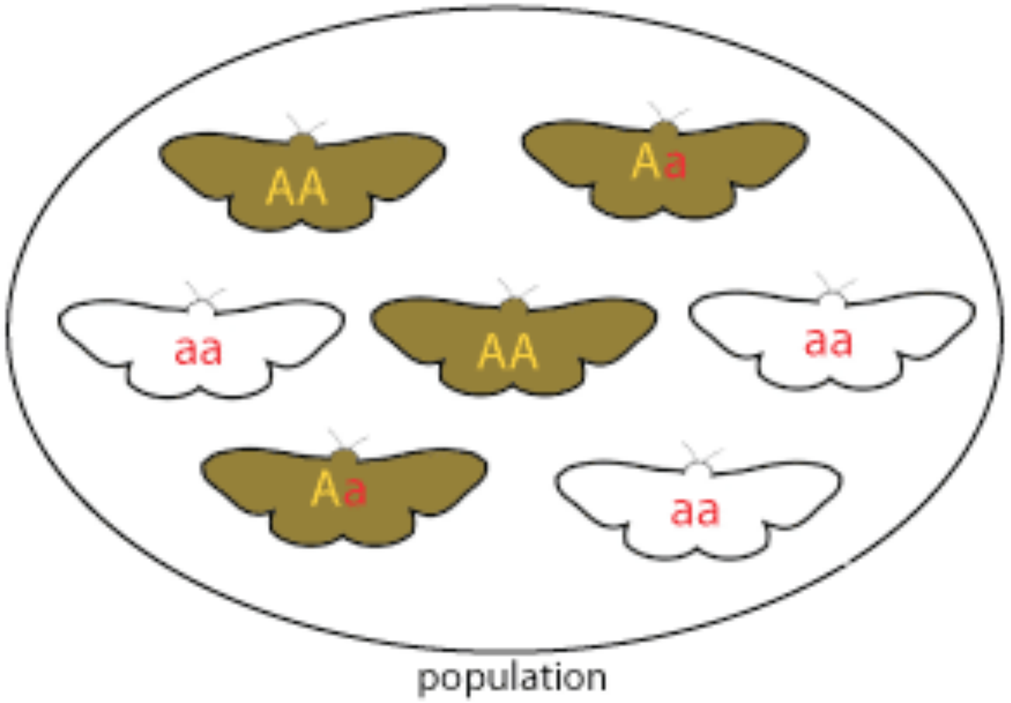
- **Locus**: a specific “spot” in the genome (could be a single base-pair or an entire gene or region, depending on context)
- **Allele**: 1 of the alternative forms of a locus that exist in a population
- **Polymorphism:** a locus with 2 or more alleles segregating in a population

# Terminology

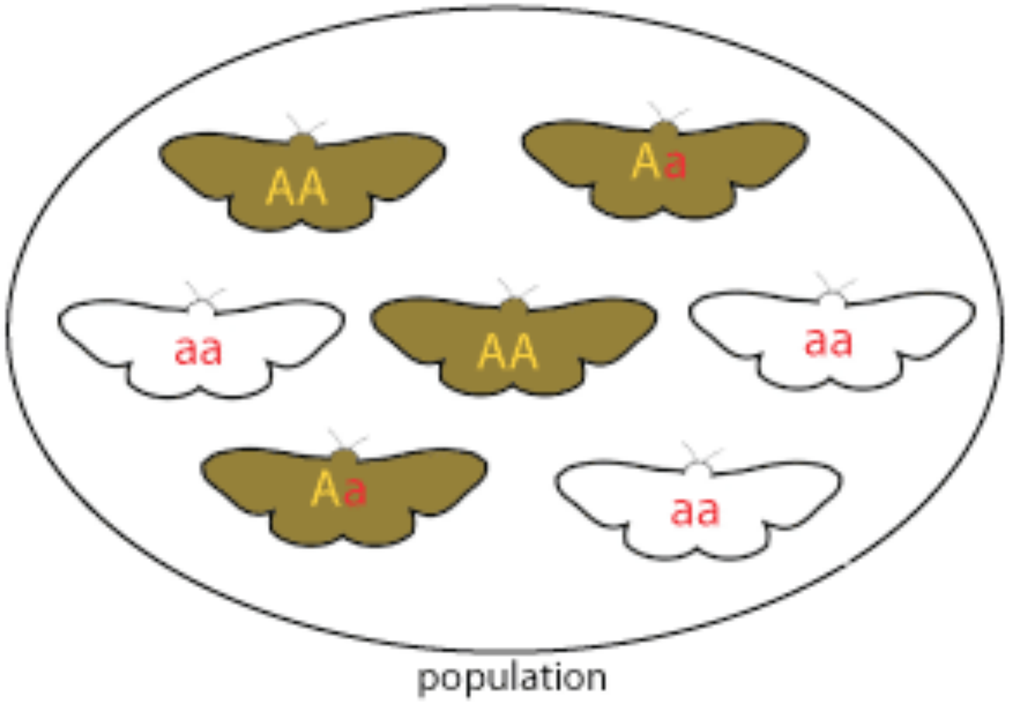
- **Locus**: a specific “spot” in the genome (could be a single base-pair or an entire gene or region, depending on context)
- **Allele**: 1 of the alternative forms of a locus that exist in a population
- **Polymorphism:** a locus with 2 or more alleles segregating in a population
- **Genotype**: the set of alleles present at a locus in a particular organism ( set size = 2 if the organism is diploid)



# Terminology



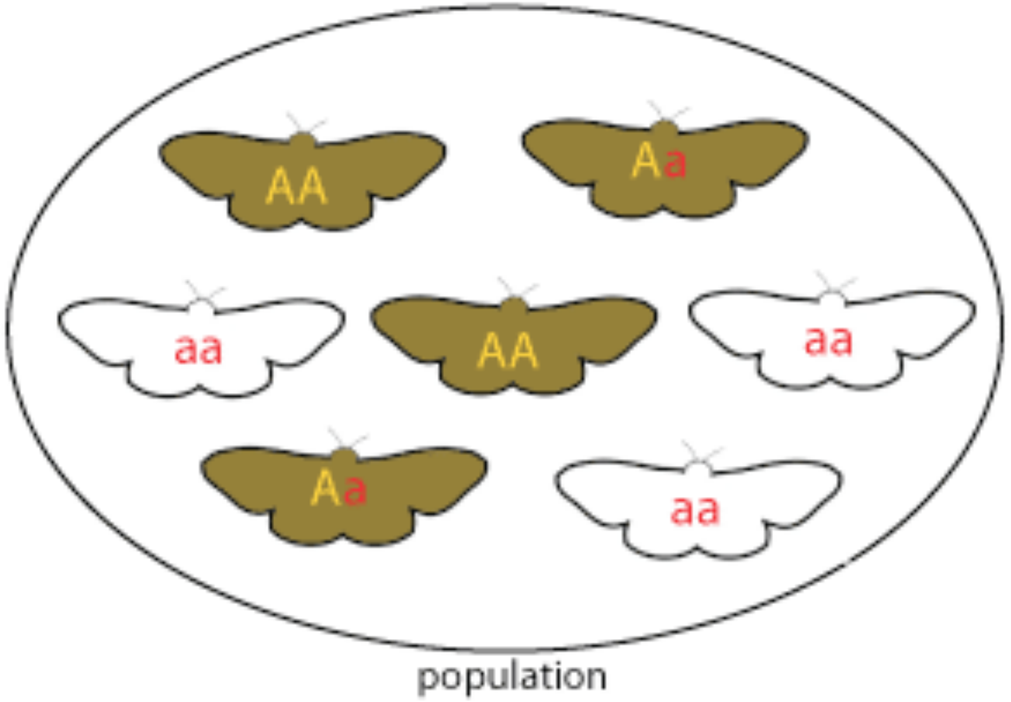
# Terminology




## Allele frequencies

- f(A) = 6/14
- f(a) = 8/14

# Terminology



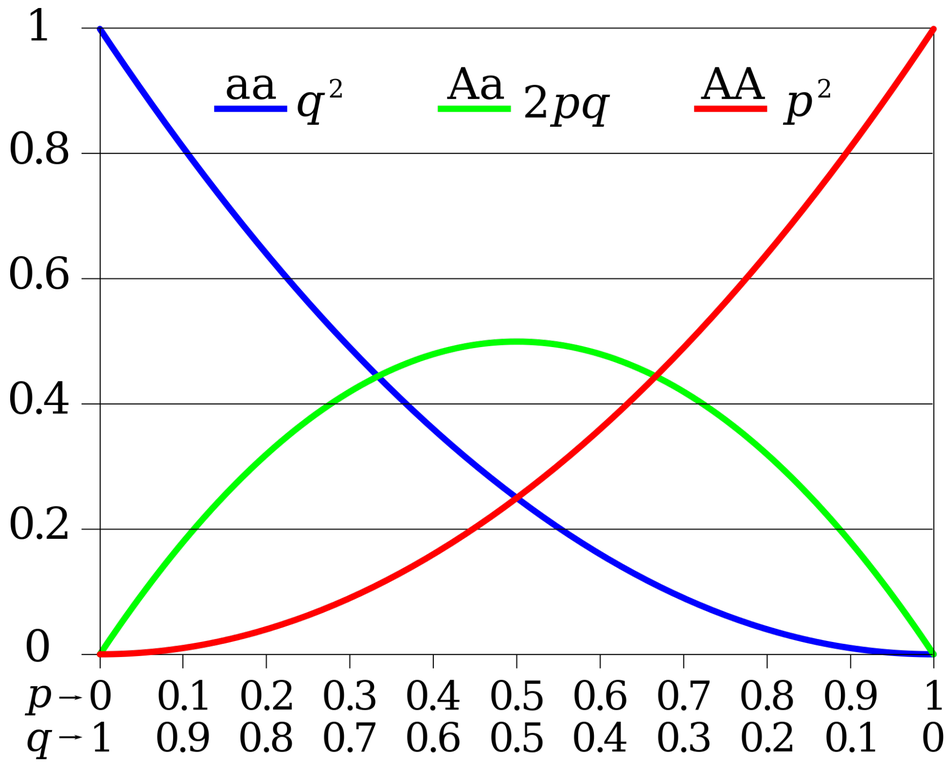

## Allele frequencies

- f(A) = 6/14
- f(a) = 8/14

## Genotype frequencies

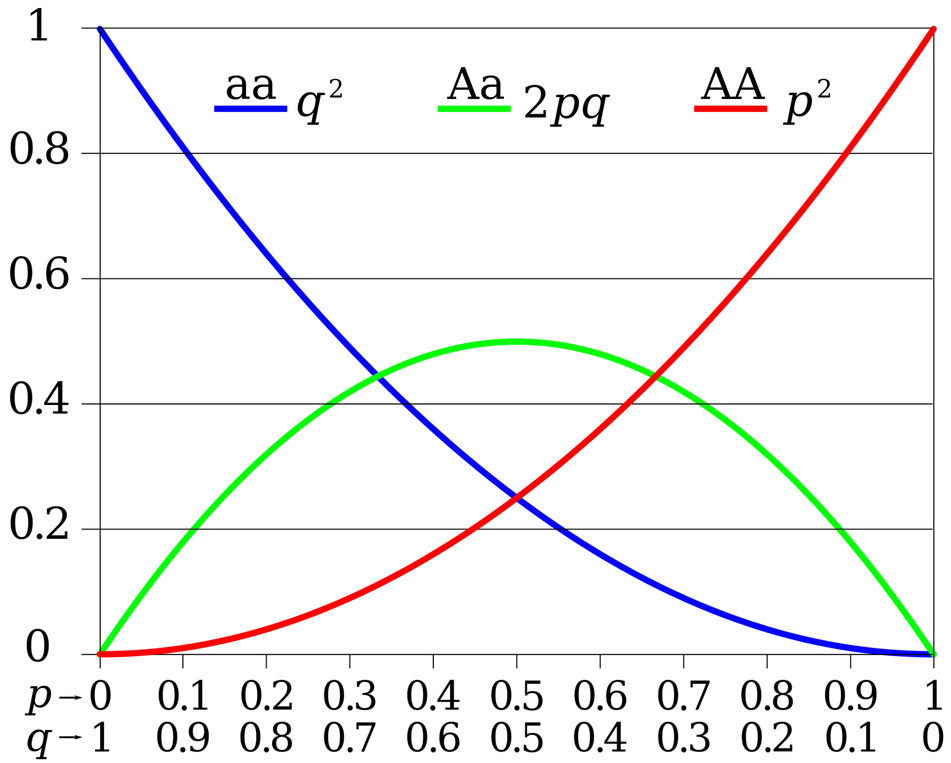
- f(AA) = 2/7
- f(Aa) = 2/7
- f(aa) = 3/7

# Hardy-Weinberg Equilibrium




**Allele frequencies** → **Genotype frequencies**

# Hardy-Weinberg Equilibrium

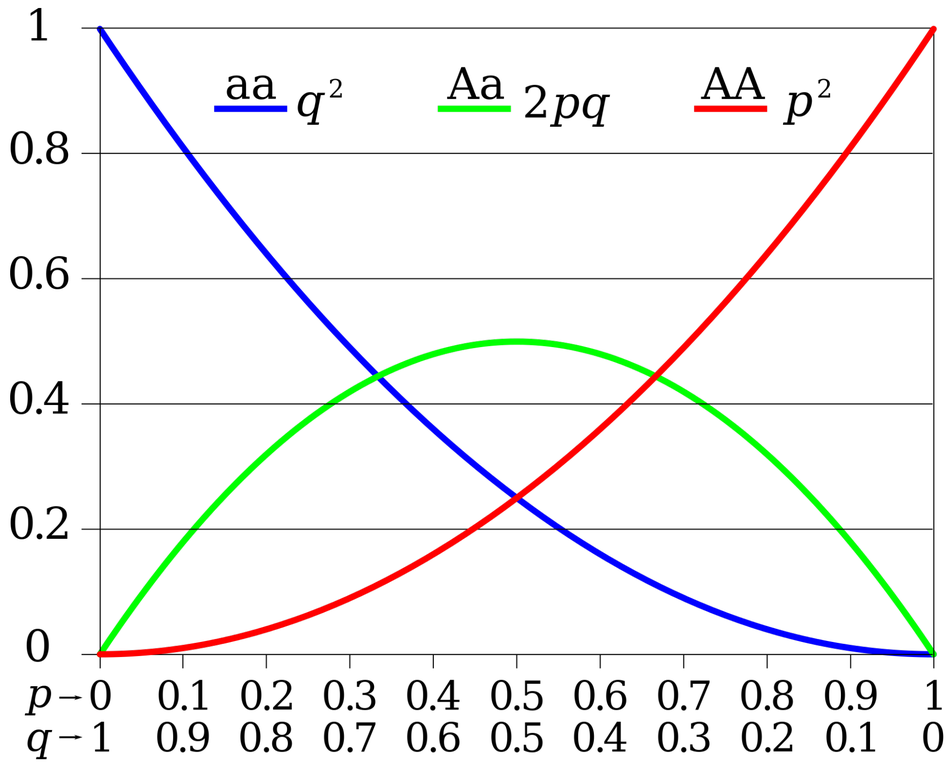



**Assuming:**

- Infinite population size
- No migration
- No mutation
- No selection
- Random mating
- Non-overlapping generations

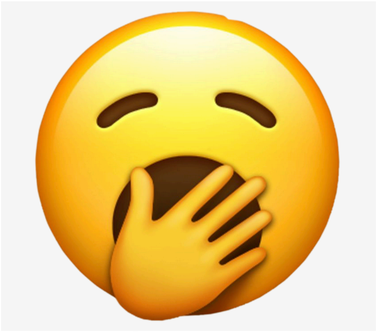
**Allele frequencies** → **Genotype frequencies**

# Hardy-Weinberg Equilibrium



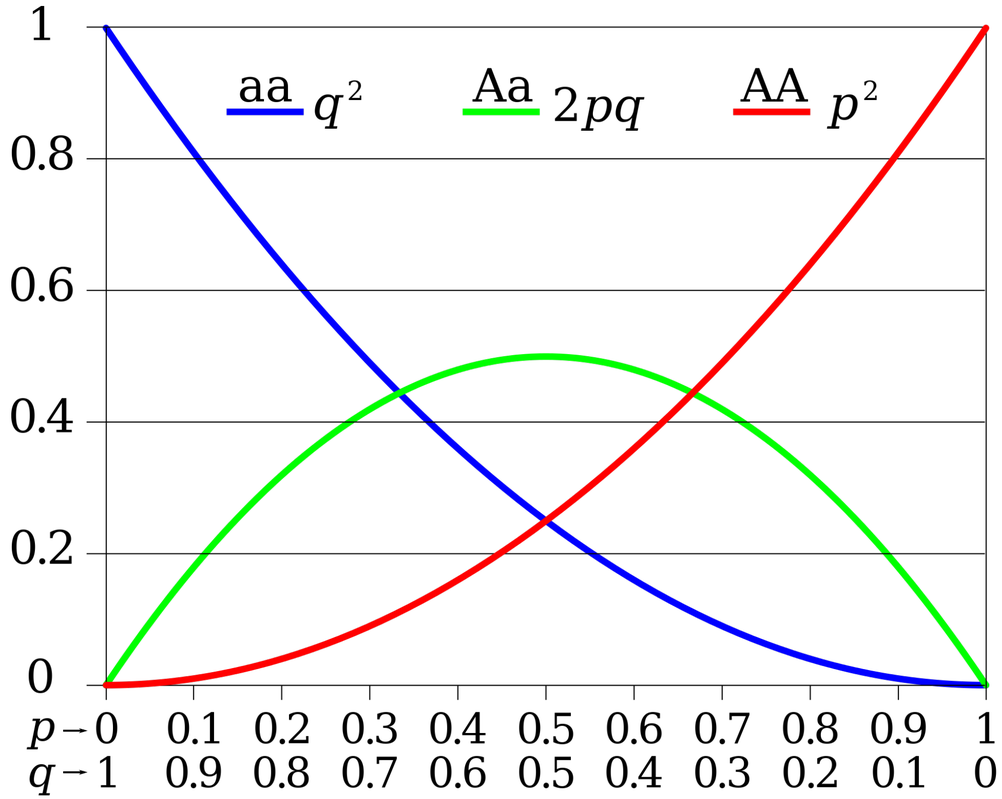

**Assuming:**

- Infinite population size
- No migration
- No mutation
- No selection
- Random mating
- Non-overlapping generations



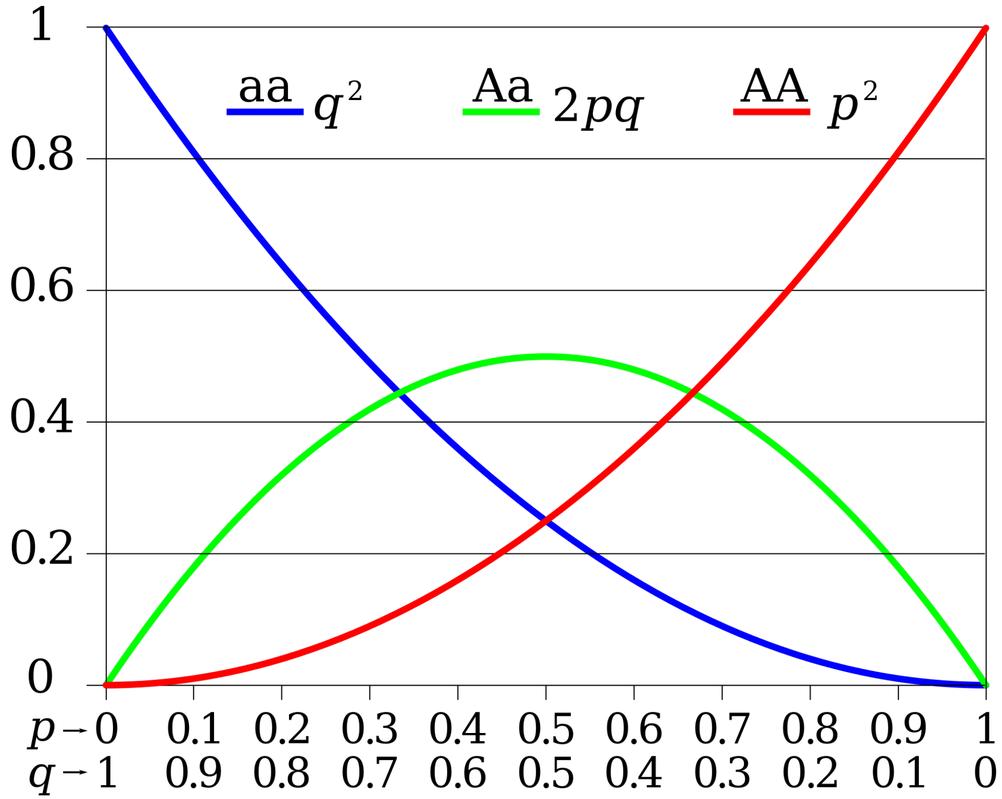
**Allele frequencies** → **Genotype frequencies**

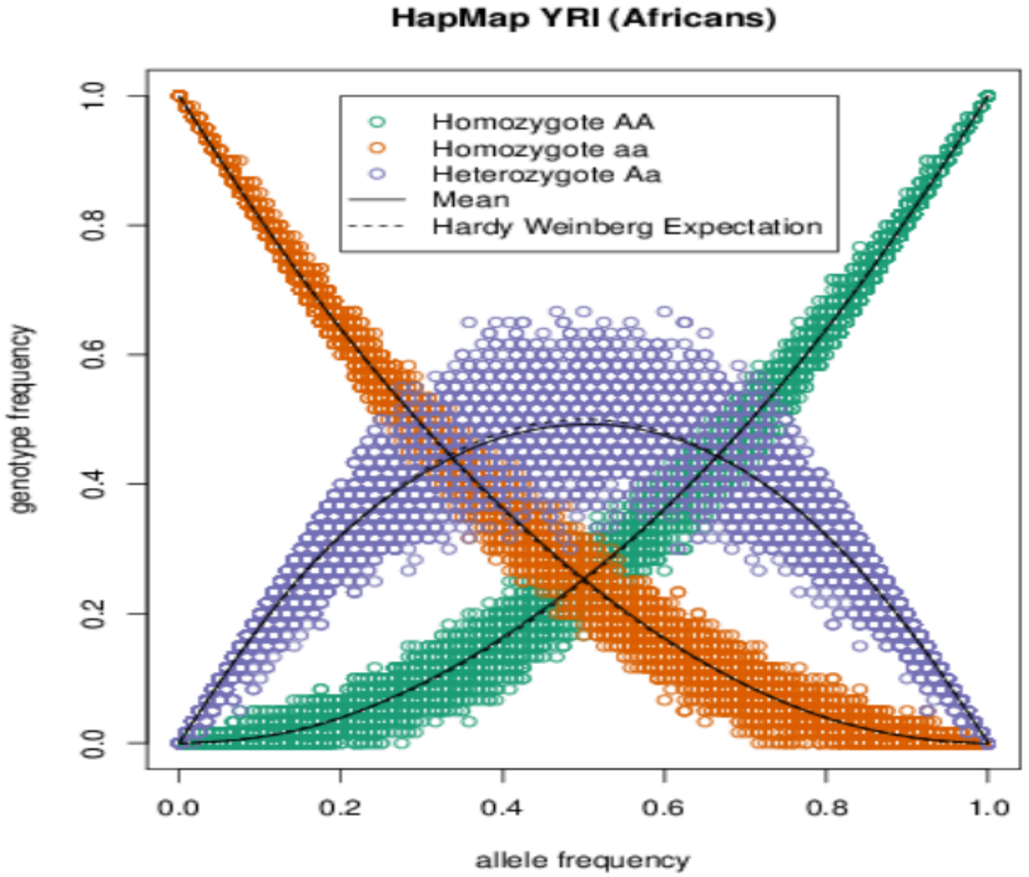
# That being said…




https://cooplab.github.io/popgen-notes/

# That being said…




https://cooplab.github.io/popgen-notes/

# Polymorphisms

# Polymorphisms

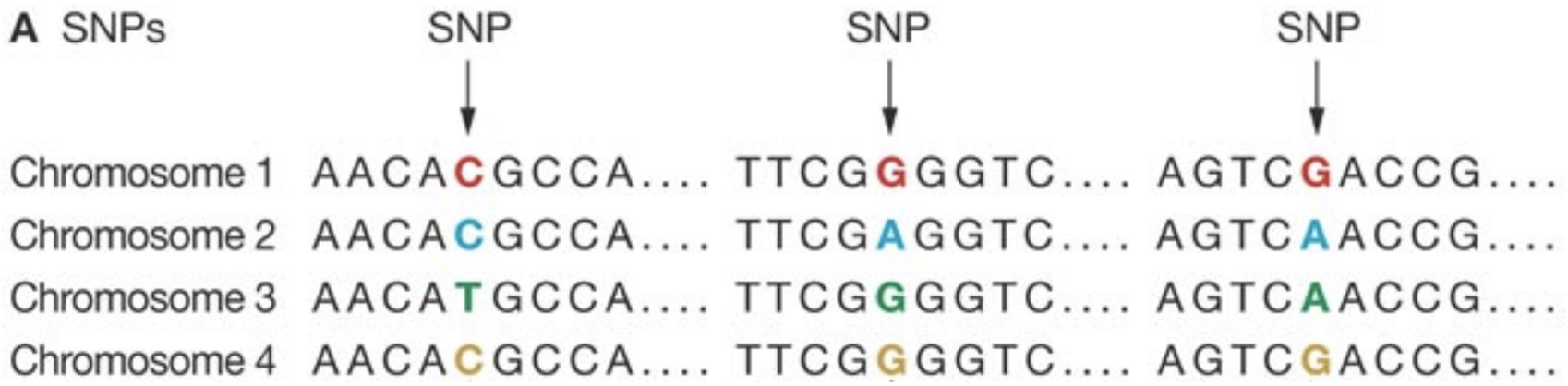
- Types of polymorphisms: insertions, deletions, translocations, copy number variants, etc.

# Polymorphisms

- Types of polymorphisms: insertions, deletions, translocations, copy number variants, etc.
- **SNP** (single-nucleotide polymorphism): a single nucleotide differs among members of the population

# Polymorphisms

- Types of polymorphisms: insertions, deletions, translocations, copy number variants, etc.
- **SNP** (single-nucleotide polymorphism): a single nucleotide differs among members of the population



# Linkage


Haplotype 1
A A G T A C G G T T C A G G C A
Haplotype 2
T T G C G C A A C A G T A A T A
Haplotype 3
A T C T G T G A T A ? T G G T G
Haplotype 4
T T C C G C G G T T ? A G A C A

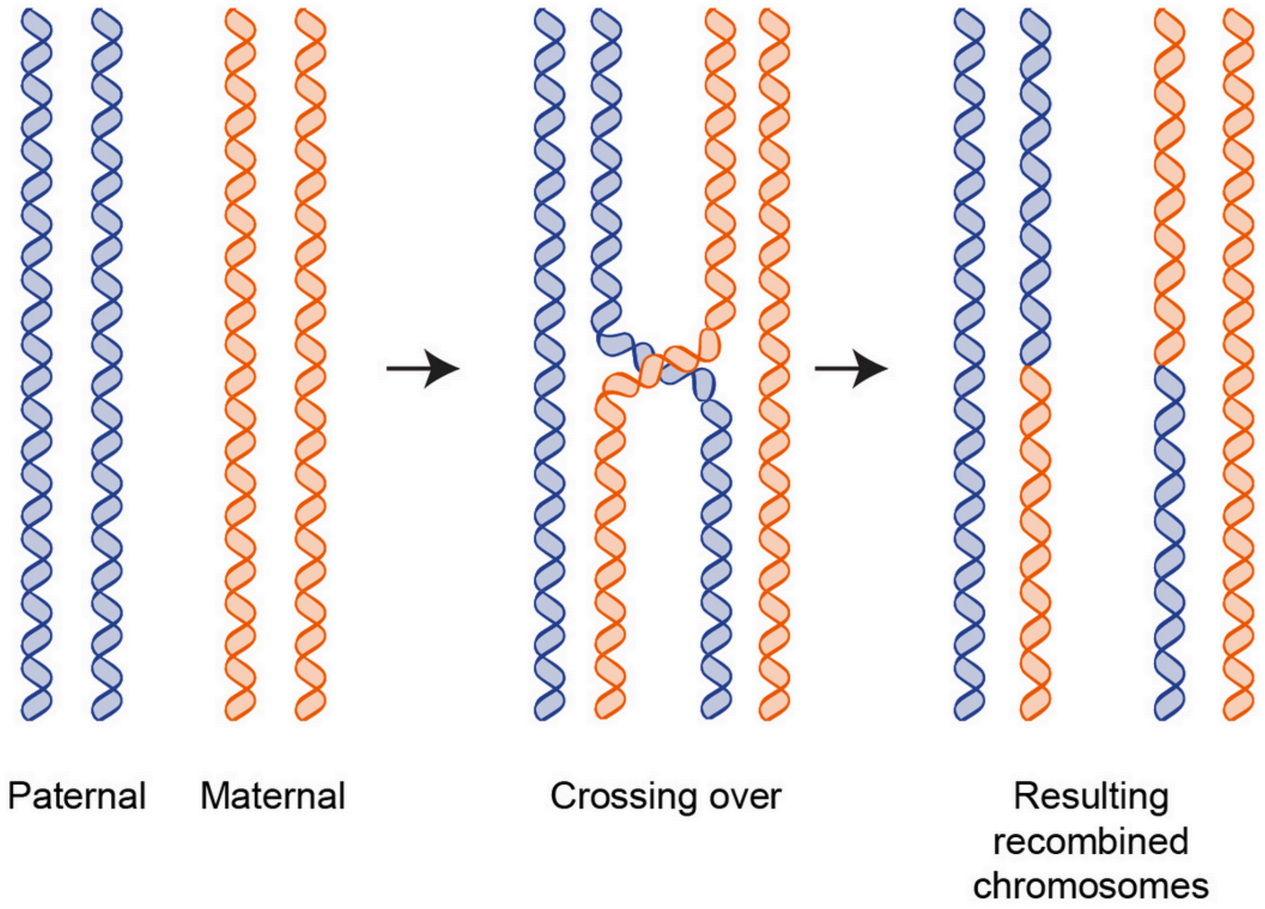
# Linkage


Haplotype 1
Haplotype 2
Haplotype 3
Haplotype 4
A A G T A C G G T T C A G G C A
T T G C G C A A C A G T A A T A
A T C T G T G A T A C T G G T G
T T C C G C G G T T ? A G A C A

# Linkage


Haplotype 1
Haplotype 2
Haplotype 3
Haplotype 4
A A G T A C G G T T C A G G C A
T T G C G C A A C A G T A A T A
A T C T G T G A T A C T G G T G
T T C C G C G G T T C A G A C A

# Recombination breaks apart correlations between SNPs




https://www.genome.gov/genetics-glossary/homologous-recombination

# Independent segregation

# Independent segregation

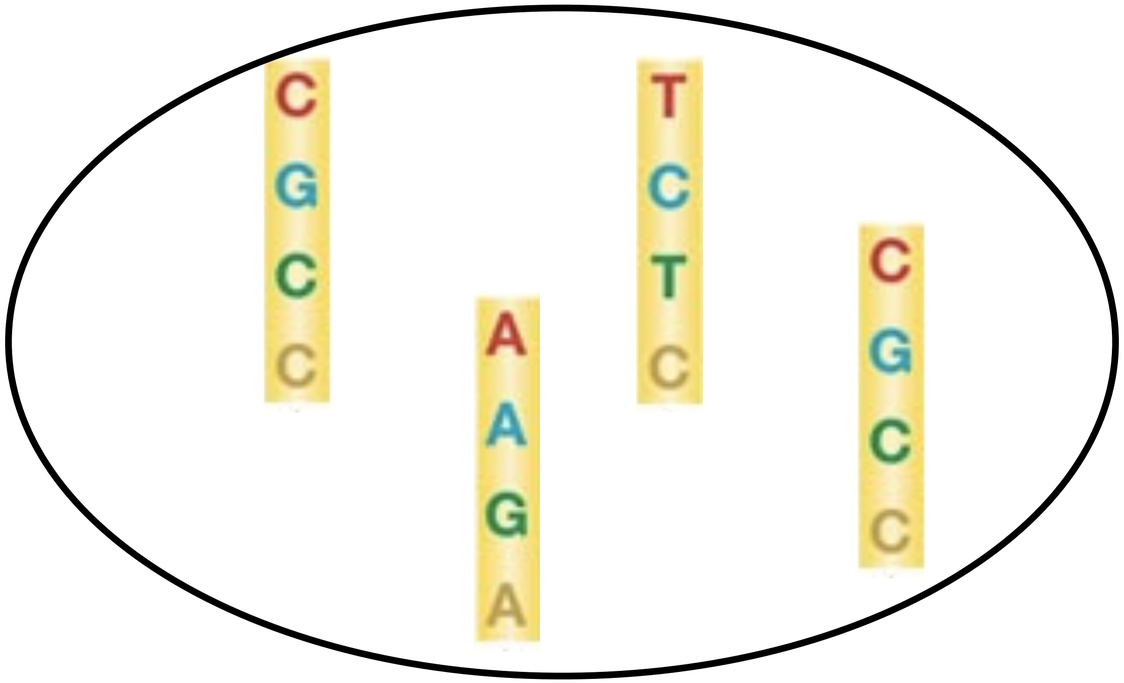
- We'll start by treating SNPs as if they **exist in a void**

# Independent segregation

- We’ll start by treating SNPs as if they **exist in a void**

# Independent segregation

- We’ll start by treating SNPs as if they **exist in a void**



# Today

- Terminology
- Wright-Fisher Model and Genetic Drift
- Intro to coalescent theory
- Mutations and the coalescent

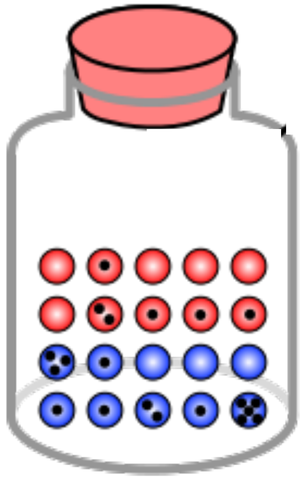
# Today

- Terminology
- Wright-Fisher Model and Genetic Drift
- Intro to coalescent theory
- Mutations and the coalescent

# Wright-Fisher Model

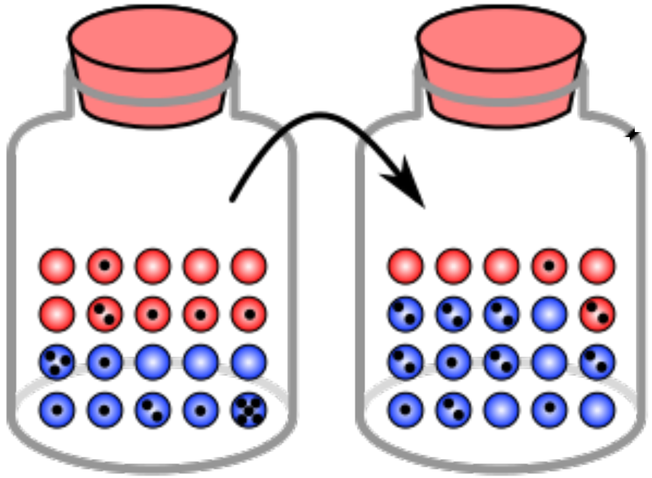
10 red

10 blue



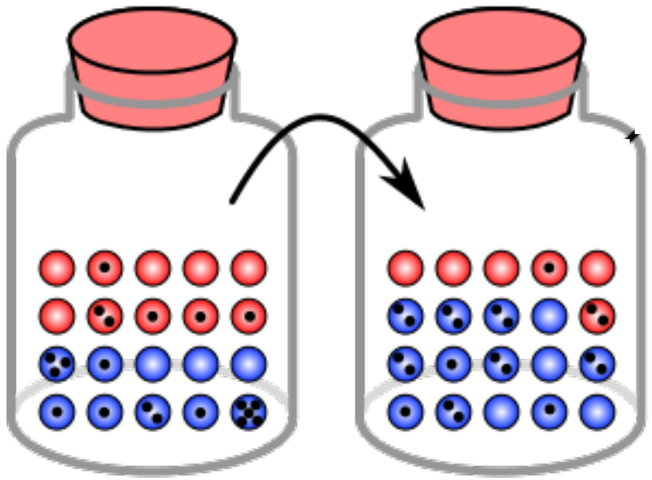
20 marbles

# Wright-Fisher Model



20 marbles  20 marbles

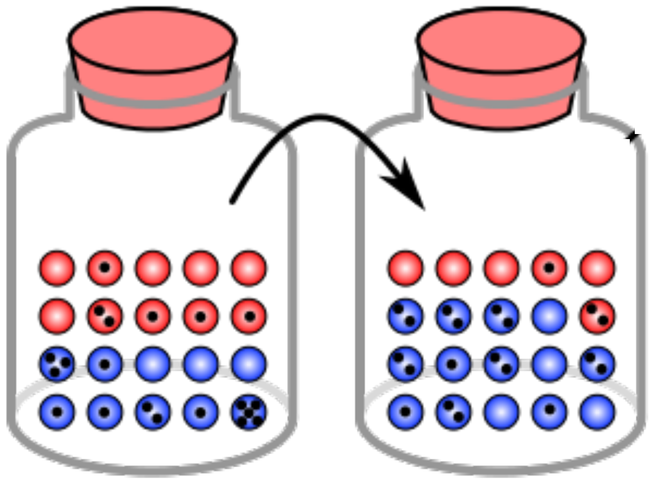
# Wright-Fisher Model



20 marbles  20 marbles

- Sample **with replacement** to fill the next jar

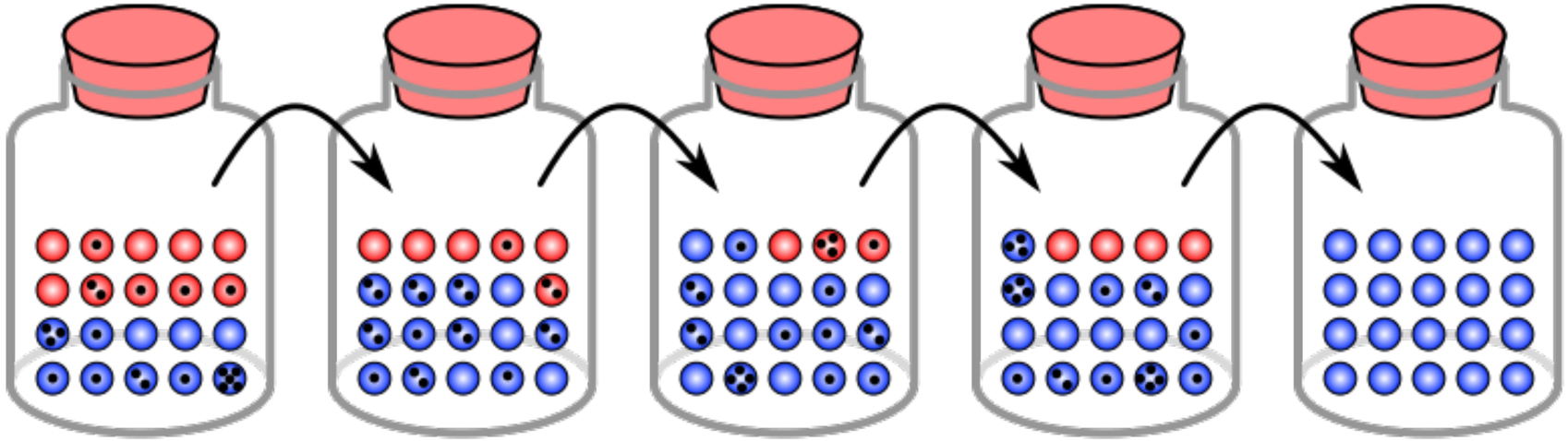
# Wright-Fisher Model



20 marbles  20 marbles

- Sample **with replacement** to fill the next jar
- The total number of marbles in each jar is the same (N = 20)

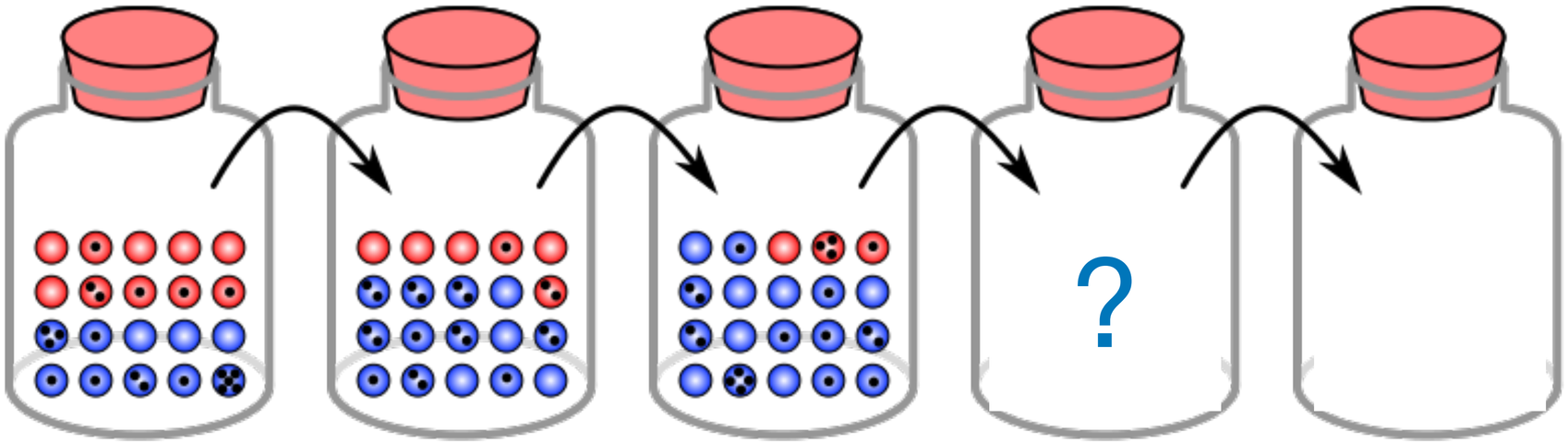
# Wright-Fisher Model



https://en.wikipedia.org/wiki/Genetic_drift#/media/File:Random_sampling_genetic_drift.svg

# Wright-Fisher Model

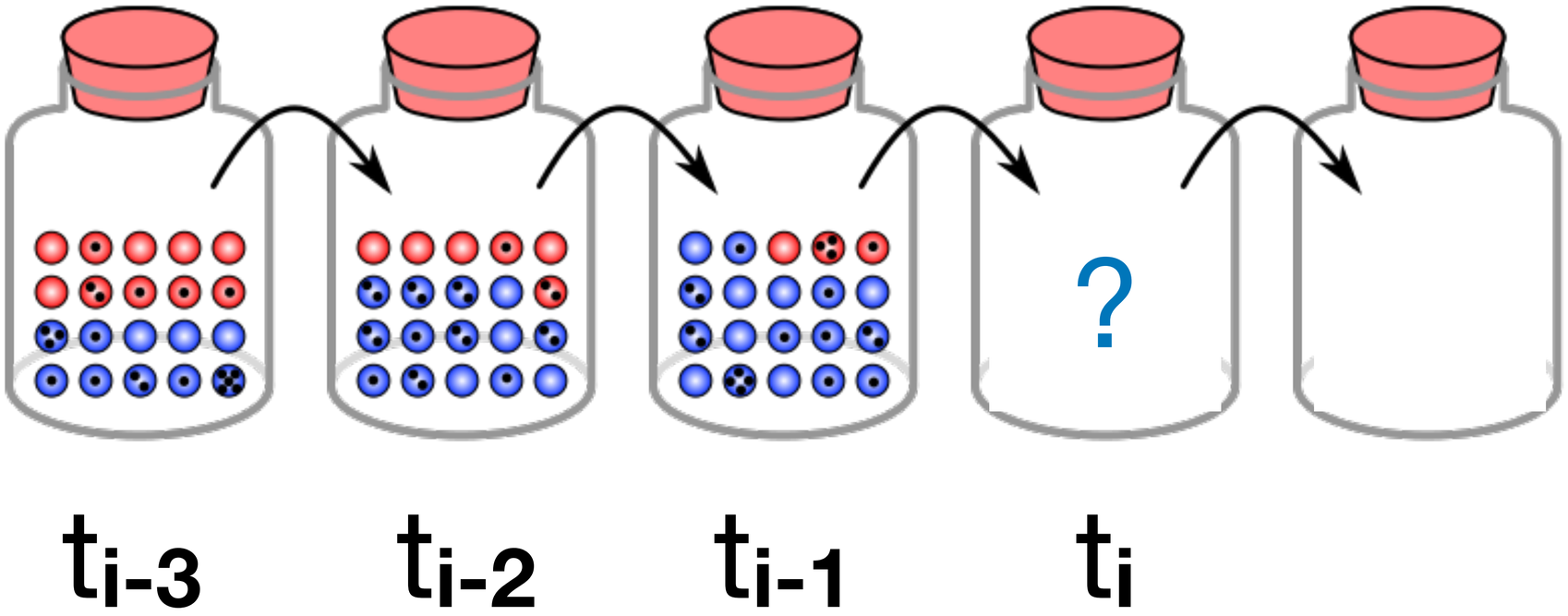
- Let's say we're interested in predicting the number of blue marbles at **time $t_i$**




$t_{i-3}$ $t_{i-2}$ $t_{i-1}$ $t_i$

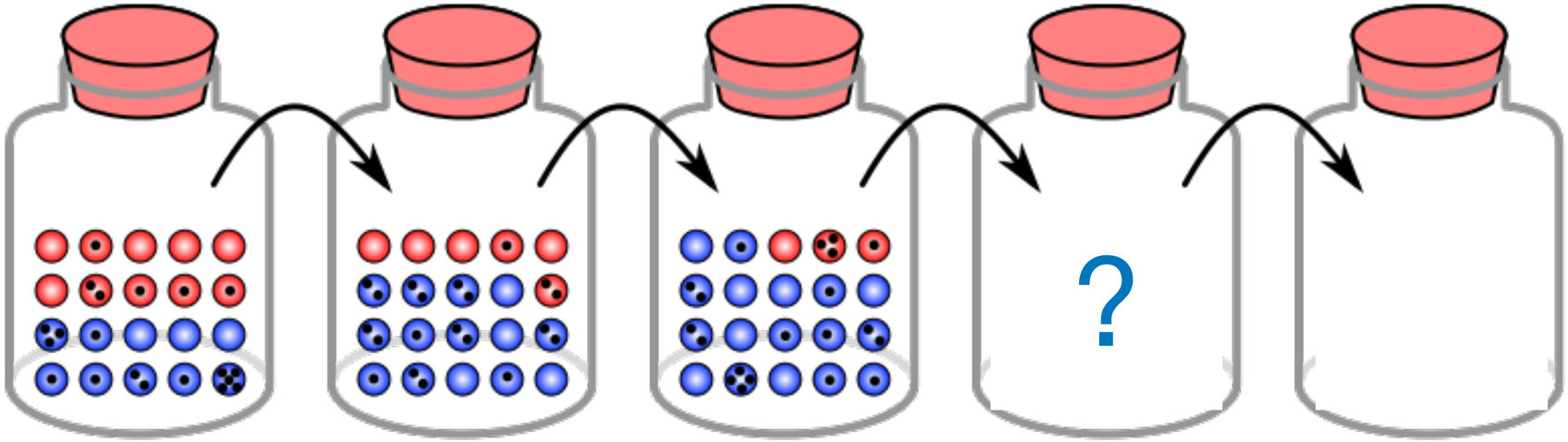
# Wright-Fisher Model

- Let $f(t_i)$ be the frequency of blue marbles at time $t_i$



# Wright-Fisher Model

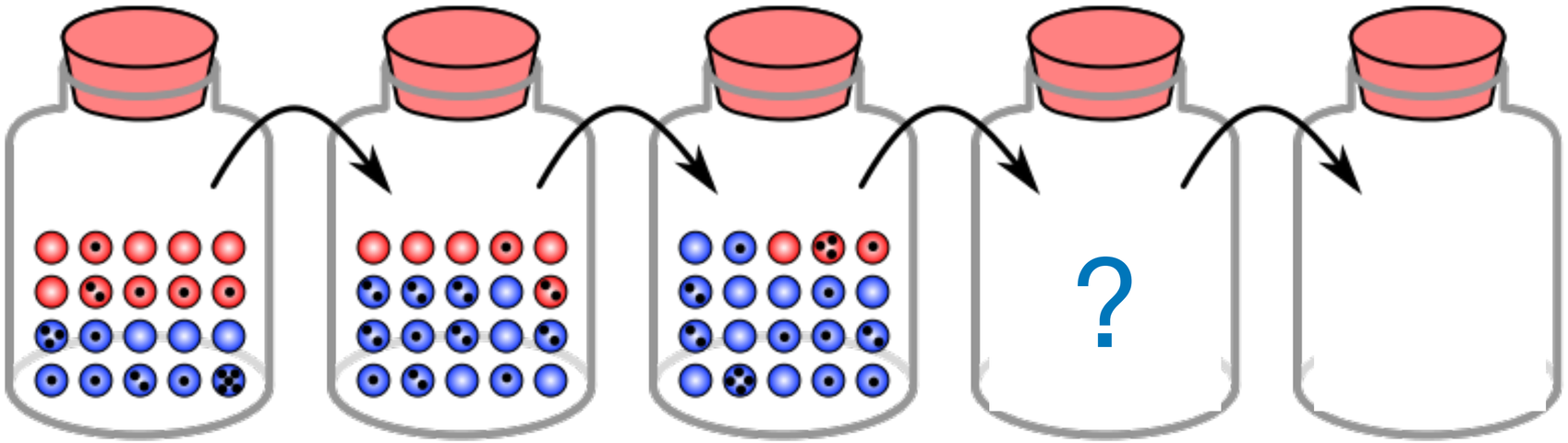
- Let $f(t_i)$ be the frequency of blue marbles at time $t_i$
- P[ # blue marbles = k | $f(t_{i-1})$, $f(t_{i-2})$, $f(t_{i-3})$, …] = P[ # blue marbles = k | $f(t_{i-1})$]




$t_{i-3}$ $t_{i-2}$ $t_{i-1}$ $t_i$

# Wright-Fisher Model

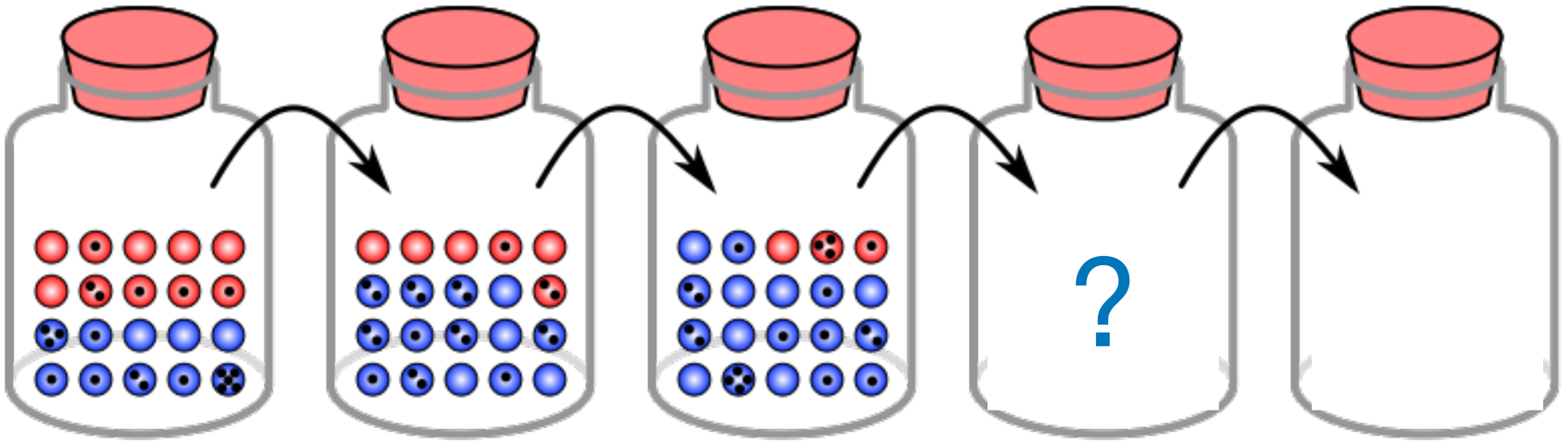
- P[ # blue marbles = k | $f(t_{i-1})$ ] = $\binom{N}{k} f(t_{i-1})^k (1 - f(t_{i-1}))^{N-k}$



# Wright-Fisher Model

Binomial distribution

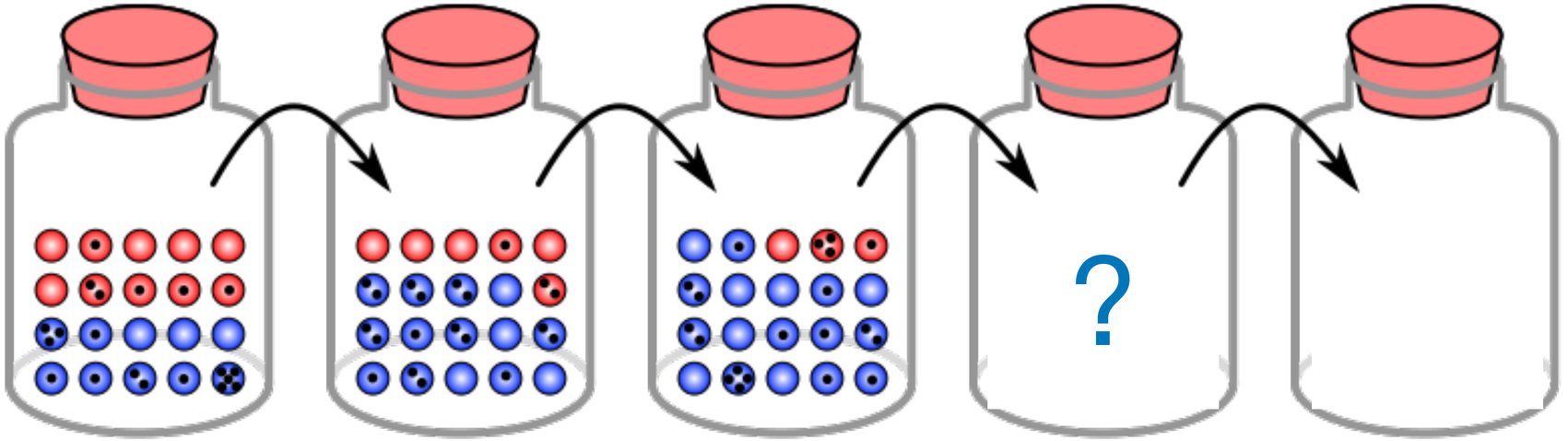
- P[ # blue marbles = k | $f(t_{i-1})$ ] = $\binom{N}{k} f(t_{i-1})^k (1 - f(t_{i-1}))^{N-k}$




$t_{i-3}$ $t_{i-2}$ $t_{i-1}$ $t_i$

# Wright-Fisher Model

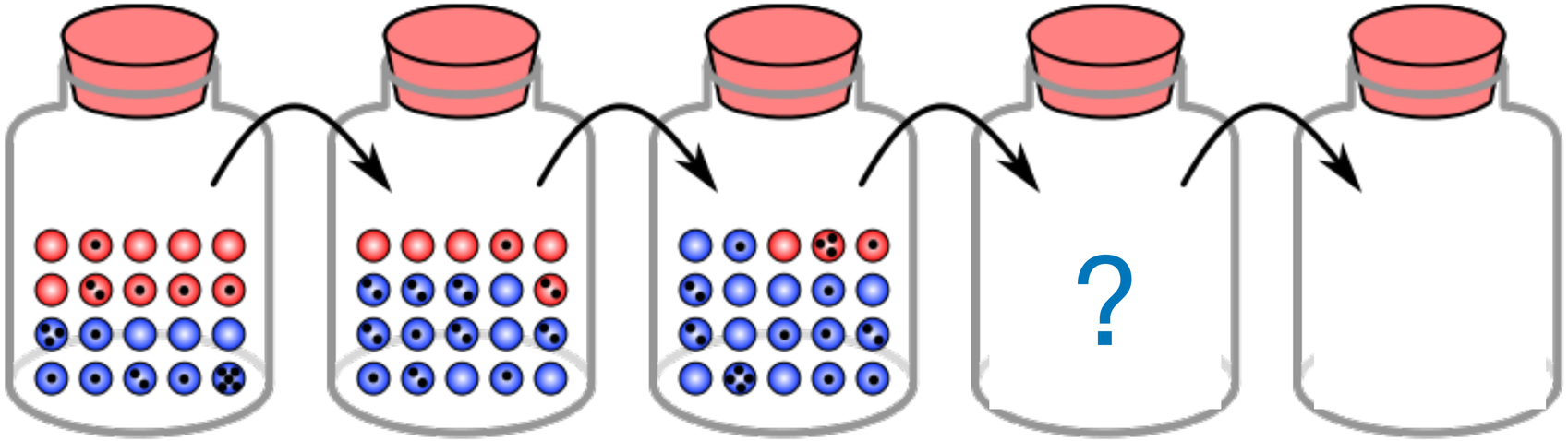
- P[ # blue marbles = k | $f(t_{i-1})$ ] = $\binom{N}{k} f(t_{i-1})^k (1 - f(t_{i-1}))^{N-k}$



# Wright-Fisher Model

We draw a blue marble
k times

- P[ # blue marbles = k | $f(t_{i-1})$ ] = $\binom{N}{k} f(t_{i-1})^k (1 - f(t_{i-1}))^{N-k}$



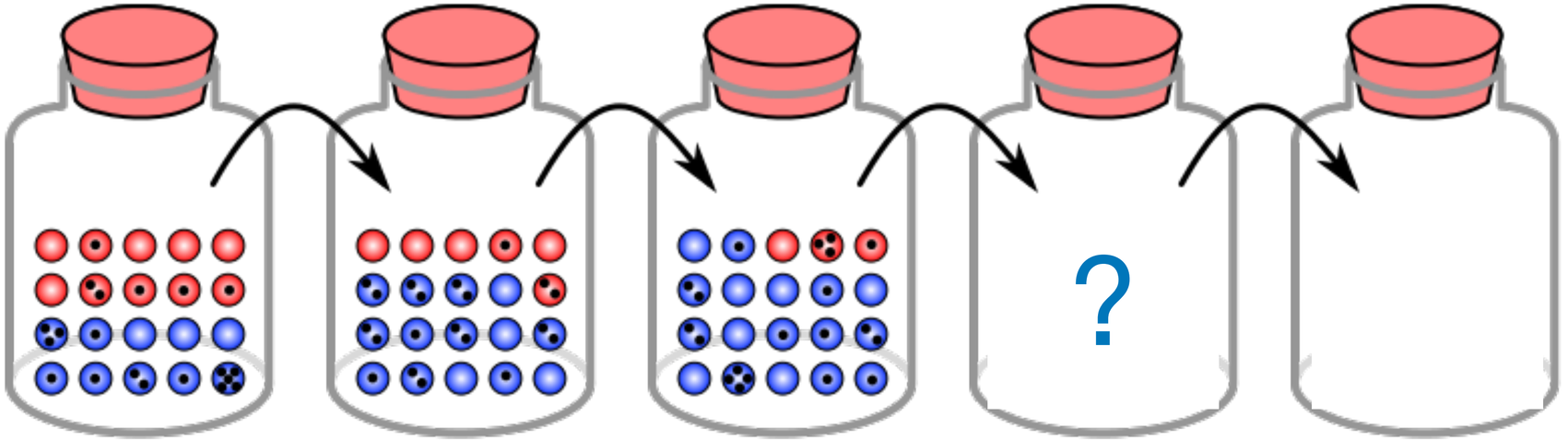

$t_{i-3}$ $t_{i-2}$ $t_{i-1}$ $t_i$

# Wright-Fisher Model

We draw a blue marble k times

We draw a red marble n-k times

- P[ # blue marbles = k | $f(t_{i-1})$ ] = $\binom{N}{k} f(t_{i-1})^k (1 - f(t_{i-1}))^{N-k}$




$t_{i-3}$ $t_{i-2}$ $t_{i-1}$ $t_i$

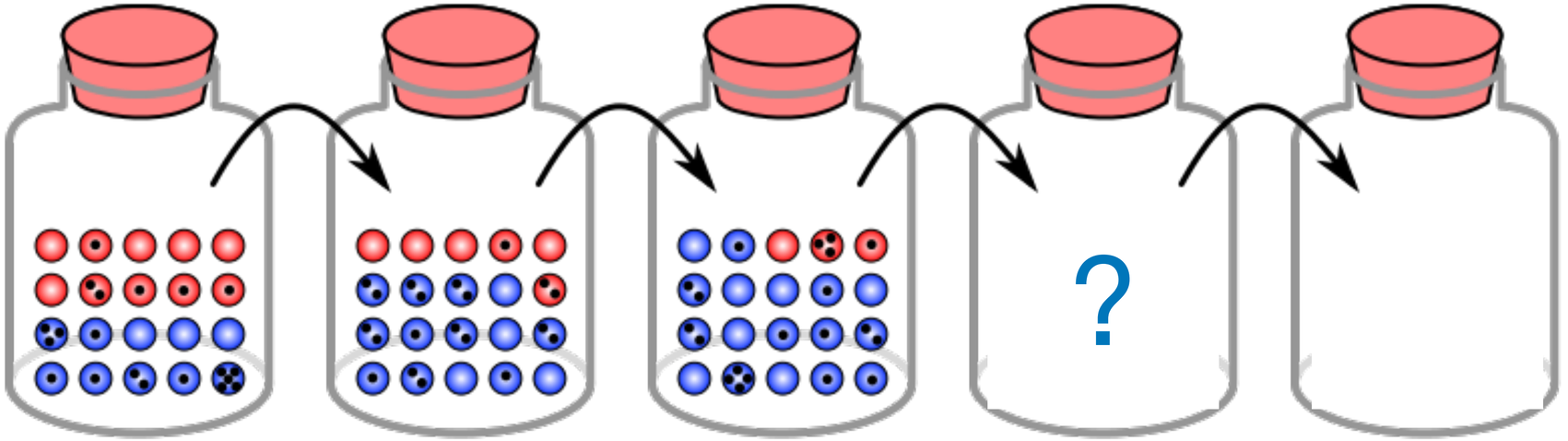
# Wright-Fisher Model

There are these many ways to order **k** blue marbles in a set of **n marbles**

We draw a blue marble k times

We draw a red marble n-k times

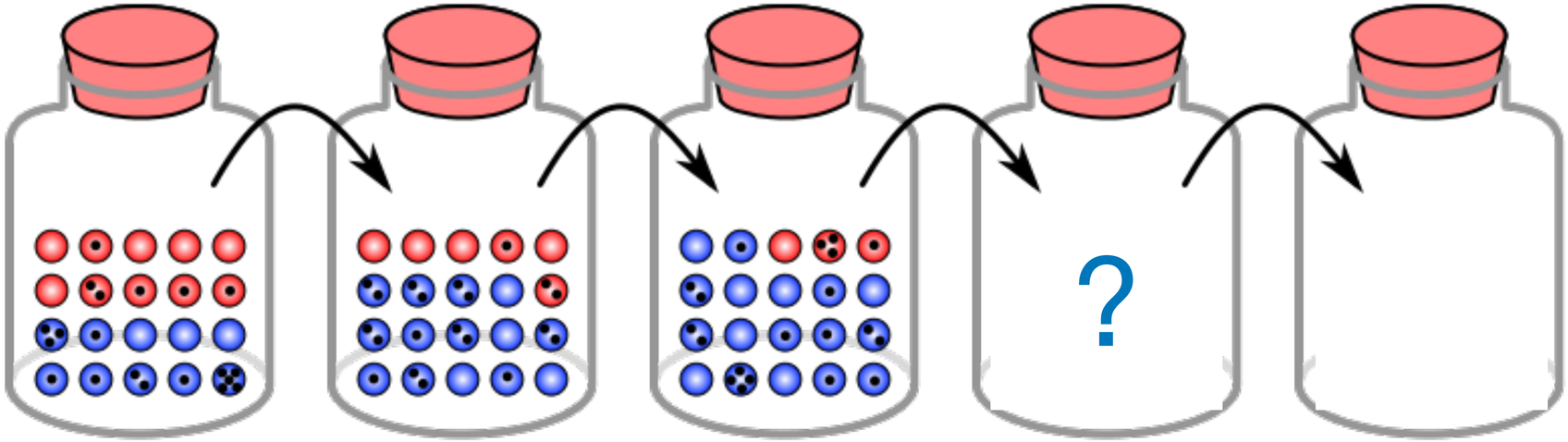
- P[ # blue marbles = k | $f(t_{i-1})$ ] = $\binom{N}{k} f(t_{i-1})^k (1 - f(t_{i-1}))^{N-k}$




$t_{i-3}$ $t_{i-2}$ $t_{i-1}$ $t_i$

# Wright-Fisher Model

- In statistics lingo, $\#\ blue\ marbles \mid f(t_{i-1}) \sim Bin(N, f(t_{i-1}))$



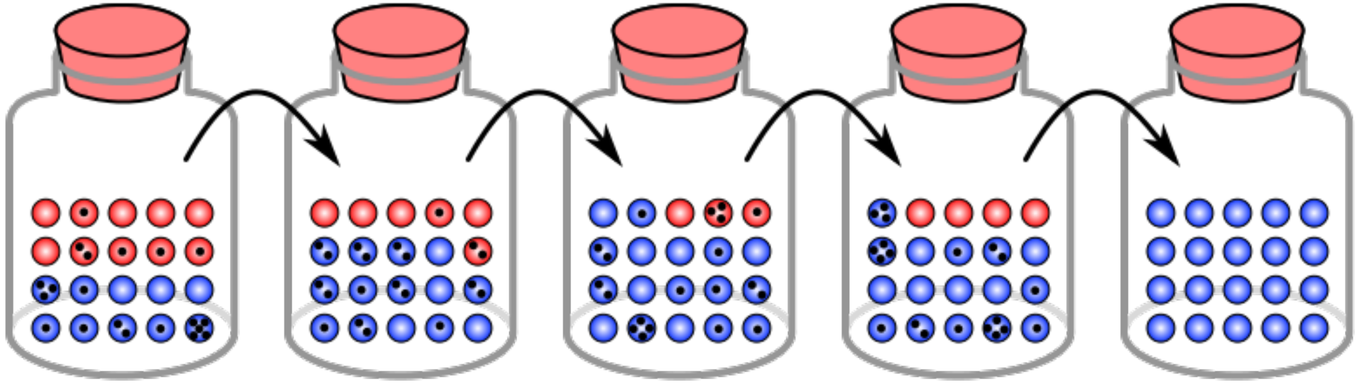
# Wright-Fisher Model

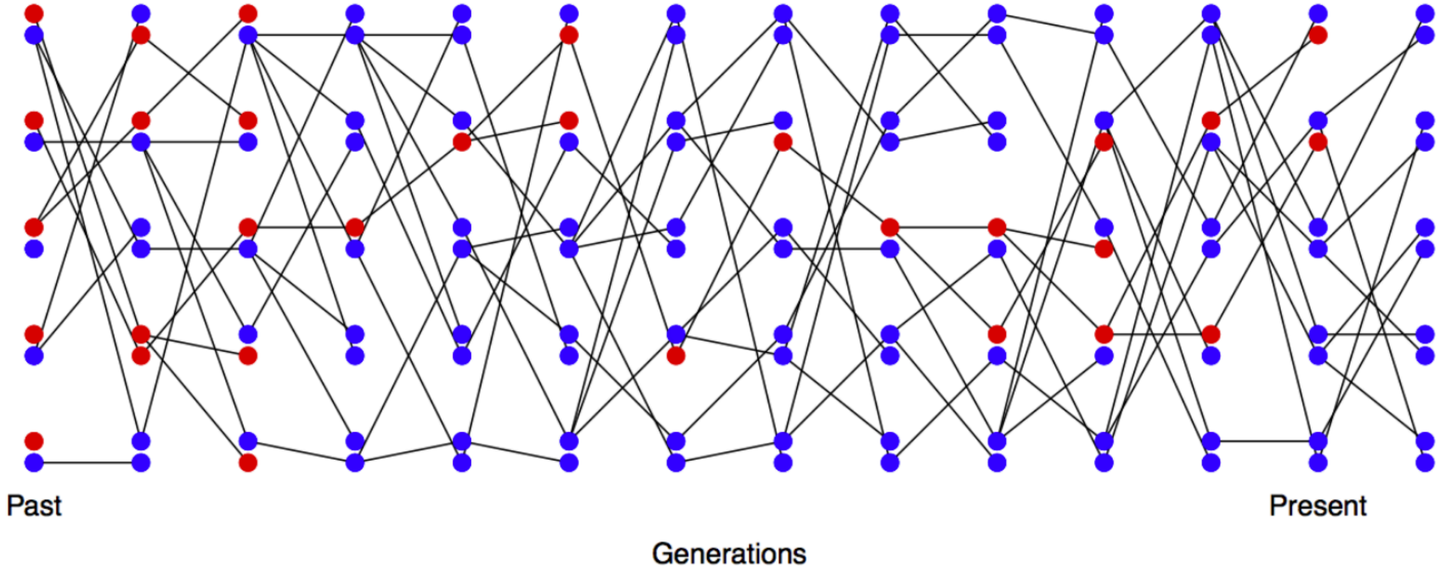
blue marbles =
blue alleles

red marbles =
red alleles

**N** total marbles =
population size **N**

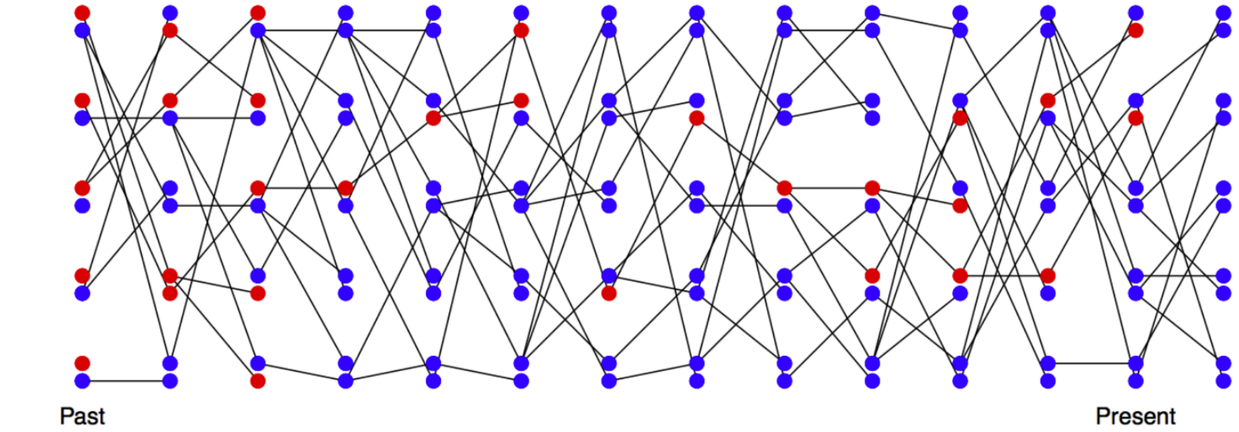
Jar $\mathbf{t_i}$ =
Generation $\mathbf{t_i}$

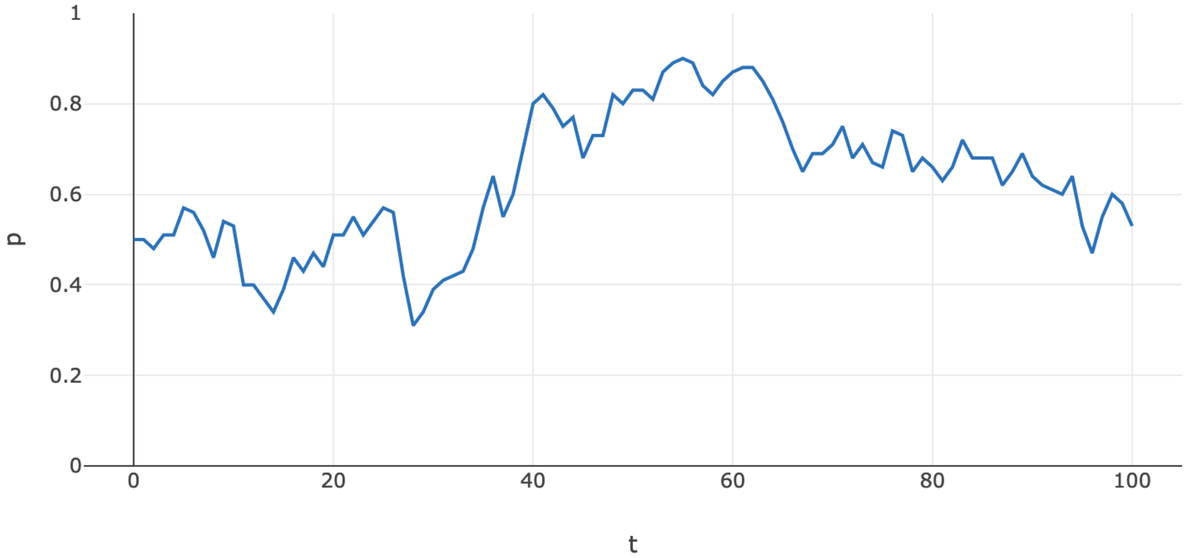





https://cooplab.github.io/popgen-notes/

# Wright-Fisher Model




https://cooplab.github.io/popgen-notes/

# Assumptions of the Wright-Fisher Model

- Constant population size
- Individuals reproduce asexually and randomly
- No migration
- No selection
- No mutation
- Non-overlapping generations

# Assumptions of the Wright-Fisher Model

- Constant population size
- Individuals reproduce asexually and randomly
- No migration
- No selection
- No mutation
- Non-overlapping generations

What assumption of the Hardy-Weinberg model are we now relaxing?

# Assumptions of the Wright-Fisher Model

- Constant population size
- Individuals reproduce asexually and randomly
- No migration
- No selection
- No mutation
- Non-overlapping generations

- **N** if we are studying a population of N haploid organisms.
- **2N** if we are studying a population of N diploid organisms (which we treat as if they were equivalent to a haploid population twice its size)

What assumption of the Hardy-Weinberg model are we now relaxing?

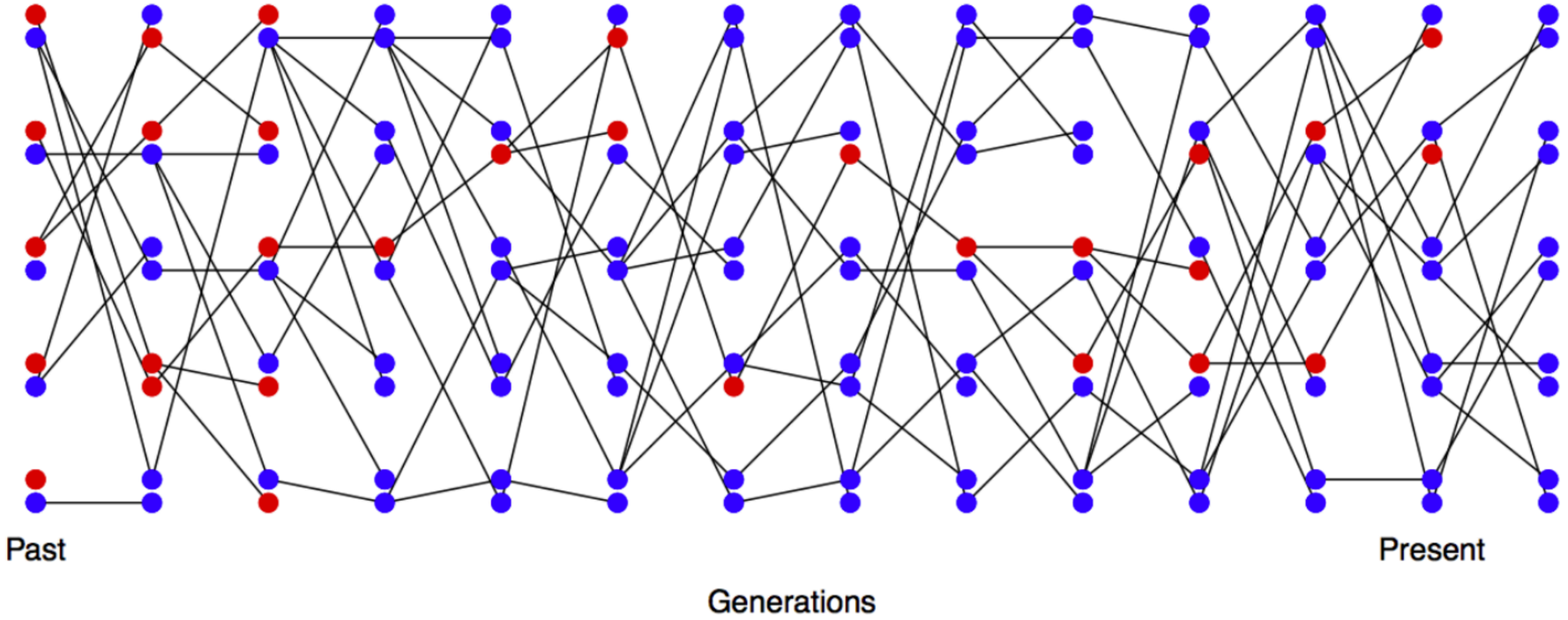
# Assumptions of the Wright-Fisher Model

- Constant population size
- Individuals reproduce asexually and randomly
- No migration
- No selection
- No mutation
- Non-overlapping generations

- **N** if we are studying a population of N haploid organisms.
- **2N** if we are studying a population of N diploid organisms (which we treat as if they were equivalent to a haploid population twice its size)

What assumption of the Hardy-Weinberg model are we now relaxing?

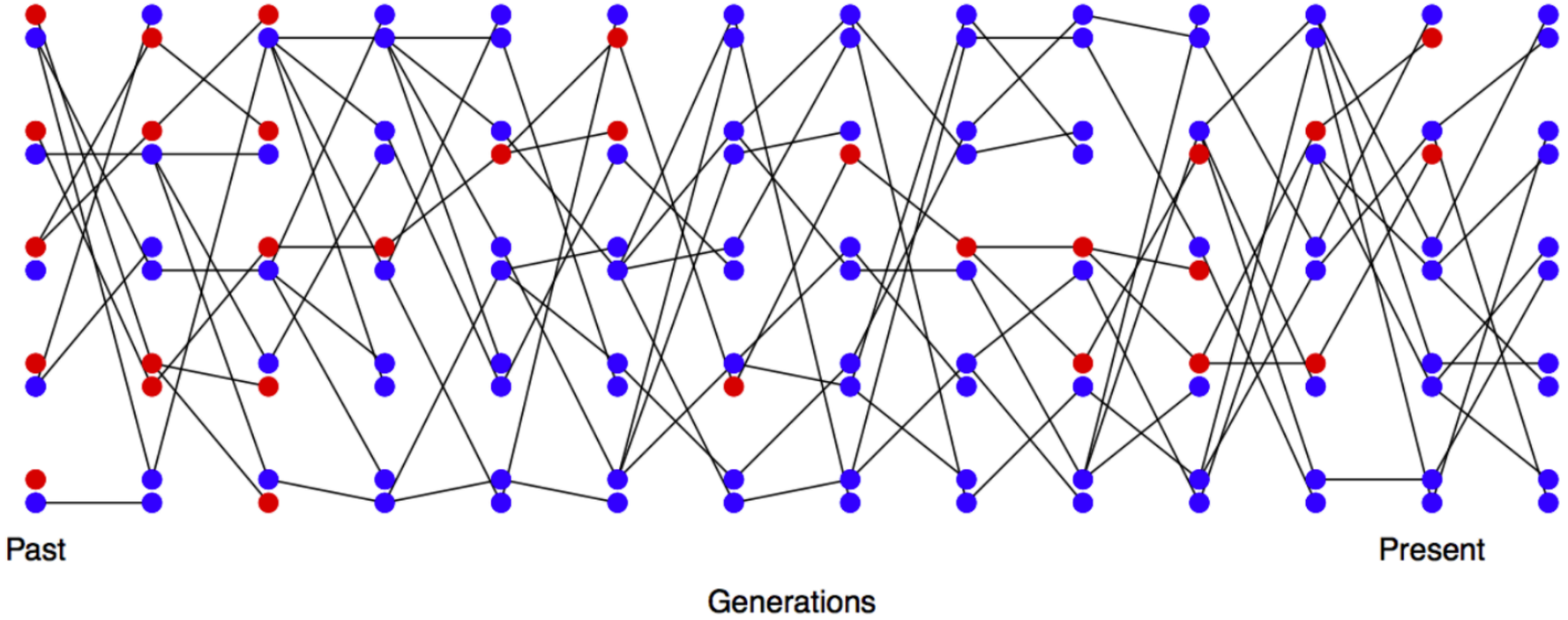
# Wright-Fisher Model




https://cooplab.github.io/popgen-notes/

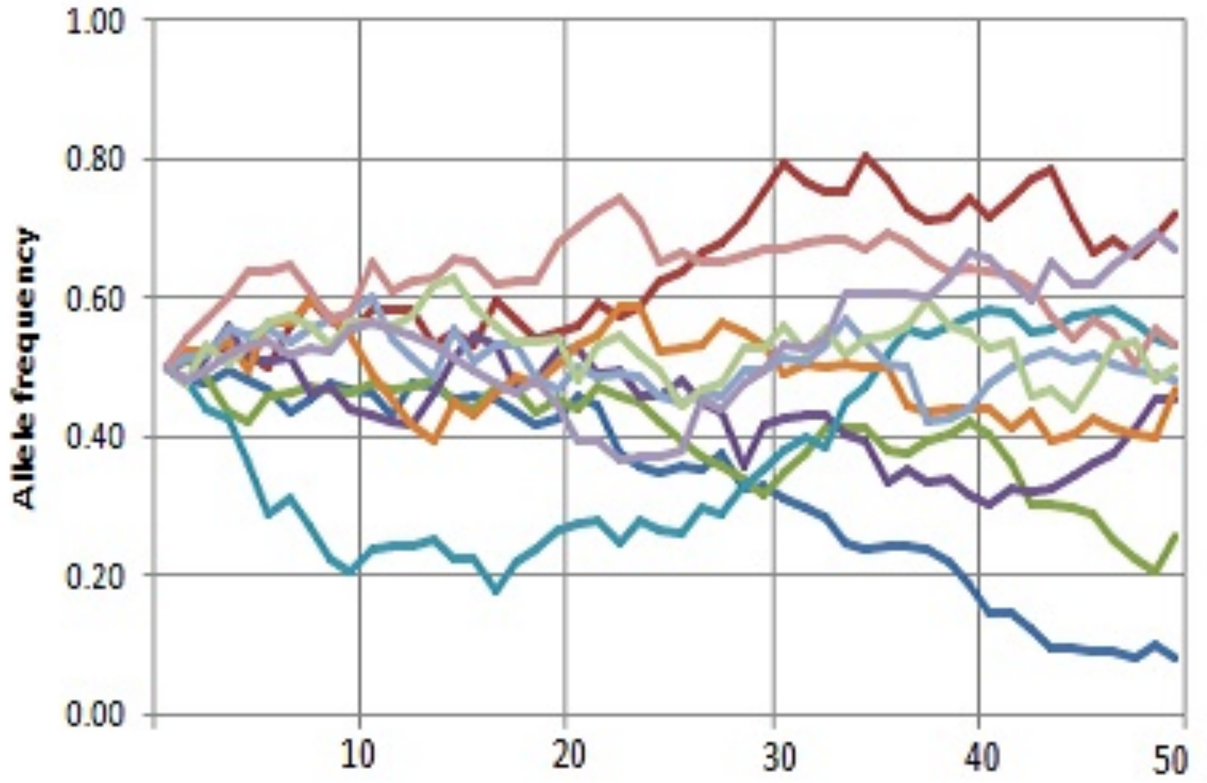
# Wright-Fisher Model

5 diploid organisms in each generation, so **2N = 10**



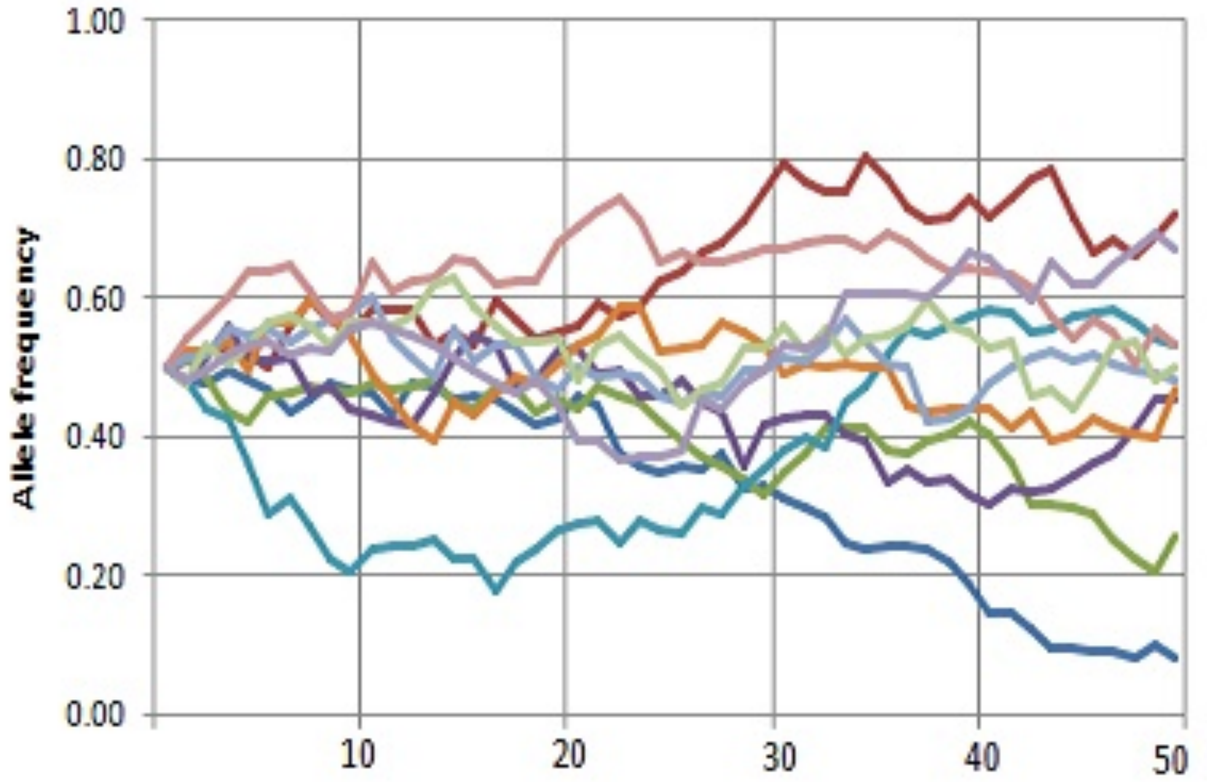

https://cooplab.github.io/popgen-notes/

# Wright-Fisher Model



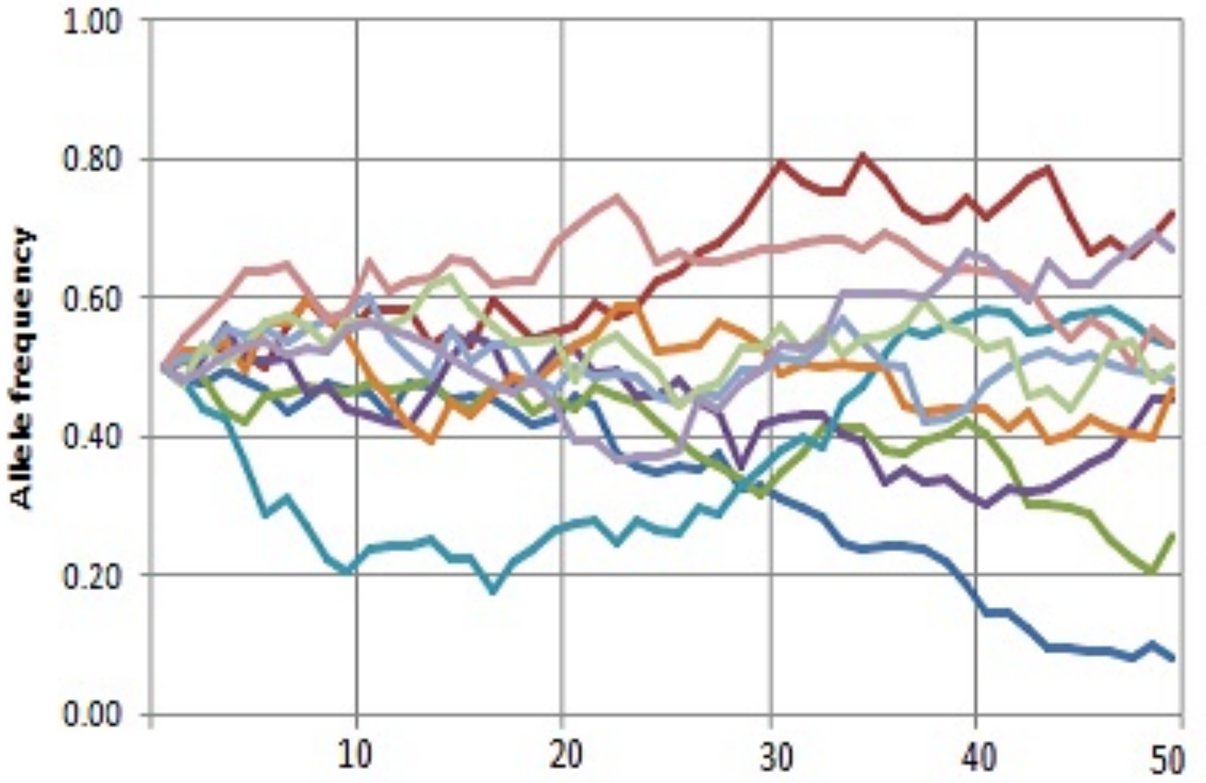
# Wright-Fisher Model

- The Wright-Fisher model is a stochastic process



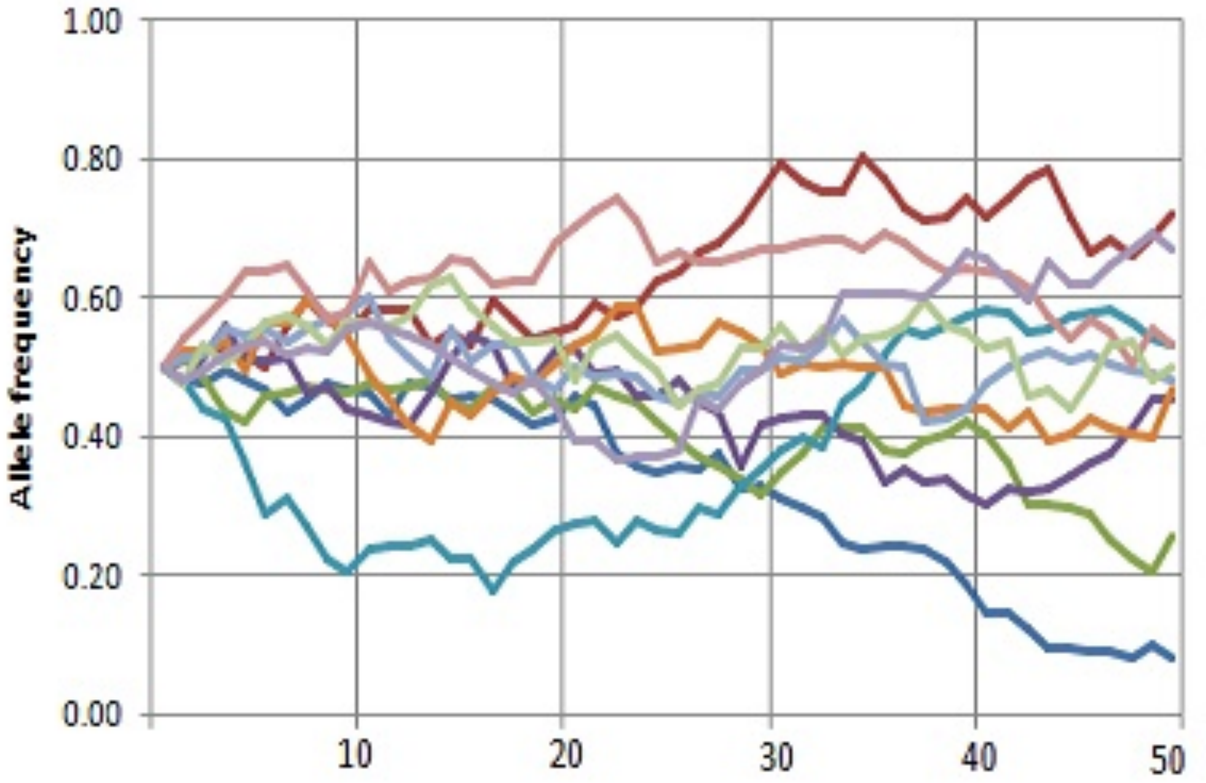
# Wright-Fisher Model

- The Wright-Fisher model is a stochastic process
- If you start from the same allele frequency, you might end up at a different frequency **purely by chance**.



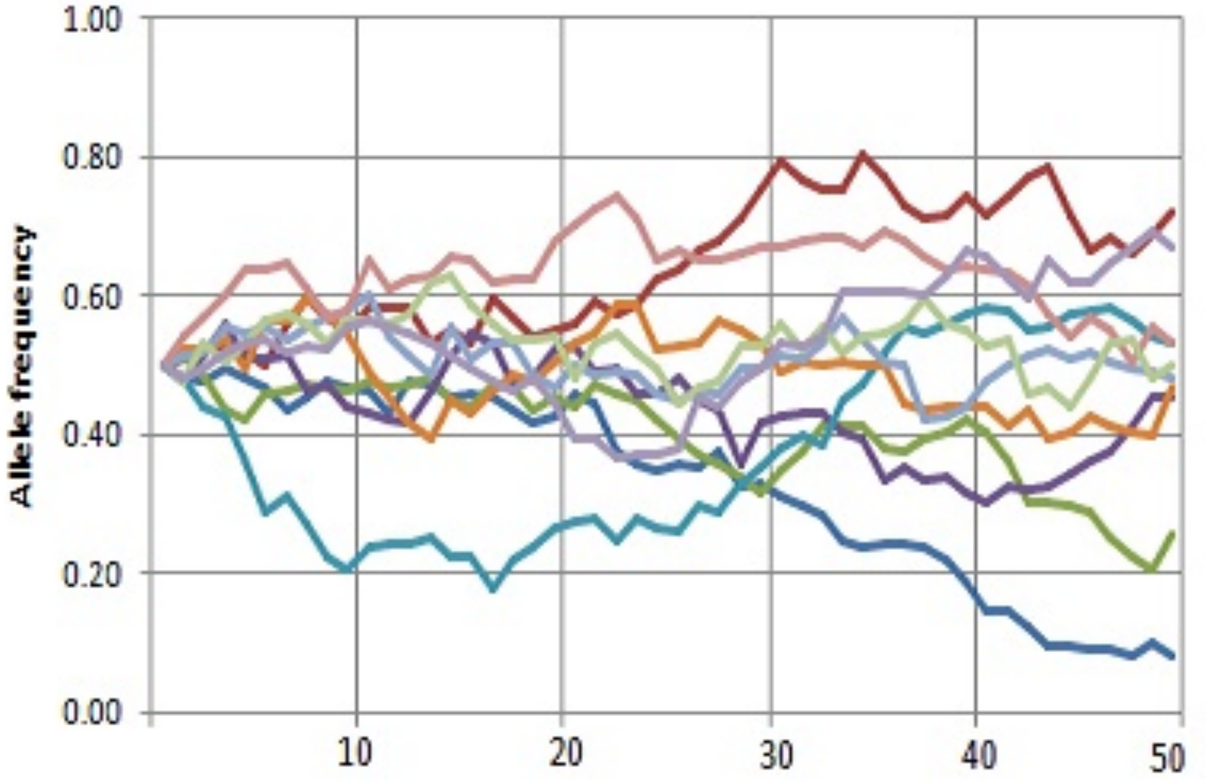
# Wright-Fisher Model

- The Wright-Fisher model is a stochastic process
- If you start from the same allele frequency, you might end up at a different frequency **purely by chance**.
- **Genetic drift** is the change in allele frequencies over time due to **random sampling.**

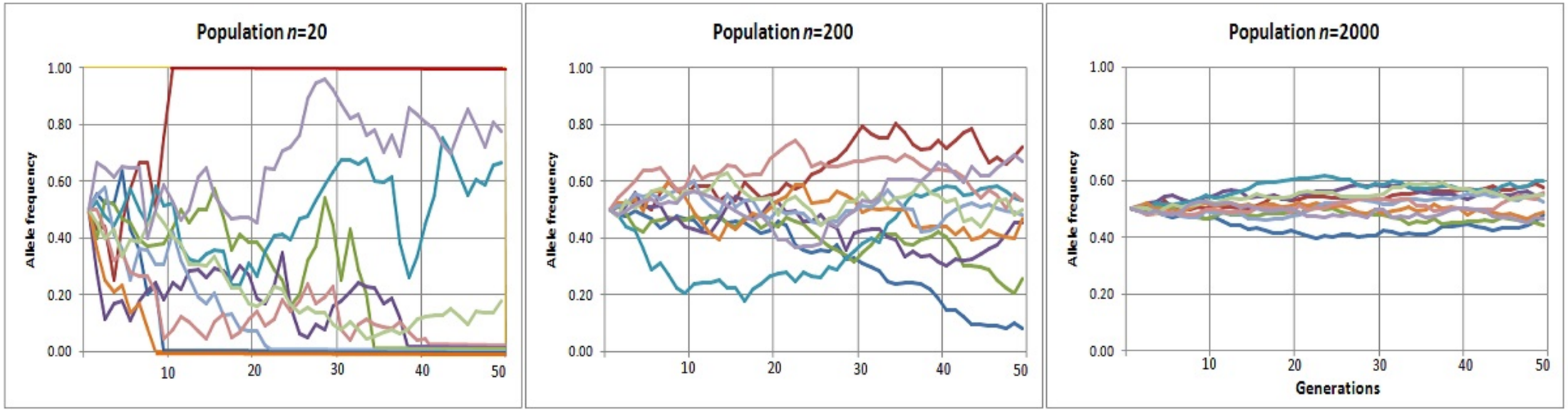


# Wright-Fisher Model

- The Wright-Fisher model is a stochastic process
- If you start from the same allele frequency, you might end up at a different frequency **purely by chance**.
- **Genetic drift** is the change in allele frequencies over time due to **random sampling.**
- No allele has any special advantage over the others.

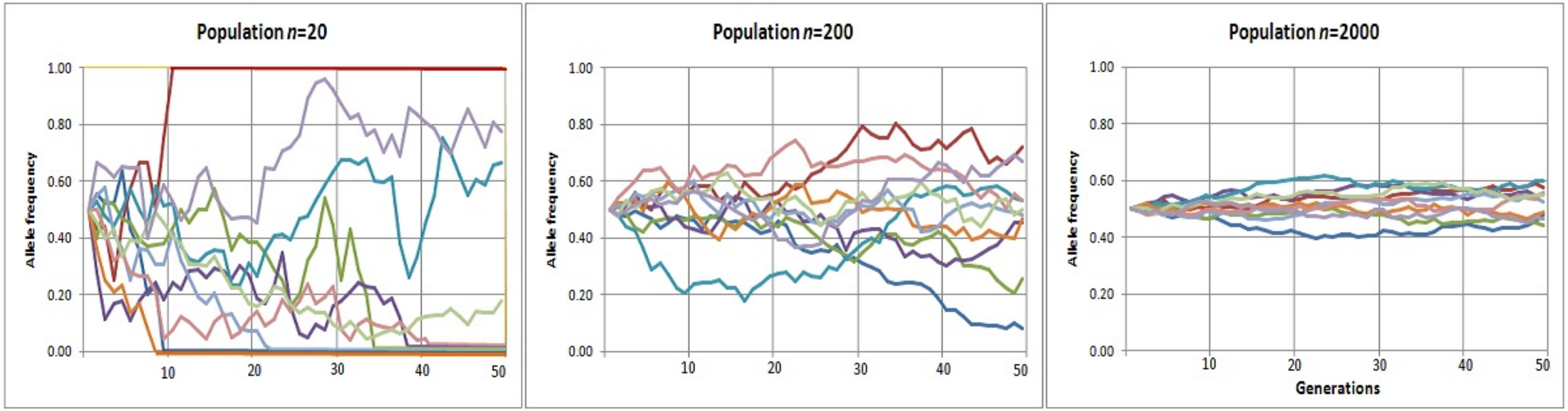


# Wright-Fisher Model



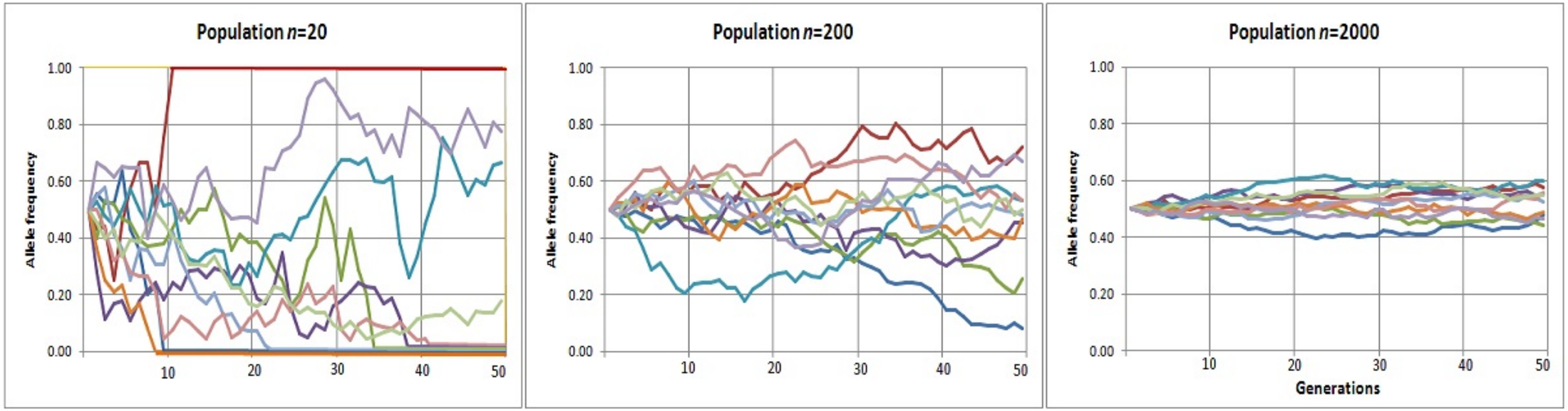
# Wright-Fisher Model

- Drift is stronger in small than in large populations

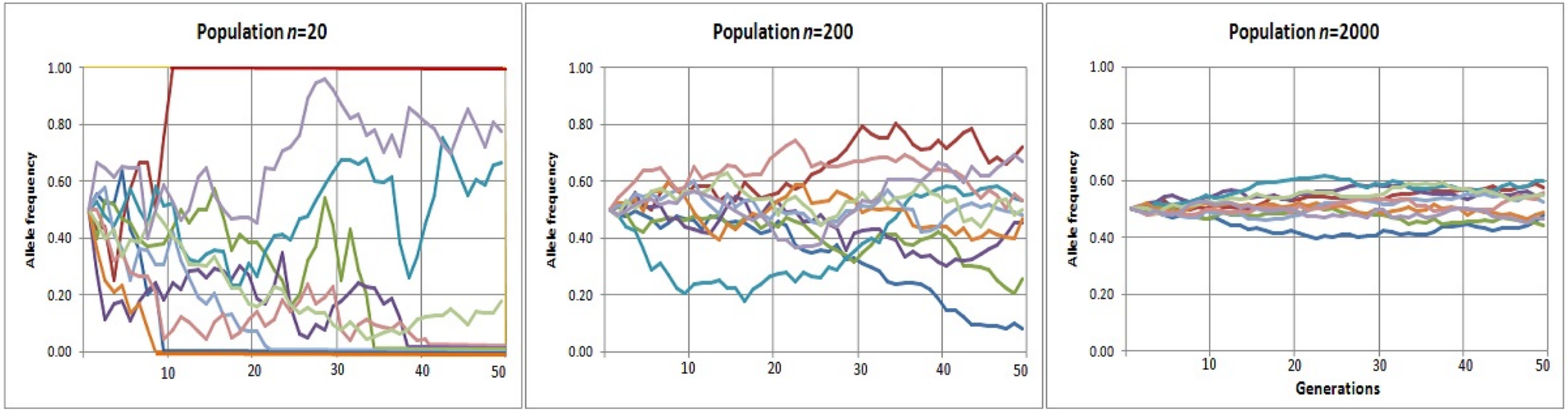


# Wright-Fisher Model

- Drift is stronger in small than in large populations

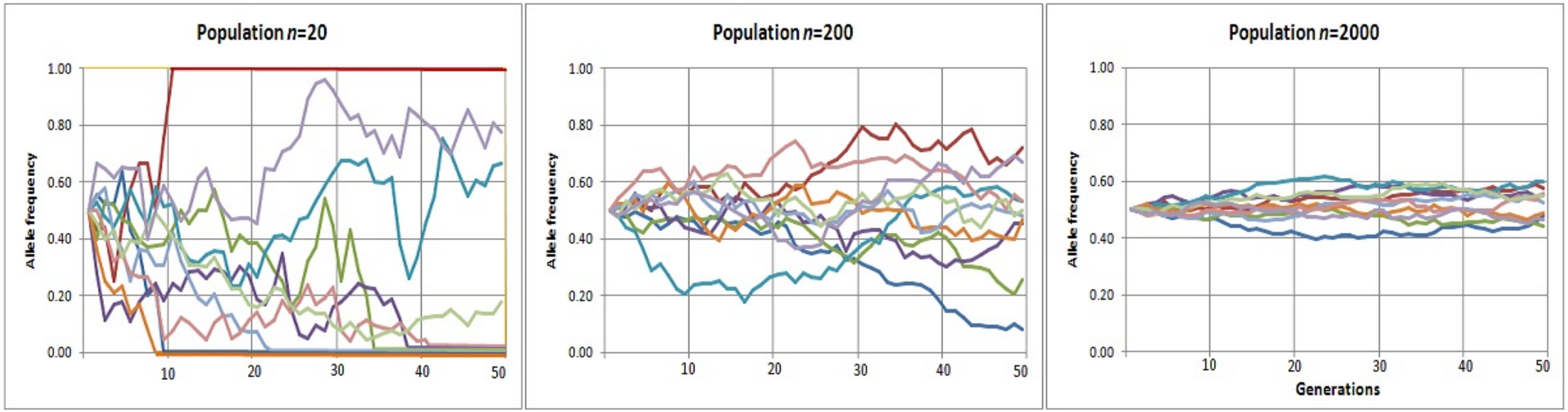


# Wright-Fisher Model



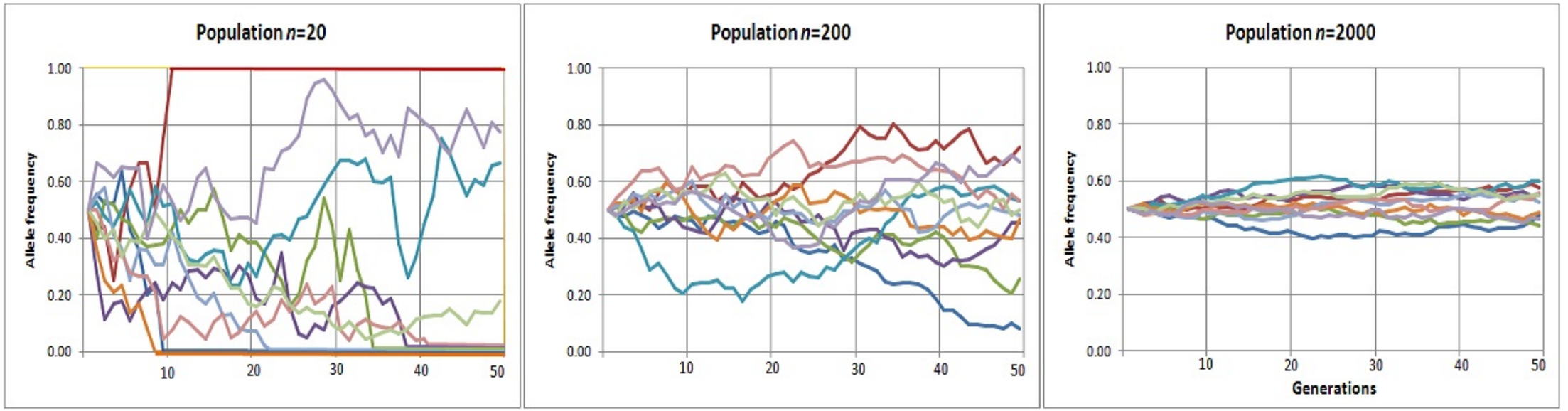
# Wright-Fisher Model

- The expected allele frequency is the same at each generation, and equal to the starting frequency p.



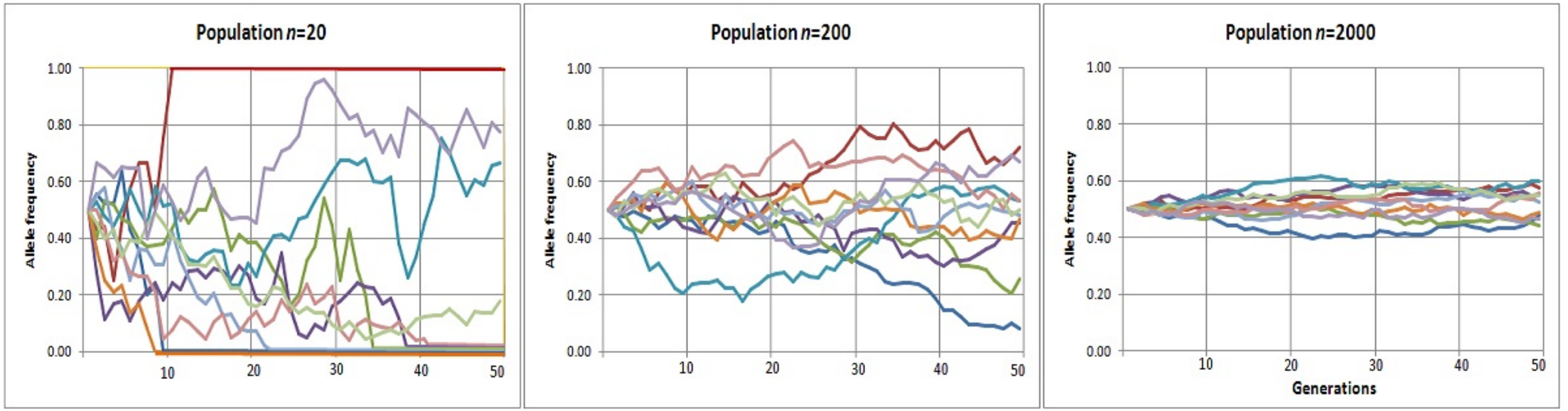
# Wright-Fisher Model

- The expected allele frequency is the same at each generation, and equal to the starting frequency p.
- If X ~ Binomial(n,p), then E[X] = np and therefore **E[X/n] = p** in each generation

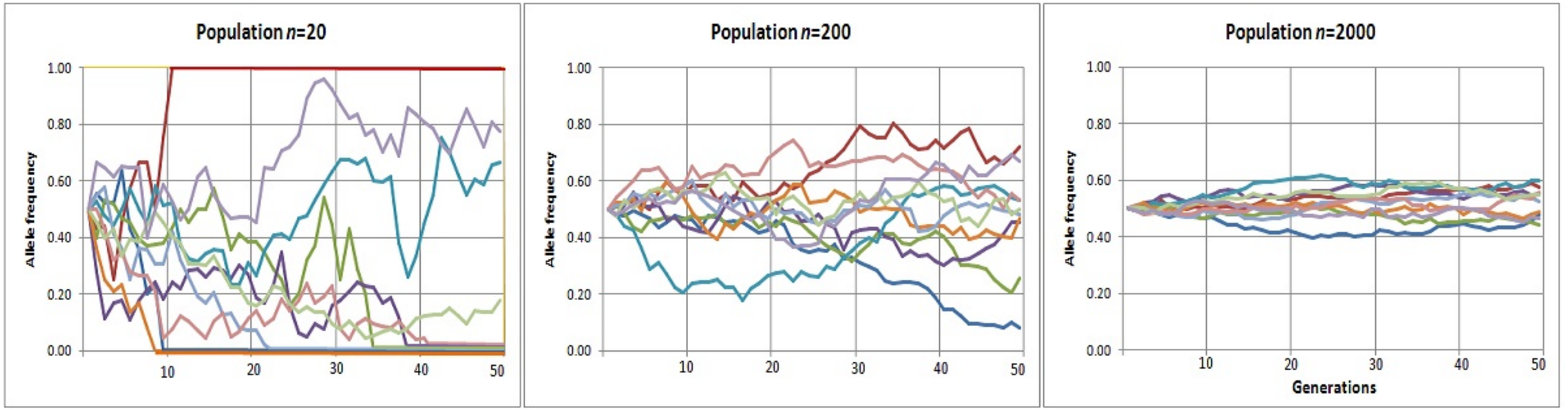


# Wright-Fisher Model

- The expected allele frequency is the same at each generation, and equal to the starting frequency p.
- If X ~ Binomial(n,p), then E[X] = np and therefore **E[X/n] = p** in each generation

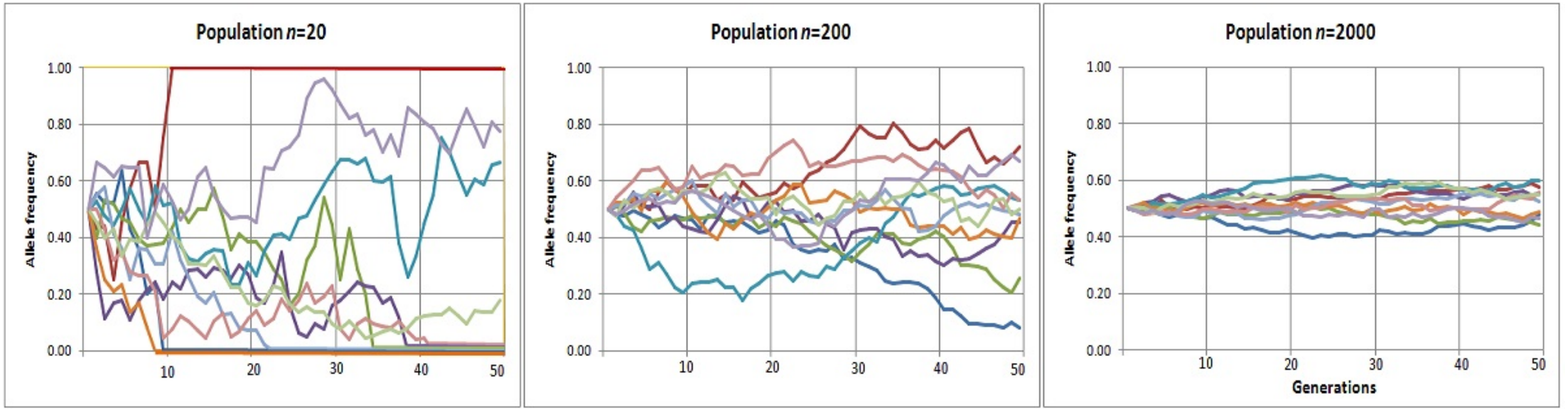


# Wright-Fisher Model



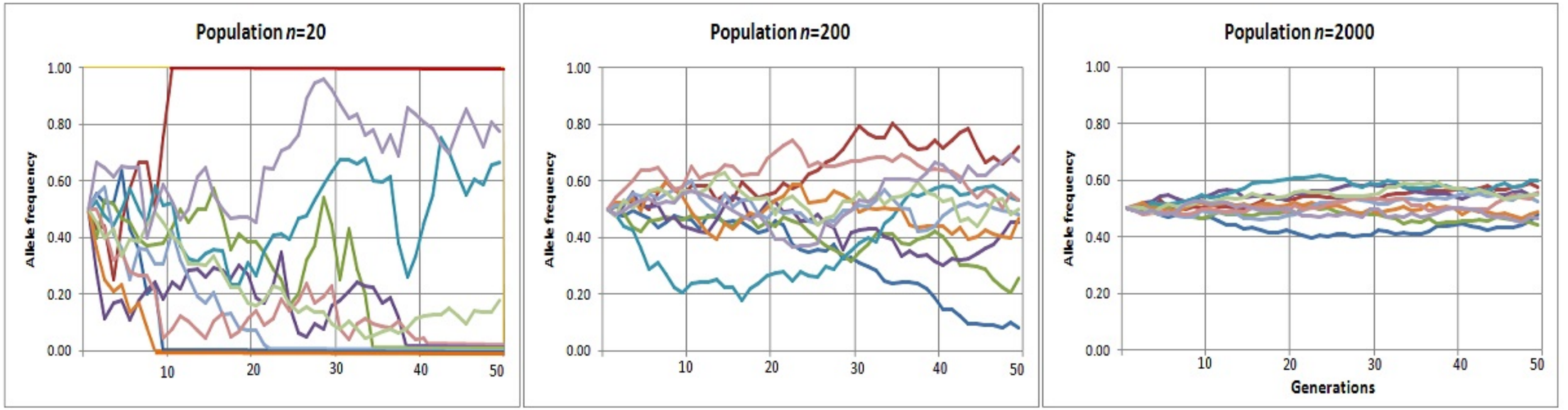
# Wright-Fisher Model

- When an allele reaches a frequency of 1, we say it has **fixed.**



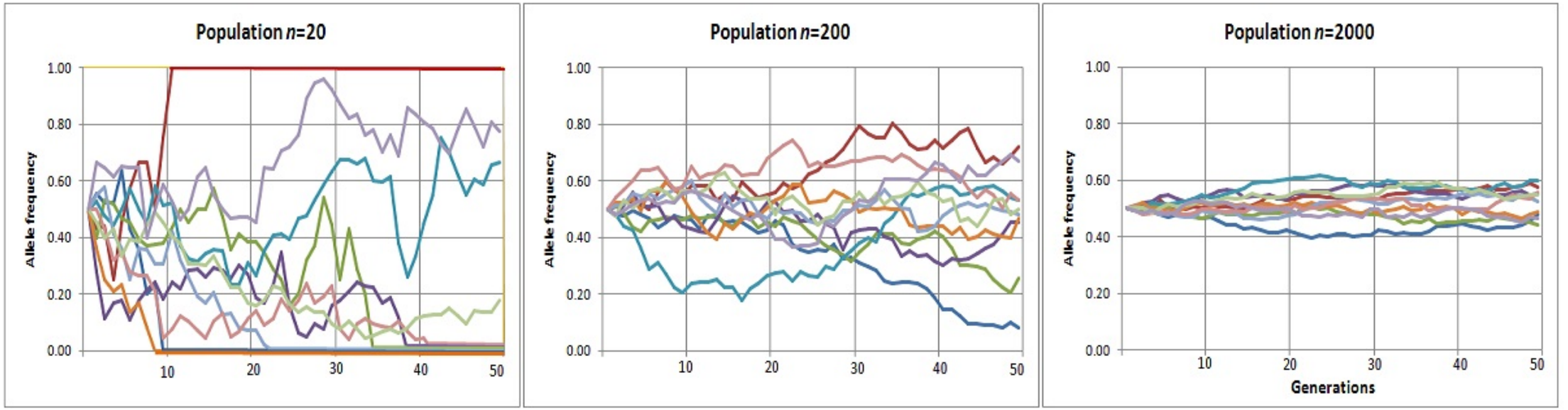
# Wright-Fisher Model

- When an allele reaches a frequency of 1, we say it has **fixed**.
- When an allele reaches a frequency of 0, we say it has **gone extinct**.



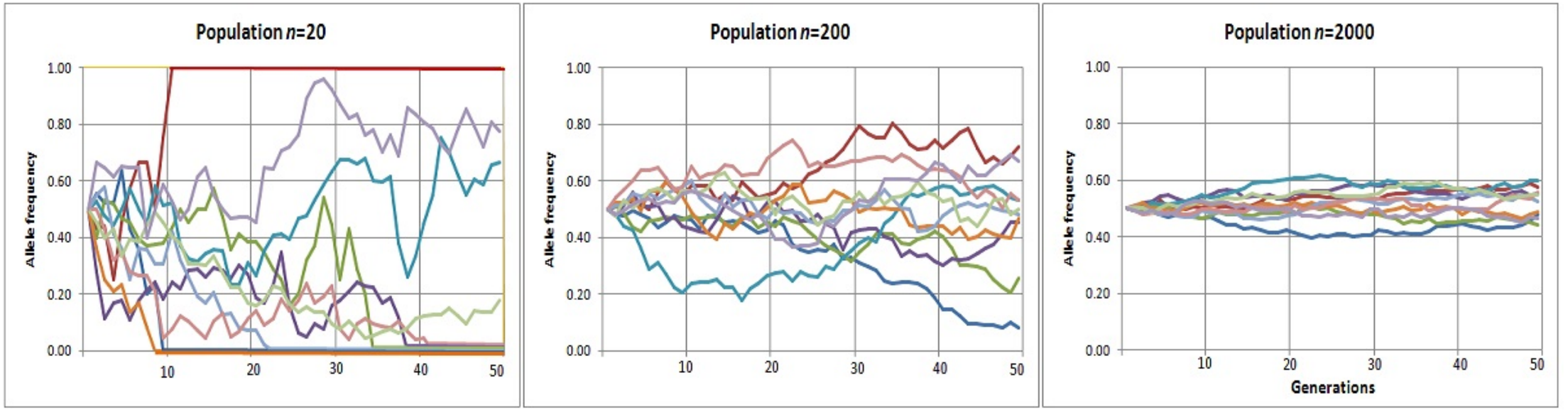
# Wright-Fisher Model

- When an allele reaches a frequency of 1, we say it has **fixed.**
- When an allele reaches a frequency of 0, we say it has **gone extinct.**
- **Assuming no recurrent mutation**, alleles that fix or go extinct remain so forever



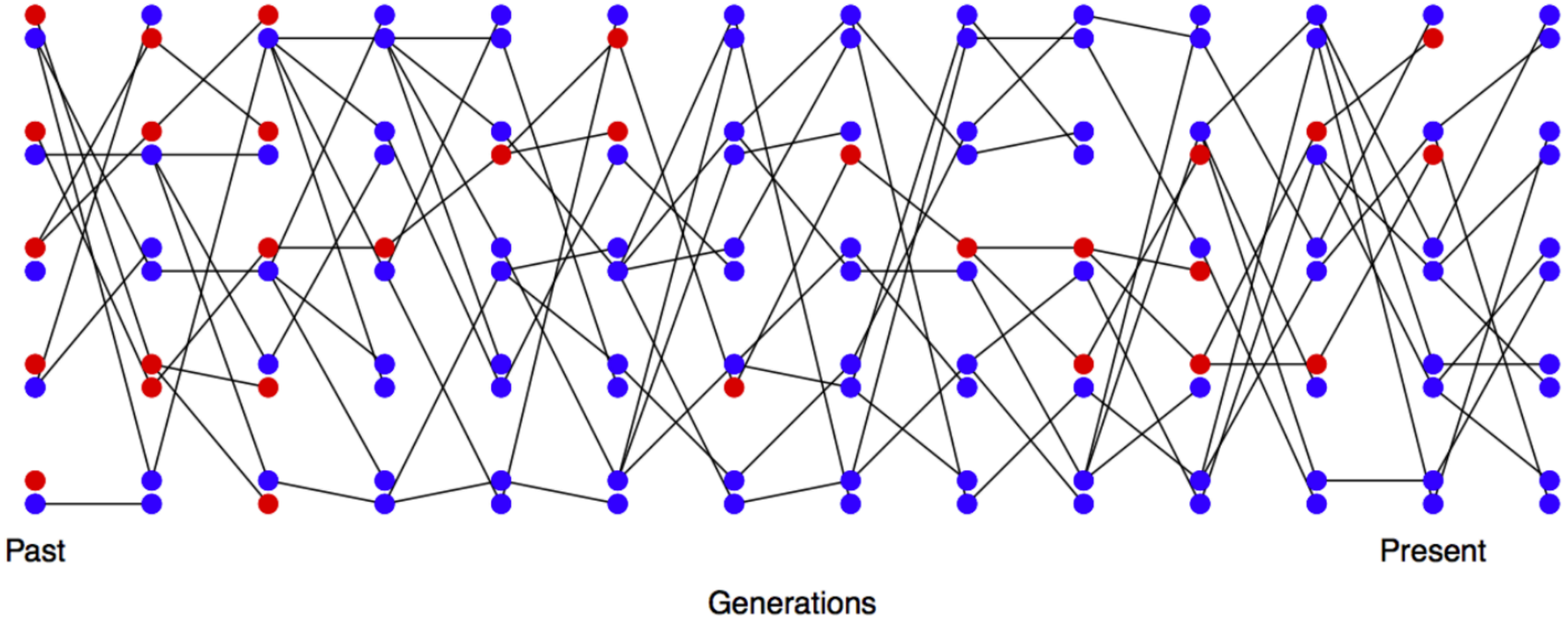
# Wright-Fisher Model

- When an allele reaches a frequency of 1, we say it has **fixed**.
- When an allele reaches a frequency of 0, we say it has **gone extinct**.
- **Assuming no recurrent mutation**, alleles that fix or go extinct remain so forever
- All alleles must eventually fix or go extinct, given enough time.



# Thinking forwards in time in time

# Thinking forwards in time in time

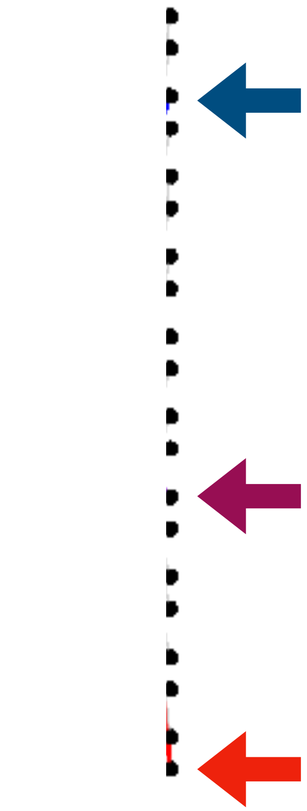


# Thinking backwards in time



Present-day population

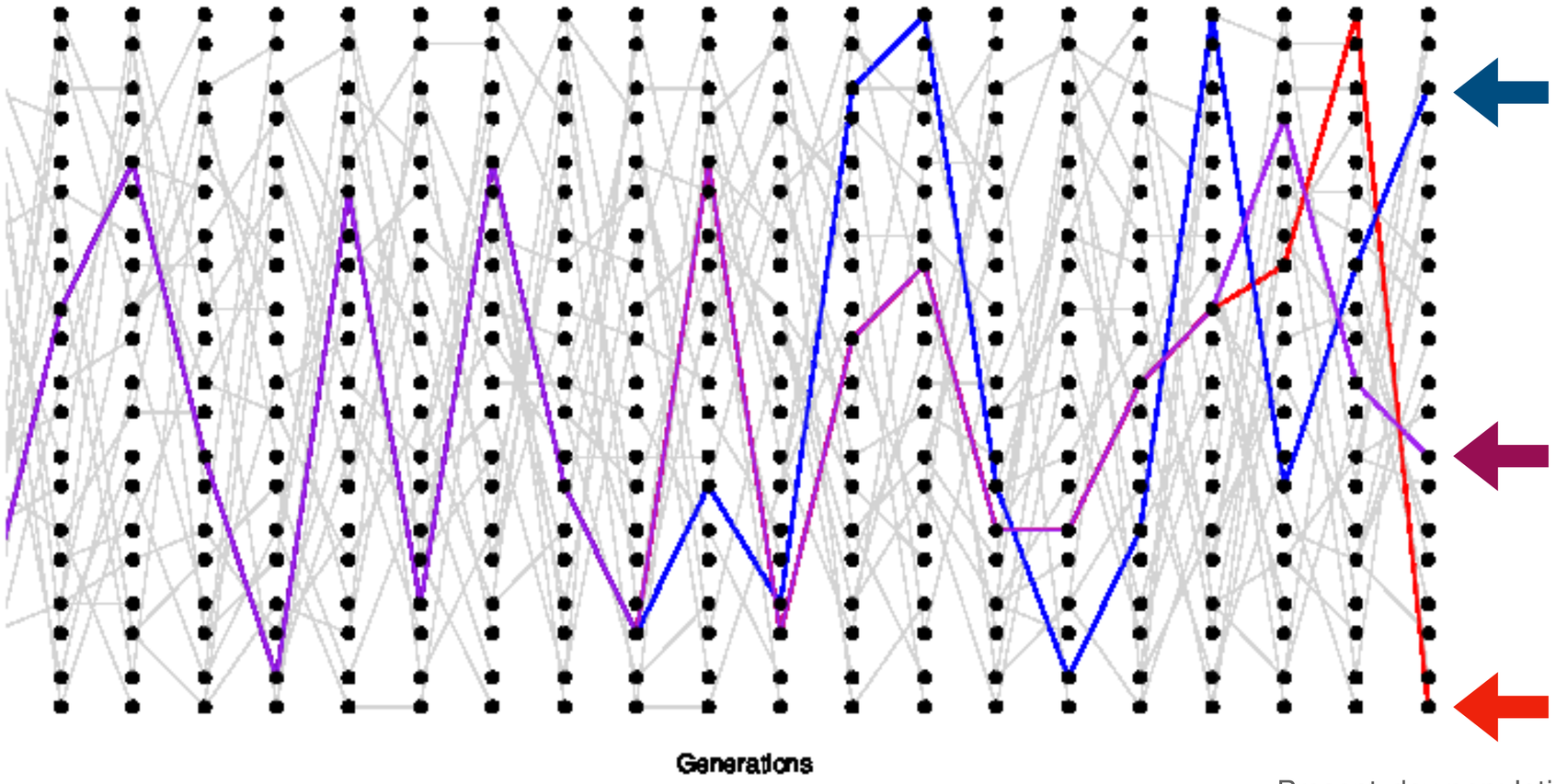
# Thinking backwards in time



We sample n=3 alleles in a population of size 2N=20

Present-day population

# Thinking backwards in time

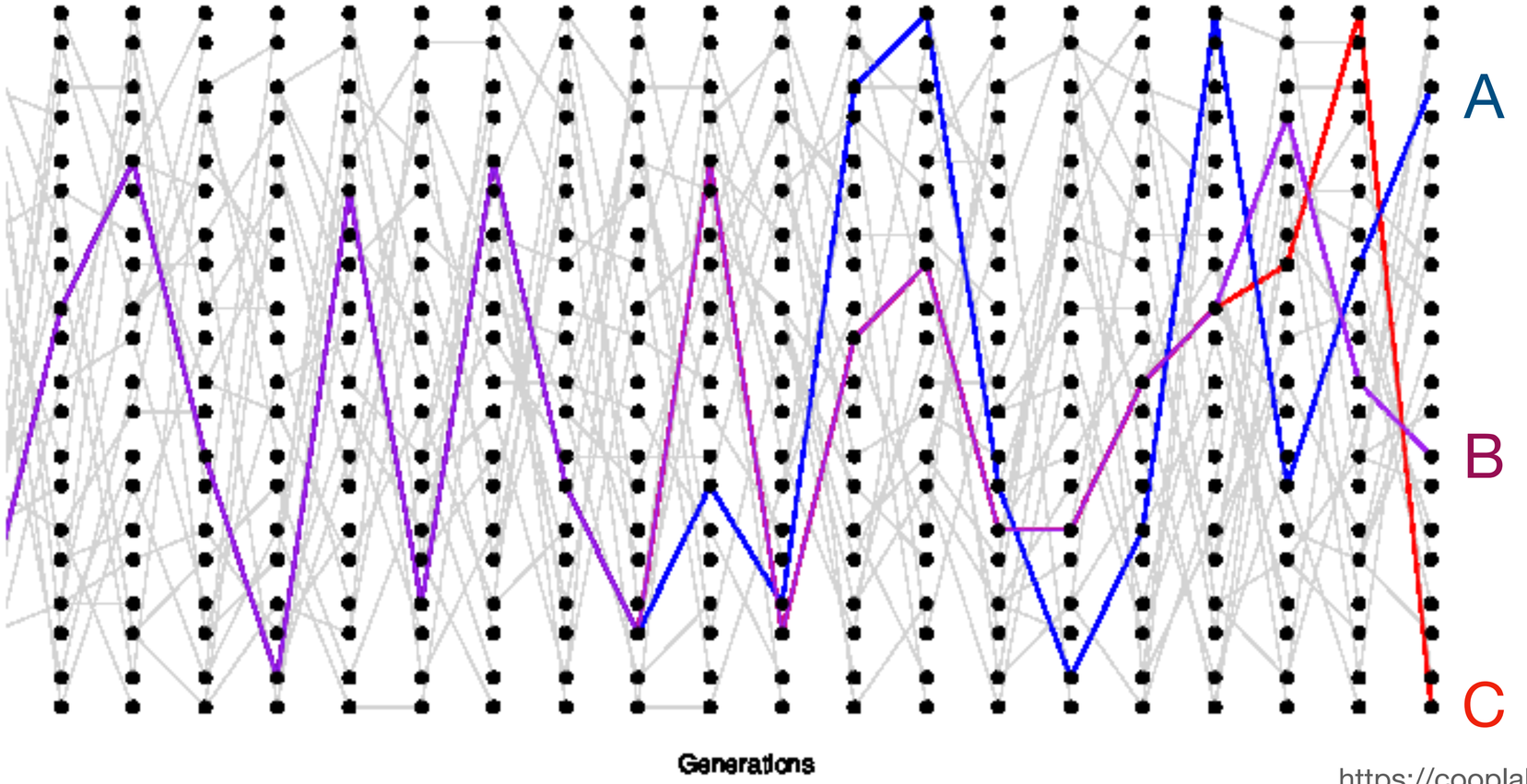



We sample n=3 alleles in a population of size 2N=20

Present-day population

We trace their lineages into the past

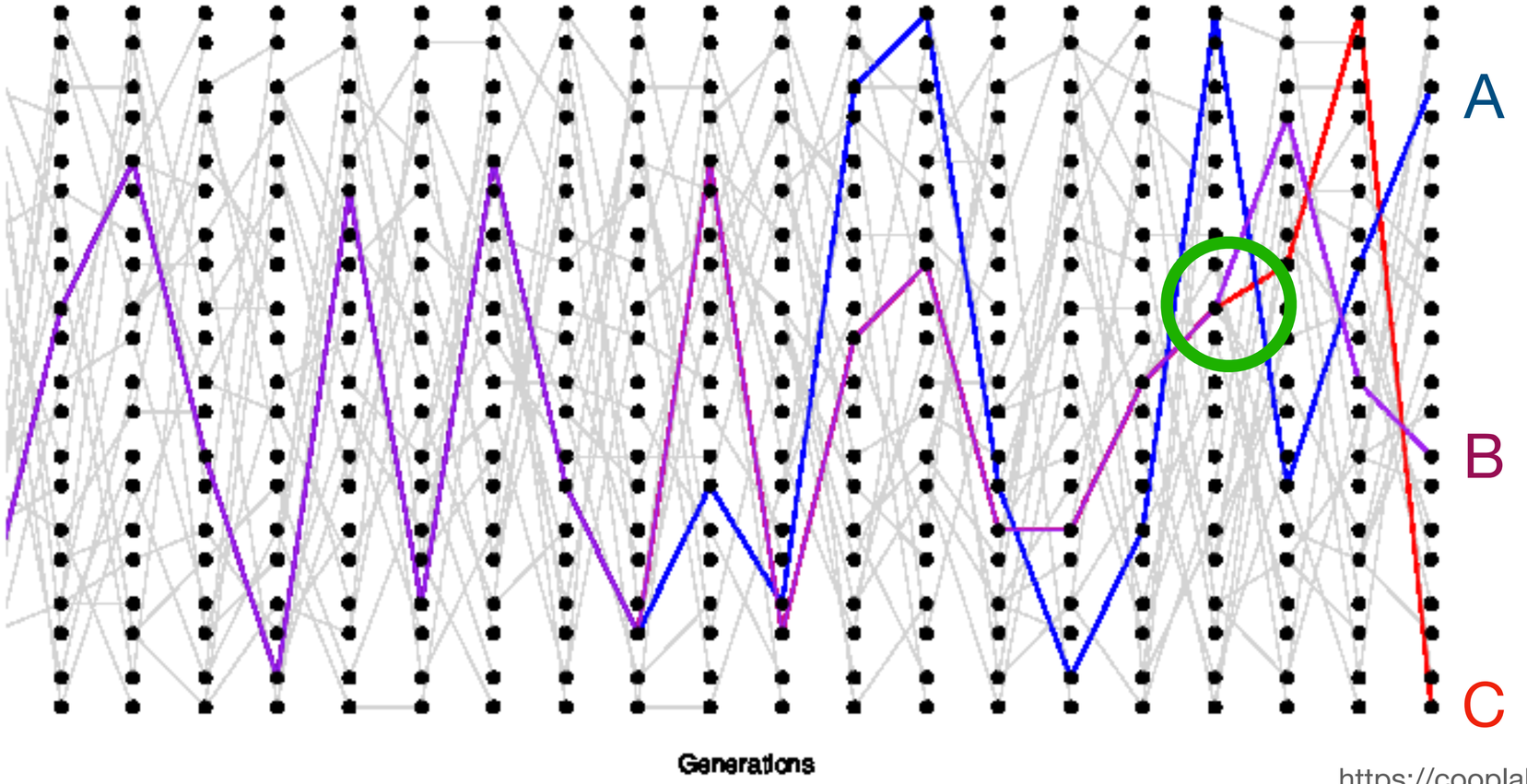
# Thinking backwards in time




https://cooplab.github.io/popgen-notes/

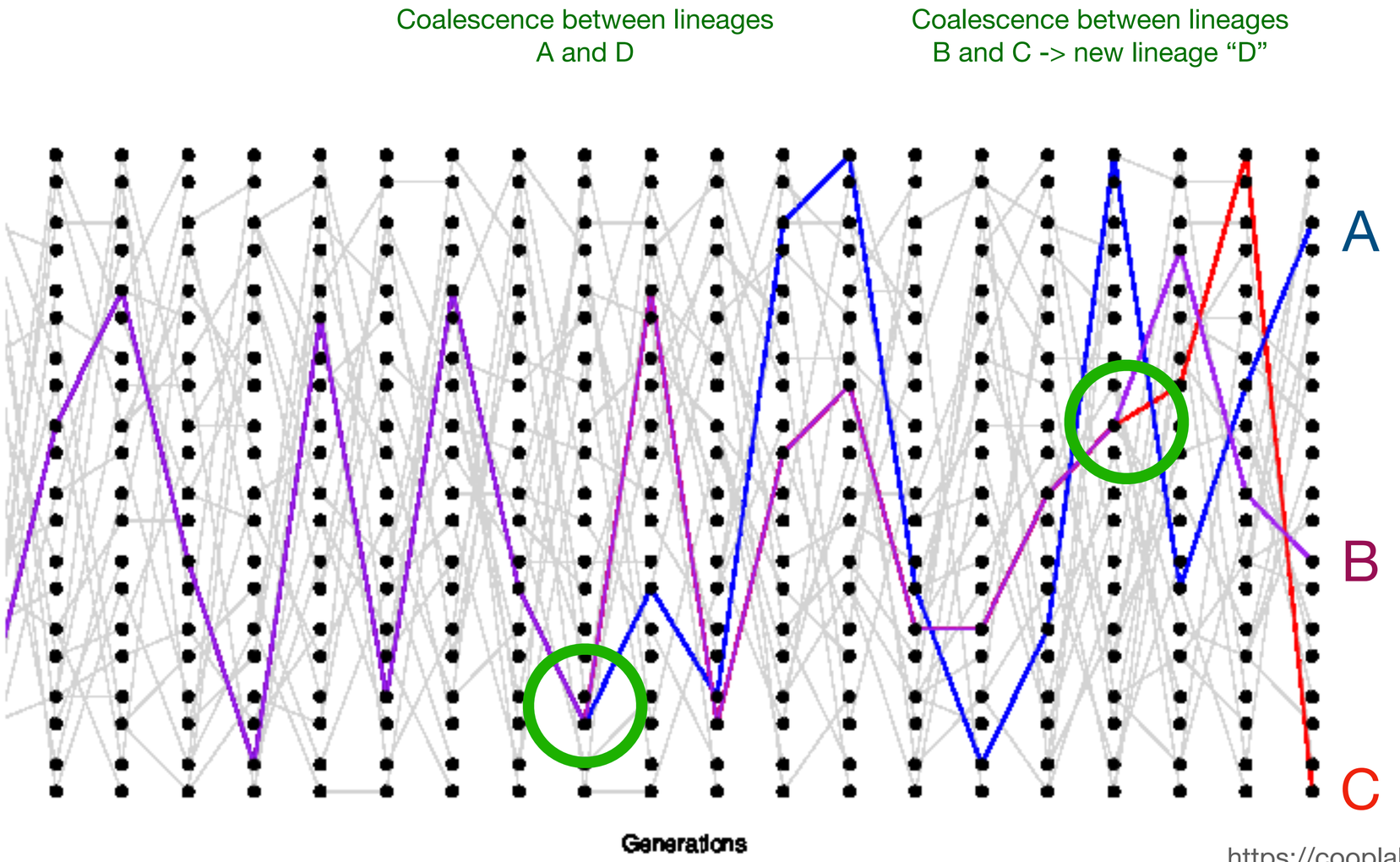
# Thinking backwards in time

Coalescence between lineages
B and C -> new lineage “D”



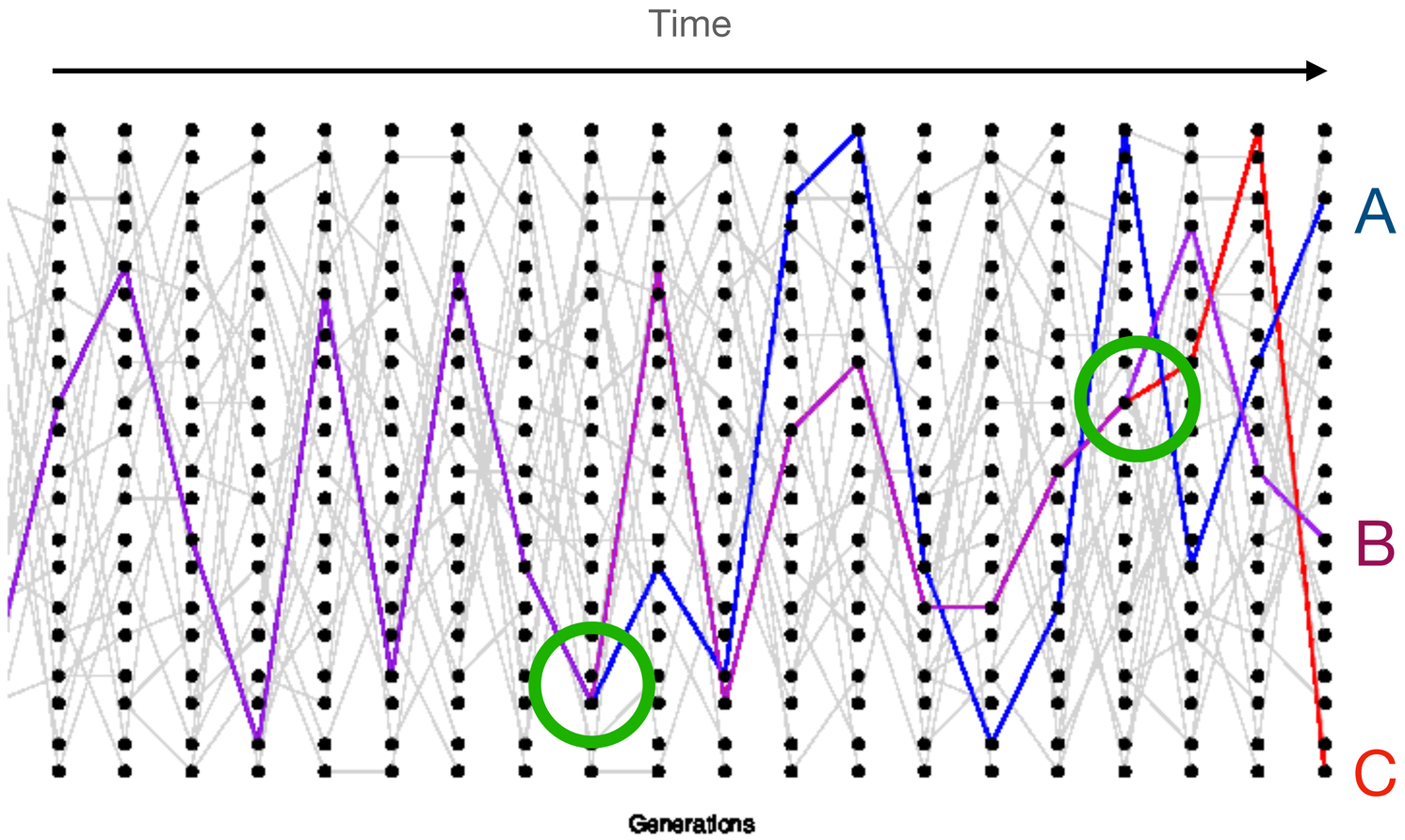

https://cooplab.github.io/popgen-notes/

# Thinking backwards in time



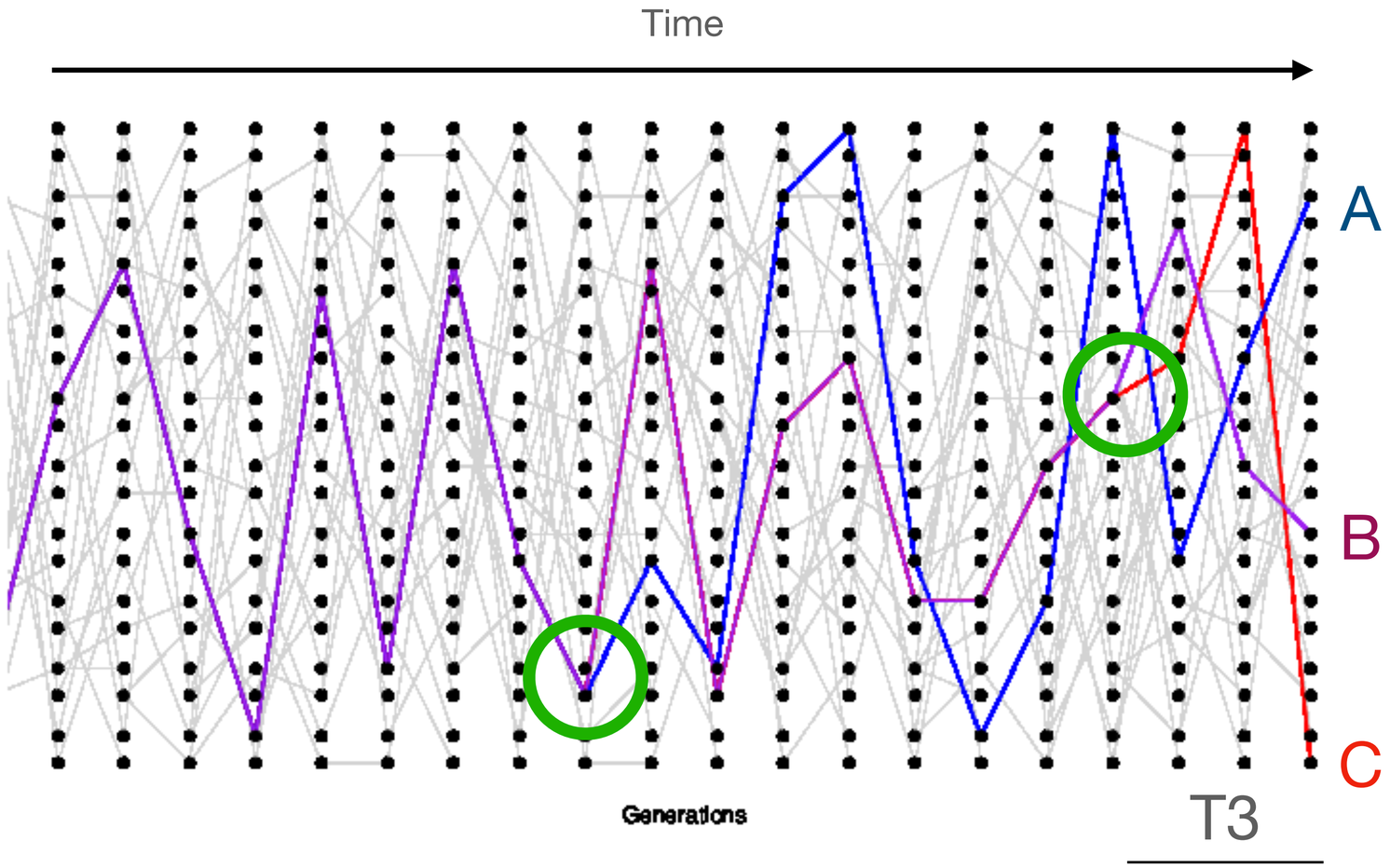

https://cooplab.github.io/popgen-notes/

# Thinking backwards in time



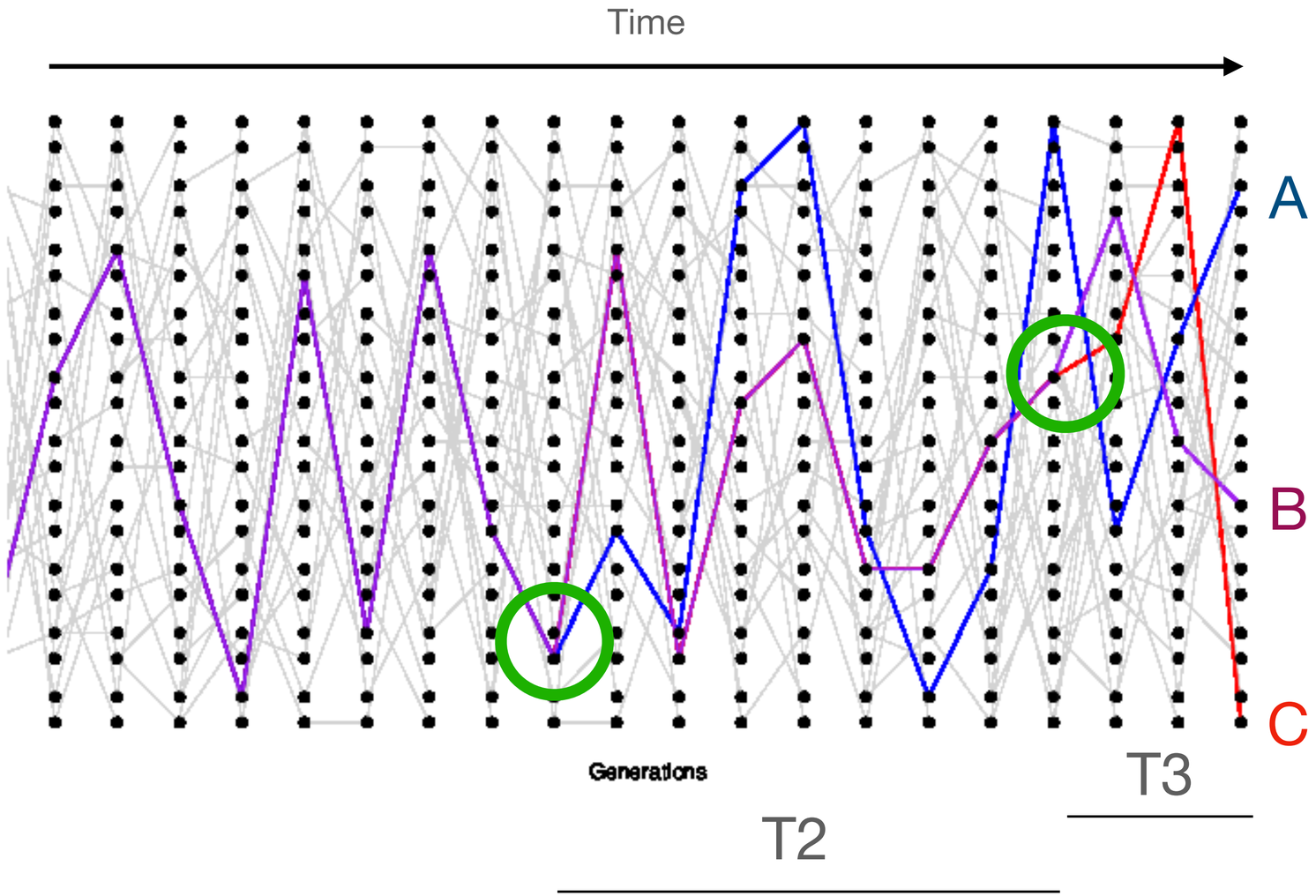

https://cooplab.github.io/popgen-notes/

# Thinking backwards in time



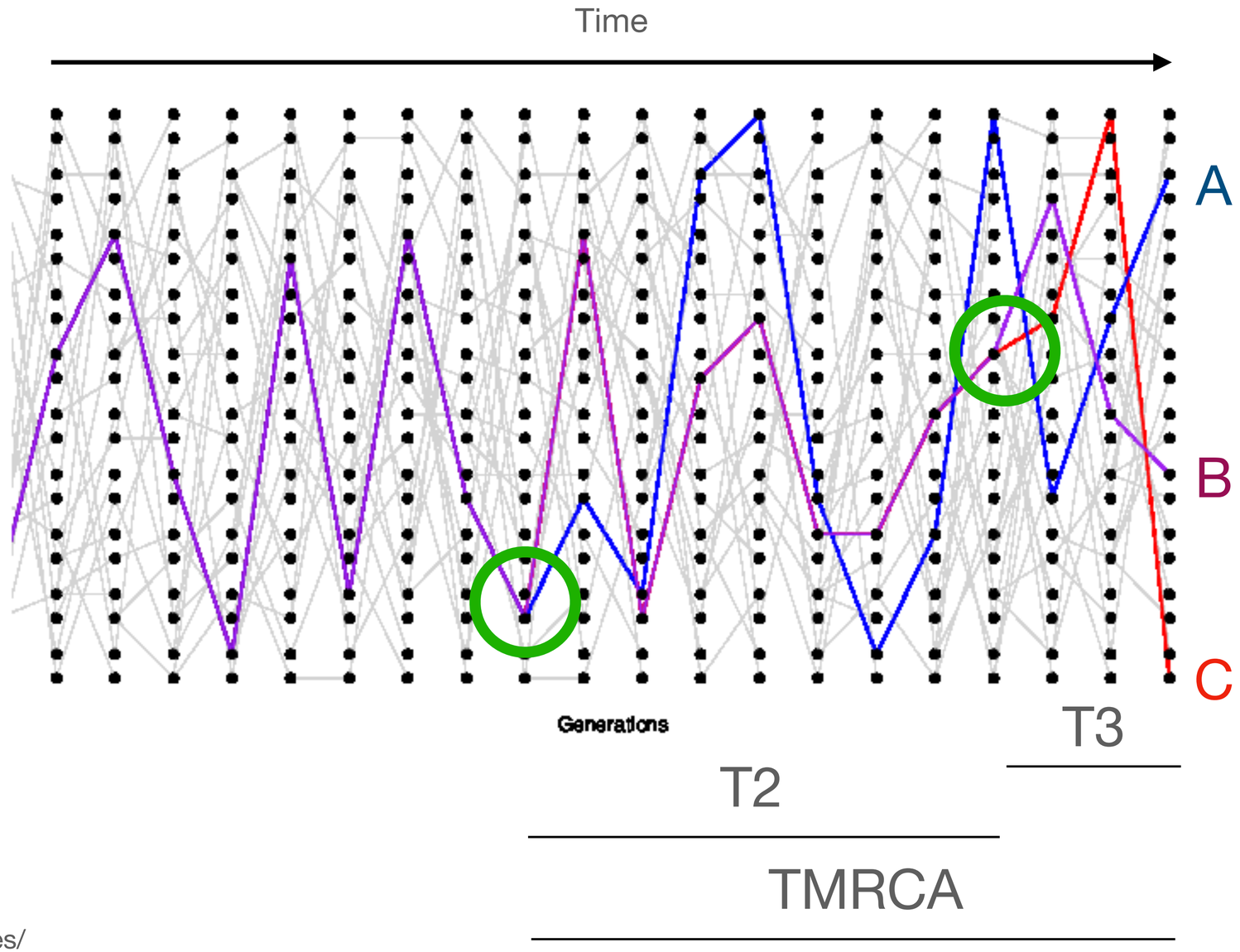

https://cooplab.github.io/popgen-notes/

# Thinking backwards in time




https://cooplab.github.io/popgen-notes/

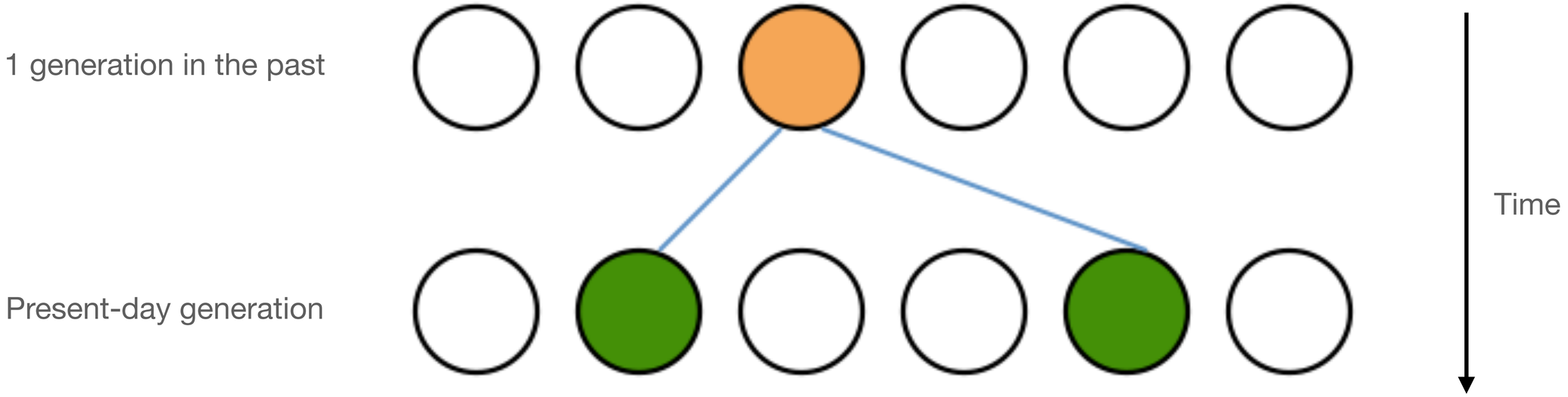
# Thinking backwards in time




https://cooplab.github.io/popgen-notes/

# Coalescence in a sample of two sequences

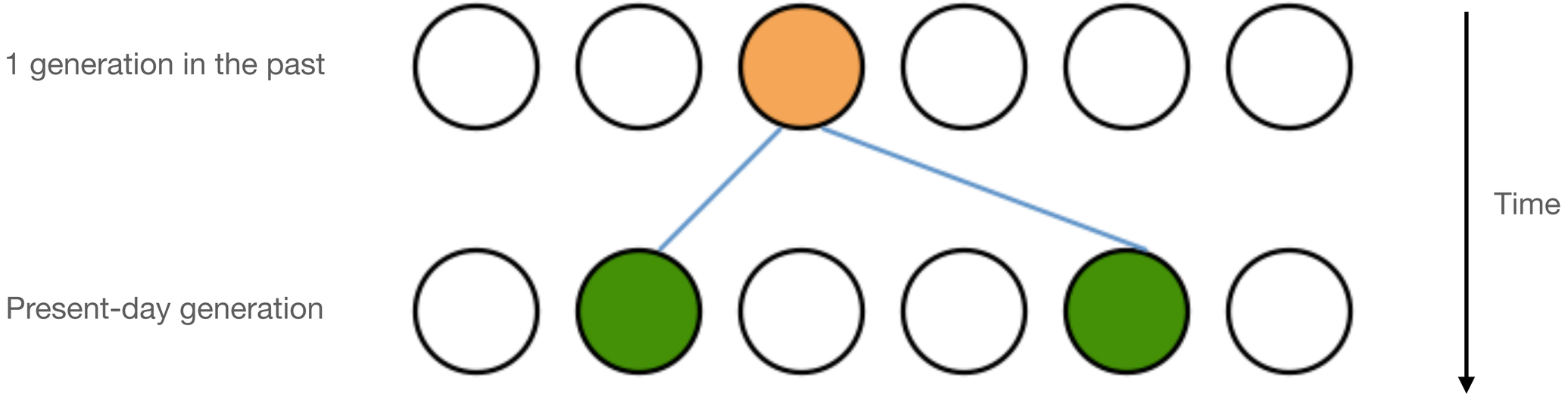
P[ 2 samples have the same parent in the previous generation ] =



# Coalescence in a sample of two sequences

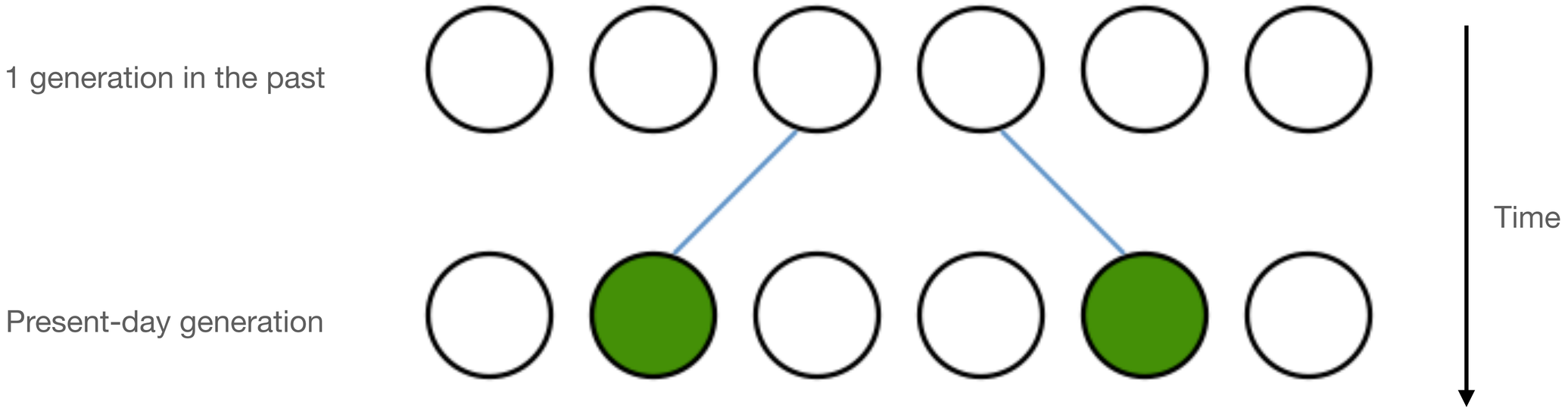
P[ 2 samples have the same parent in the previous generation ] =

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



# Coalescence in a sample of two sequences

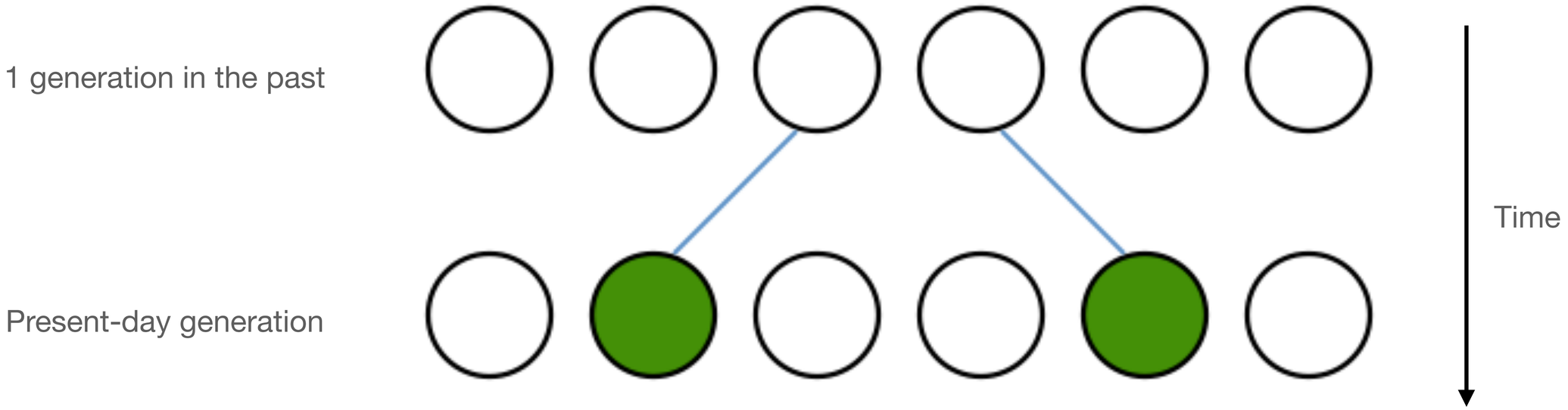
P[ 2 samples do not have the same parent in the previous generation ] =



# Coalescence in a sample of two sequences

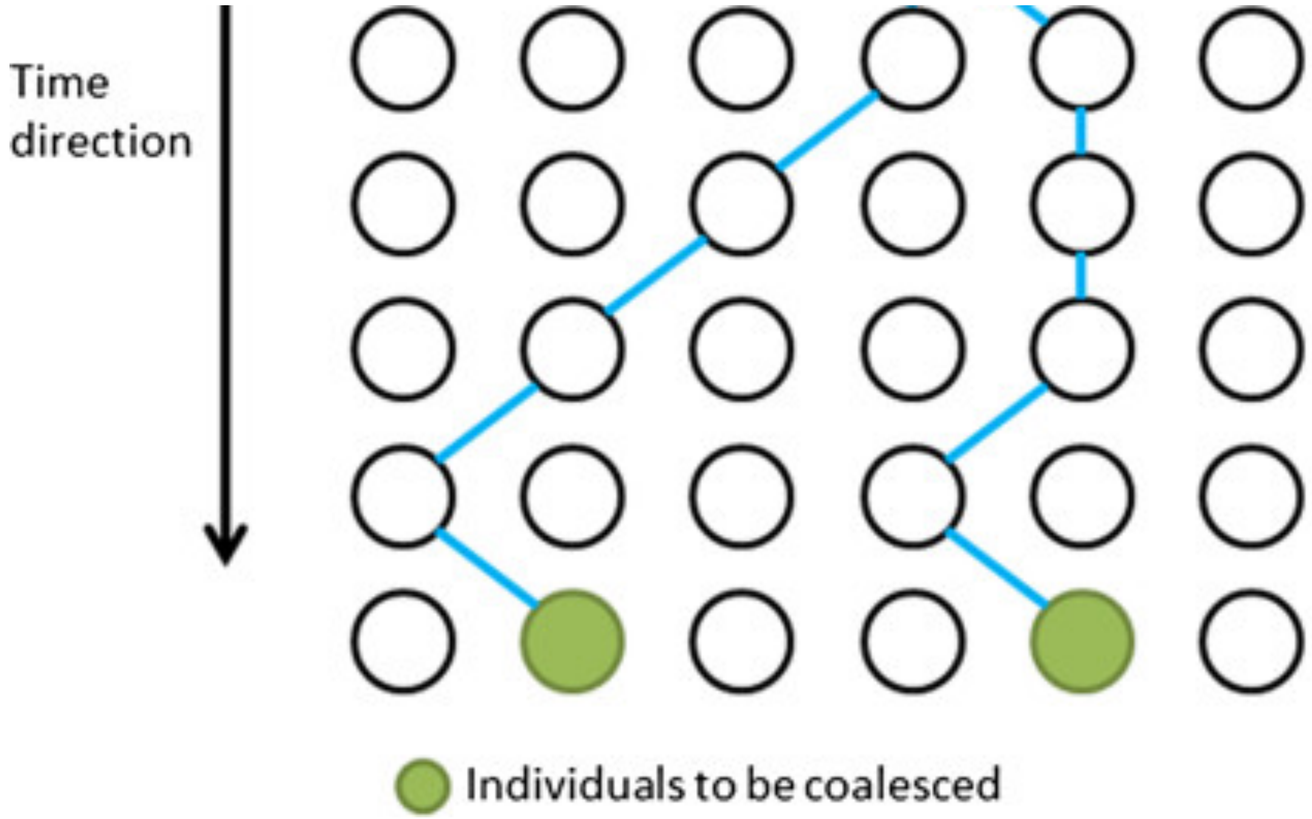
P[ 2 samples do not have the same parent in the previous generation ] =

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



# Coalescence in a sample of two sequences

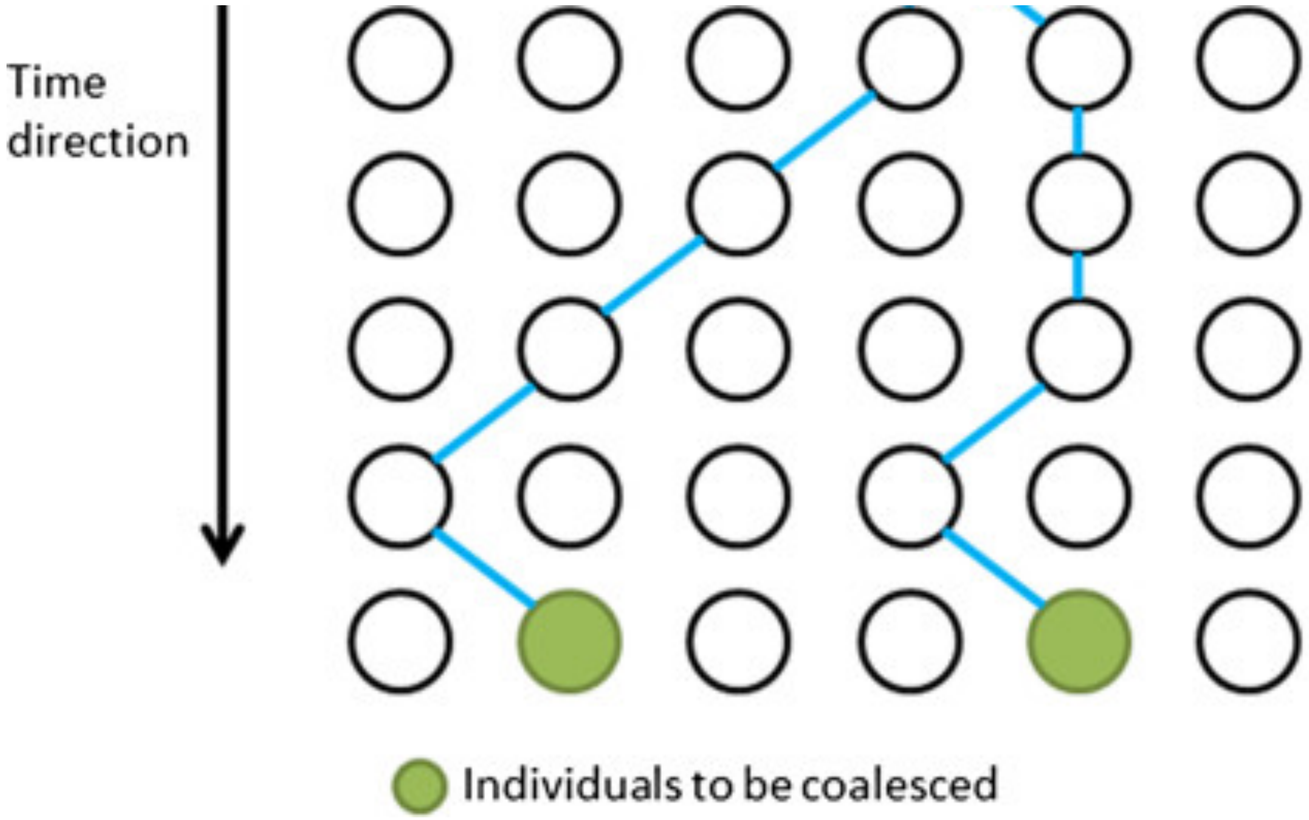
P[ 2 samples do not find a common ancestor in **r** generations ] =



# Coalescence in a sample of two sequences

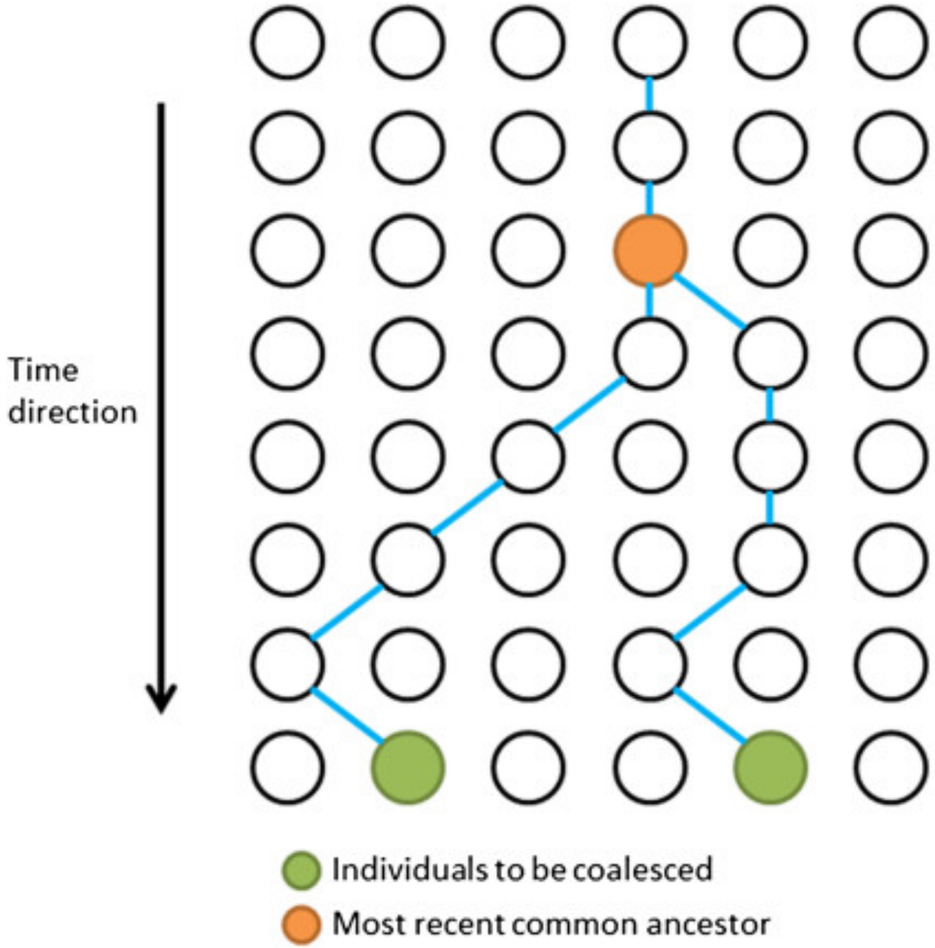
P[ 2 samples do not find a common ancestor in **r** generations ] =

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



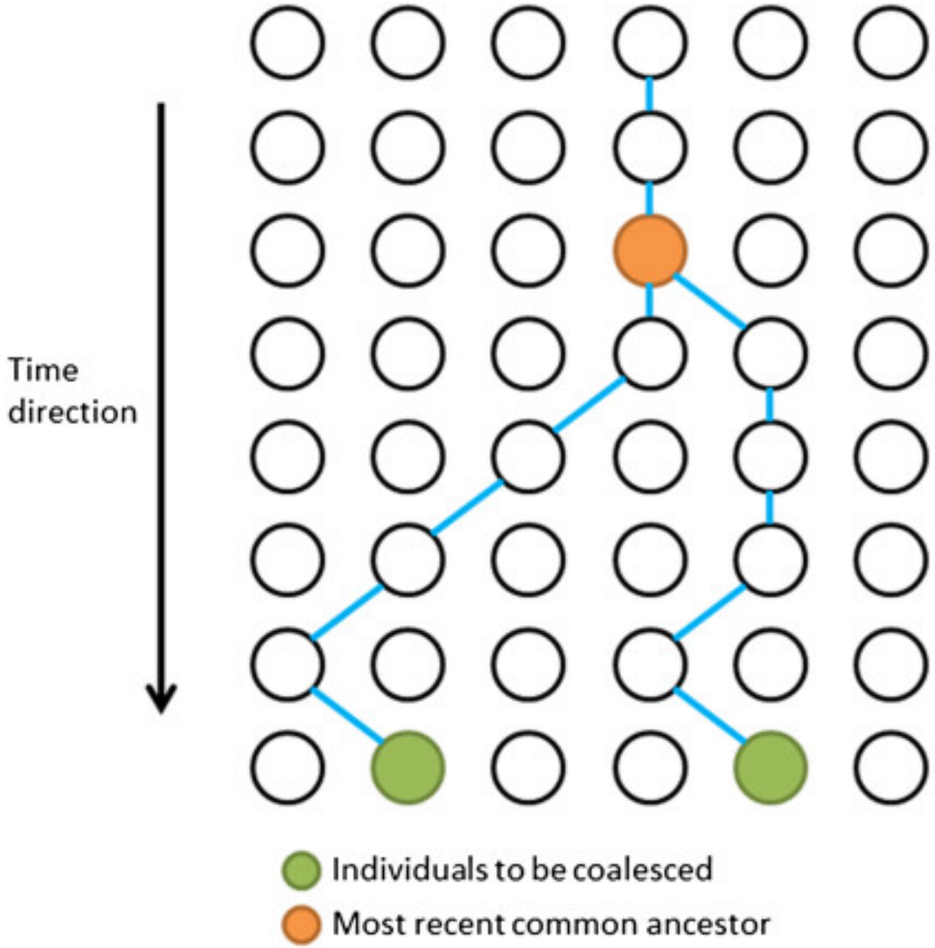
# Coalescence in a sample of two sequences

P[ 2 samples find a common ancestor in **exactly r** generations ] =



# Coalescence in a sample of two sequences

P[ 2 samples find a common ancestor in **exactly r** generations ] =

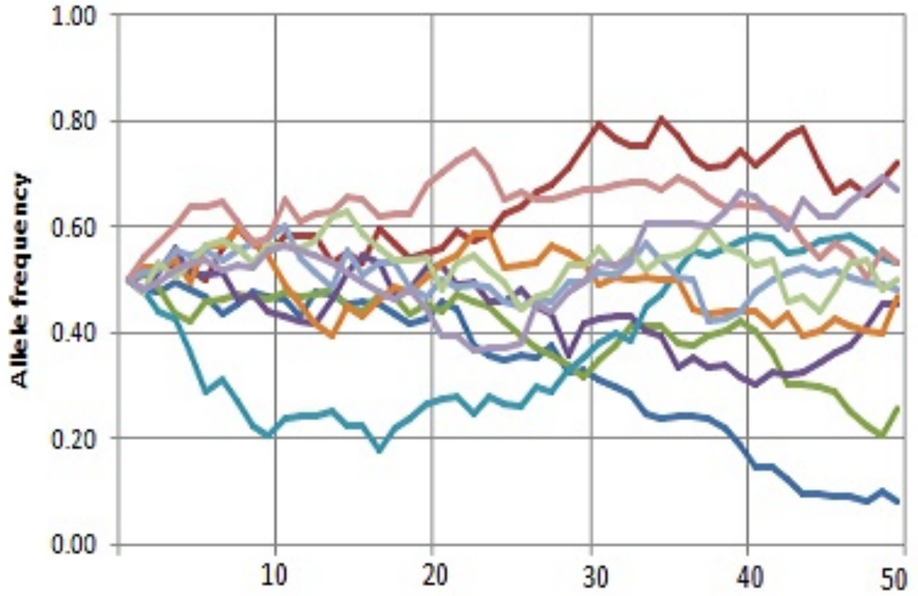



$$\left(1-\frac{1}{2N}\right)^{r-1}\frac{1}{2N}$$

# Wright-Fisher exercises

Follow the instructions in the Wright-Fisher exercise prompt:

https://github.com/FerRacimo/CopenhagenTutorial/blob/master/WrightFisherTutorial.md



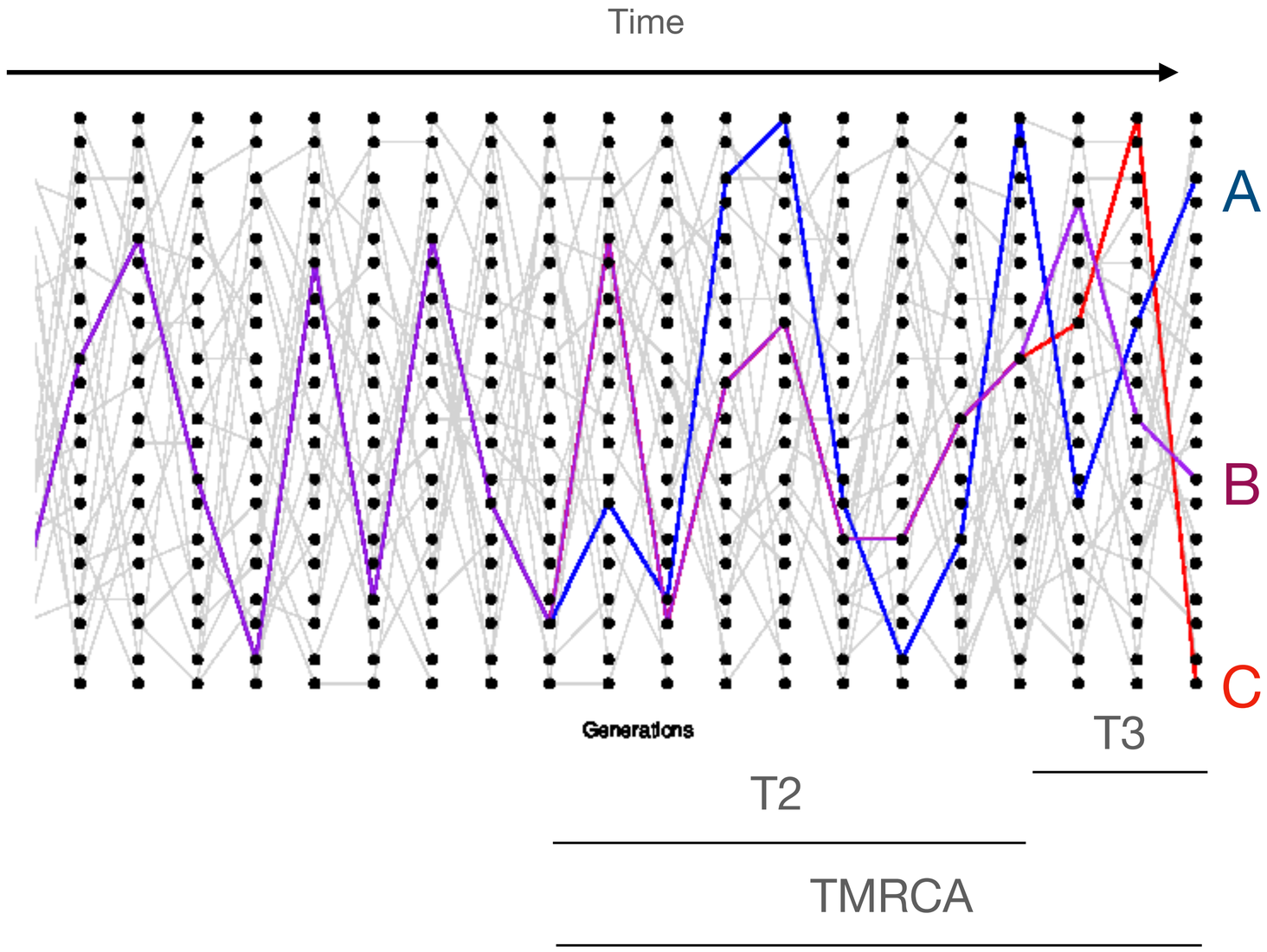
# Today

- Terminology
- Wright-Fisher Model and Genetic Drift
- Intro to coalescent theory
- Mutations and the coalescent

# Today

- Terminology
- Wright-Fisher Model and Genetic Drift
- Intro to coalescent theory
- Mutations and the coalescent

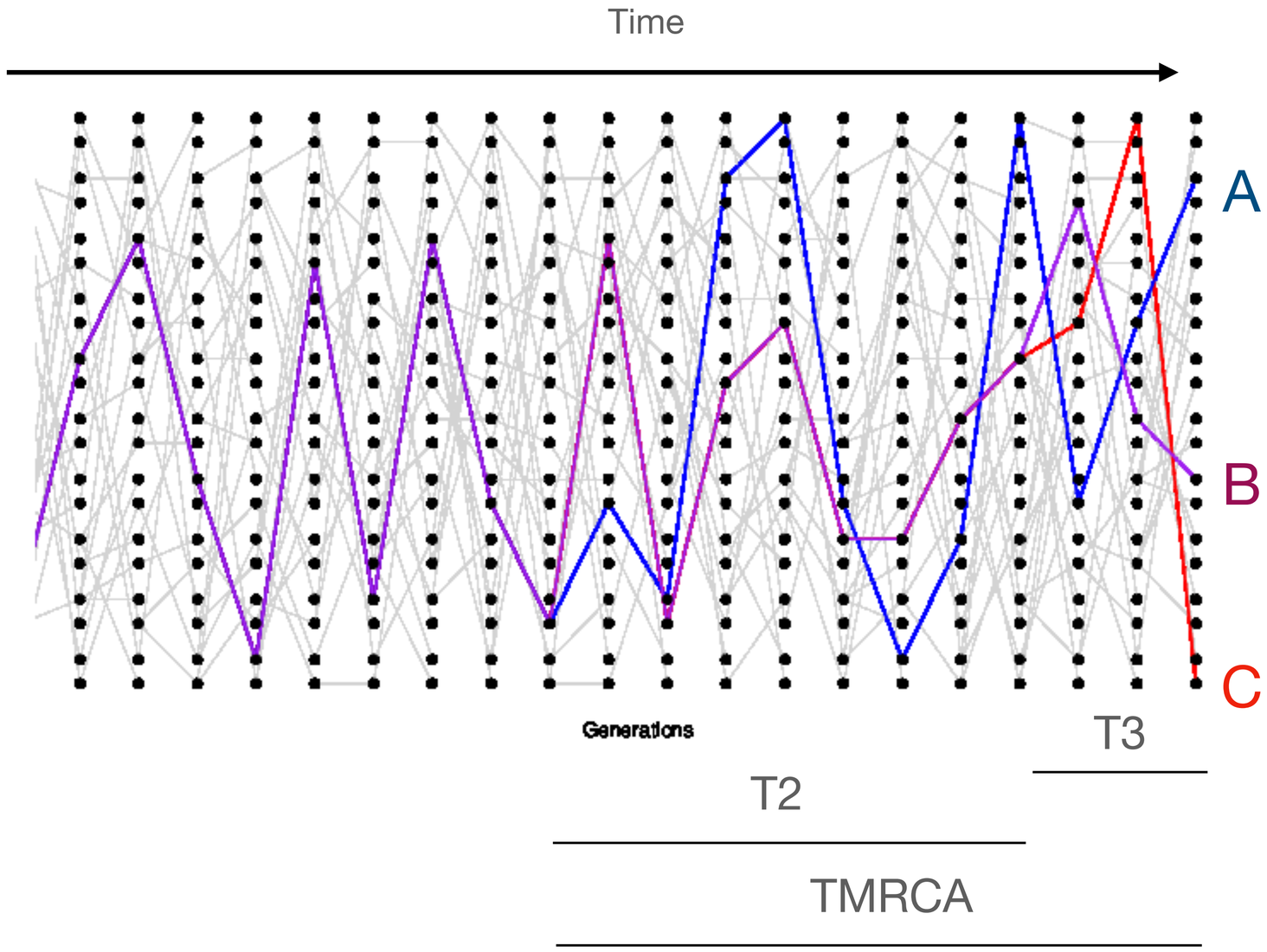
# Motivation: tractability




https://cooplab.github.io/popgen-notes/

# Motivation: tractability

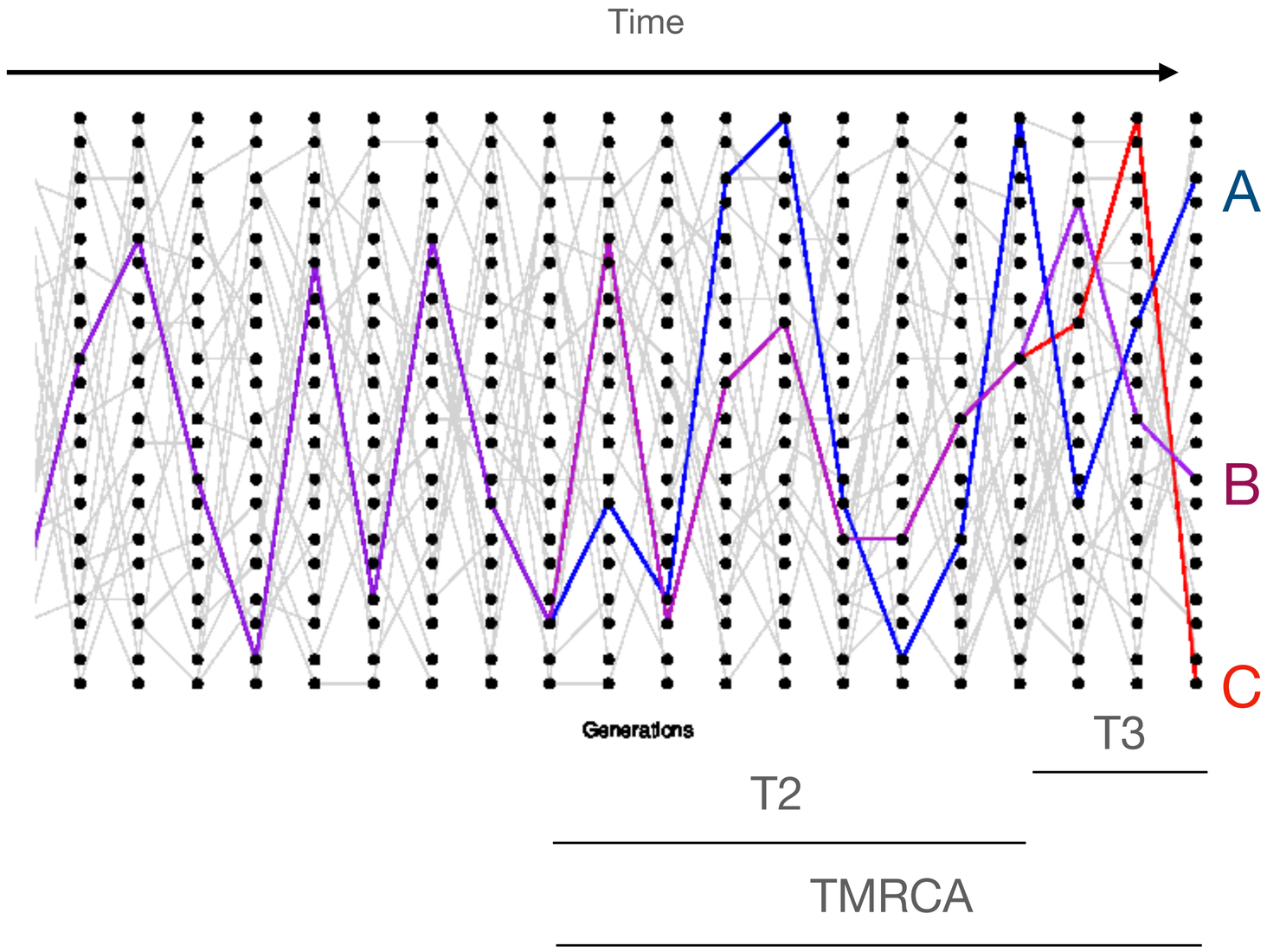
- Hard to **keep track of all alleles at each time step**




https://cooplab.github.io/popgen-notes/

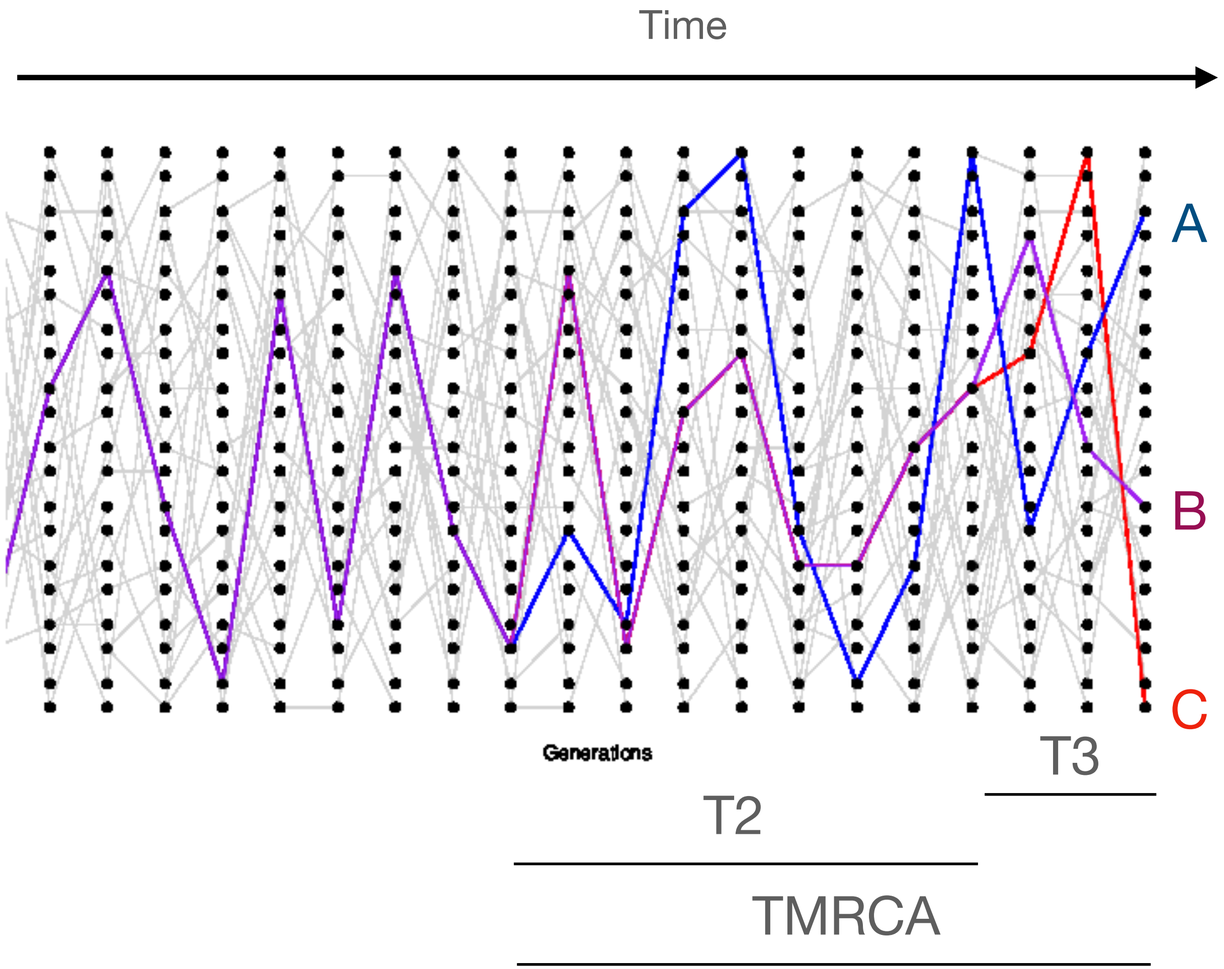
# Motivation: tractability

- Hard to **keep track of all alleles at each time step**
- Example: given that we sample 3 individuals in a population of size 2N=20, what is the **expected time till all 3 individuals find a common ancestor**?



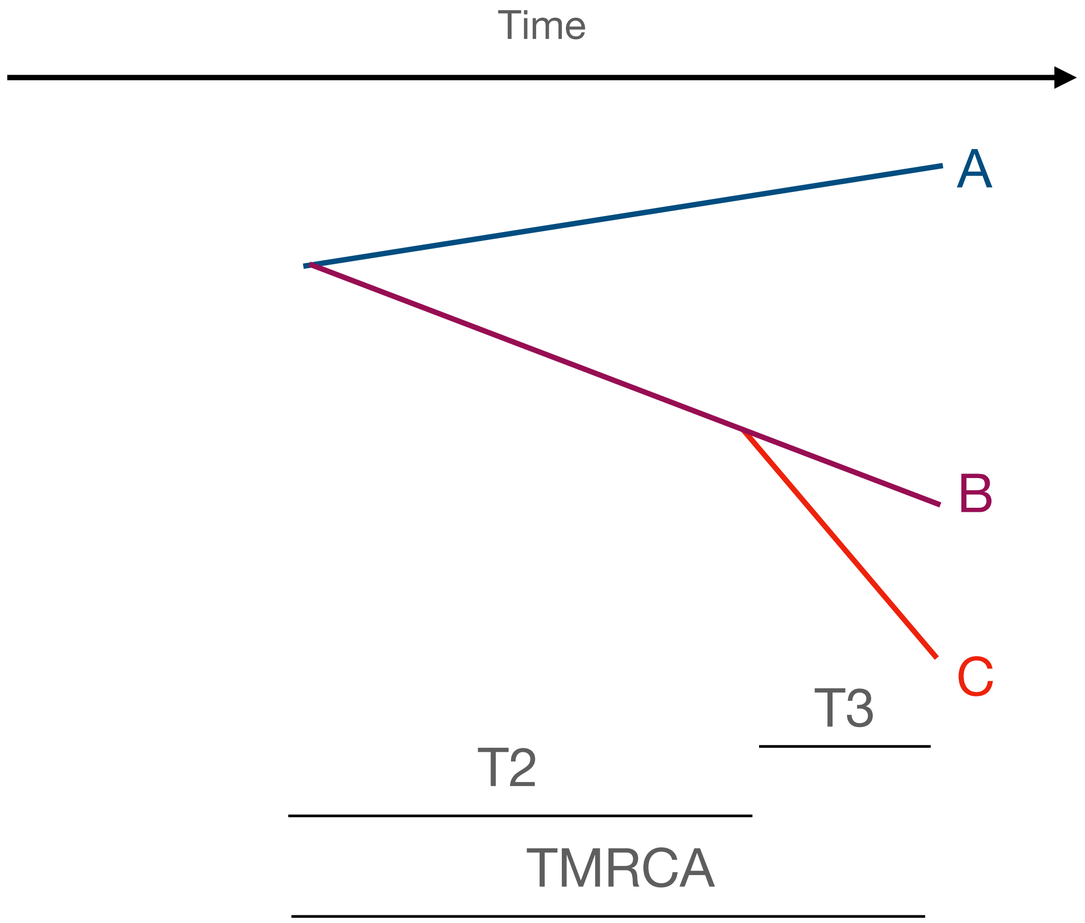

https://cooplab.github.io/popgen-notes/

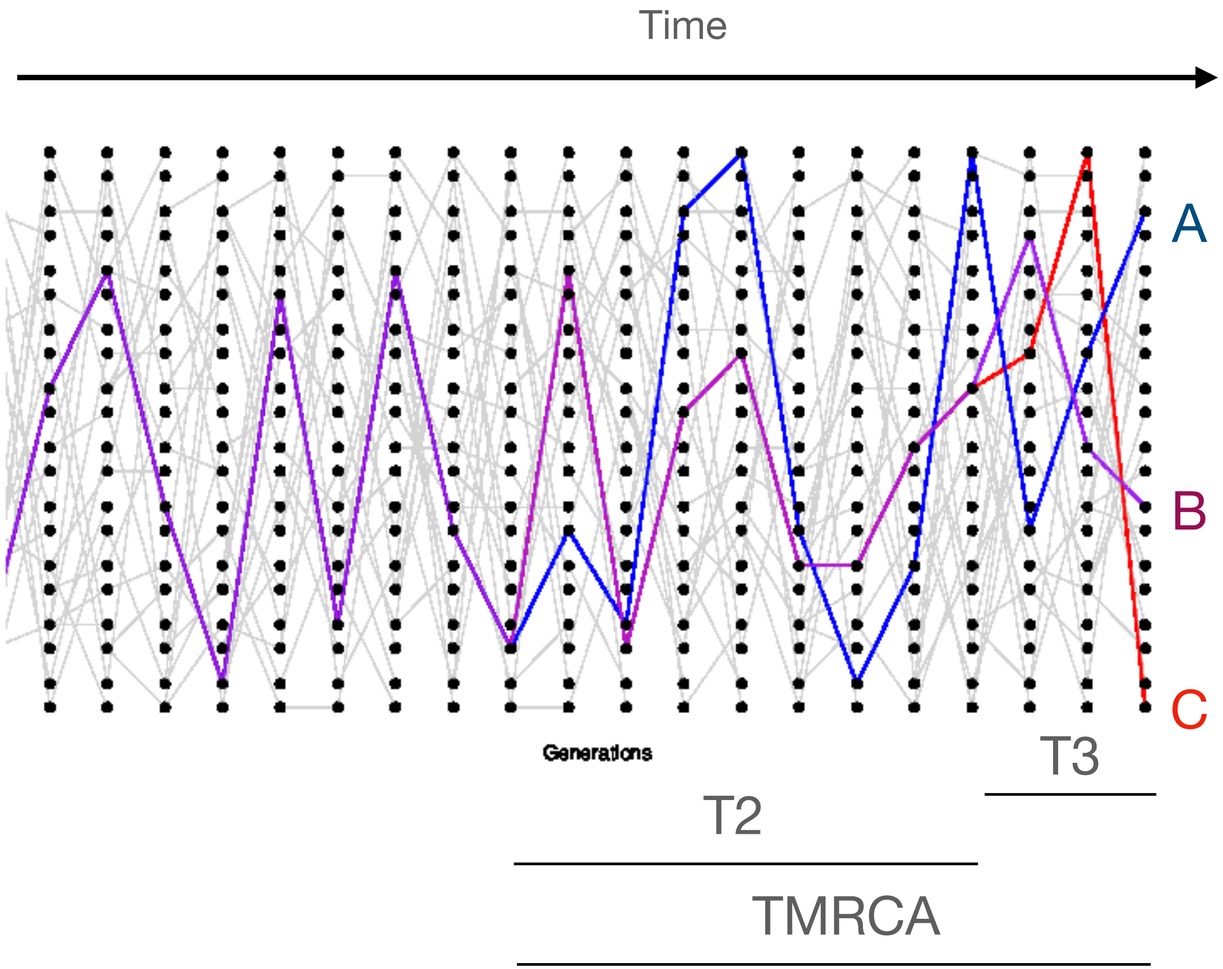
# Solution: keep track of sampled lineages only




https://cooplab.github.io/popgen-notes/

# Solution: keep track of sampled lineages only




https://cooplab.github.io/popgen-notes/

# Coalescent theory

# Coalescent theory

- A way to model how the genealogies of a sample behave over long time periods.

# Coalescent theory

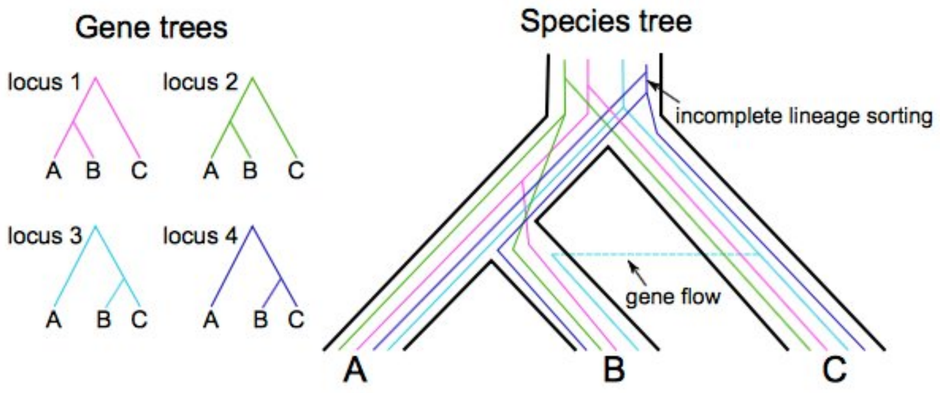
- A way to model how the genealogies of a sample behave over long time periods.
- Basis of many **inference** and **simulation** tools in population genetics today

# Coalescent theory

- A way to model how the genealogies of a sample behave over long time periods.
- Basis of many **inference** and **simulation** tools in population genetics today

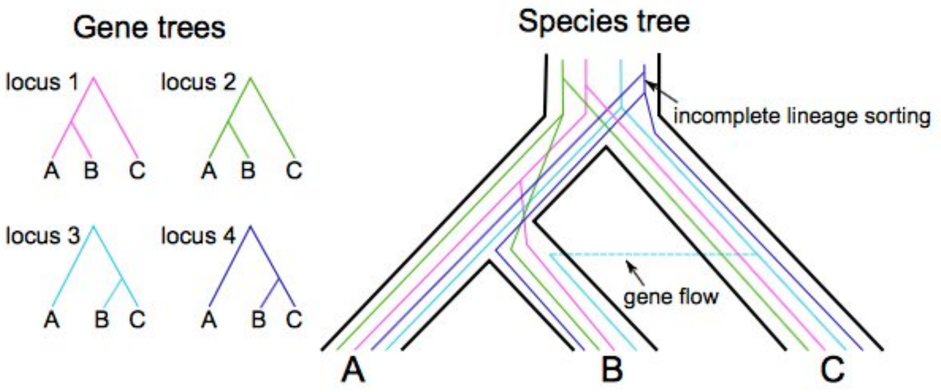
# Coalescent theory

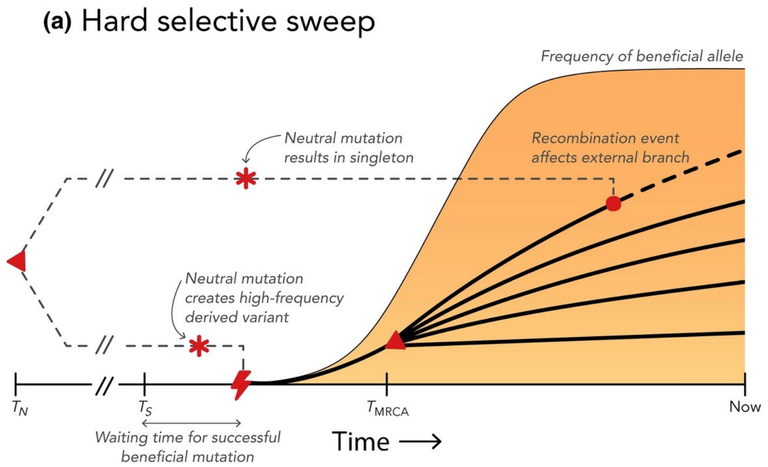
- A way to model how the genealogies of a sample behave over long time periods.
- Basis of many **inference** and **simulation** tools in population genetics today

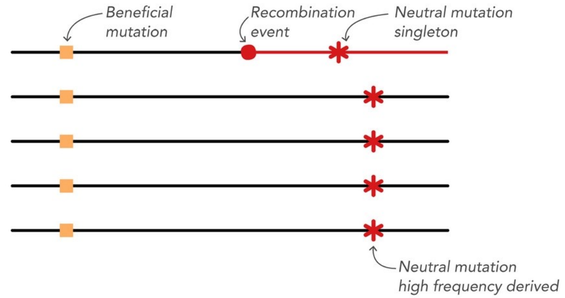



# Coalescent theory

- A way to model how the genealogies of a sample behave over long time periods.
- Basis of many **inference** and **simulation** tools in population genetics today

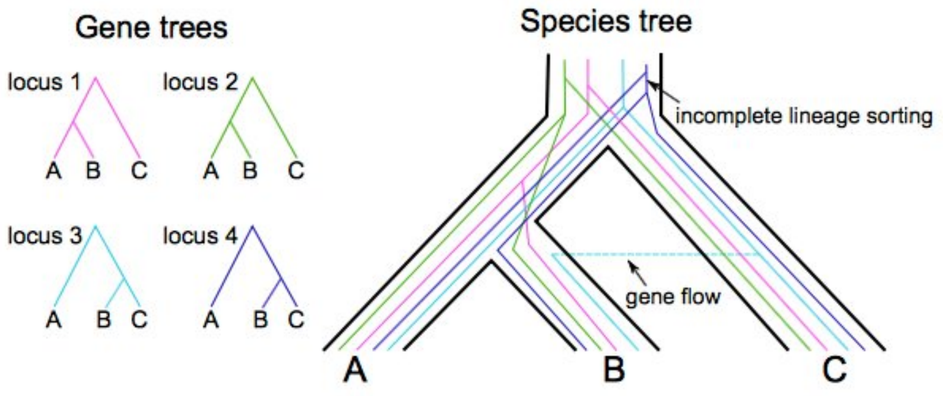



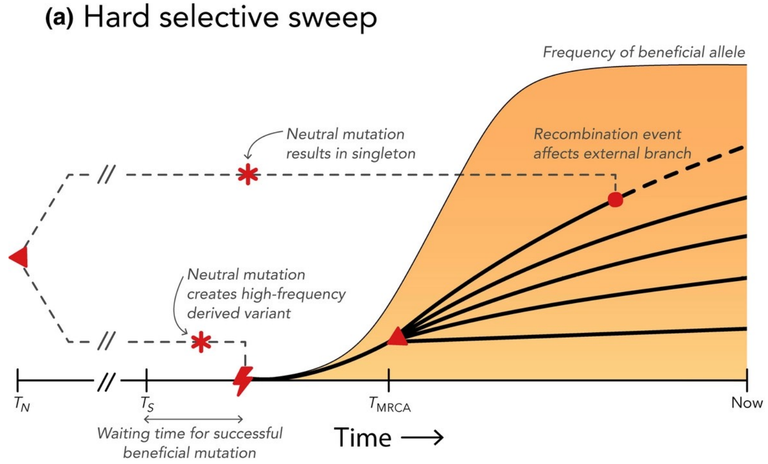



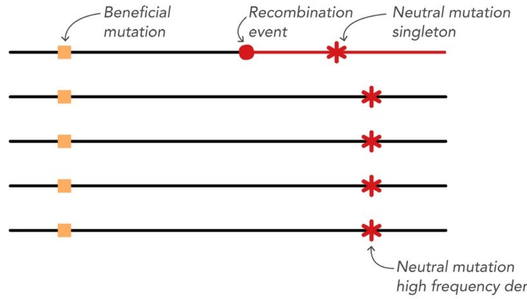
# Coalescent theory

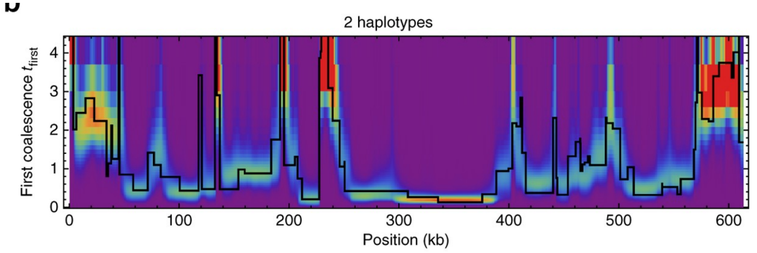
- A way to model how the genealogies of a sample behave over long time periods.
- Basis of many **inference** and **simulation** tools in population genetics today

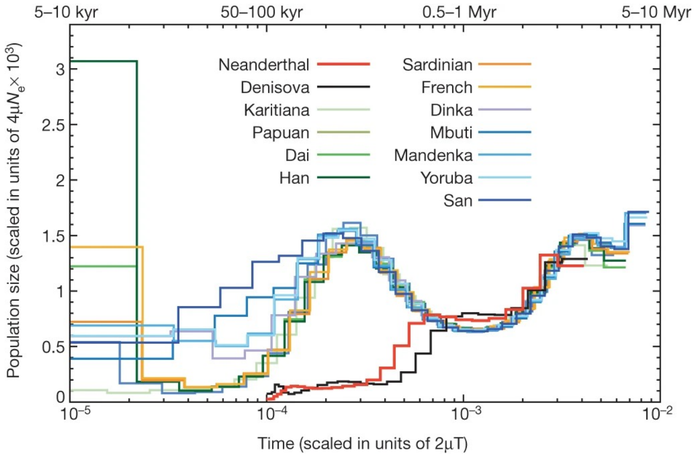
ms prime



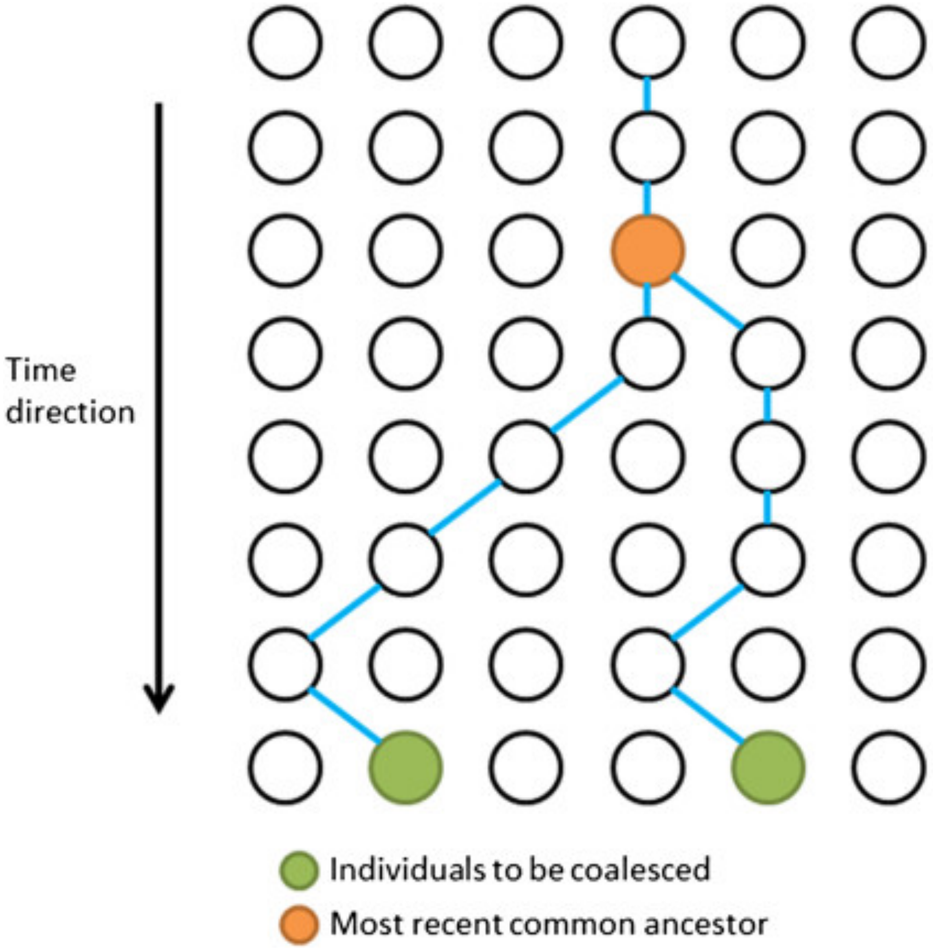



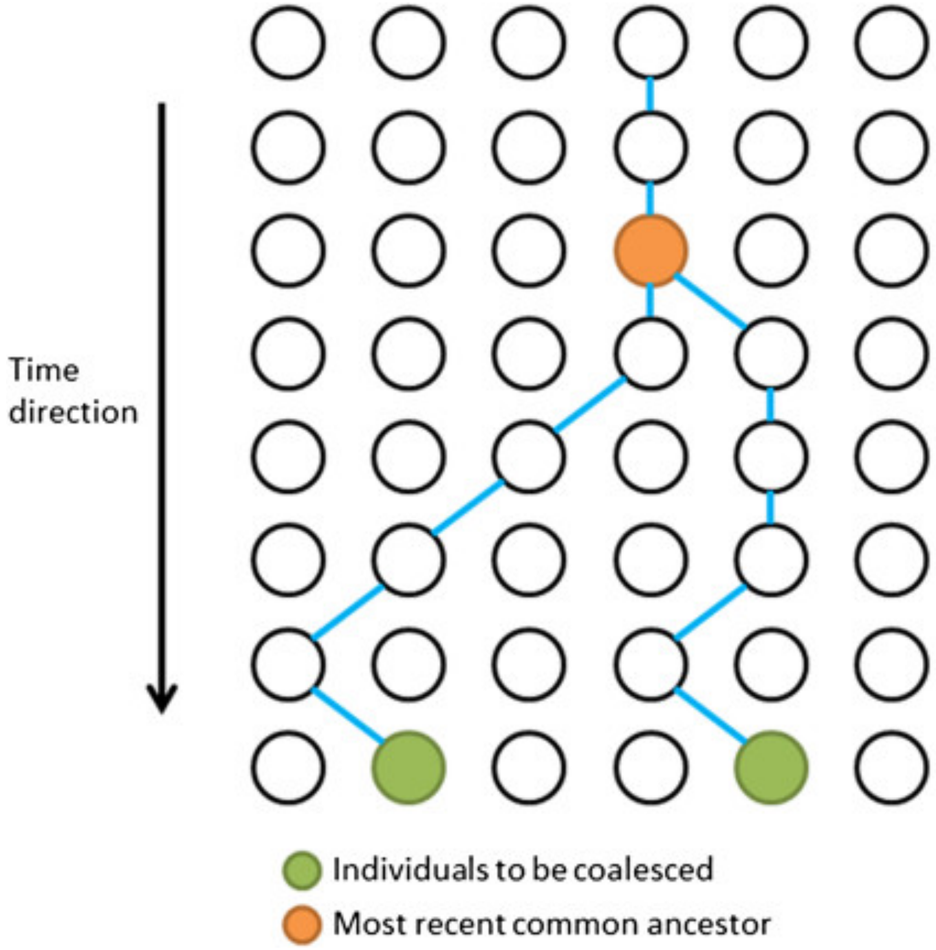
# Coalescence in a sample of two sequences

P[ 2 samples find a common ancestor in **exactly r** generations ] =



# Coalescence in a sample of two sequences

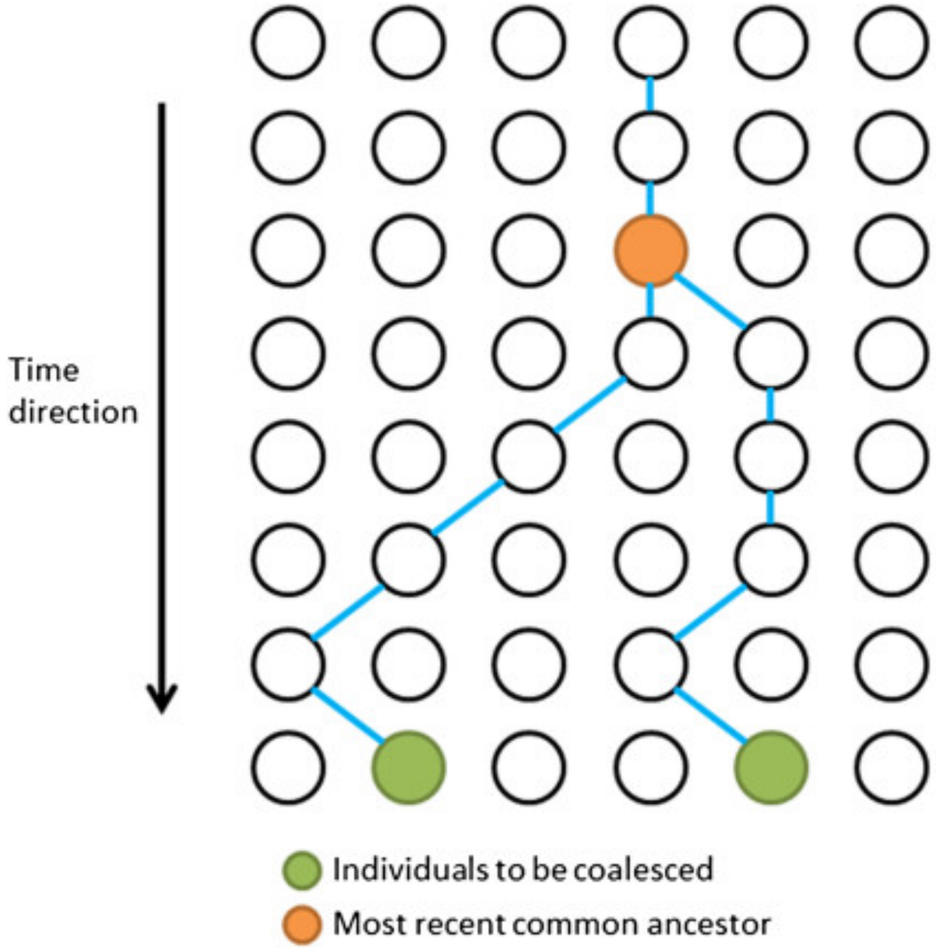
P[ 2 samples find a common ancestor in **exactly r** generations ] =




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

# Coalescence in a sample of two sequences

P[ 2 samples find a common ancestor in **exactly r** generations ] =



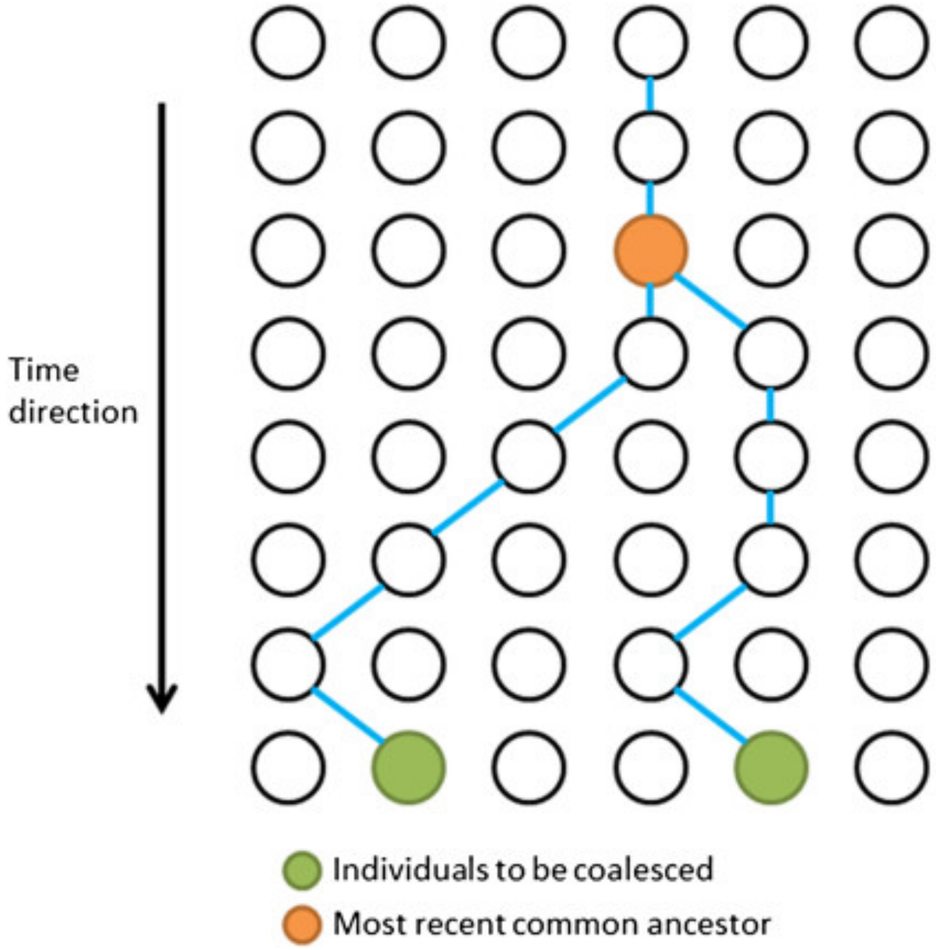

$$\left(1-\frac{1}{2N}\right)^{r-1}\frac{1}{2N}$$

$$\approx \frac{1}{2N}e^{-r/(2N)}$$

This approximation is very good when the population size is large (2N > ~1000)

# Coalescence in a sample of two sequences

P[ 2 samples find a common ancestor in **exactly r** generations ] =




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

$$\approx \frac{1}{2N}e^{-r/(2N)}$$

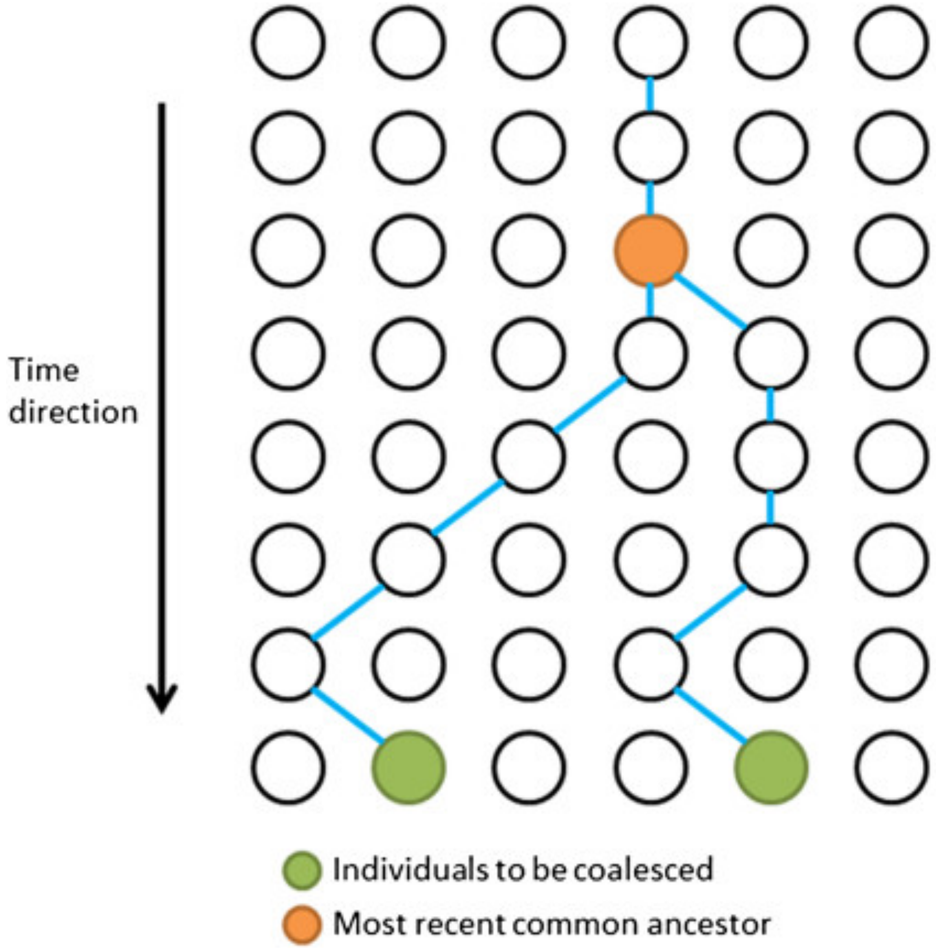
This approximation is very good when the population size is large (2N > ~1000)

$$= e^{-t}$$

If we measure time in units of 2N generations ( 1 t = 2N r )

# Coalescence in a sample of two sequences

P[ 2 samples find a common ancestor in **exactly r** generations ] =




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

$$\approx \frac{1}{2N}e^{-r/(2N)}$$

This approximation is very good when the population size is large (2N > ~1000)

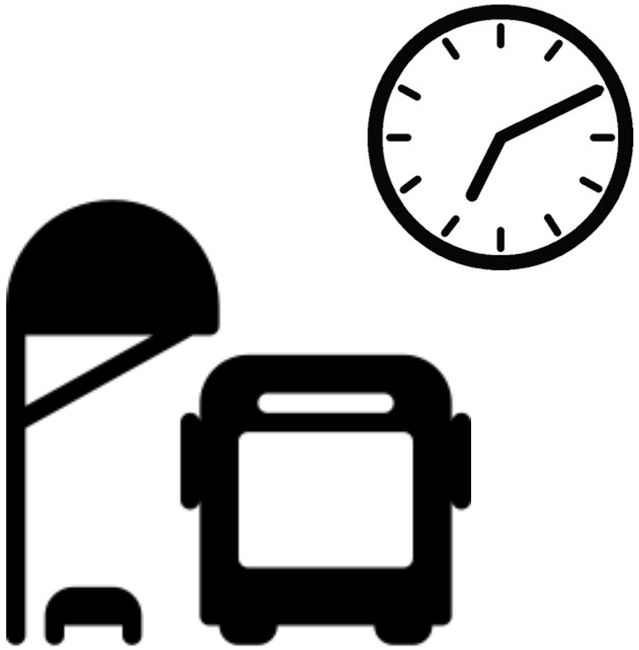
$$= e^{-t}$$

If we measure time in units of 2N generations ( 1 t = 2N r )

This is an **exponential distribution** with rate 1

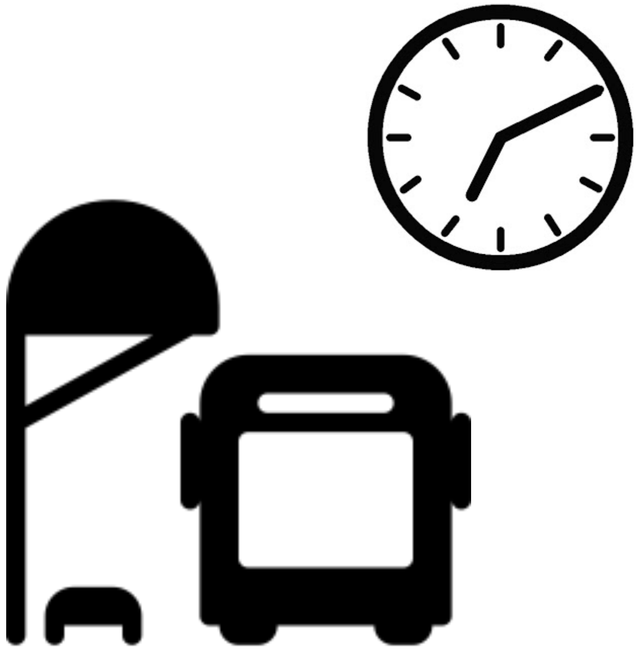
# The Exponential Distribution

- Used to model **waiting times**
  - “What is the probability that I have to wait less than 30 minutes till the next bus arrives?”



# The Exponential Distribution

- Used to model **waiting times**
  - “What is the probability that I have to wait less than 30 minutes till the next bus arrives?”



- One parameter: rate ($\lambda$)
  - The higher the rate, the less I will have to wait

# The Exponential Distribution

One parameter: $\lambda$

$$E[T] = \frac{1}{\lambda}$$

# The Exponential Distribution

One parameter: $\lambda$

$$E[T] = \frac{1}{\lambda}$$

The expected waiting time is the inverse of the rate

# The Exponential Distribution

One parameter: $\lambda$

$$E[T] = \frac{1}{\lambda}$$

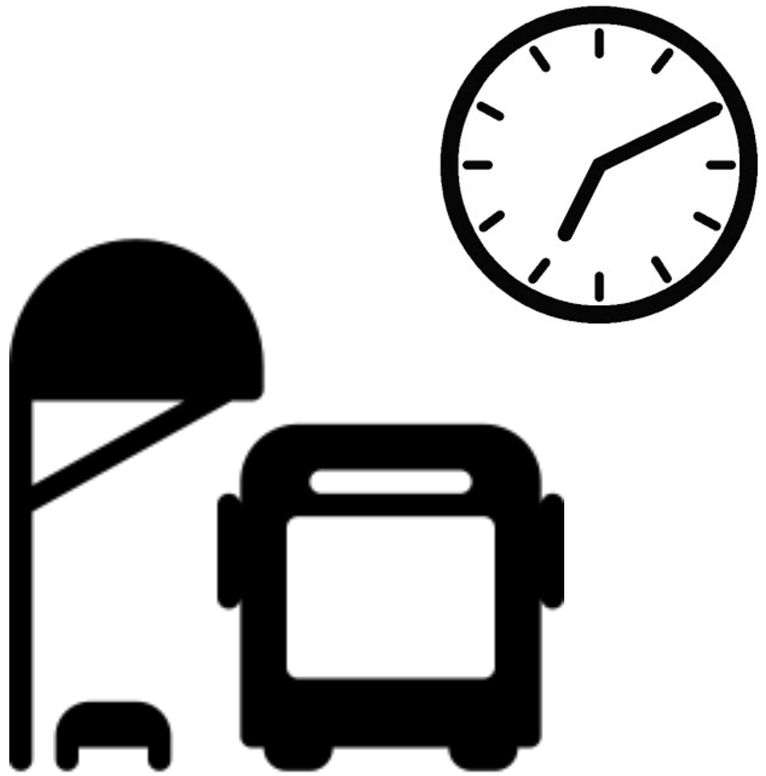
The expected waiting time is the inverse of the rate

If buses arrive at an average rate of $\lambda$=4 per hour,
I expect to wait 1/4 of an hour for the next bus, on average

# The Expected Waiting Time

"Buses arrive at a rate of $\lambda$ per hour"

$$f(t) = \lambda e^{-\lambda t}$$



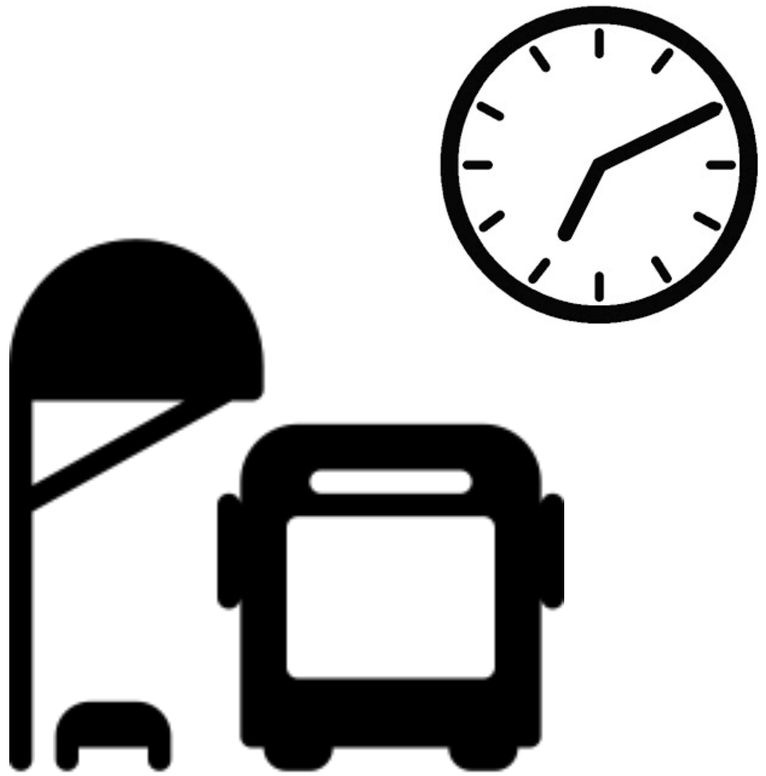
# The Expected Waiting Time

"Buses arrive at a rate of $\lambda$ per hour"

$$f(t) = \lambda e^{-\lambda t}$$

"The expected waiting time for the next bus is $1/\lambda$ hours"

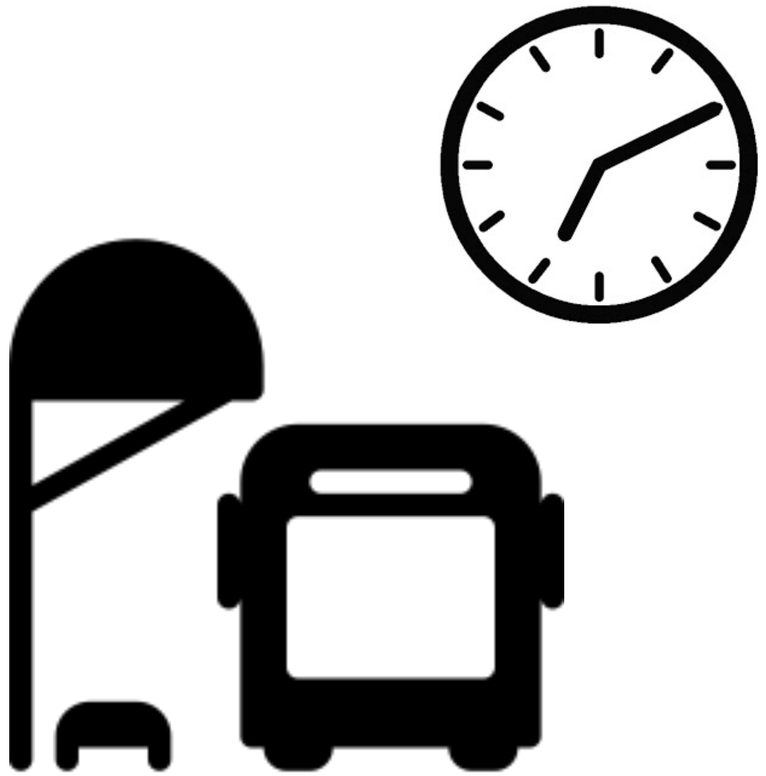
$$E[T] = \frac{1}{\lambda}$$



# The Expected Waiting Time

"Buses arrive at a rate of $\lambda$=4 per hour"

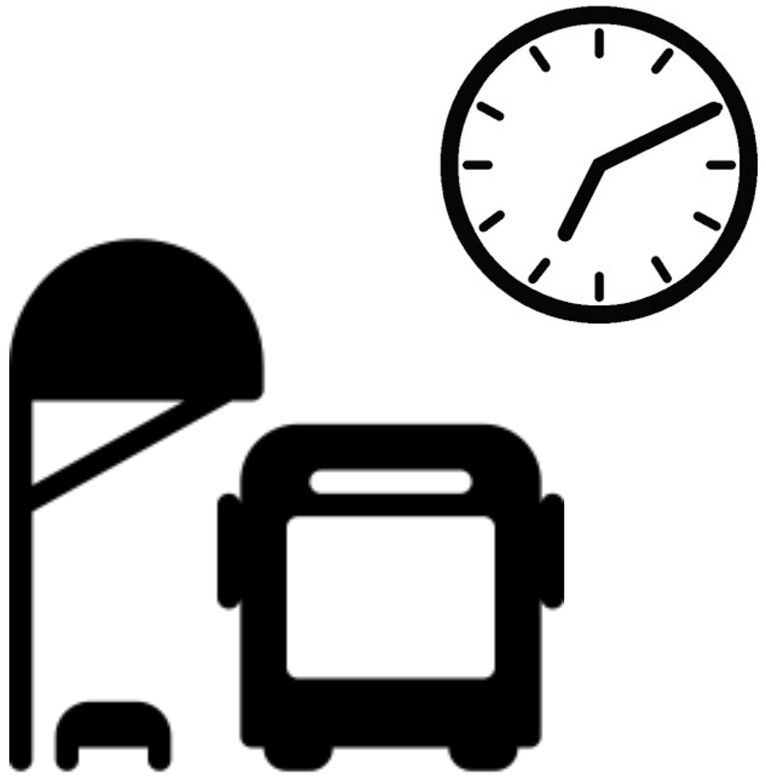
$$f(t) = 4e^{-4t}$$



# The Expected Waiting Time

"Buses arrive at a rate of $\lambda$=4 per hour"

$$f(t) = 4e^{-4t}$$



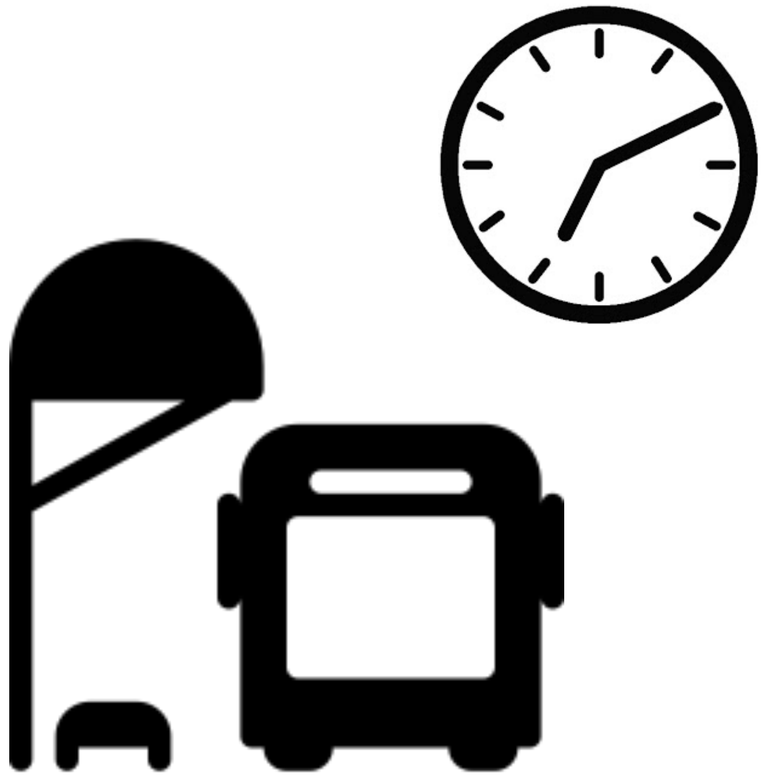
"The expected waiting time for the next bus is 1/4 hours"

$$E[T] = \frac{1}{4}$$

# The Expected Waiting Time

"Buses arrive at a rate of $\lambda$ =4/60=0.0666… per minute"

$$f(t) = \frac{4}{60} e^{-\frac{4}{60}t}$$



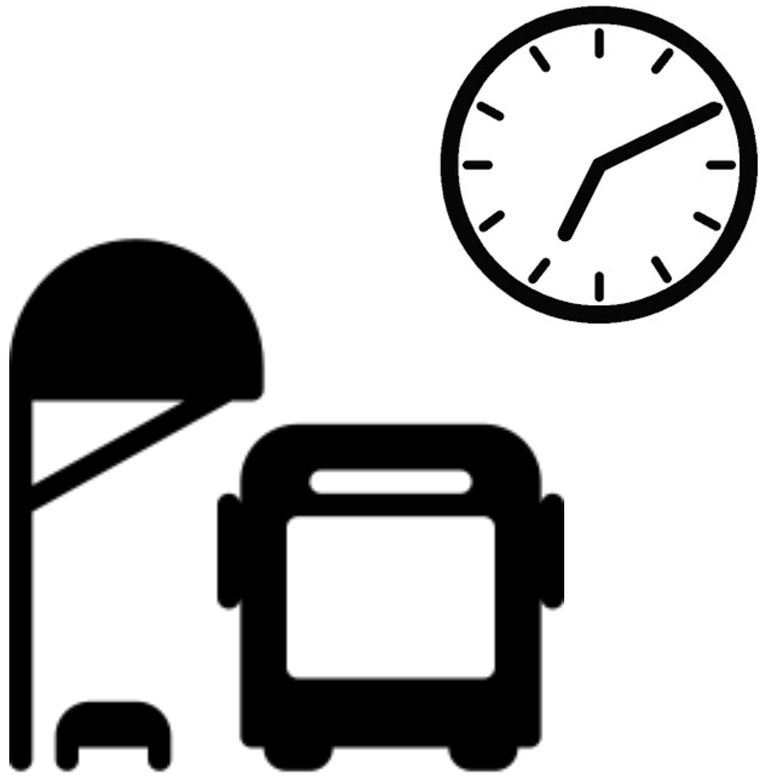
# The Expected Waiting Time

"Buses arrive at a rate of $\lambda$ =4/60=0.0666… per minute"

$$f(t) = \frac{4}{60} e^{-\frac{4}{60}t}$$

"The expected waiting time for the next bus is 1/0.0666… = 15 minutes"

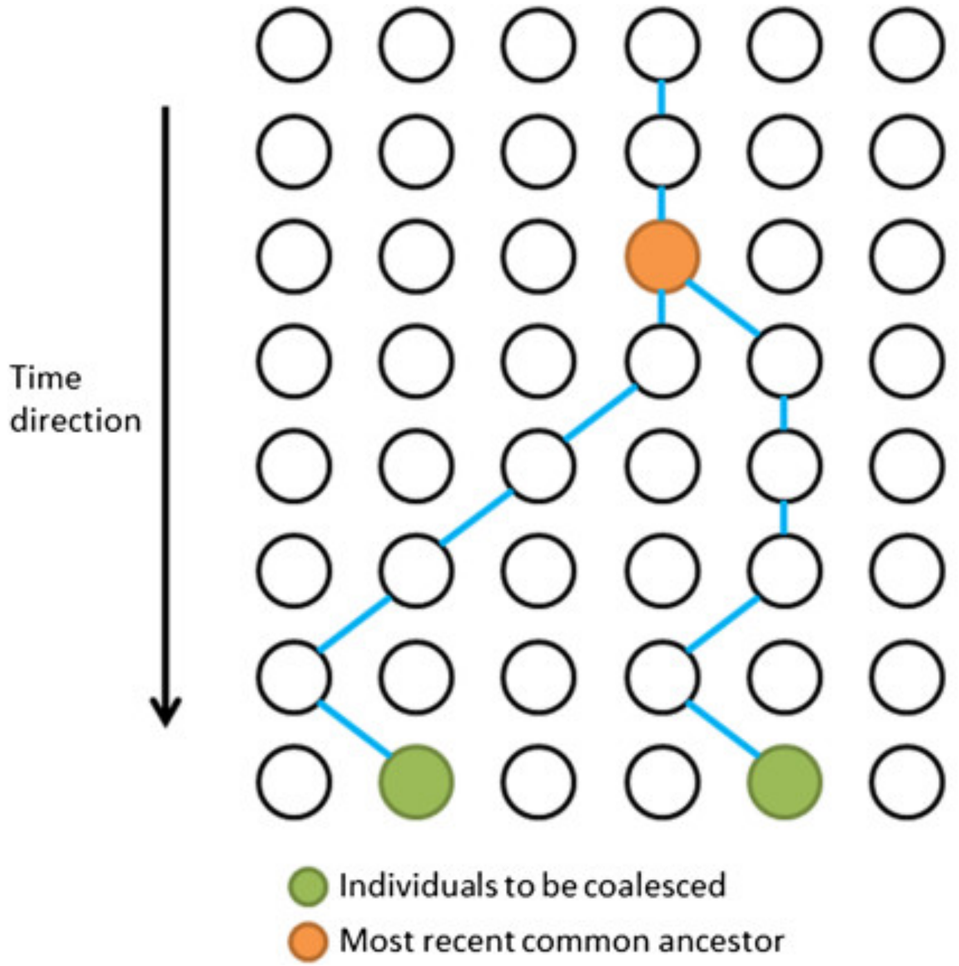
$$E[T] = 15$$



# The Expected Waiting Time

"A coalescent event between 2 lineages occurs at a rate of 1/2N per generation"

$$f(t) = \frac{1}{2N} e^{-t/(2N)}$$



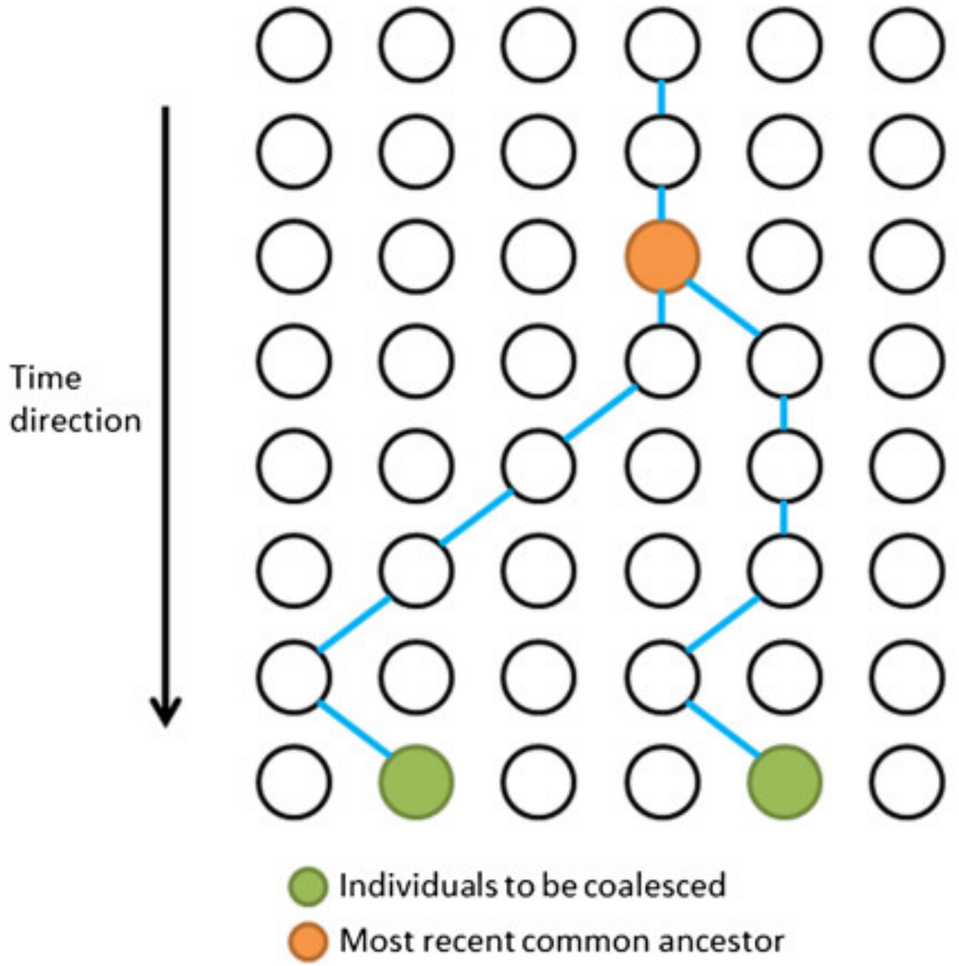
# The Expected Waiting Time

"A coalescent event between 2 lineages occurs at a rate of 1/2N per generation"

$$f(t) = \frac{1}{2N} e^{-t/(2N)}$$

"The expected waiting time for a coalescent event is 2N generations"

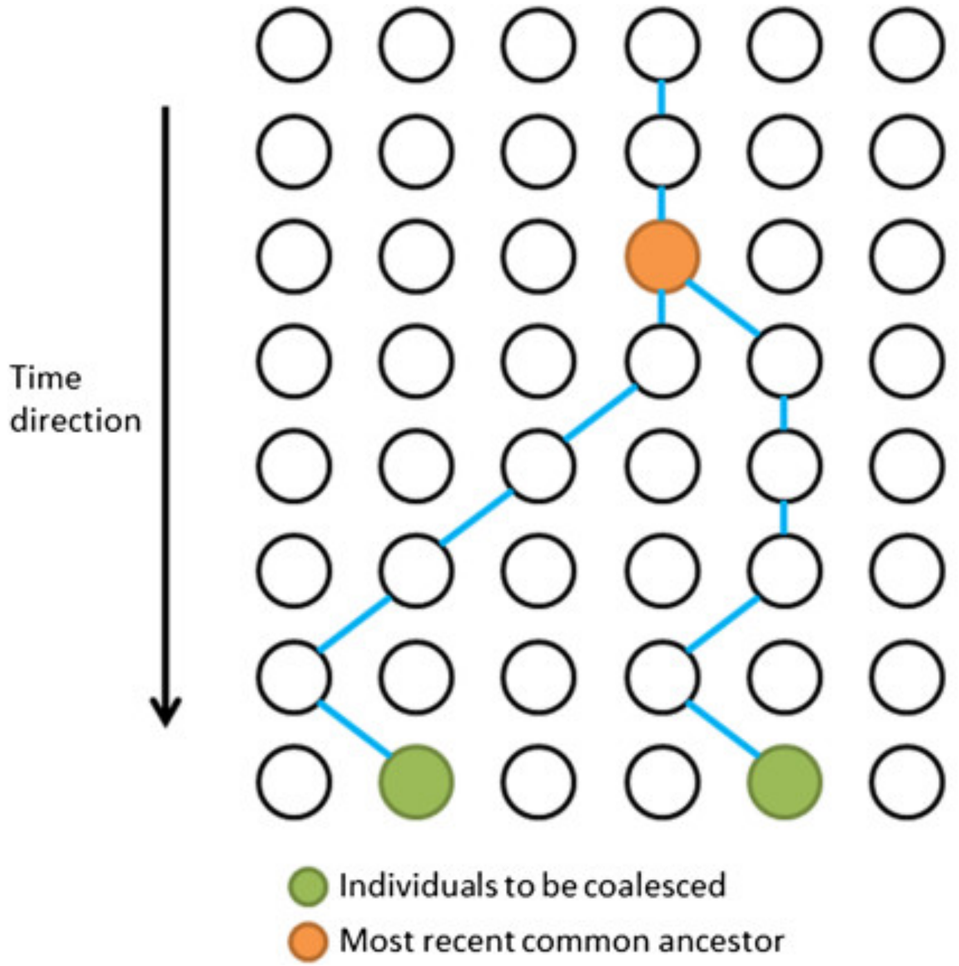
$$E[T] = 2N$$



# The Expected Waiting Time

"A coalescent event between 2 lineages occurs at a rate of 1 per coalescent unit (1 unit = 2N generations)."

$$f(t) = e^{-t}$$



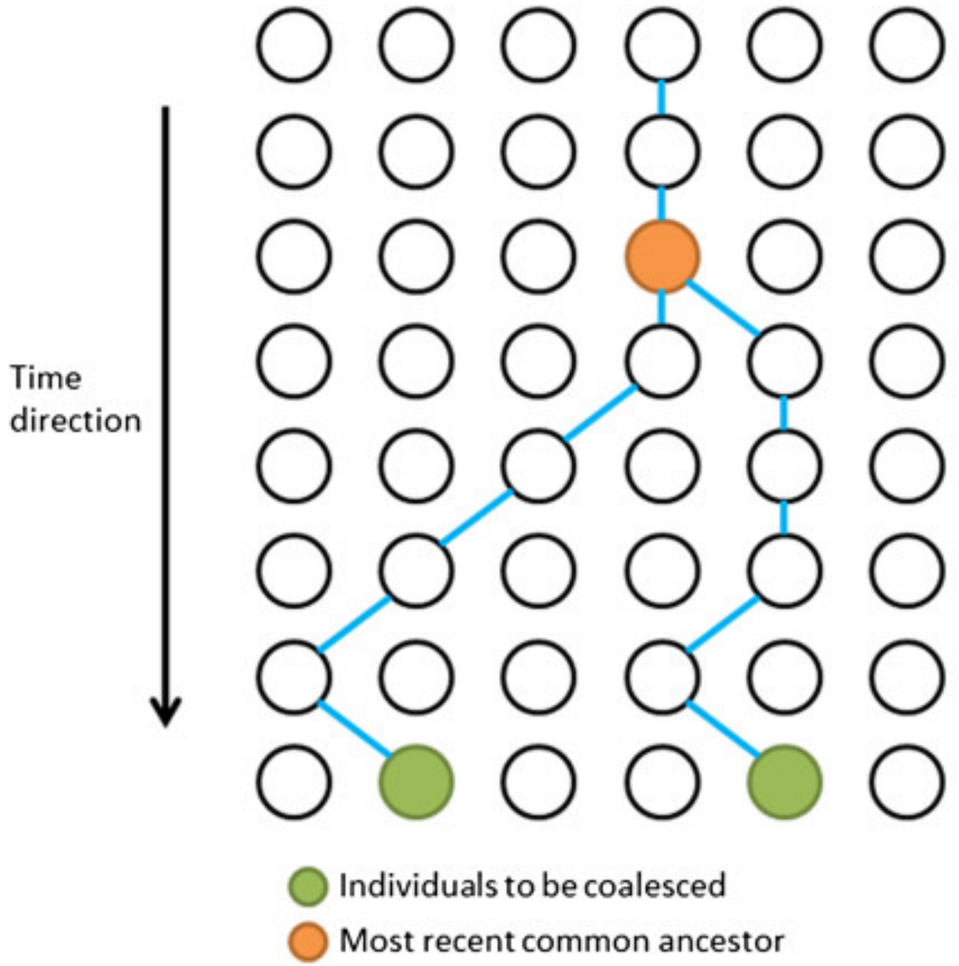
# The Expected Waiting Time

"A coalescent event between 2 lineages occurs at a rate of 1 per coalescent unit (1 unit = 2N generations)."

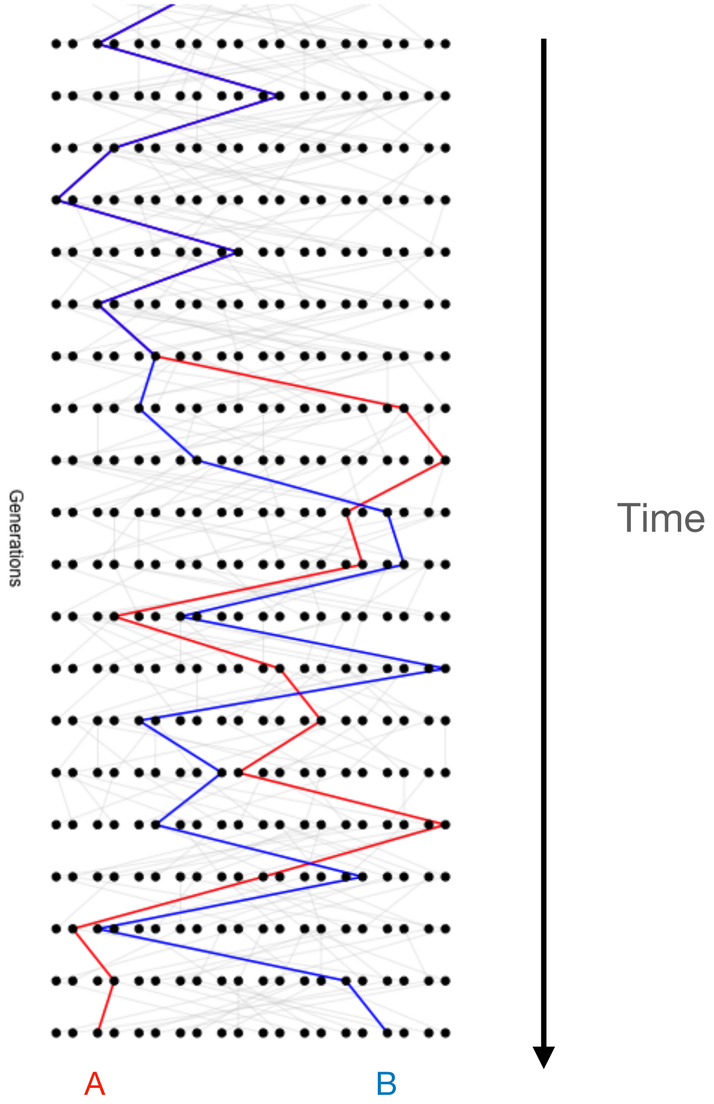
$$f(t) = e^{-t}$$

"The expected waiting time for a coalescent event is 1 coalescent unit"

$$E[T] = 1$$



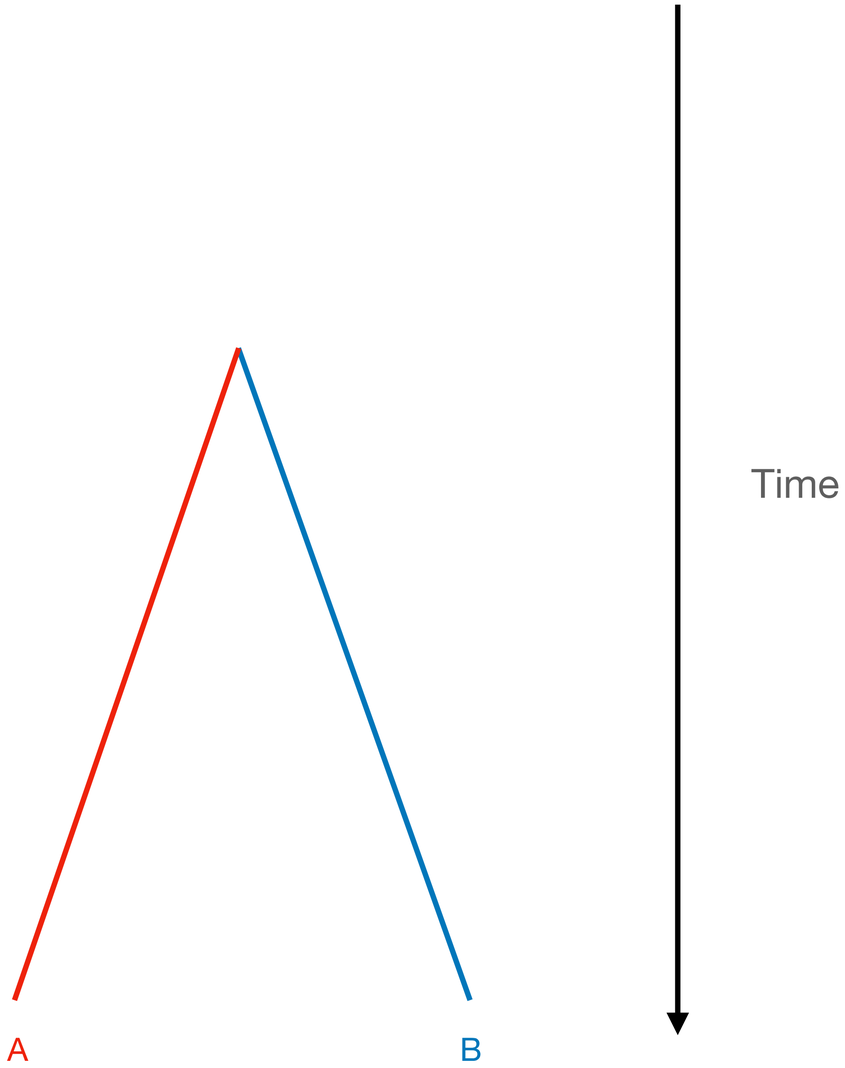
# Coalescence of 2 lineages



# Coalescence of 2 lineages

**Rate:**

$Rate = 1$ per coalescent unit

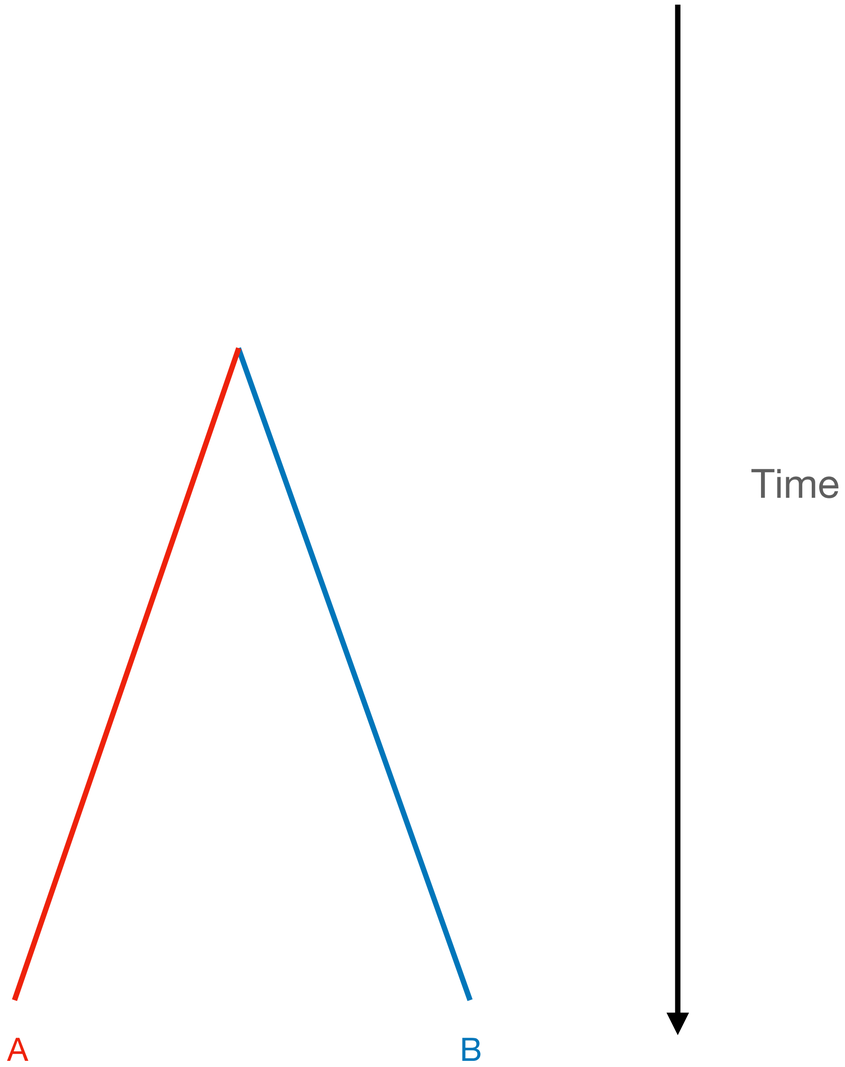


# Coalescence of 2 lineages

**Rate:**

$Rate = 1$ per coalescent unit

= 1 per 2N generations

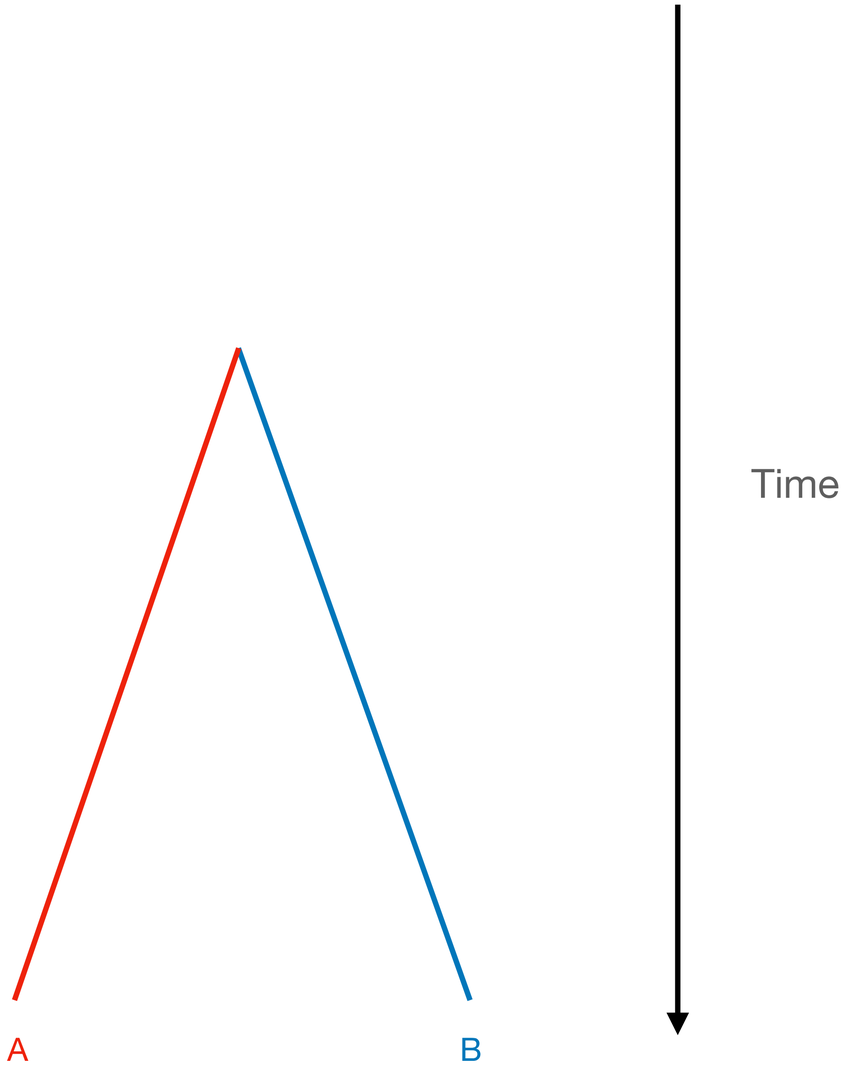


# Coalescence of 2 lineages

**Rate:**

$Rate = 1$ per coalescent unit

= 1 per 2N generations



# Coalescence of 2 lineages

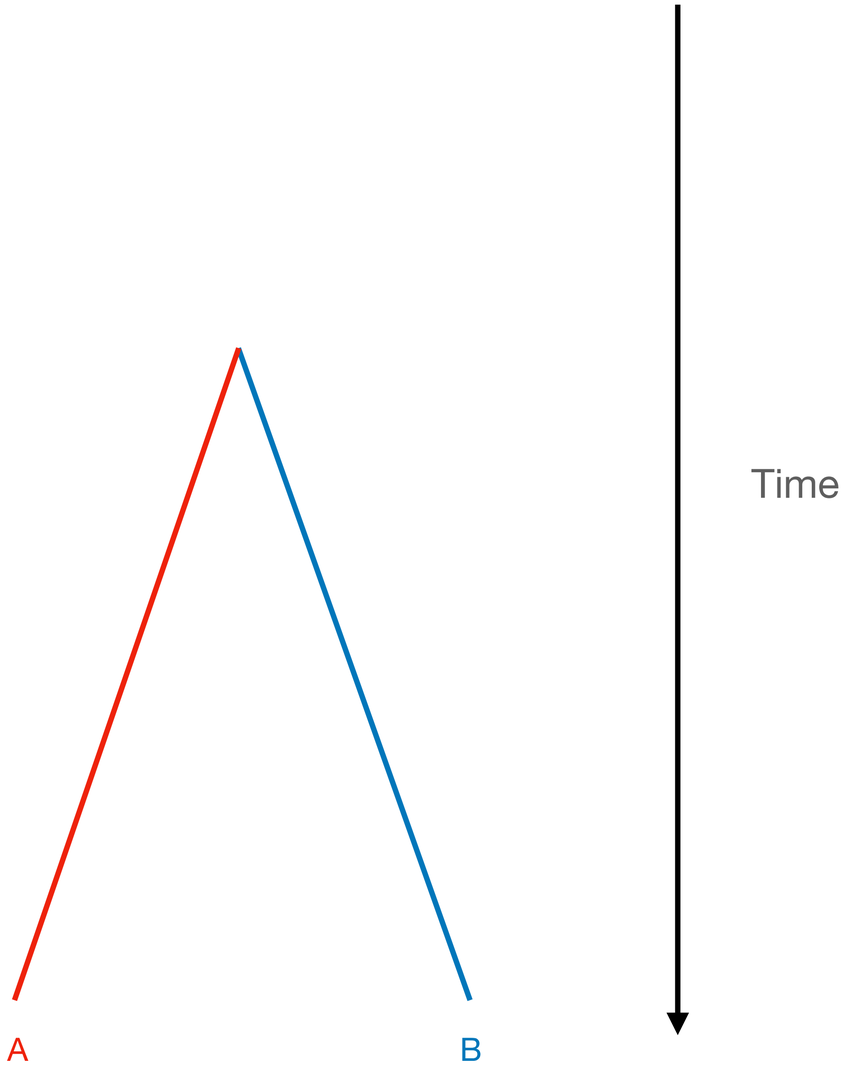
**Rate:**

$Rate = 1$ per coalescent unit

= 1 per 2N generations

**Expected waiting time:**

$E[T] = 1$ coalescent unit



# Coalescence of 2 lineages

**Rate:**

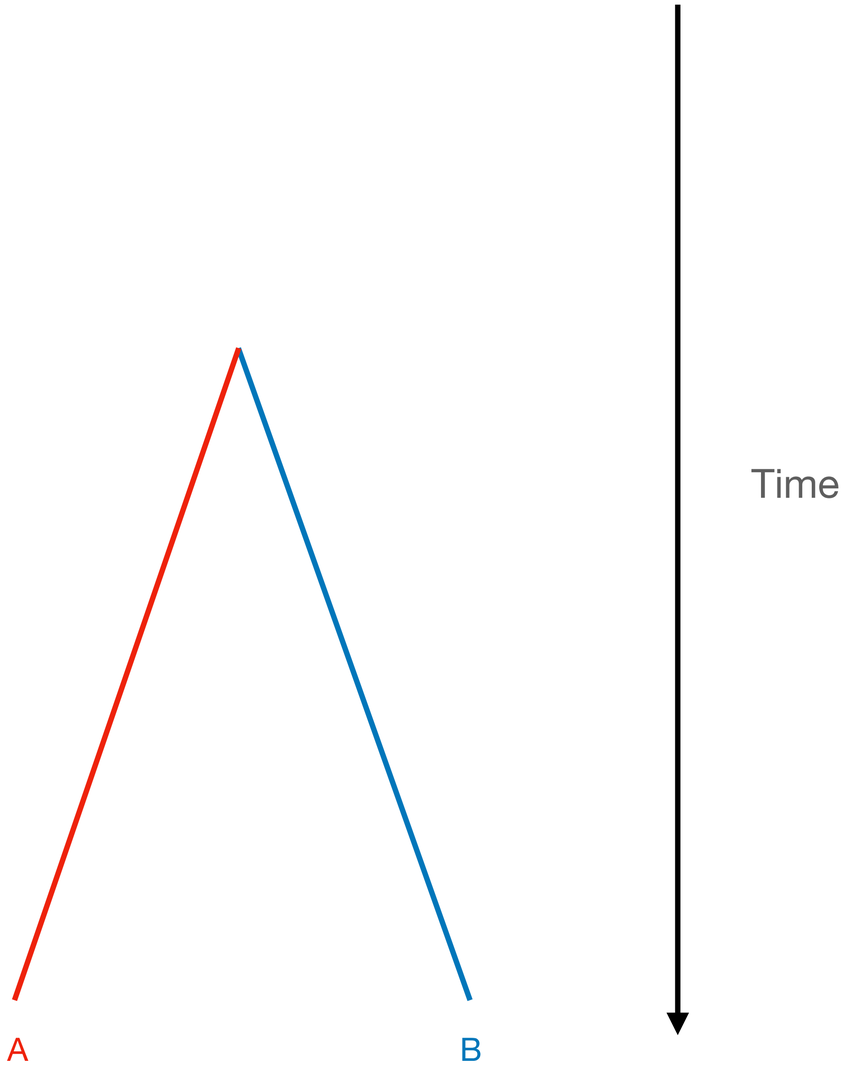
$Rate = 1$ per coalescent unit

= 1 per 2N generations

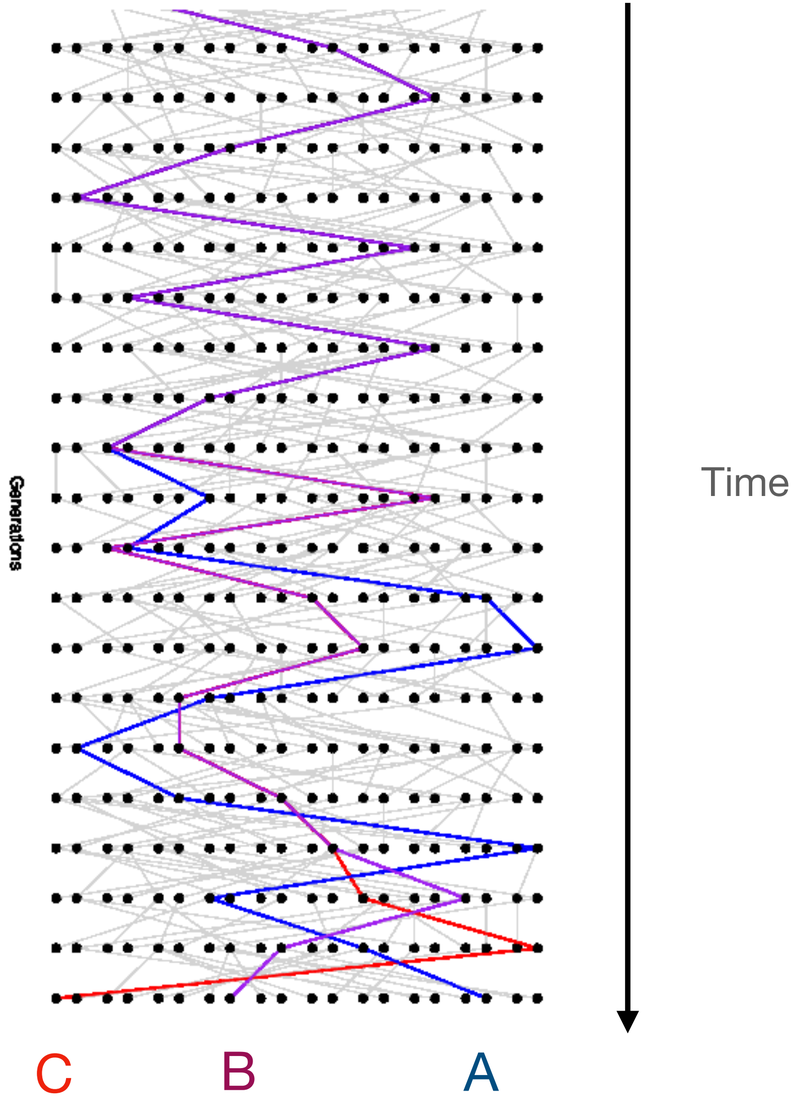
**Expected waiting time:**

$E[T] = 1$ coalescent unit

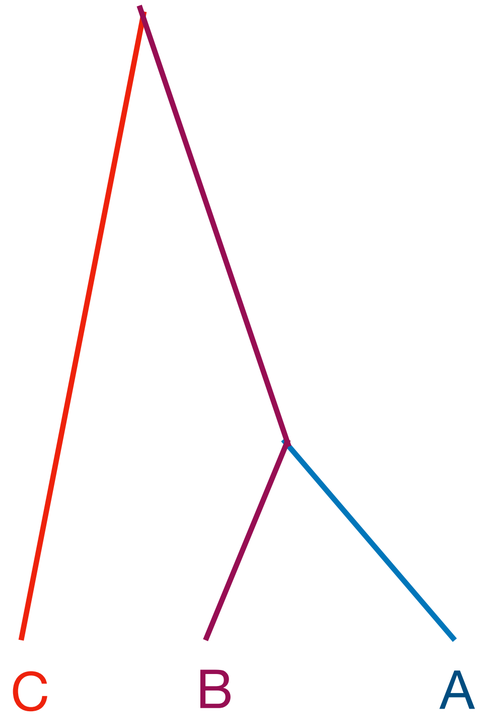
$= 2N$ generations



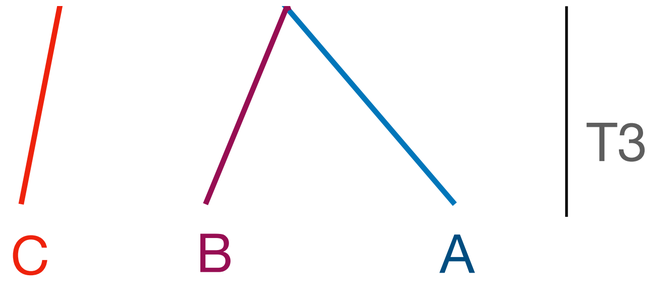
# Tree with 3 samples



# Tree with 3 samples



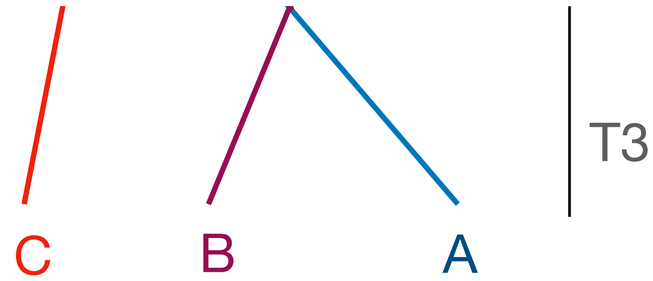
# Tree with 3 samples



# Tree with 3 samples

**T3 = time while there are 3 lineages**

**Rate:** $\lambda_{T3} = 1 * \binom{3}{2} = 3$

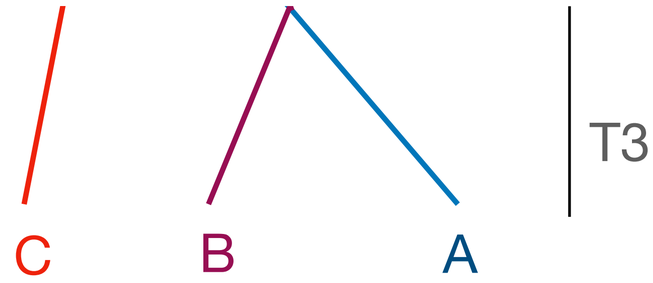


# Tree with 3 samples

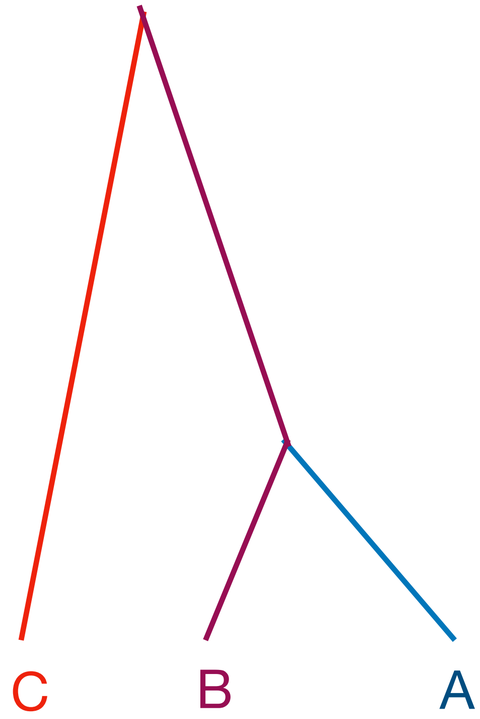
**T3 = time while there are 3 lineages**

**Rate:** $\lambda_{T3} = 1 * \binom{3}{2} = 3$

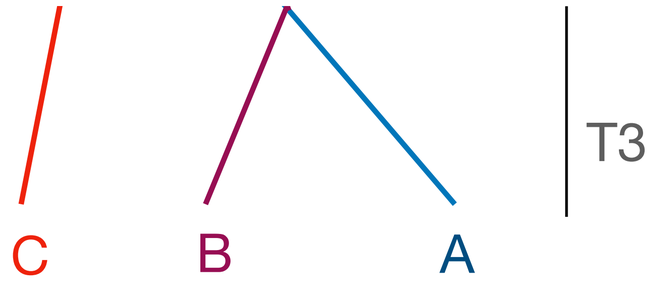
…because there are 3 “competing pairs” of lineages fighting for the opportunity to coalesce (each at rate 1): A+B, B+C, A+C



# Tree with 3 samples



# Tree with 3 samples



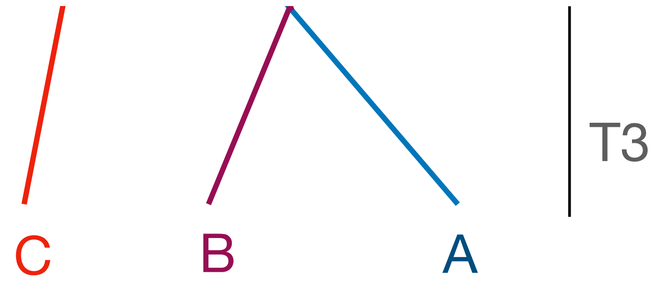
# Tree with 3 samples

T3 = time while there are 3 lineages

Rate: $\lambda_{T3} = 1 * \binom{3}{2} = 3$

…because there are 3 “competing pairs” of lineages fighting for the opportunity to coalesce (each at rate 1): A+B, B+C, A+C

Expected time: $E[T3] = \frac{1}{\binom{3}{2}} = \frac{1}{3}$



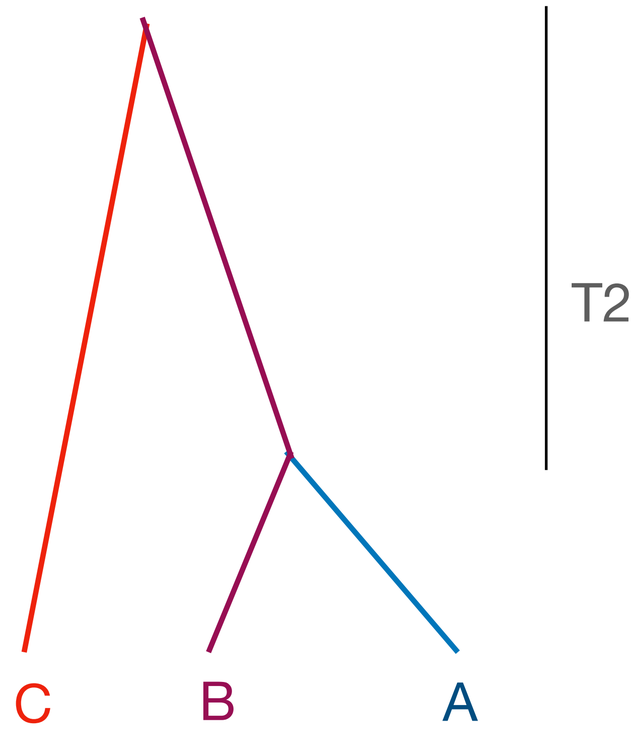
# Tree with 3 samples

**T2 = time while there are 2 lineages**

**Rate:** $\lambda_{T2} = \frac{1}{\binom{2}{2}} = 1$

…because there is just one possible pair
that can coalesce (at rate 1)

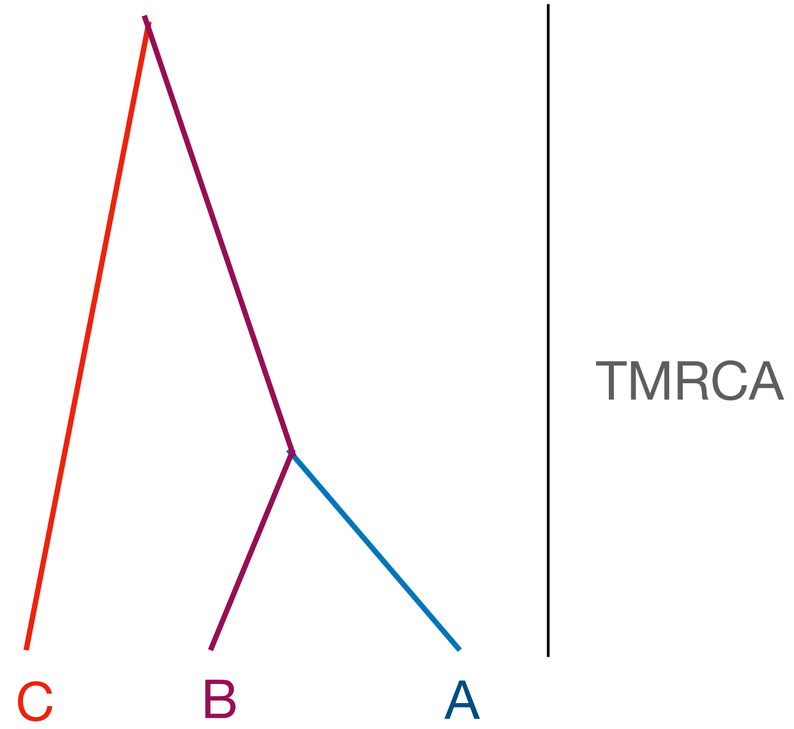
Expected time: $E[T2] = 1$



# Tree with 3 samples

$E[T_{MRCA}] = T_2 + T_3 = 1 + 1/3$ coalescent units

$E[T_{MRCA}] = (1 + 1/3) * 2N$ generations

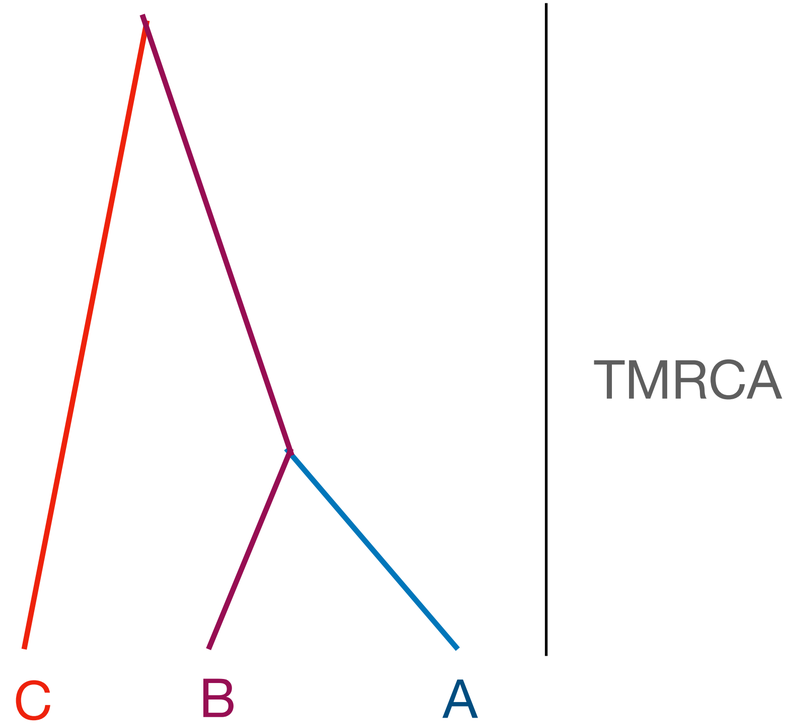


# Tree with 3 samples

$E[T_{MRCA}] = T_2 + T_3 = 1 + 1/3$ coalescent units

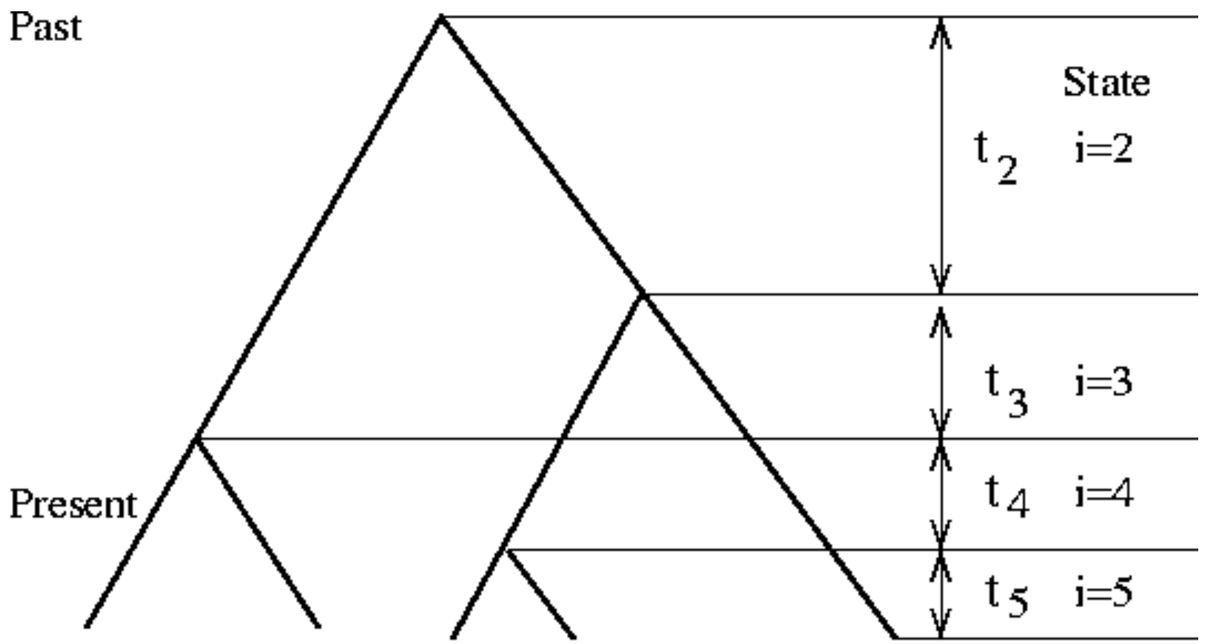
$E[T_{MRCA}] = (1 + 1/3) * 2N$ generations

Example: if 2N = 10,000, then:
$E[T_{MRCA}] \approx$ 13,333 generations



# TMRCA

$$E[T_{MRCA}] = T_2 + T_3 + \ldots + T_n$$

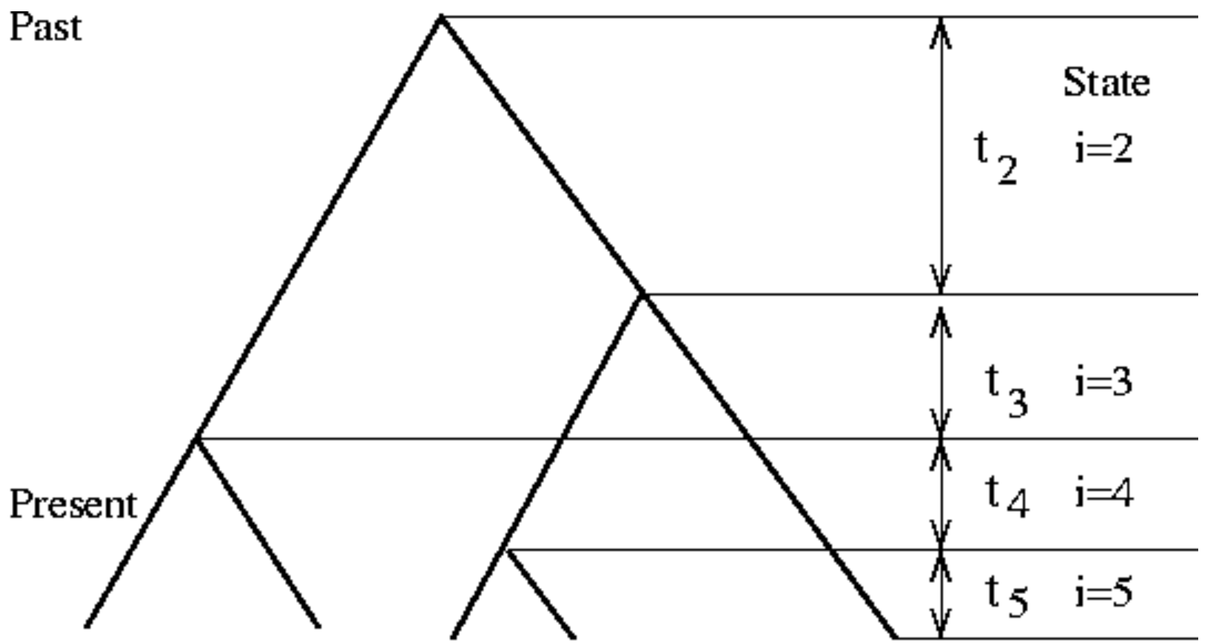


# TMRCA

$$E[T_{MRCA}] = T_2 + T_3 + \ldots + T_n$$

$$E[T_{MRCA}] = \sum_{i=2}^{n} \frac{1}{\binom{i}{2}} = 2\left(1 - \frac{1}{n}\right)$$

where n is our sample size



# Variance of an exponential distribution

One parameter: $\lambda$

$$E[T] = \frac{1}{\lambda}$$

# Variance of an exponential distribution

One parameter: $\lambda$

$$E[T] = \frac{1}{\lambda}$$

$$Var[T] = \frac{1}{\lambda^2}$$

# Variance of an exponential distribution

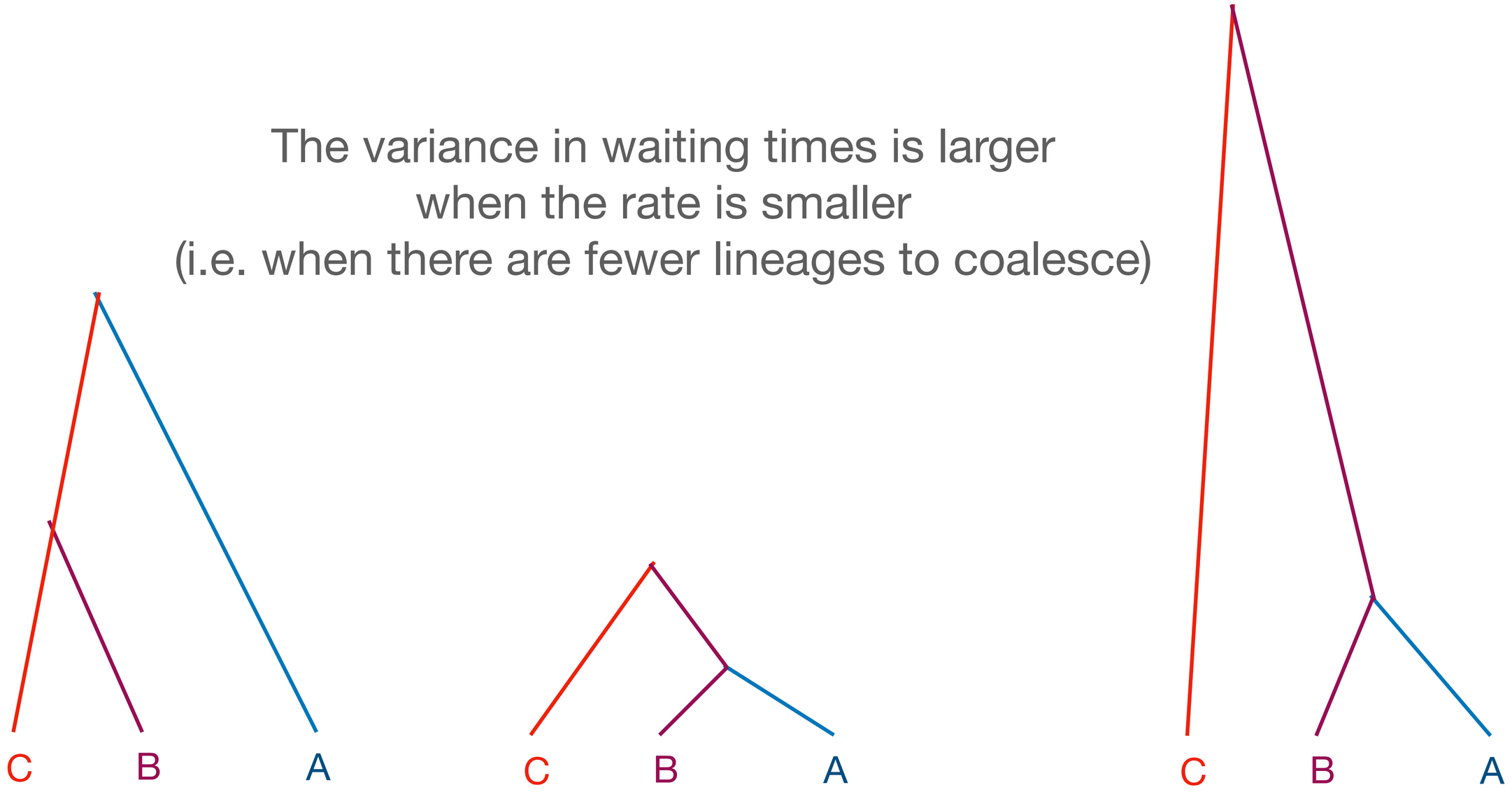
One parameter: $\lambda$

$$E[T] = \frac{1}{\lambda}$$

$$Var[T] = \frac{1}{\lambda^2} \longrightarrow$$ The variance in waiting times is larger when the rate is smaller

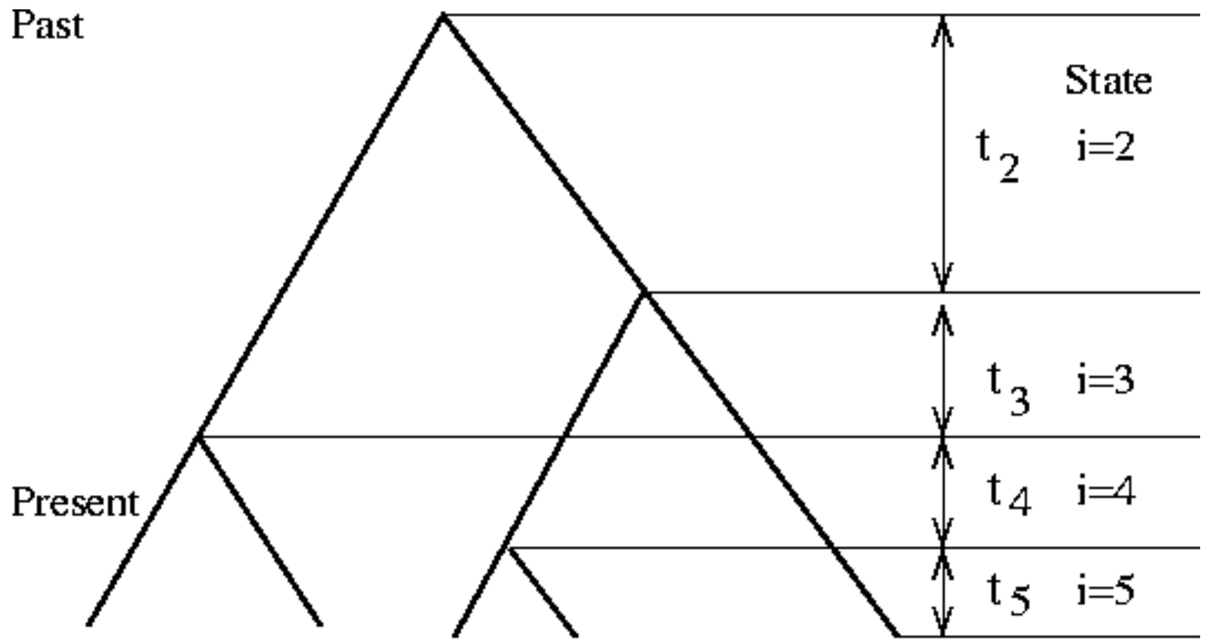
# Variance in coalescence times

The variance in waiting times is larger
when the rate is smaller
(i.e. when there are fewer lineages to coalesce)



# Take-home message

The Kingman coalescent allows us to model the **genealogy of our sample** without worrying about details of the entire unsampled population.



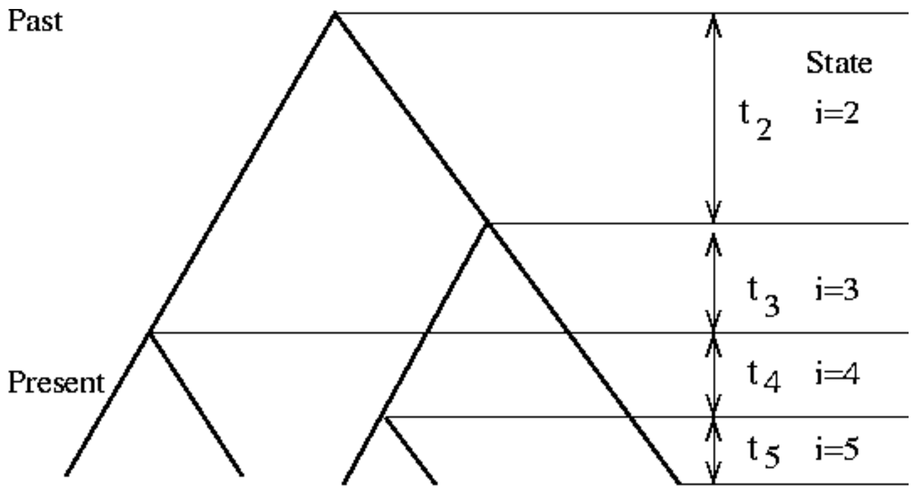
# Today

- Terminology
- Wright-Fisher Model and Genetic Drift
- Intro to coalescent theory
- Mutations and the coalescent

# Today

- Terminology
- Wright-Fisher Model and Genetic Drift
- Intro to coalescent theory
- Mutations and the coalescent

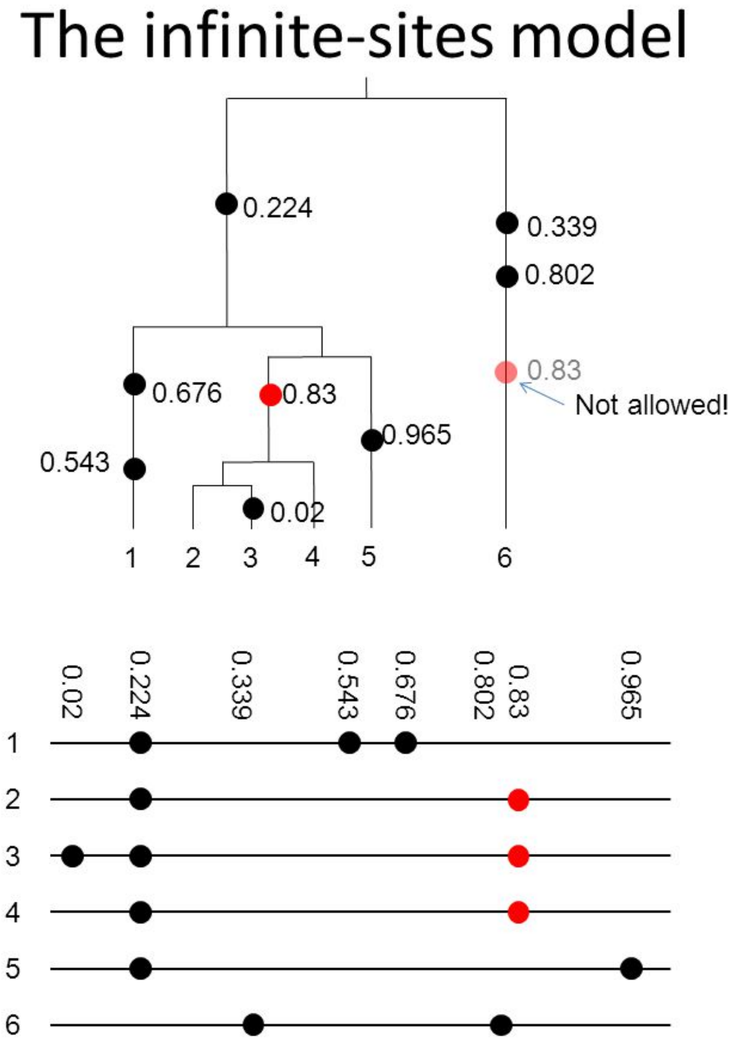
# Relating expectations about trees with data




?

# Introducing mutations into trees

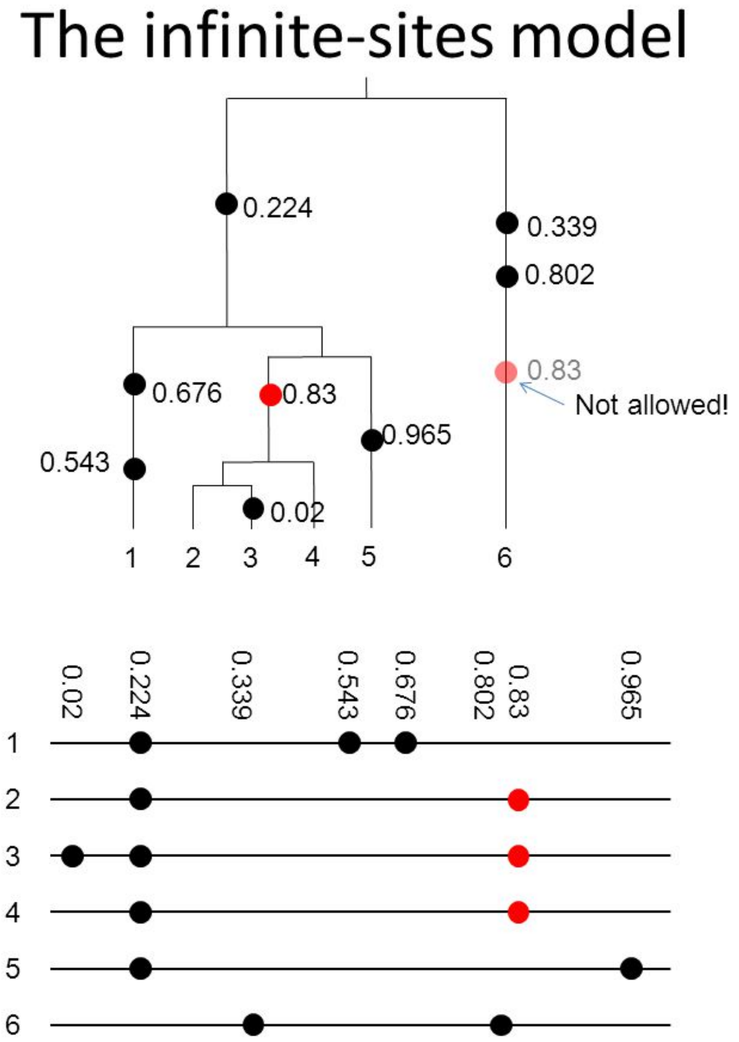
- One way of introducing mutations in a genealogy is the **infinite sites model**.
- Assume we have a sequence with infinite number of sites (for example, a real line).
- This means no two mutations can occur in the same position of our sequence.
- In other words, each mutation creates a new segregating site.




Song, Y. 2021

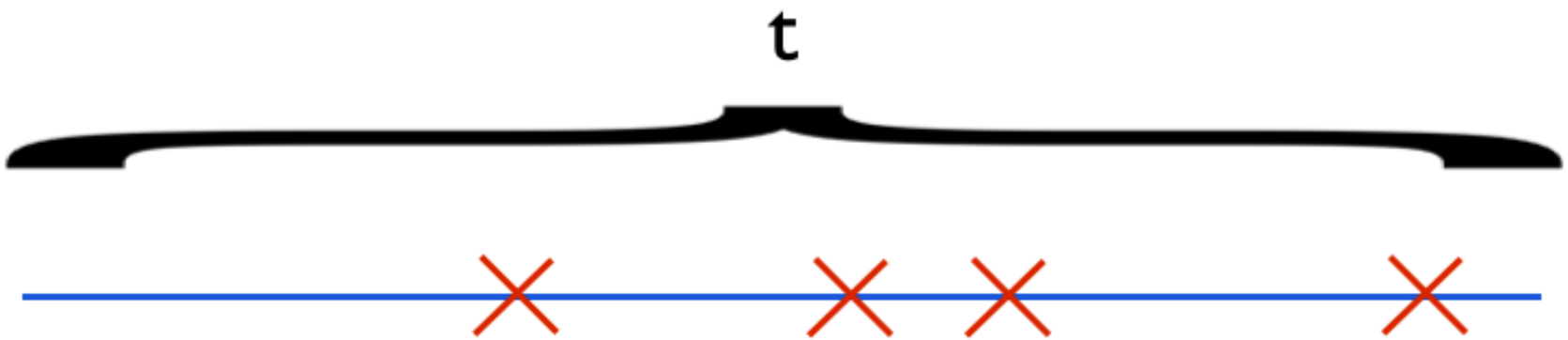
# Introducing mutations into trees

- Allows us to assume that each segregating site is **due to a single mutation** in the past
- Ignores double-substitutions and back-mutations
- Valid as long as the **sequence we are studying is long** and the **mutation rate is low**
- Good for short time-scales (population genetics) but not long time-scales (phylogenetics)



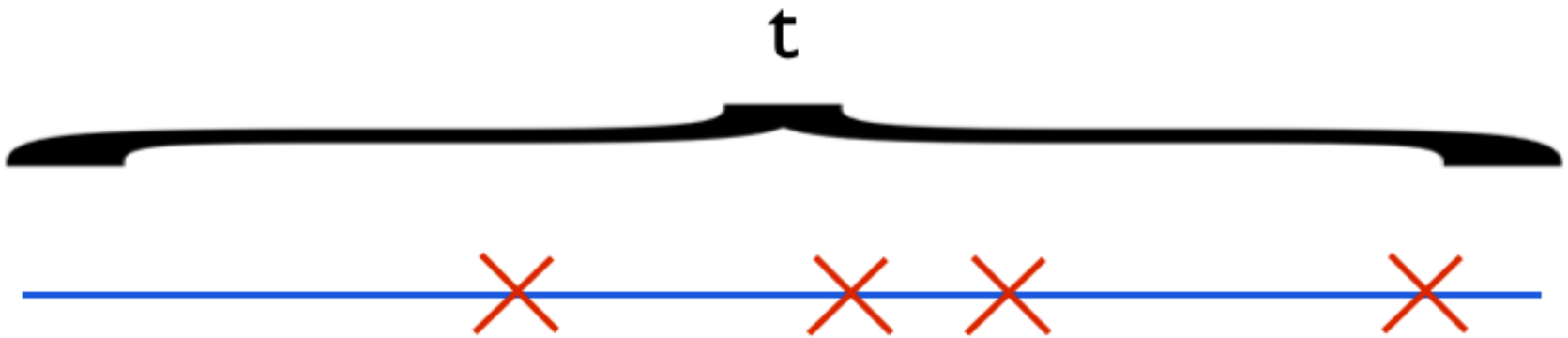

Song, Y. 2021

# Introducing mutations into a lineage



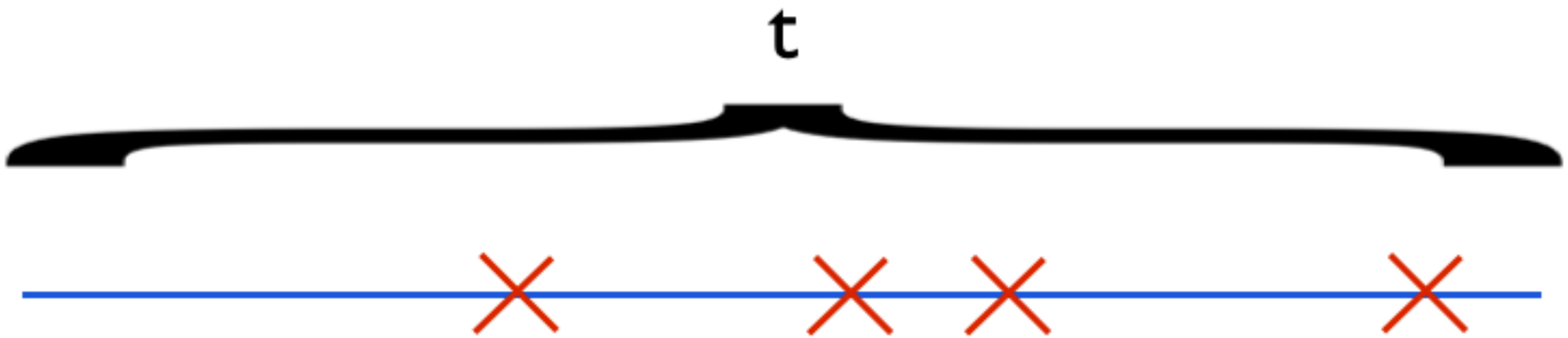
# Introducing mutations into a lineage

- Mutations occur at rate **u per generation**



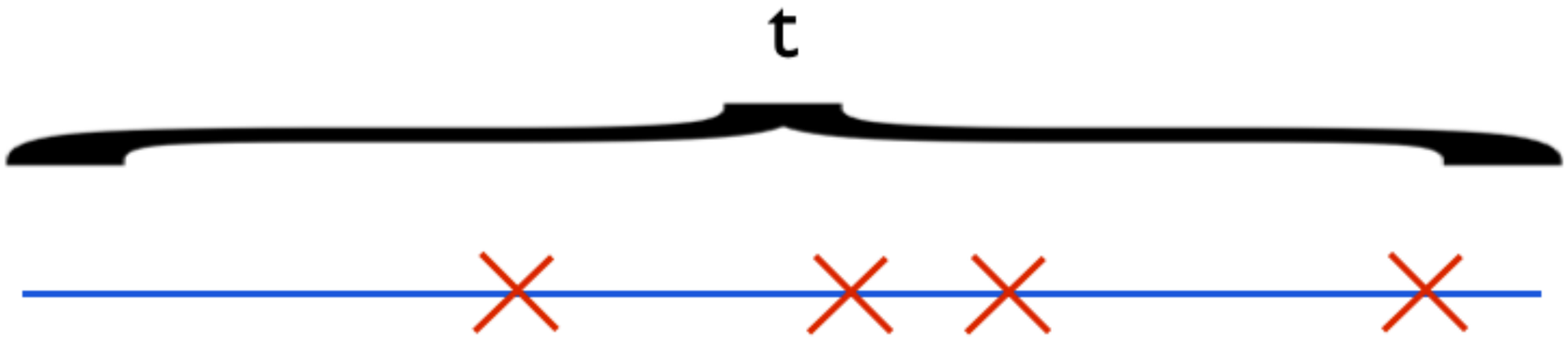
# Introducing mutations into a lineage

- Mutations occur at rate **u per generation**



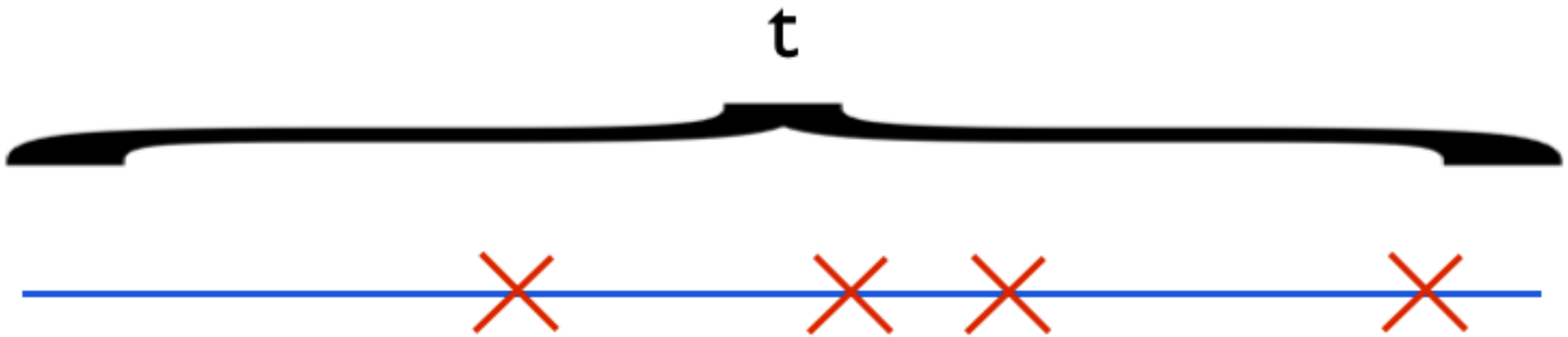
# Introducing mutations into a lineage

- Mutations occur at rate **u per generation**
- In other words, we expect u*r mutations in r generations



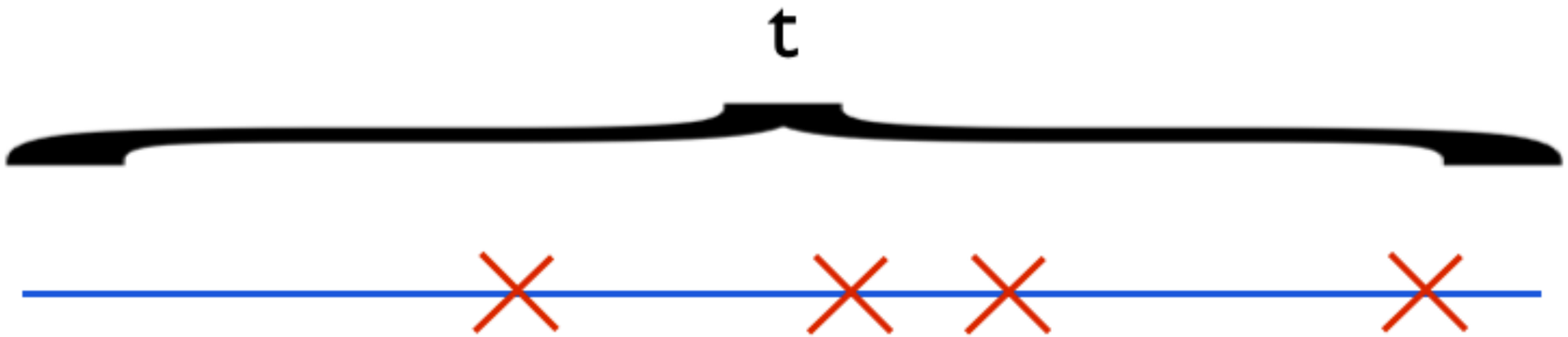
# Introducing mutations into a lineage

- Mutations occur at rate **u per generation**
- In other words, we expect u*r mutations in r generations

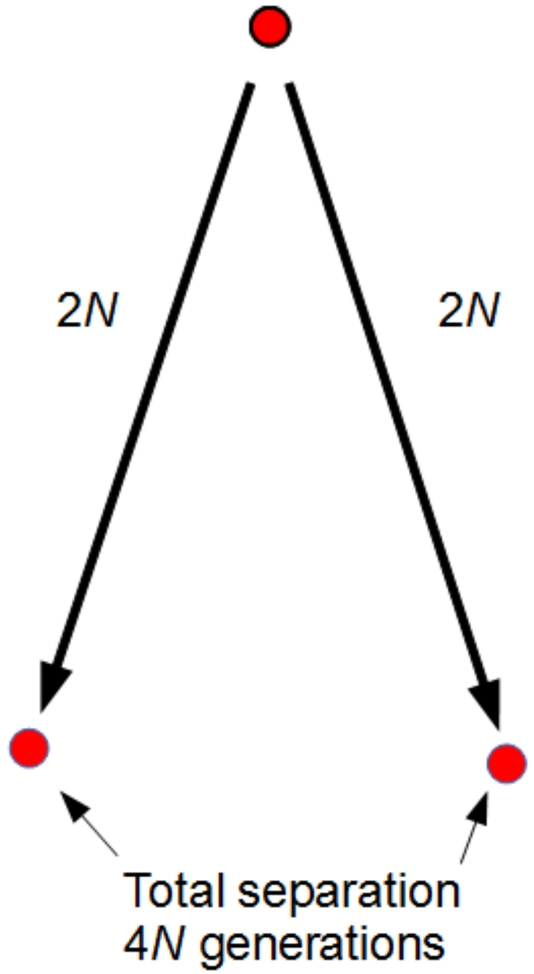


# Introducing mutations into a lineage

- Mutations occur at rate **u per generation**
- In other words, we expect u*r mutations in r generations
- If we measure time in units t of 2N generations, **we expect 2N*t*u mutations in t units of time**

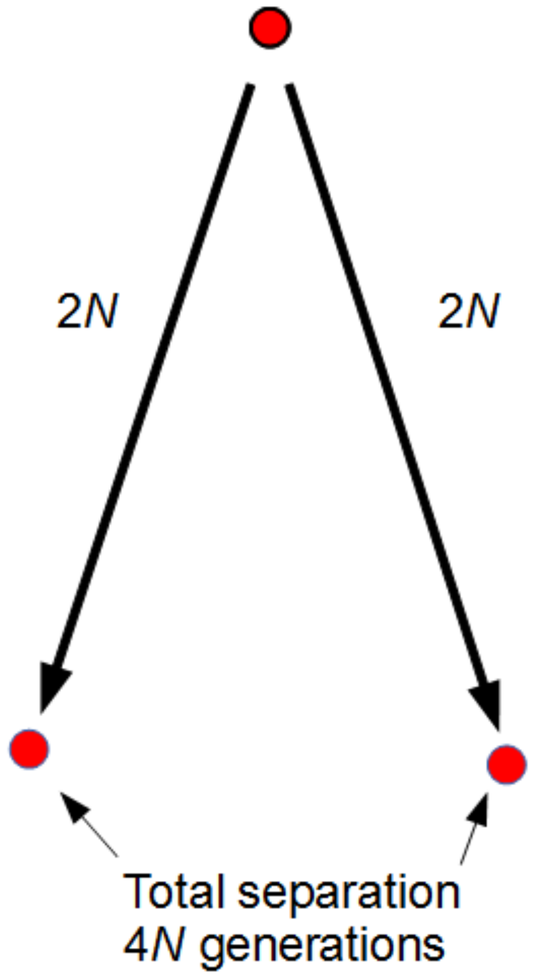


# Introducing mutations into a lineage



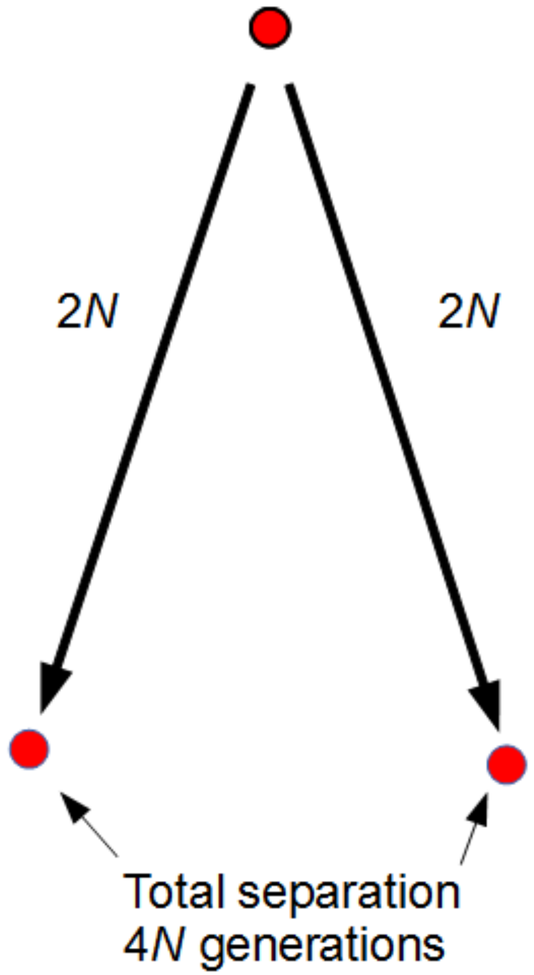
# Introducing mutations into a lineage

- Total expected length = 4N generations = **2 units of coalescent time**



# Introducing mutations into a lineage

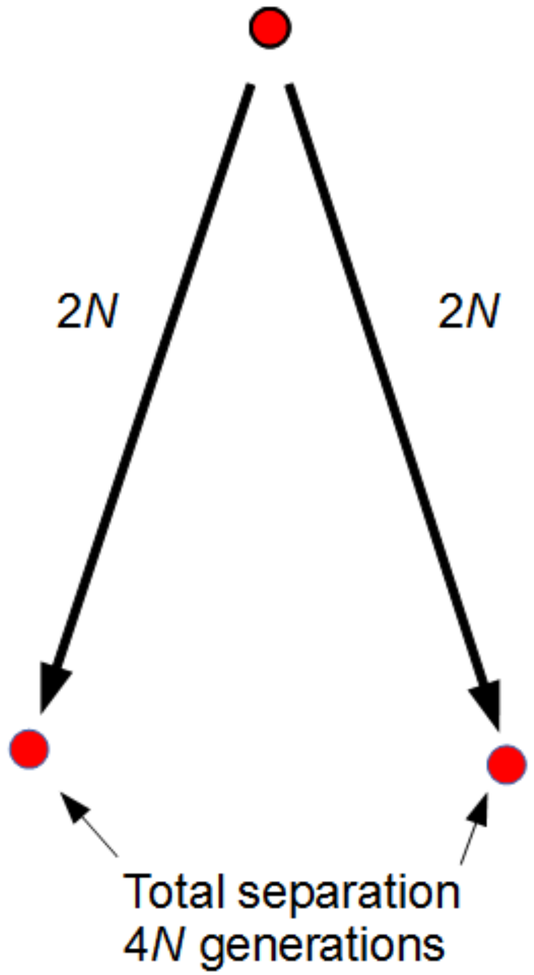
- Total expected length = 4N generations = **2 units of coalescent time**



# Introducing mutations into a lineage

- Total expected length = 4N generations = **2 units of coalescent time**

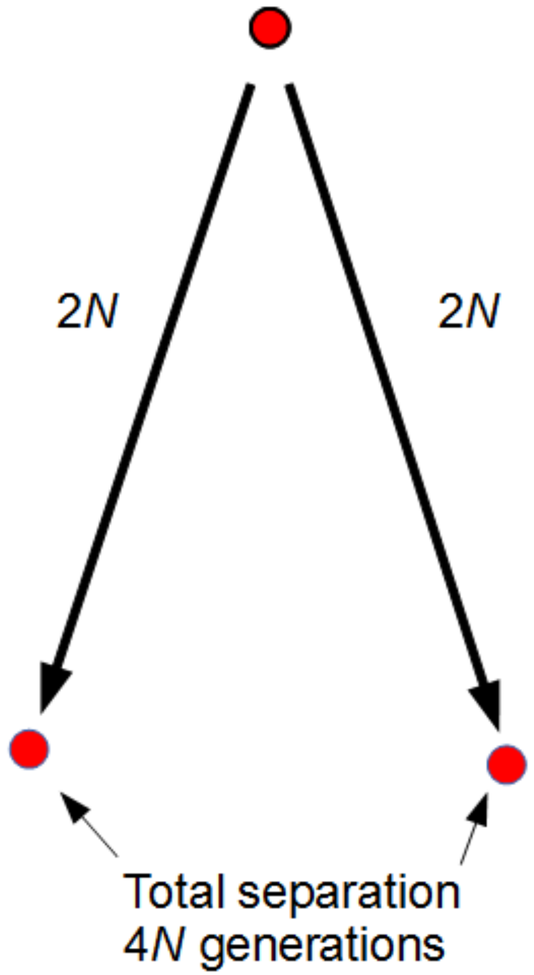
- Therefore, the total expected number of mutations is 2N*t*u = 4Nu, which is also conveniently labeled as θ



# Introducing mutations into a lineage

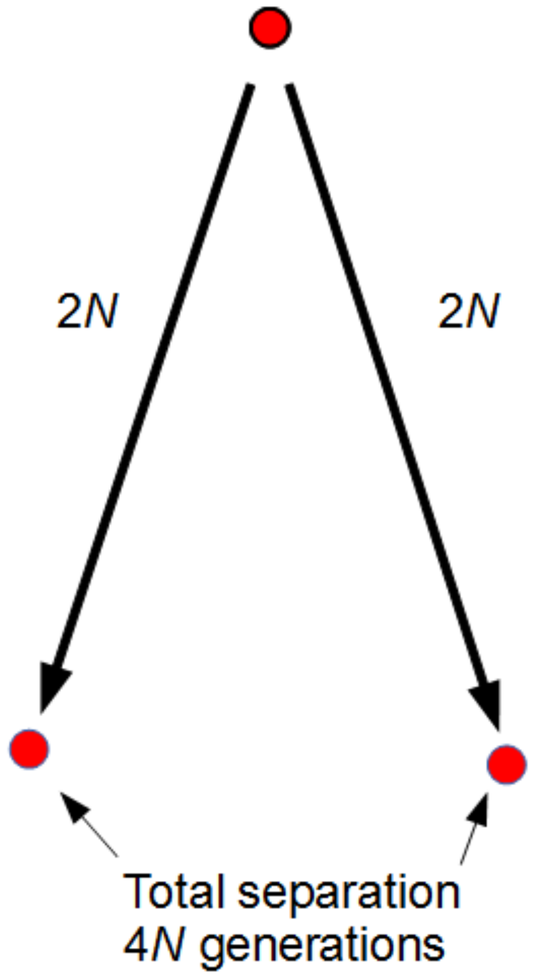
- Total expected length = 4N generations = **2 units of coalescent time**

- Therefore, the total expected number of mutations is 2N*t*u = 4Nu, which is also conveniently labeled as θ

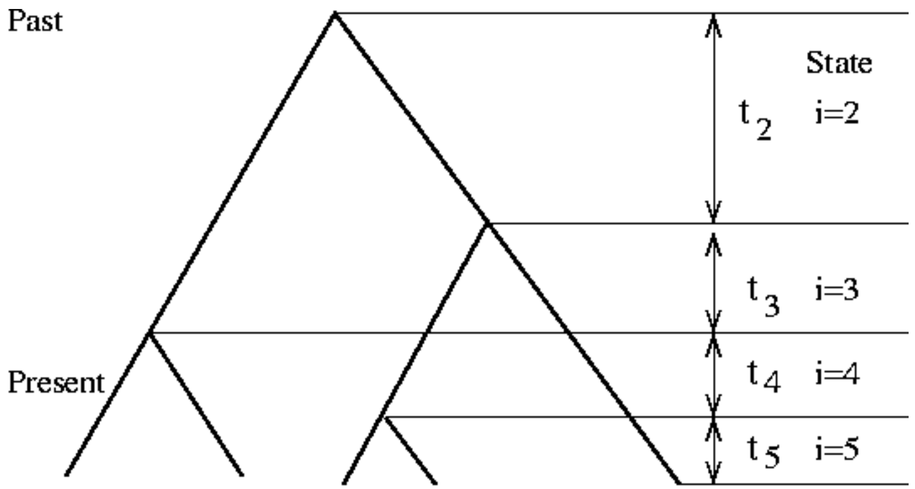


# Introducing mutations into a lineage

- Total expected length = 4N generations = **2 units of coalescent time**
- Therefore, the total expected number of mutations is 2N*t*u = 4Nu, which is also conveniently labeled as θ
- In other words, **mutations occur at rate θ/2 per unit of coalescent time**



# Relating expectations about trees to data




infinite sites model


coalescent model

data

# Relating expectations about trees to data

Hidden properties of populations

infinite sites model


data

# Relating expectations about trees to data

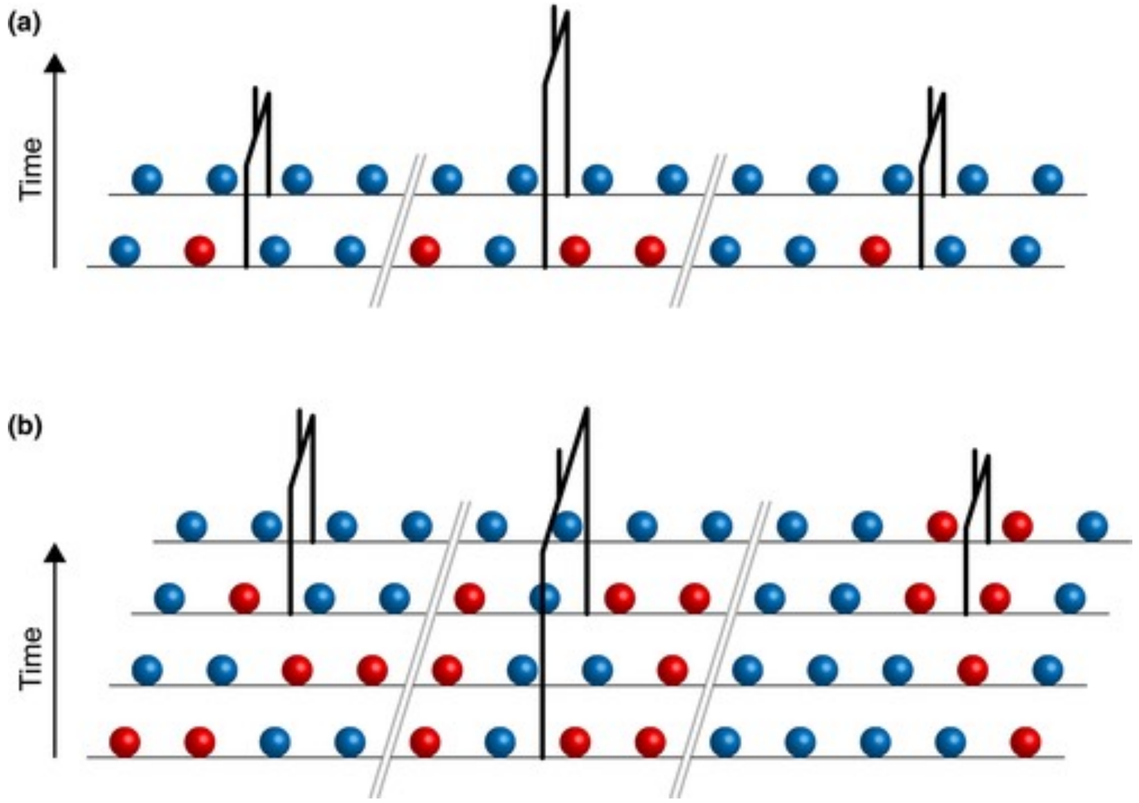
infinite sites model

Hidden properties of populations


Example: population size 2N

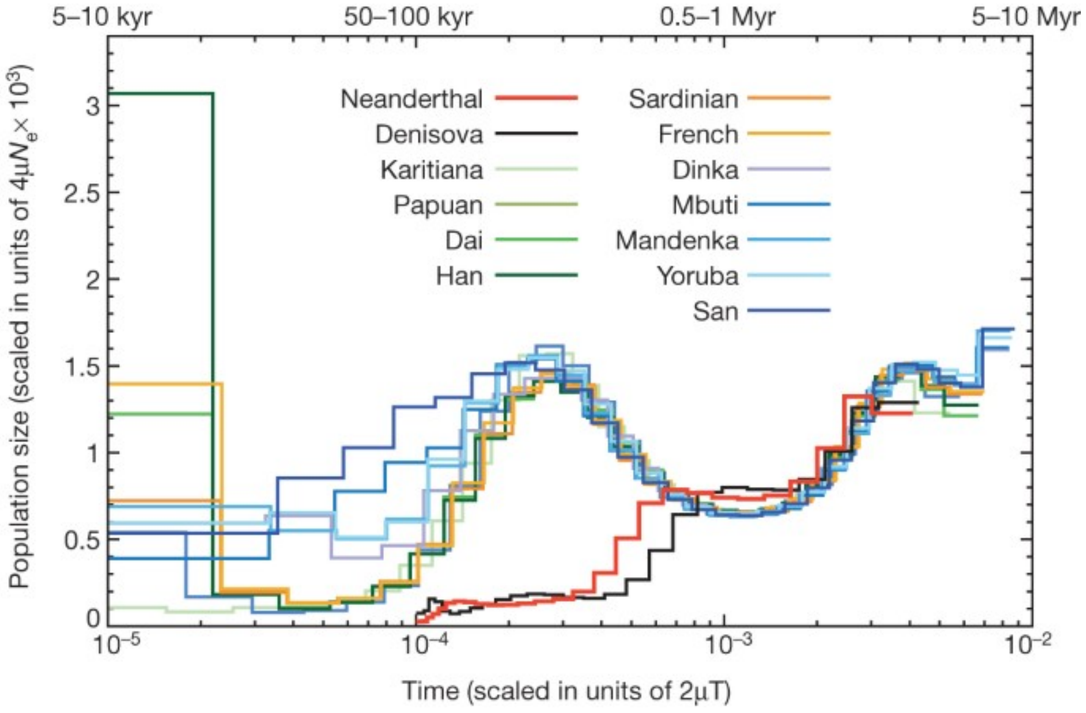
data

# Relating expectations about trees to data




**PSMC, MSMC, SMC++, etc.**

Mather et al. (2019)




Prüfer et al. (2014)

# Relating expectations about trees to data

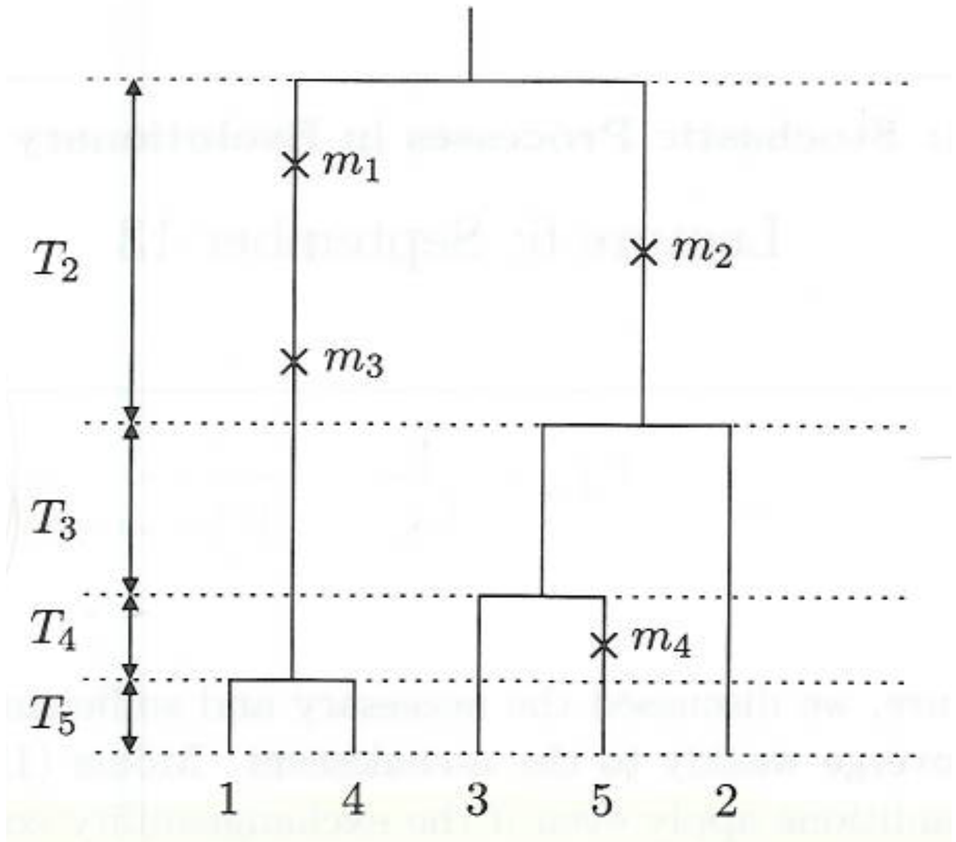
infinite sites model

Hidden properties of populations


data

Example: population size 2N, **assuming it has stayed constant over time**

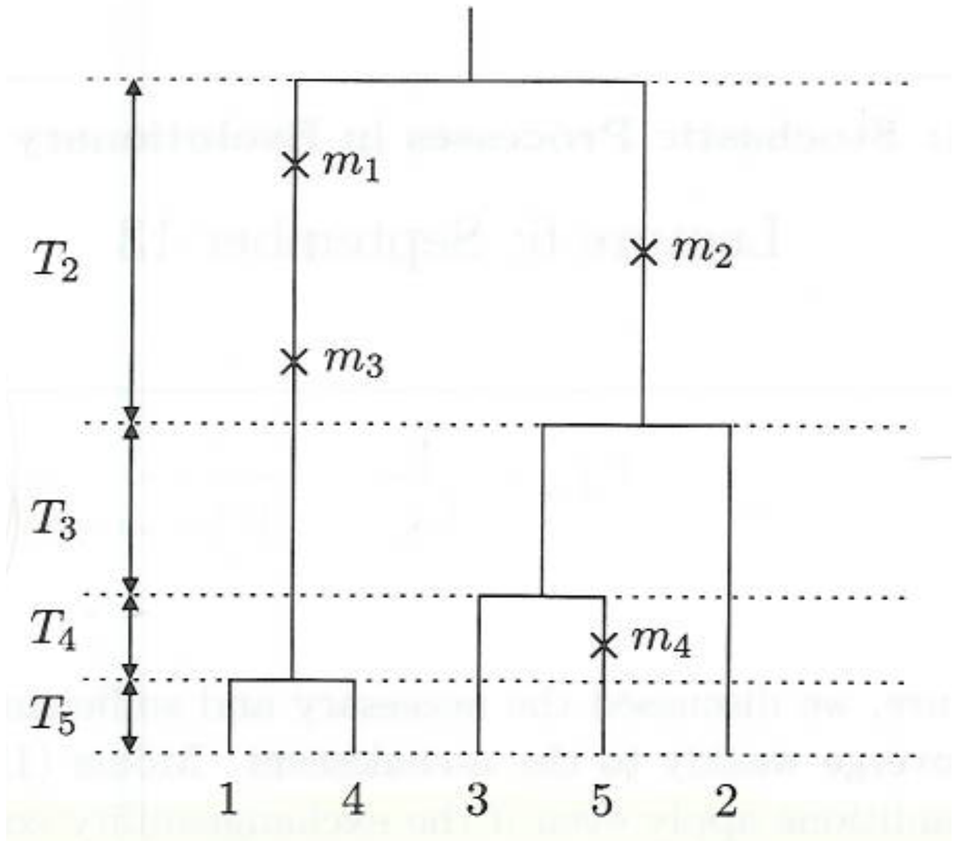
# Relating expectations about trees to data




Song, Y. 2021

# Relating expectations about trees to data

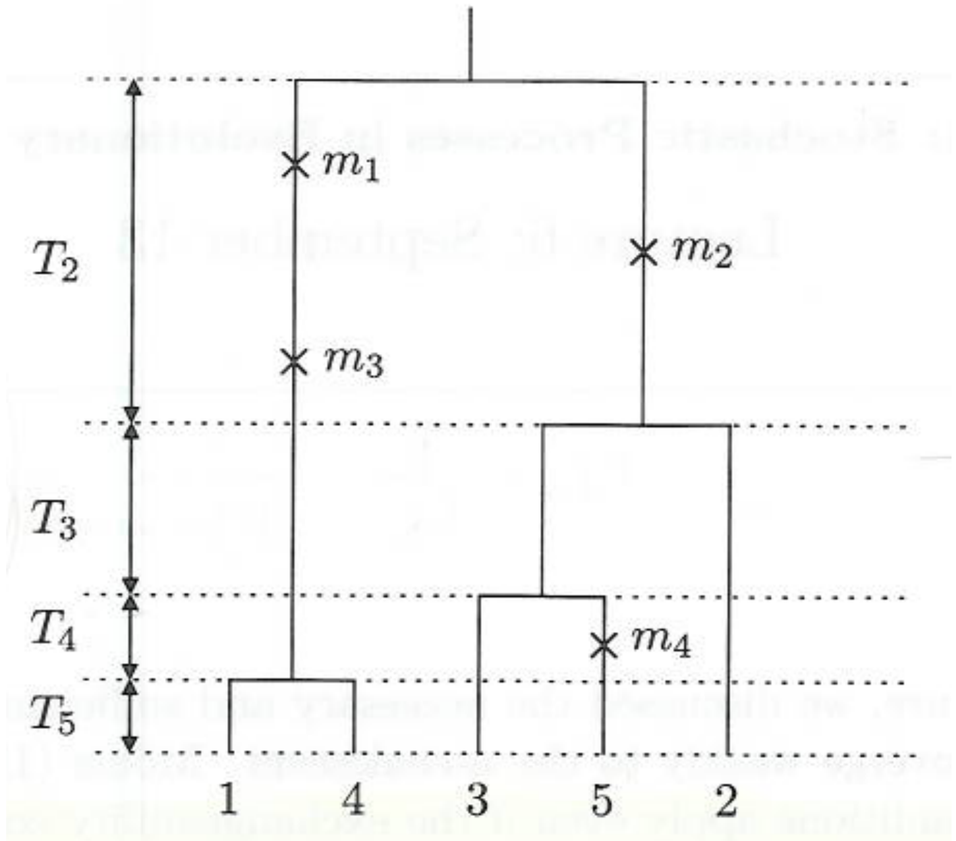
- We can use statistics computed on data to estimate hidden parameters, e.g. θ = 4Nu




Song, Y. 2021

# Relating expectations about trees to data

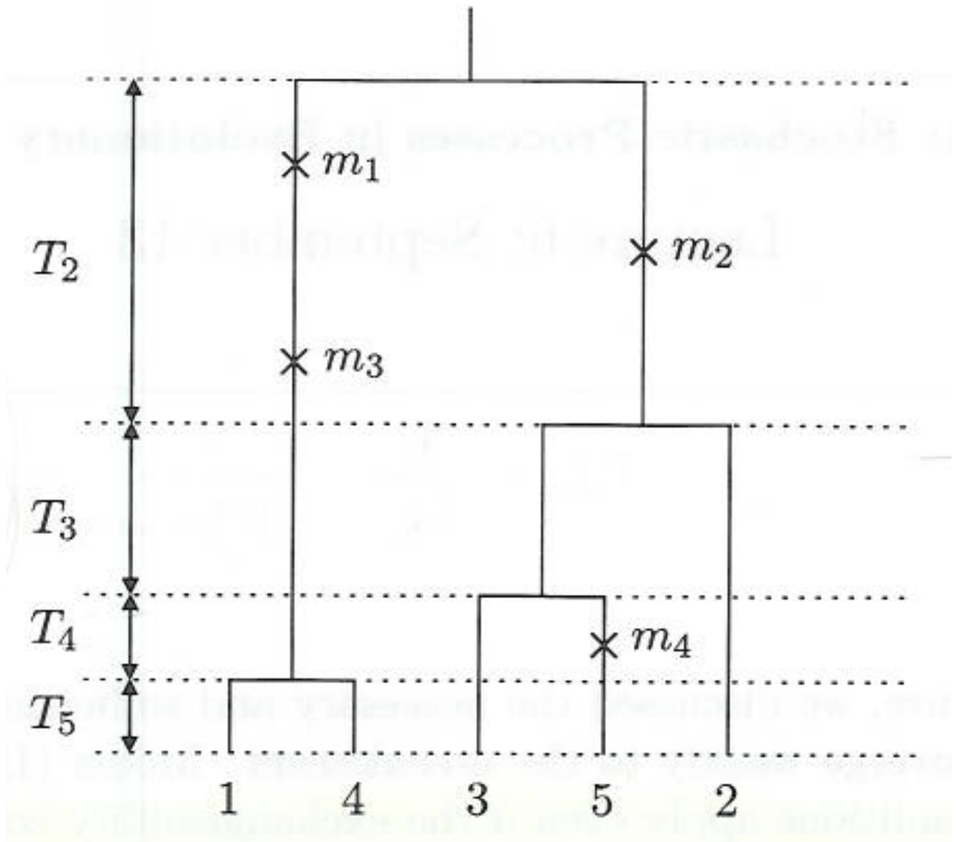
- We can use statistics computed on data to estimate hidden parameters, e.g. θ = 4Nu




Song, Y. 2021

# Relating expectations about trees to data

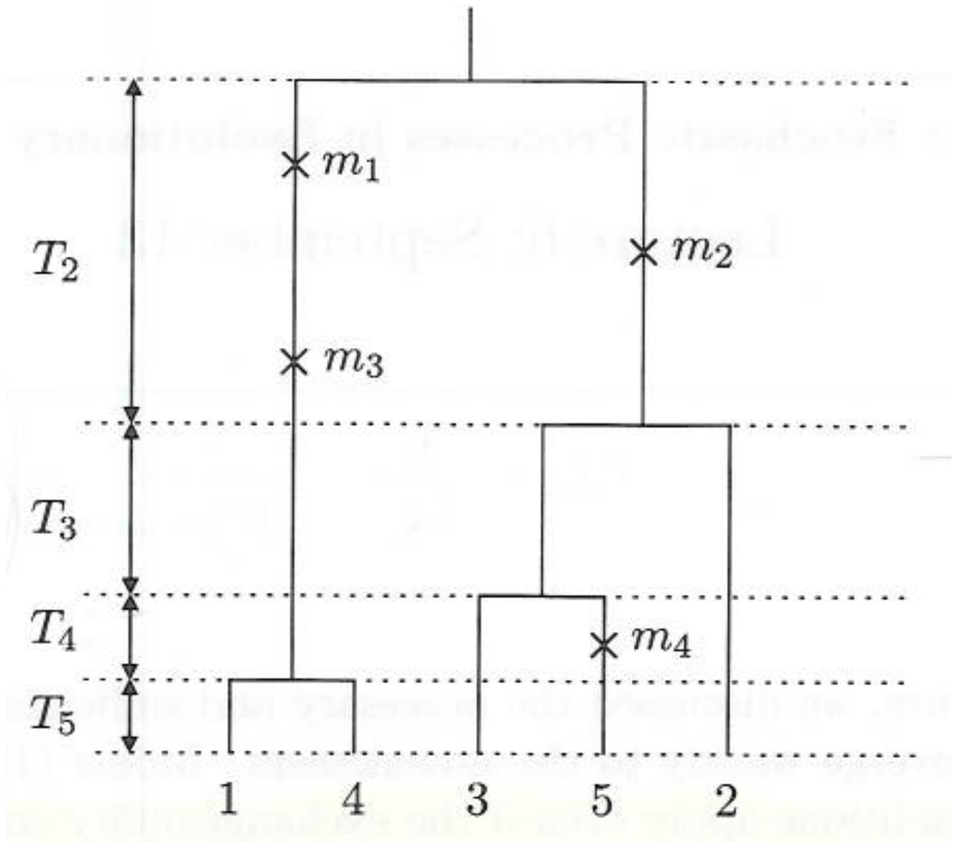
- We can use statistics computed on data to estimate hidden parameters, e.g. θ = 4Nu
- If we know “u” a priori (mutation rate at the studied locus), we can solve for 2N to obtain an **estimate of population size**




Song, Y. 2021

# Relating expectations about trees to data

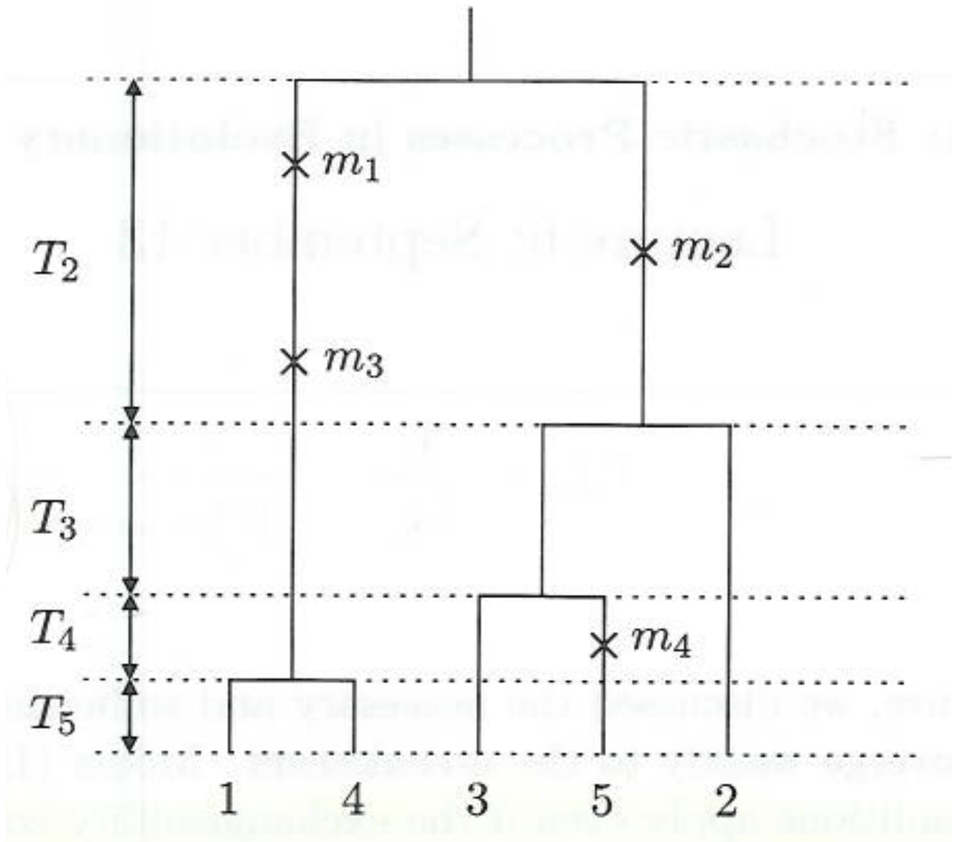
- We can use statistics computed on data to estimate hidden parameters, e.g. θ = 4Nu
- If we know “u” a priori (mutation rate at the studied locus), we can solve for 2N to obtain an **estimate of population size**




Song, Y. 2021

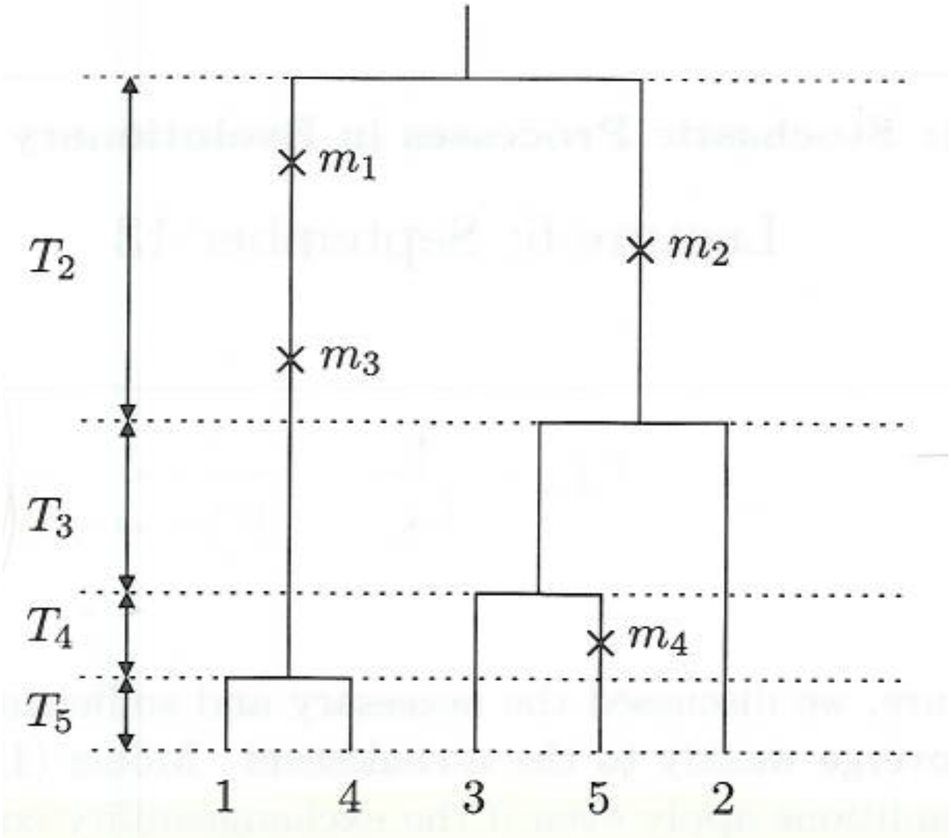
# Relating expectations about trees to data

- We can use statistics computed on data to estimate hidden parameters, e.g. θ = 4Nu
- If we know “u” a priori (mutation rate at the studied locus), we can solve for 2N to obtain an **estimate of population size**
- One statistic that is informative of θ is the **number of segregating sites in a sample**




Song, Y. 2021

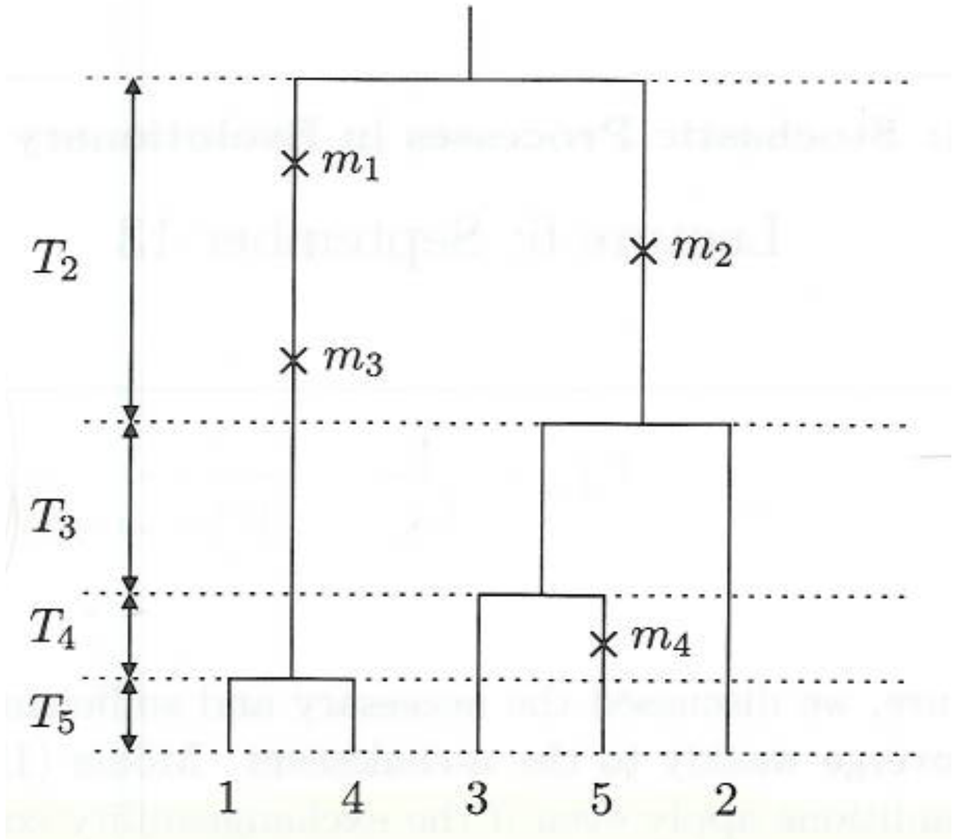
# Relating expectations about trees to data




Song, Y. 2021

# Relating expectations about trees to data

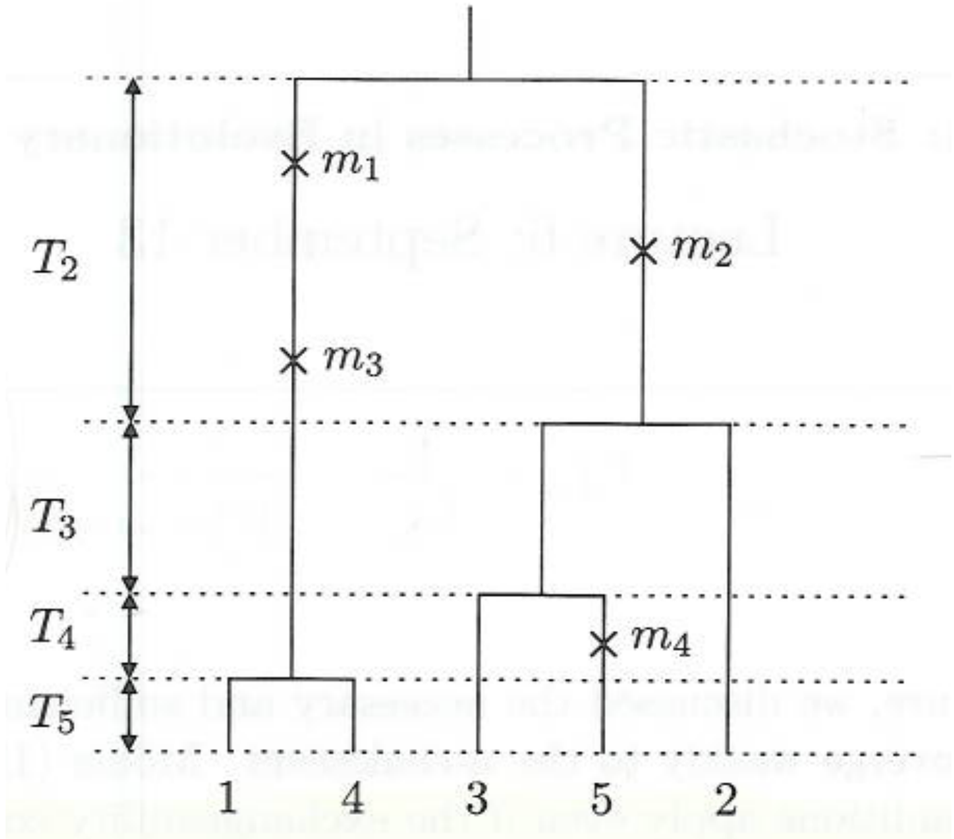
- Let S = number of segregating sites at a locus




Song, Y. 2021

# Relating expectations about trees to data

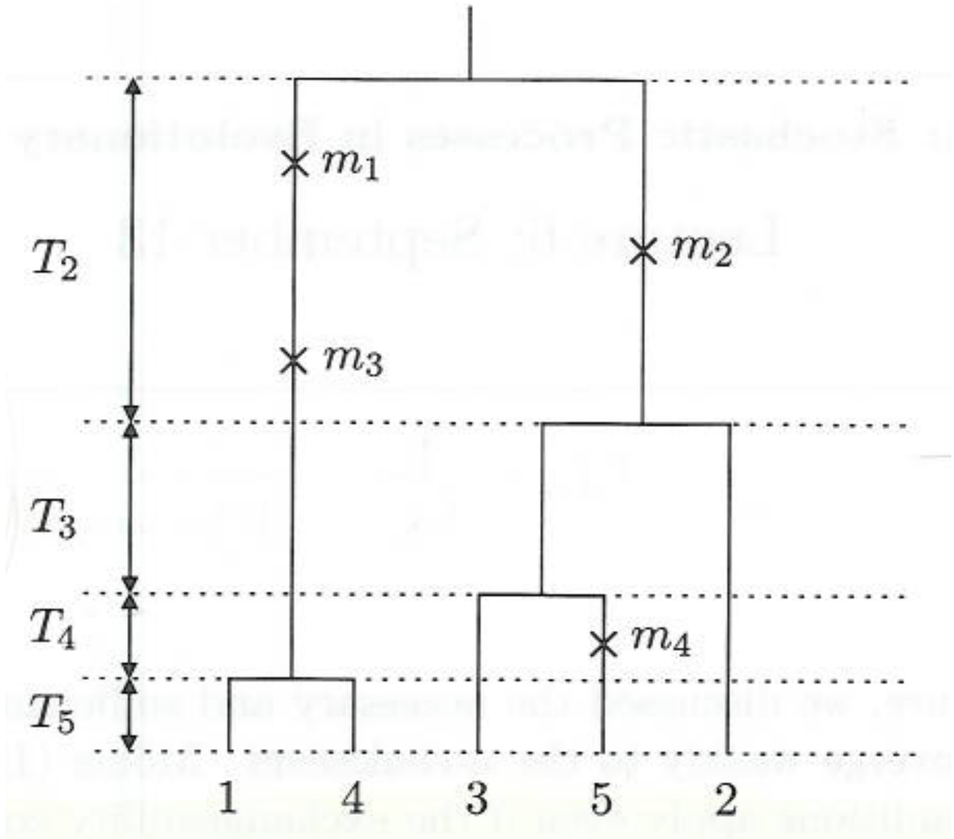
- Let S = number of segregating sites at a locus




Song, Y. 2021

# Relating expectations about trees to data

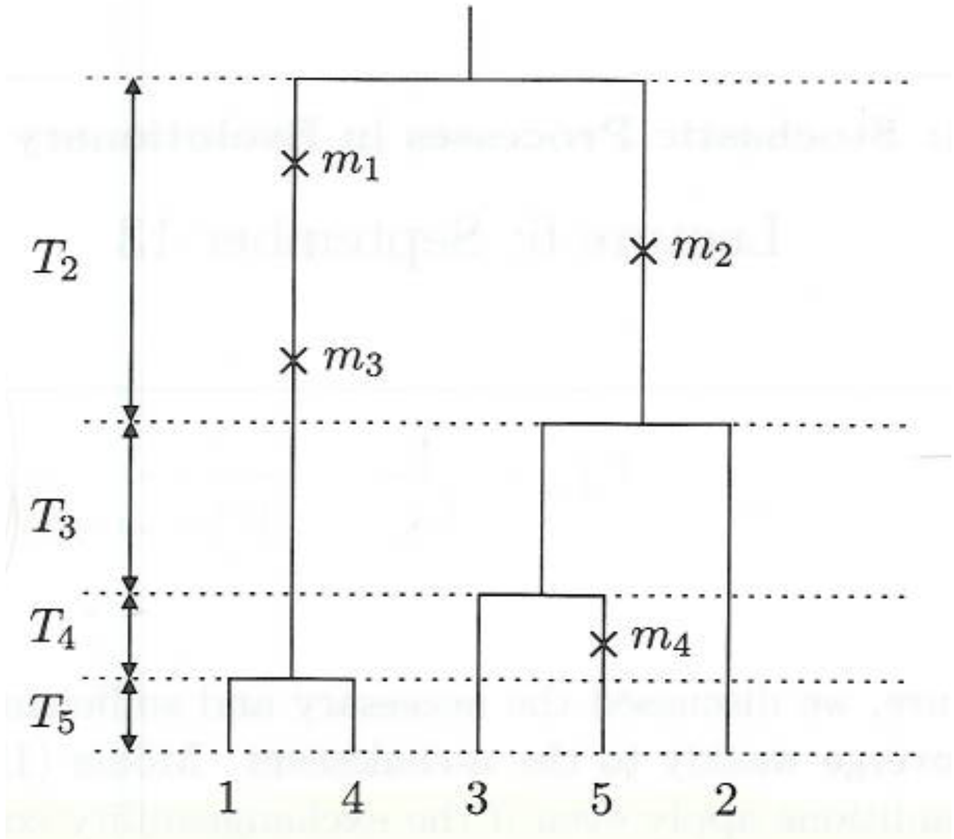
- Let S = number of segregating sites at a locus
- Under the infinite sites model, S = number of mutations in the tree at a locus = M




Song, Y. 2021

# Relating expectations about trees to data

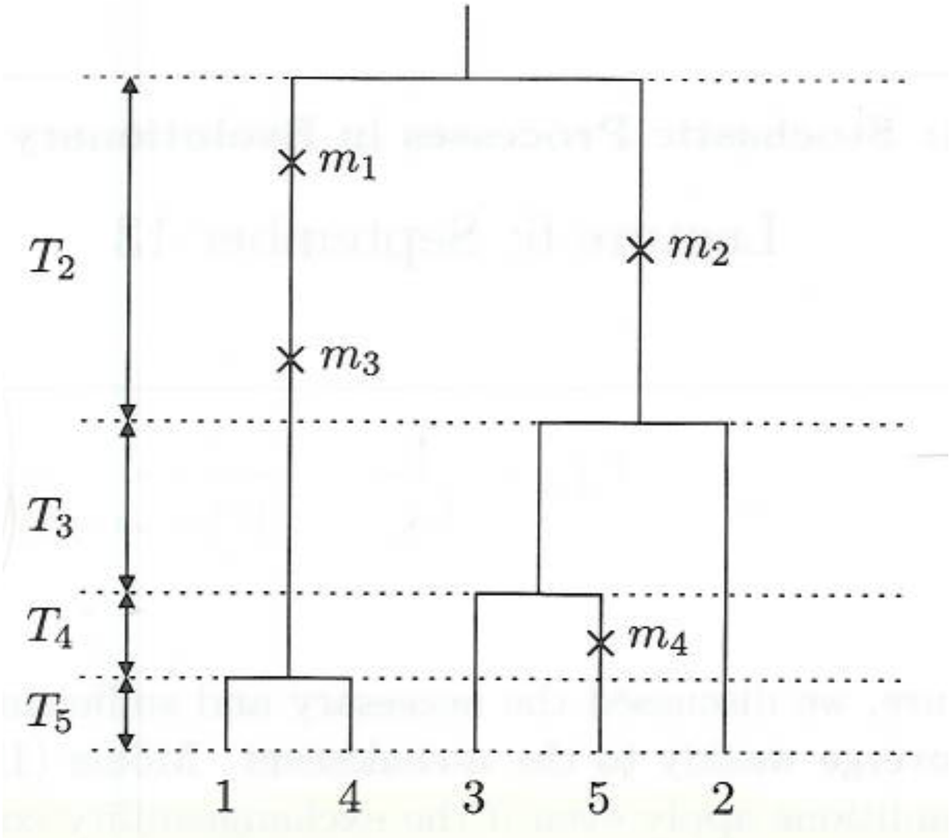
- Let S = number of segregating sites at a locus
- Under the infinite sites model, S = number of mutations in the tree at a locus = M




Song, Y. 2021

# Relating expectations about trees to data

- Let S = number of segregating sites at a locus
- Under the infinite sites model, S = number of mutations in the tree at a locus = M
- $E[S] = E[M] = \frac{\theta}{2} E[Total\ length\ of\ tree]$

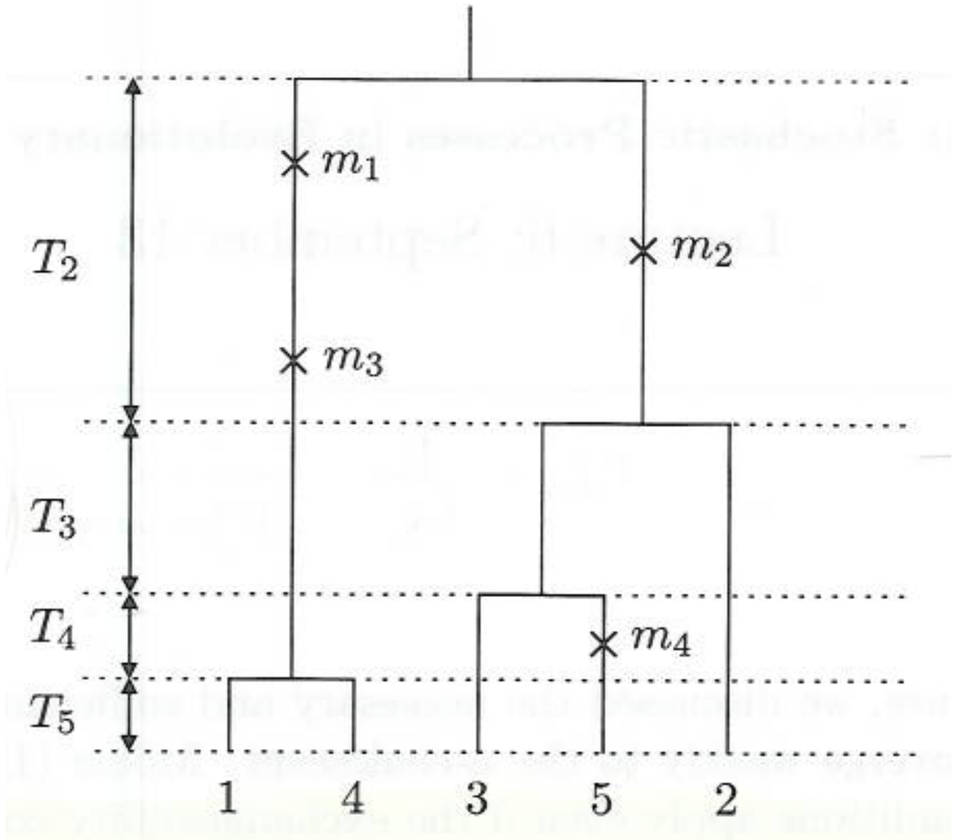



Song, Y. 2021

# Relating expectations about trees to data

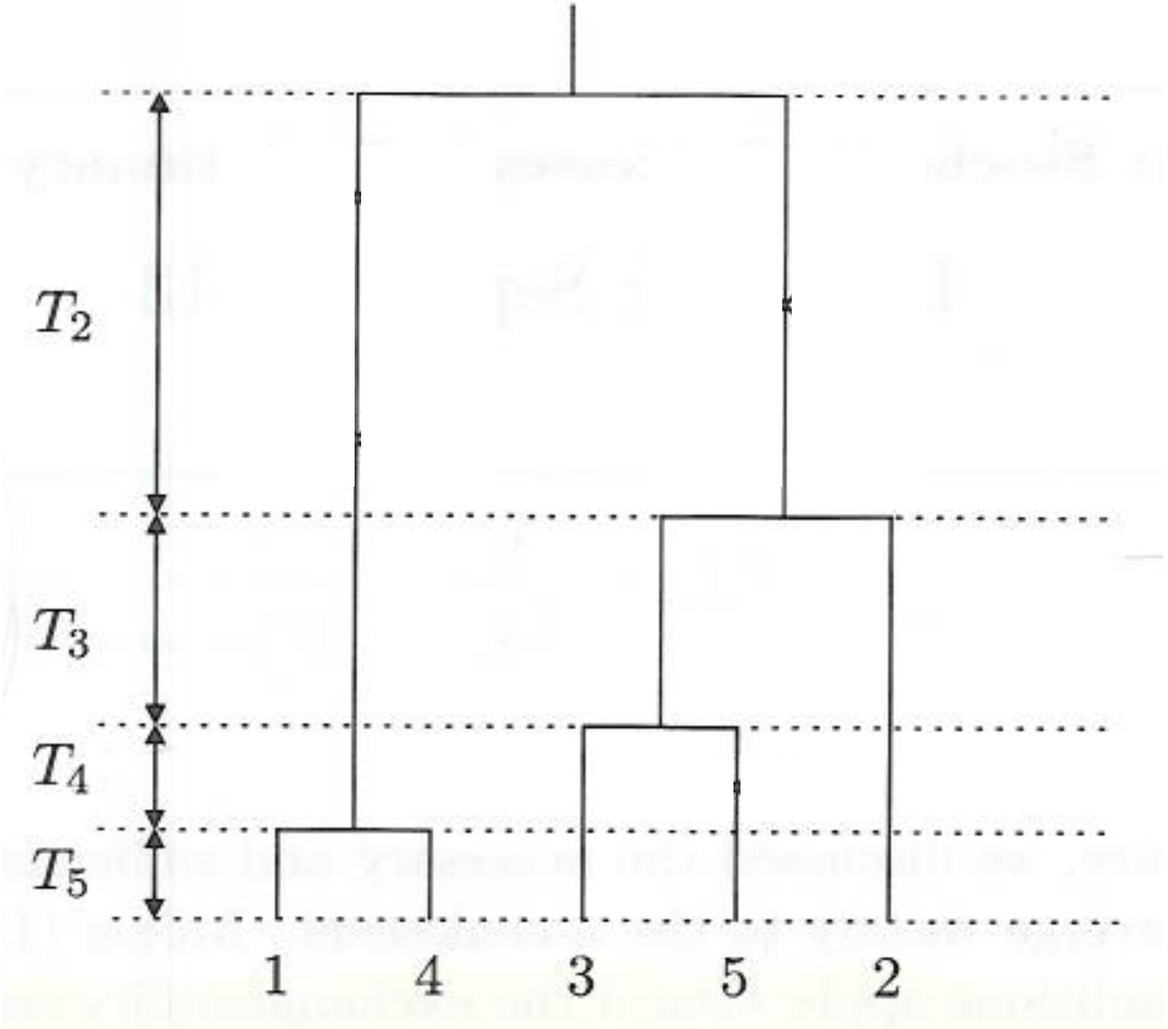
- Let S = number of segregating sites at a locus
- Under the infinite sites model, S = number of mutations in the tree at a locus = M
- $E[S] = E[M] = \frac{\theta}{2} E[Total\ length\ of\ tree]$

Rate at which mutations happen



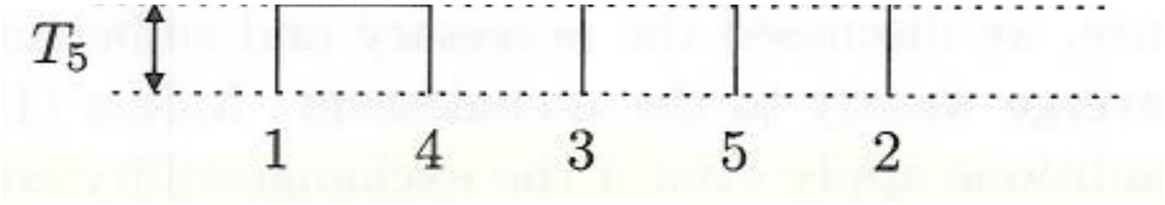

Song, Y. 2021

# Total length of tree




Song, Y. 2021

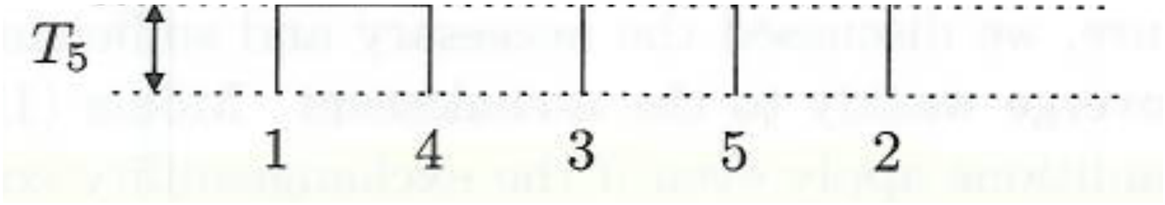
# Total length of tree




Song, Y. 2021

# Total length of tree

5 lineages

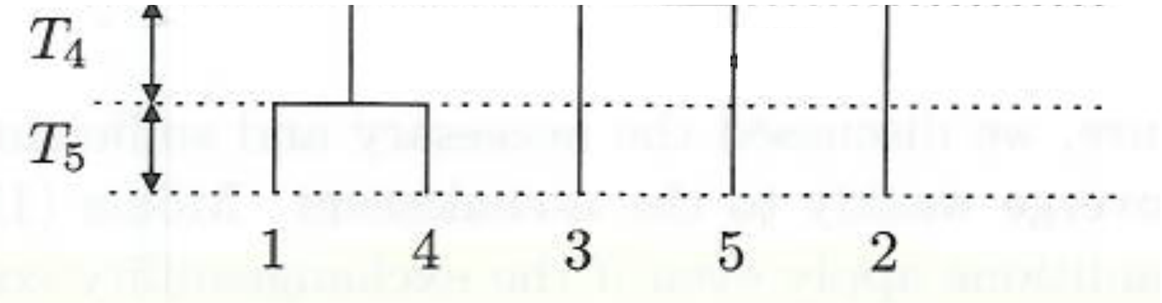



Song, Y. 2021

# Total length of tree

4 lineages

5 lineages



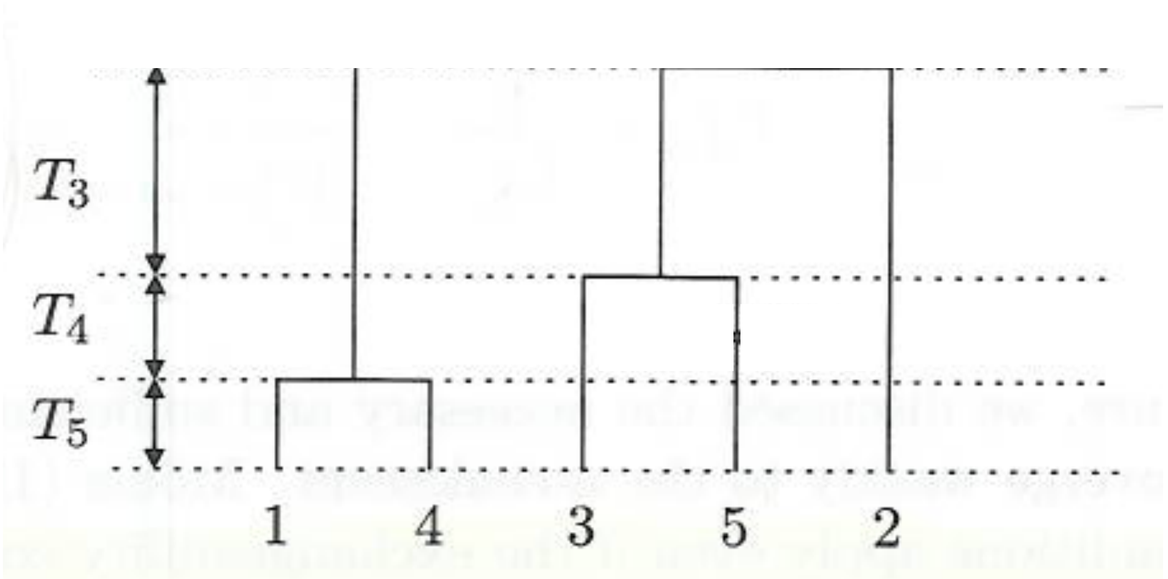

Song, Y. 2021

# Total length of tree

3 lineages

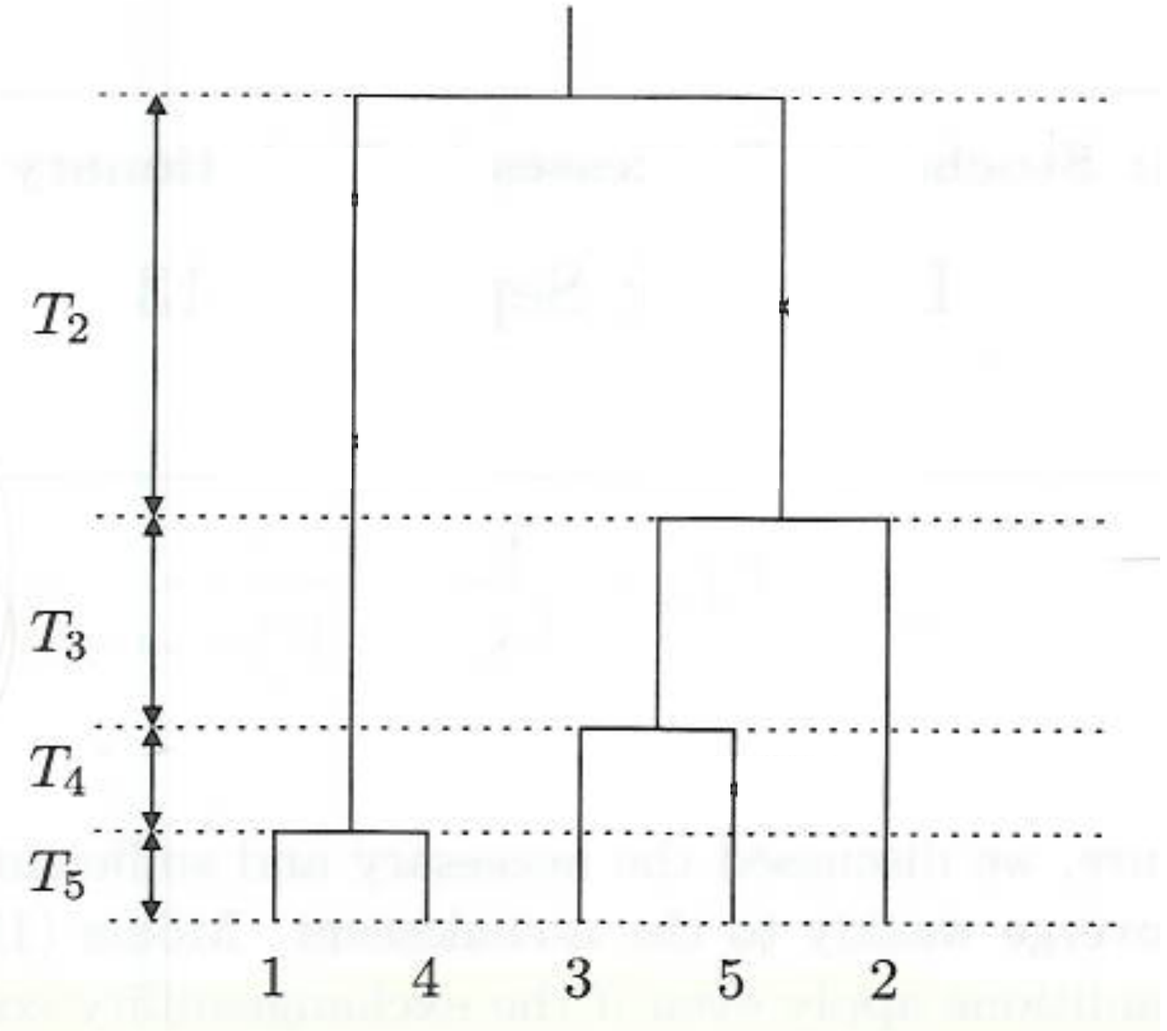
4 lineages

5 lineages



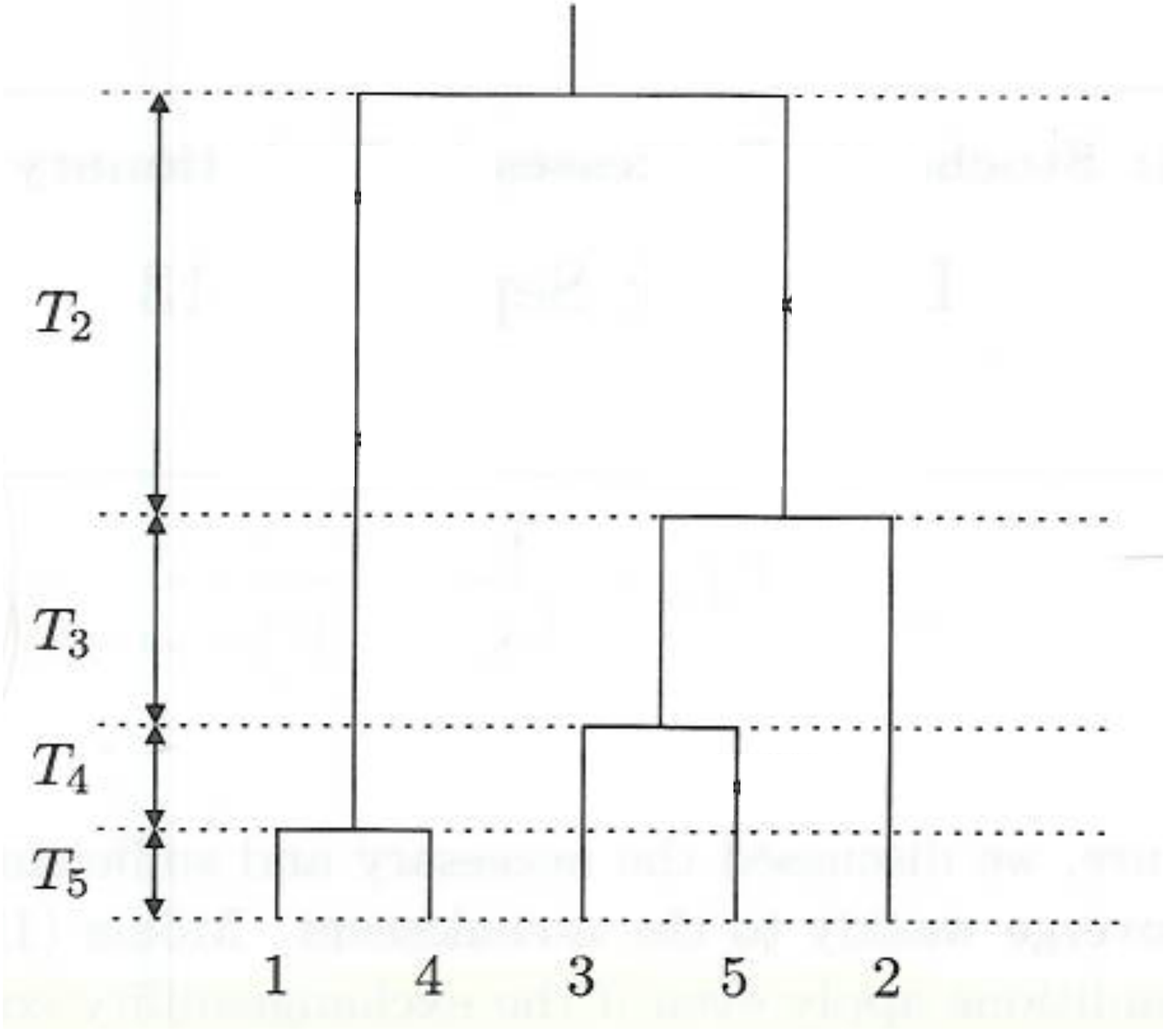

Song, Y. 2021

# Total length of tree



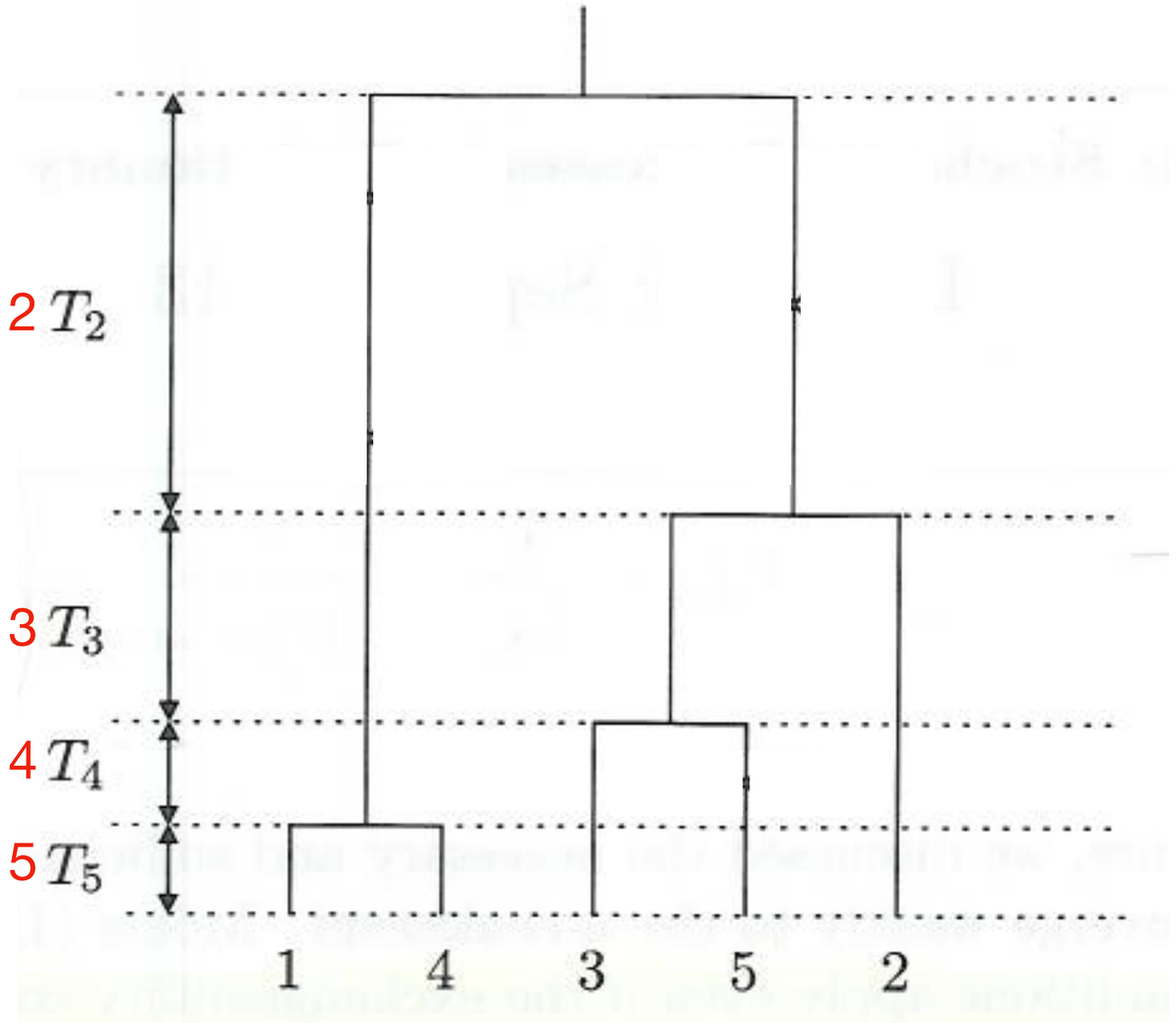

Song, Y. 2021

# Total length of tree




Song, Y. 2021

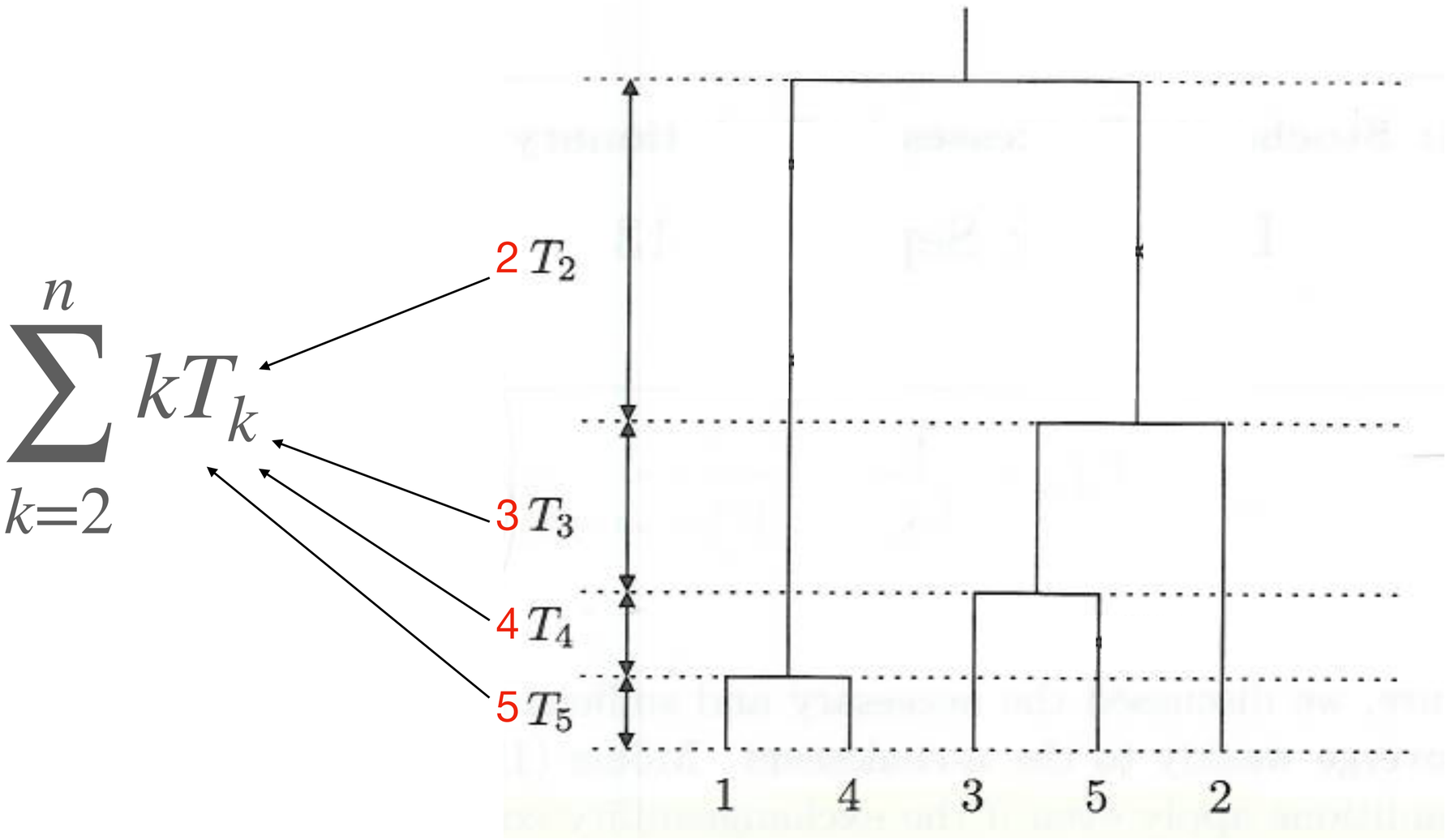
# Total length of tree




Song, Y. 2021

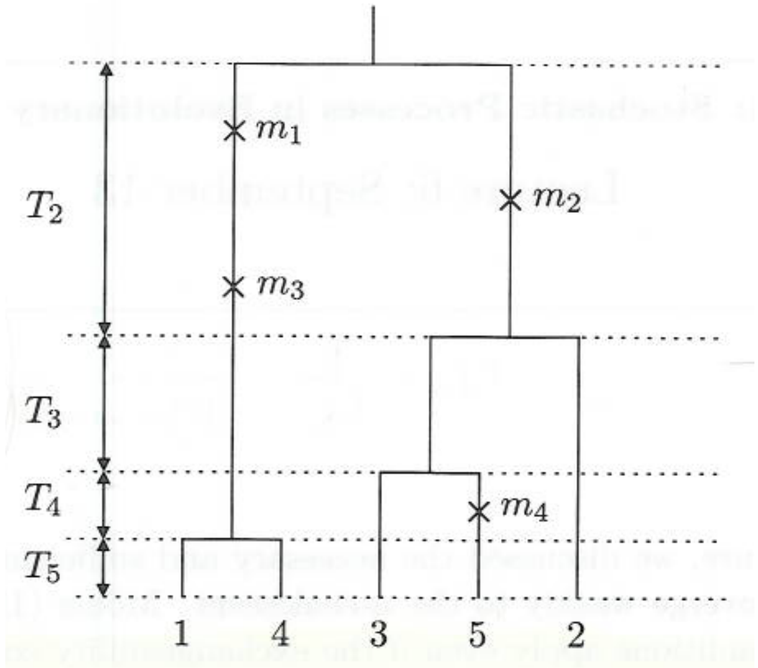
# Total length of tree

$$\sum_{k=2}^{n} kT_k$$




Song, Y. 2021

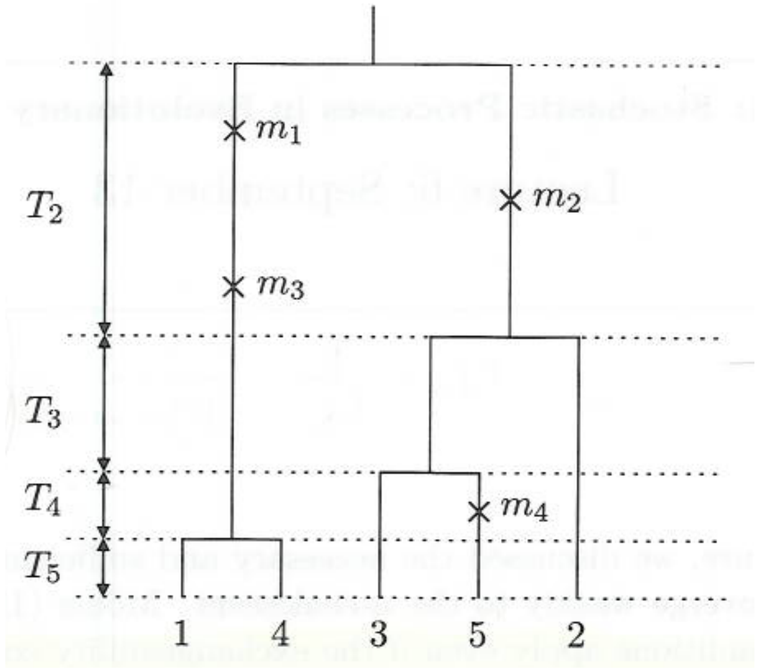
# Relating expectations about trees to data




Song, Y. 2021

# Relating expectations about trees to data

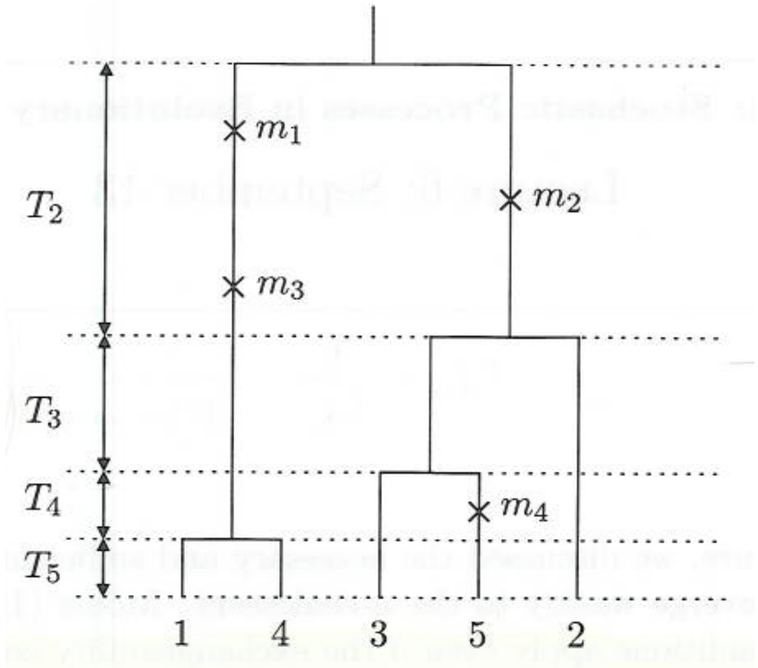
- $E[S] = E[M] = \frac{\theta}{2} E[Total\ length\ of\ tree]$




Song, Y. 2021

# Relating expectations about trees to data

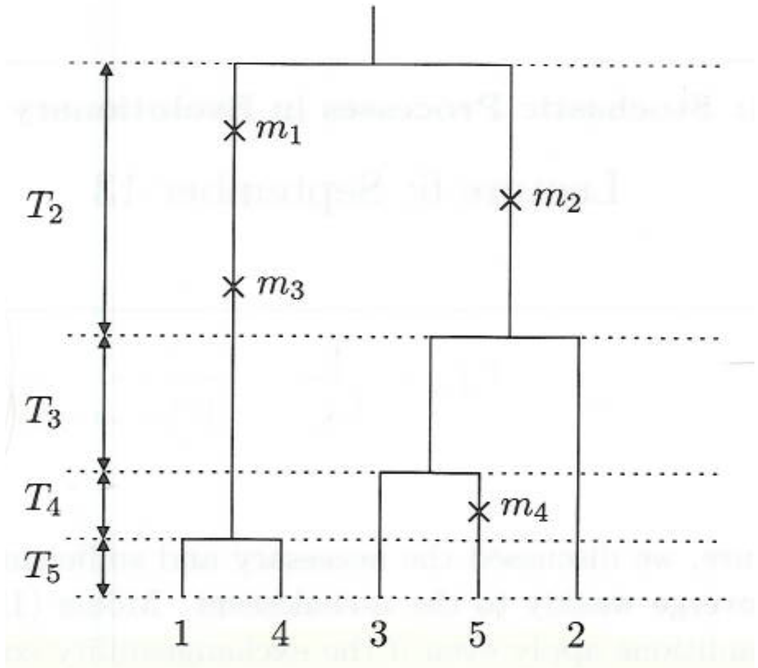
- $E[S] = E[M] = \frac{\theta}{2} E[Total\ length\ of\ tree]$




Song, Y. 2021

# Relating expectations about trees to data

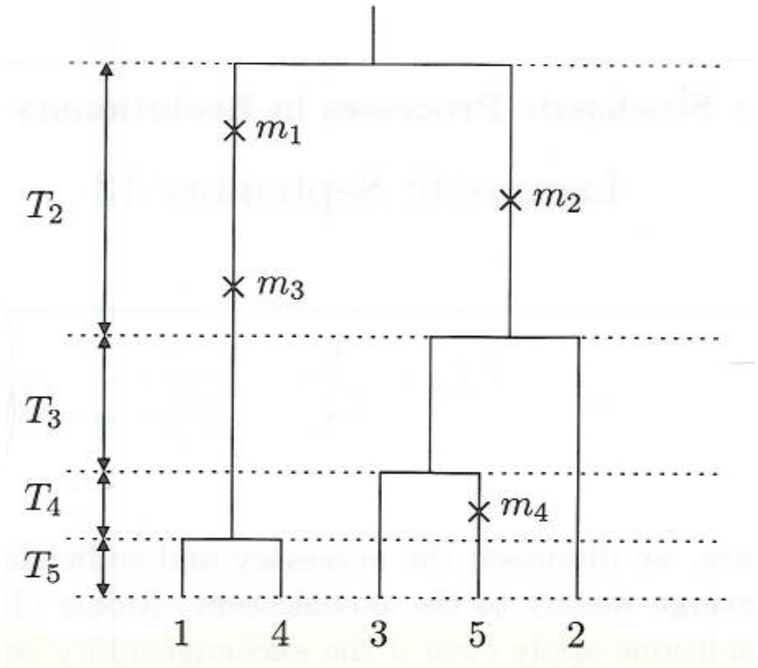
- $E[S] = E[M] = \frac{\theta}{2} E[Total\ length\ of\ tree]$
- $E[Total\ length\ of\ tree] = E\left[\sum_{k=2}^{n} kT_k\right]$




Song, Y. 2021

# Relating expectations about trees to data

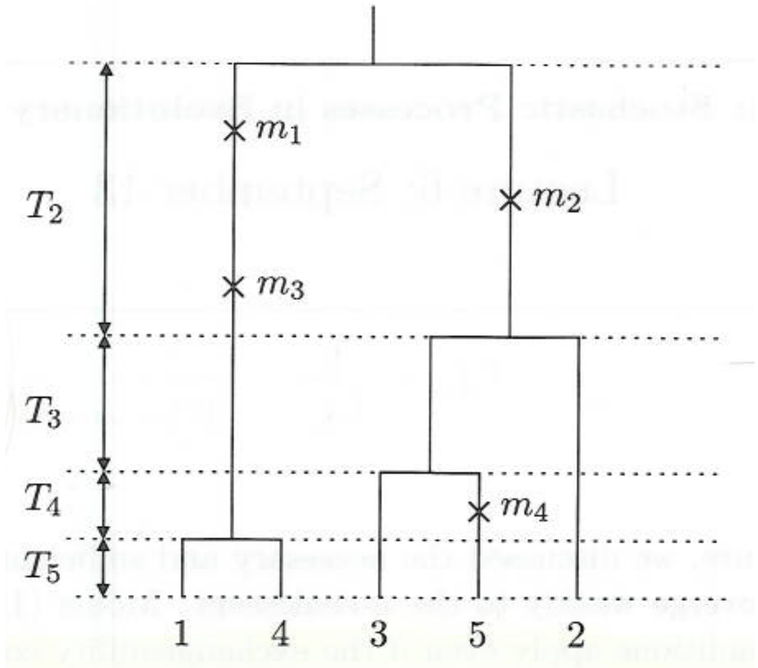
- $E[S] = E[M] = \frac{\theta}{2} E[Total\ length\ of\ tree]$
- $E[Total\ length\ of\ tree] = E\left[\sum_{k=2}^{n} kT_k\right]$




Song, Y. 2021

# Relating expectations about trees to data

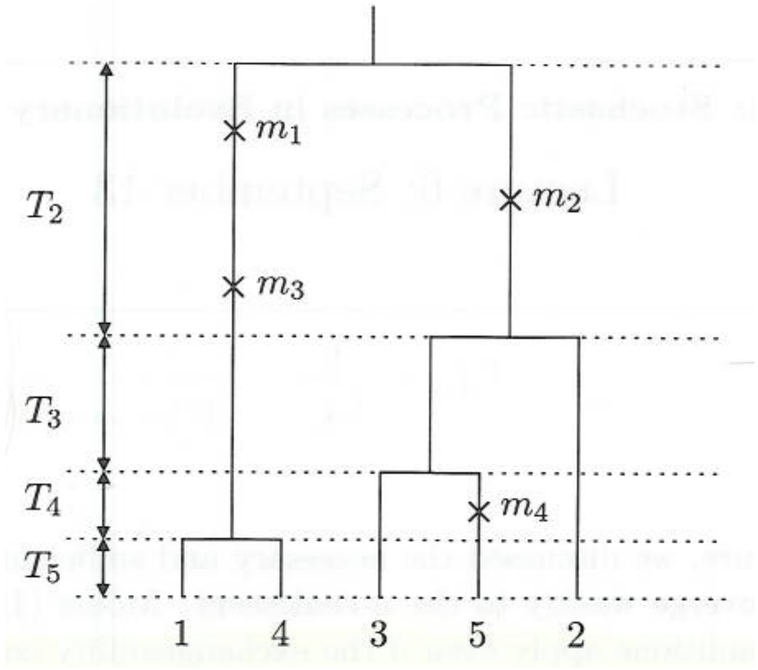
- $E[S] = E[M] = \frac{\theta}{2} E[Total\ length\ of\ tree]$
- $E[Total\ length\ of\ tree] = E\left[\sum_{k=2}^{n} kT_k\right]$




Song, Y. 2021

# Relating expectations about trees to data

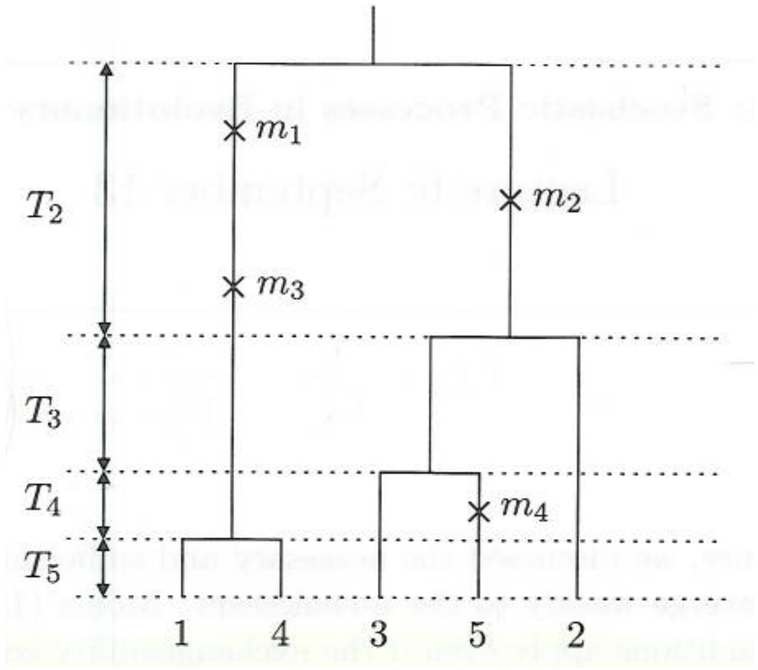
- $E[S] = E[M] = \frac{\theta}{2} E[Total\ length\ of\ tree]$
- $E[Total\ length\ of\ tree] = E\left[\sum_{k=2}^{n} kT_k\right]$
- $E\left[\sum_{k=2}^{n} kT_k\right] = \sum_{k=2}^{n} kE[T_k]$




Song, Y. 2021

# Relating expectations about trees to data

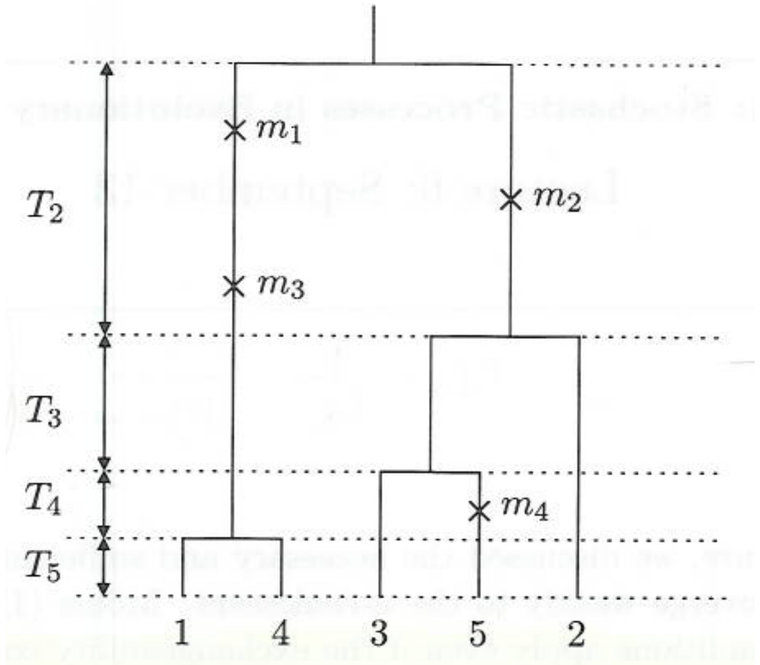
- $E[S] = E[M] = \frac{\theta}{2} E[Total\ length\ of\ tree]$
- $E[Total\ length\ of\ tree] = E\left[\sum_{k=2}^{n} kT_k\right]$
- $E\left[\sum_{k=2}^{n} kT_k\right] = \sum_{k=2}^{n} kE[T_k] = \sum_{k=2}^{n} k\frac{1}{\binom{k}{2}} = \sum_{k=2}^{n} \frac{2}{k-1}$




Song, Y. 2021

# Relating expectations about trees to data

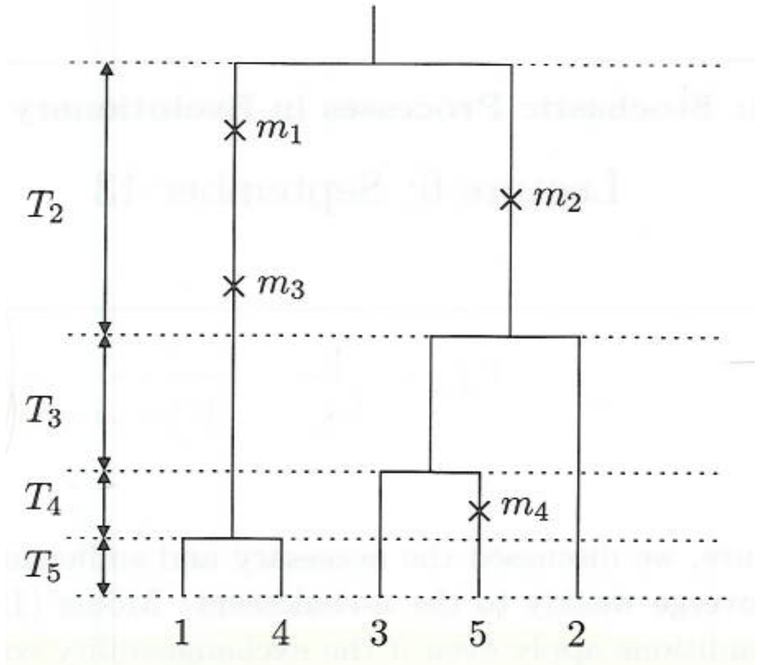
$$E[S] = \frac{\theta}{2} \sum_{k=2}^{n} \frac{2}{k-1} = \theta \sum_{k=2}^{n} \frac{1}{k-1}$$




Song, Y. 2021

# Relating expectations about trees to data

$$E[S] = \frac{\theta}{2}\sum_{k=2}^{n}\frac{2}{k-1} = \theta\sum_{k=2}^{n}\frac{1}{k-1}$$



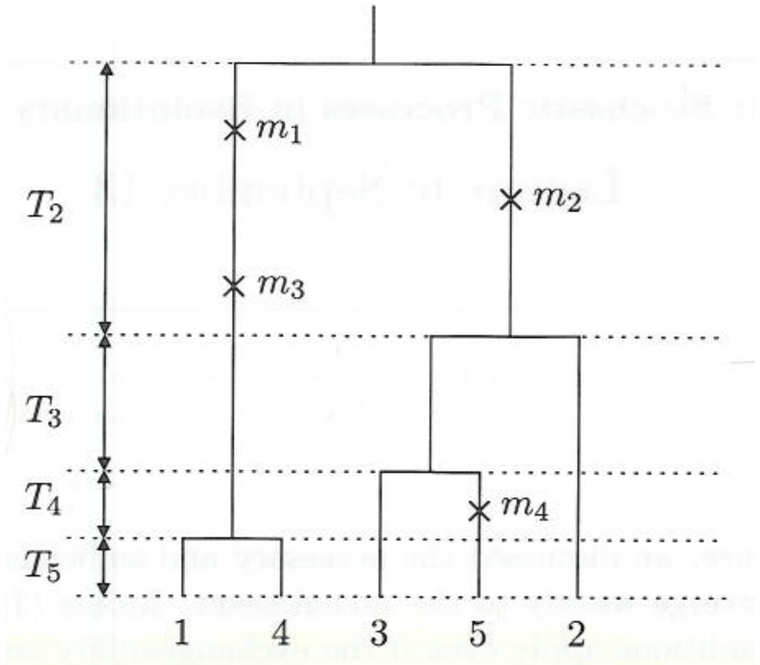

Song, Y. 2021

# Relating expectations about trees to data

$$E[S] = \frac{\theta}{2}\sum_{k=2}^{n}\frac{2}{k-1} = \theta\sum_{k=2}^{n}\frac{1}{k-1}$$

Observed

Unknown



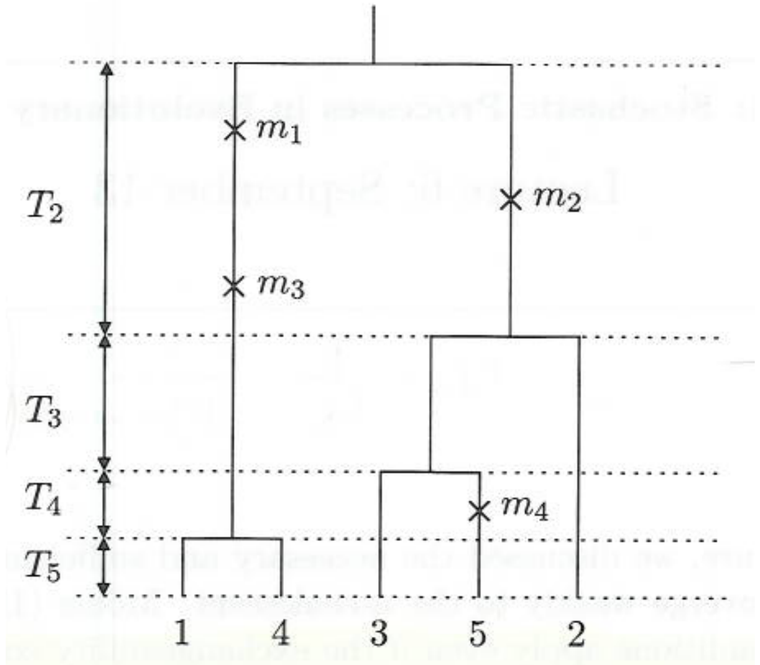

Song, Y. 2021

# Relating expectations about trees to data

$$E[S] = \frac{\theta}{2}\sum_{k=2}^{n}\frac{2}{k-1} = \theta\sum_{k=2}^{n}\frac{1}{k-1}$$

Observed

Unknown




$$\hat{\theta} = \frac{S}{\sum_{k=2}^{n}\frac{1}{k-1}}$$

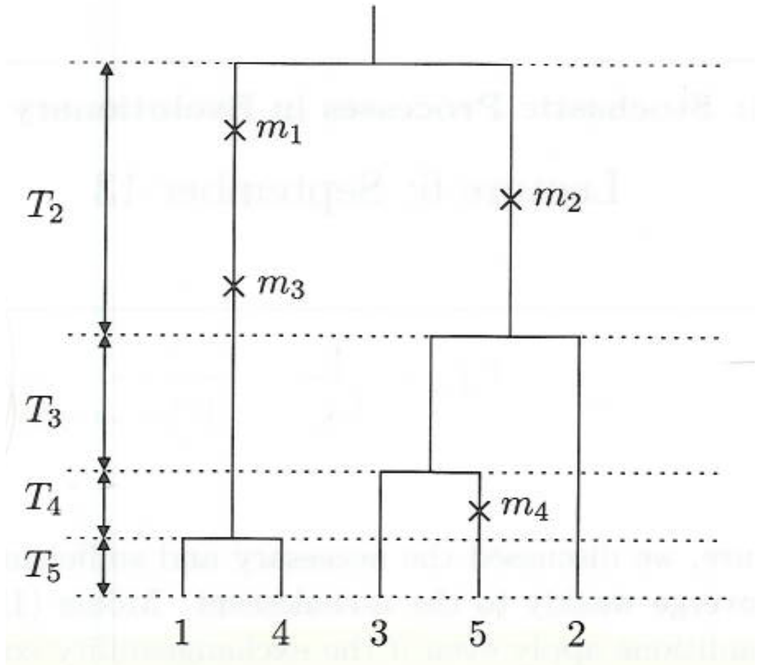
Song, Y. 2021

# Relating expectations about trees to data

$$E[S] = \frac{\theta}{2}\sum_{k=2}^{n}\frac{2}{k-1} = \theta\sum_{k=2}^{n}\frac{1}{k-1}$$

Observed

Unknown




Observed

$$\hat{\theta} = \frac{S}{\sum_{k=2}^{n}\frac{1}{k-1}}$$

"Watterson's estimate" for θ

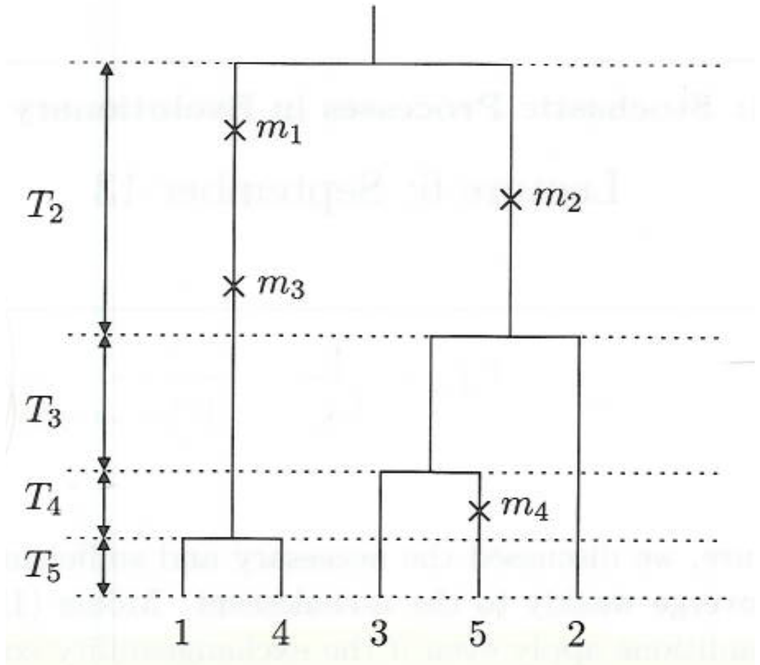
Song, Y. 2021

# Relating expectations about trees to data

$$E[S] = \frac{\theta}{2}\sum_{k=2}^{n}\frac{2}{k-1} = \theta\sum_{k=2}^{n}\frac{1}{k-1}$$

Observed

Unknown




$$\hat{\theta} = \frac{S}{\sum_{k=2}^{n}\frac{1}{k-1}}$$

Observed

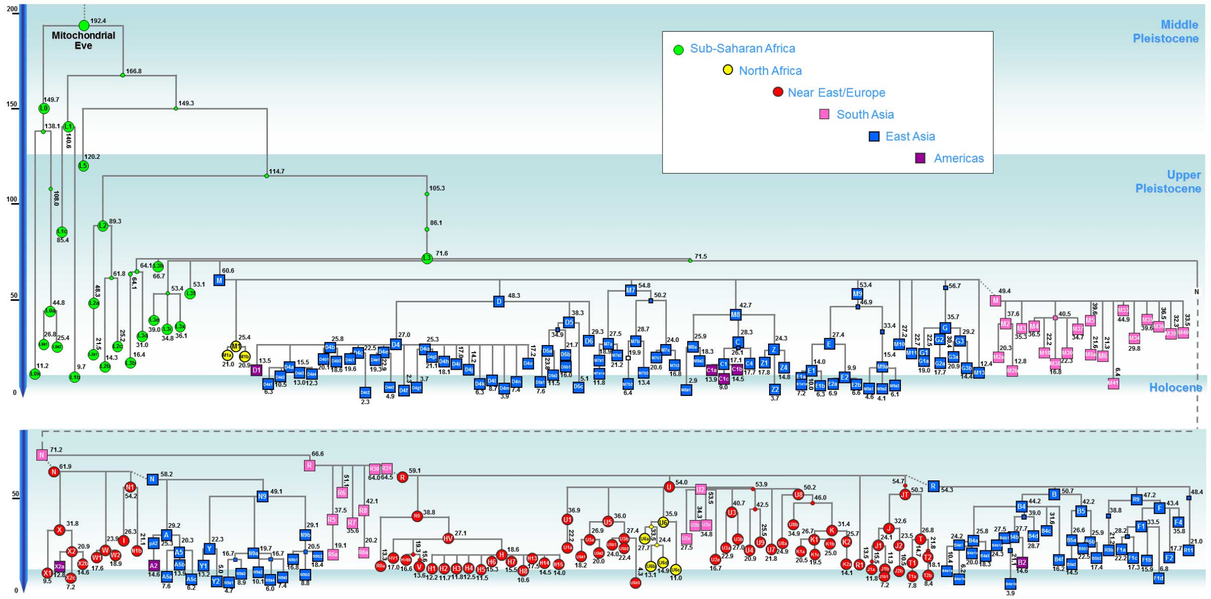
"Watterson's estimate" for θ

If we know u a priori, we can solve for N, and get an estimate of the population size:

$$2\hat{N} = \frac{\hat{\theta}}{2u}$$

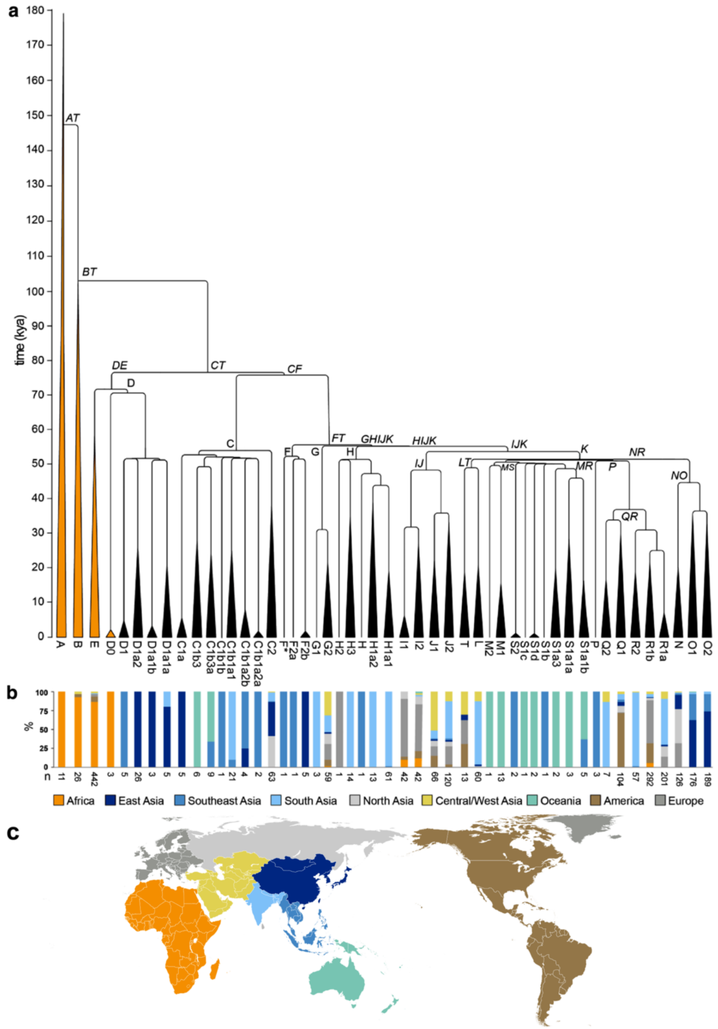
Song, Y. 2021

# Trees in non-recombining loci




mtDNA

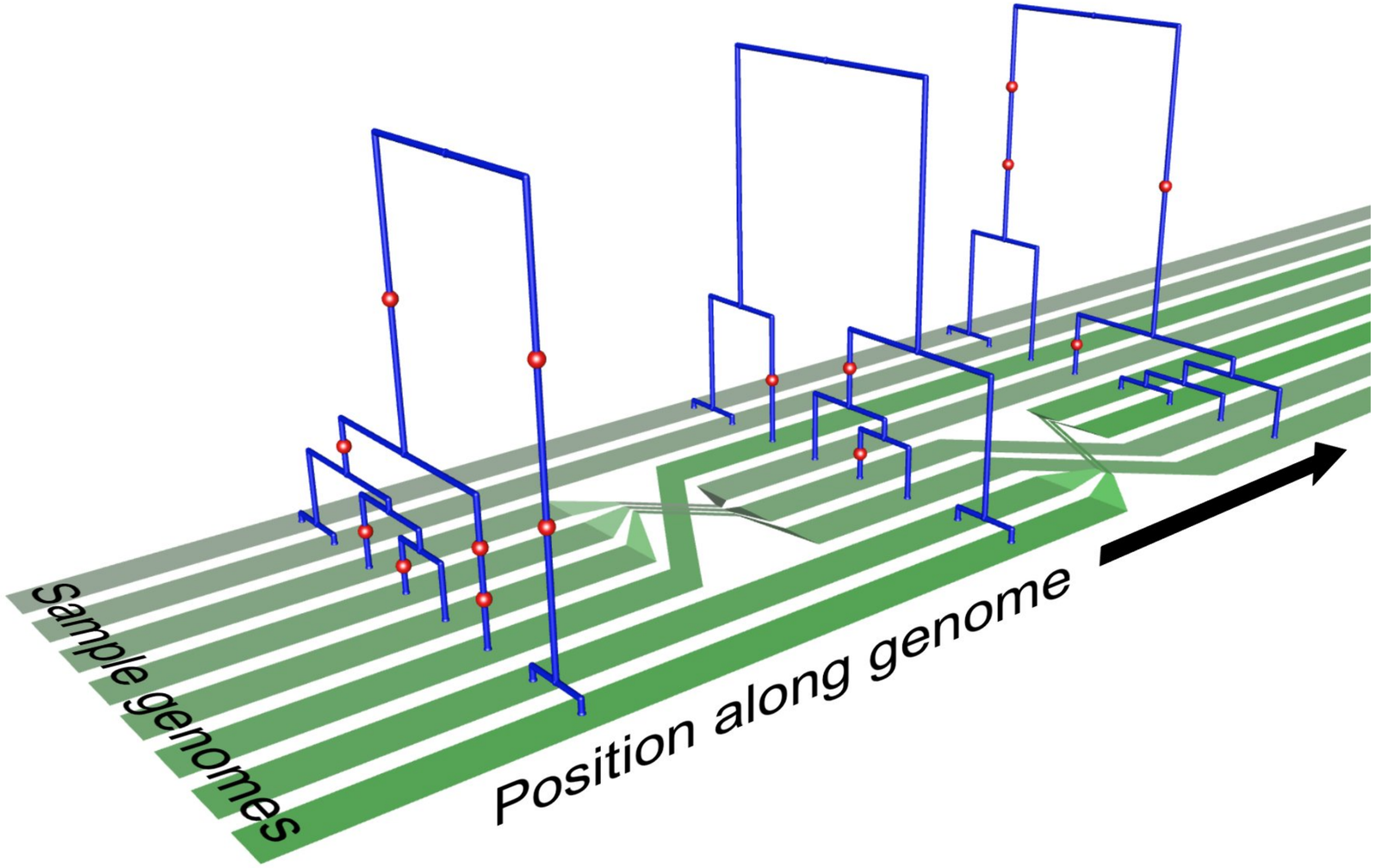
Soares et al. 2009




chrY

Agdzhoyan et al. 2020

# Trees in the rest of the genome

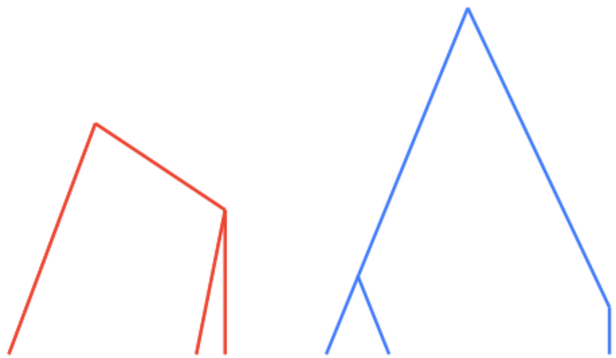



Yan Wong, 2020

# Trees contain information about population history

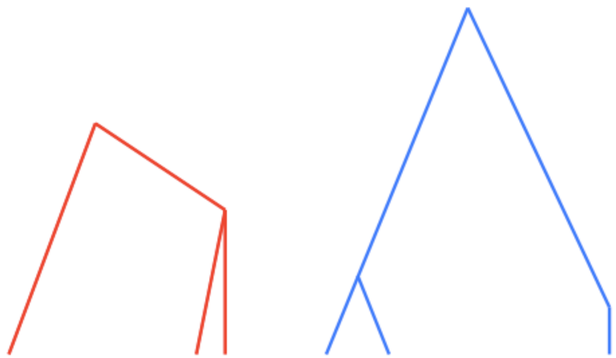
# Trees contain information about population history

- The shape of trees is random (but follows some rules)



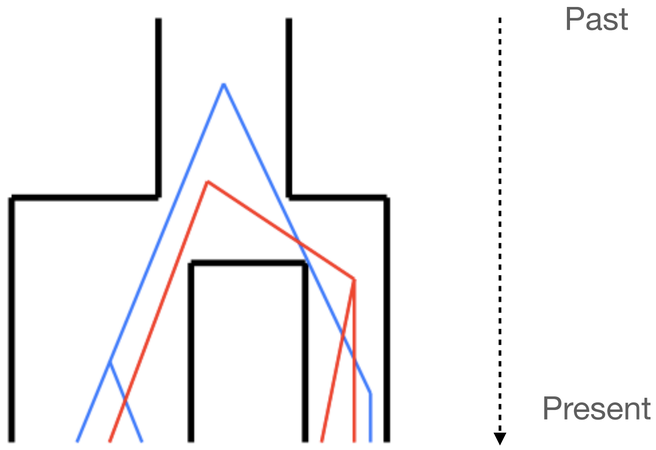
# Trees contain information about population history

- The shape of trees is random (but follows some rules)



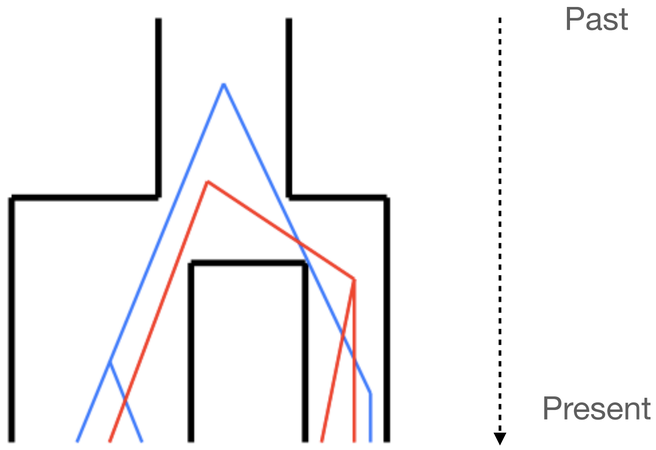
# Trees contain information about population history

- The shape of trees is random (but follows some rules)
- Two lineages can only coalesce if they exist in the same population



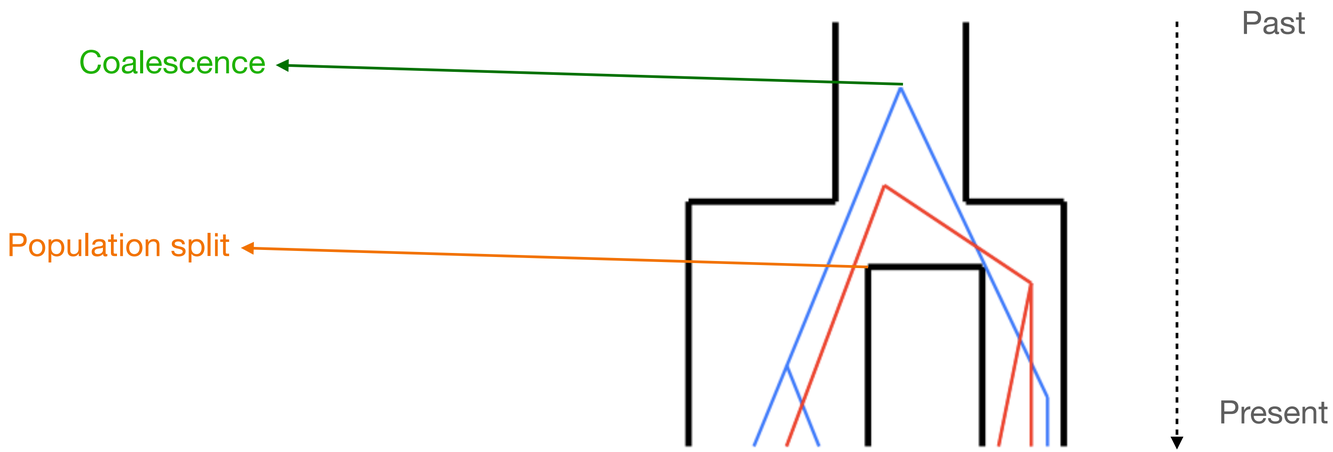
# Trees contain information about population history

- The shape of trees is random (but follows some rules)
- Two lineages can only coalesce if they exist in the same population

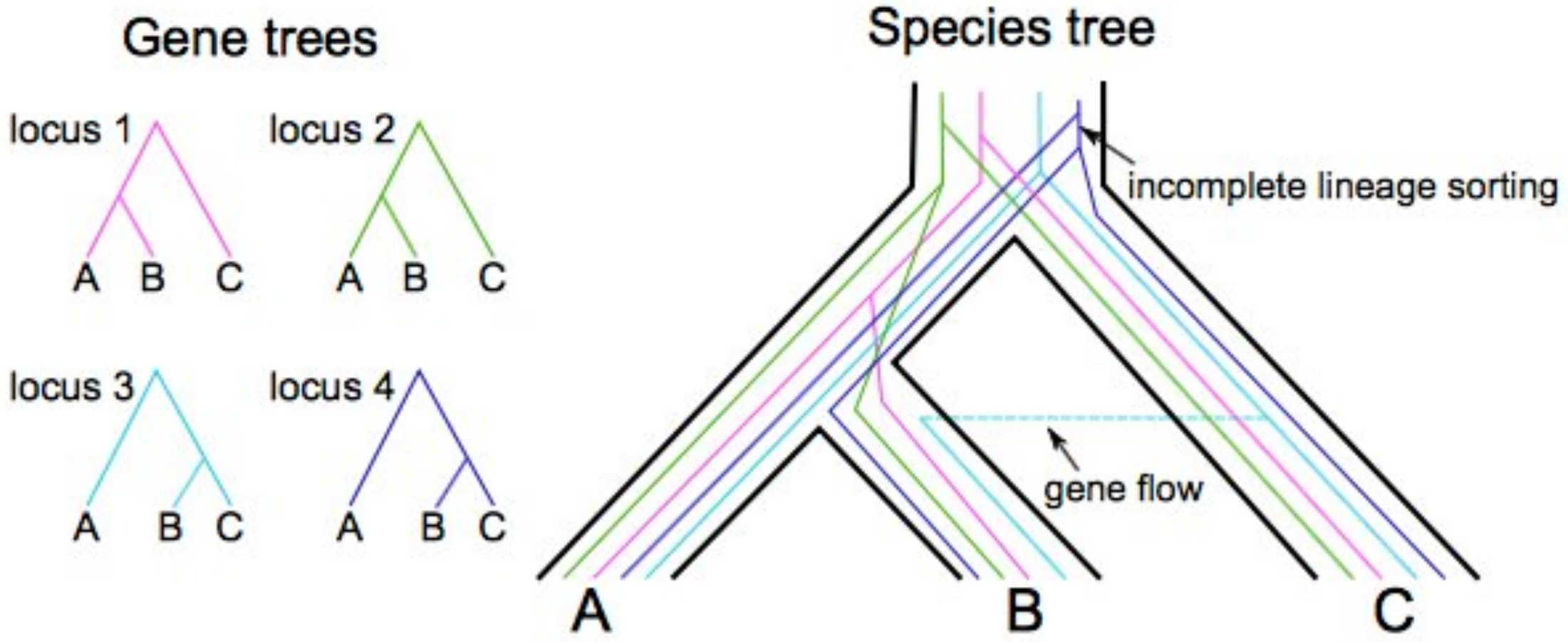


# Trees contain information about population history

- The shape of trees is random (but follows some rules)
- Two lineages can only coalesce if they exist in the same population
- Coalescent events do not represent population splits! (but they contain information about them)



# Trees contain information about population history



# Exercises

# Exercises

- Follow the instructions in this prompt: https://github.com/FerRacimo/CopenhagenTutorial/blob/master/CoalTutorial.md