Random genetic drift and effective population size

Biologia Evoluzionistica 2025/2026,
Fabrizio Mafessoni, Università di Trieste

木村 資生, *Motō Kimura (1924-1994)*

Genotypes
p^2, $2pq$, q^2

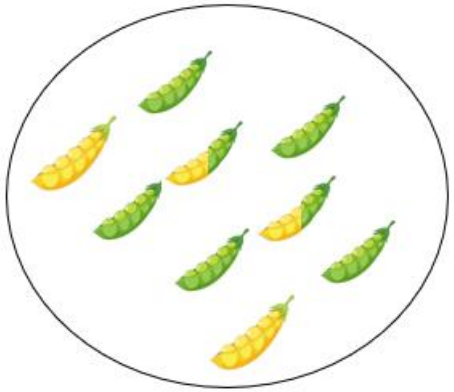
Allele/gametes p,q

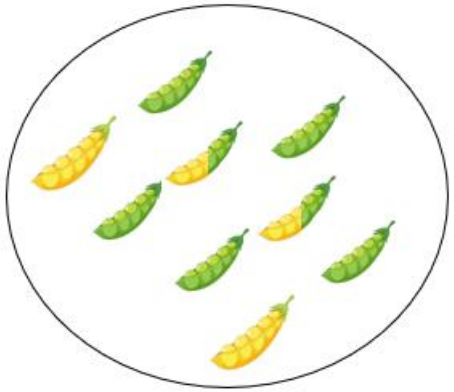
gametes

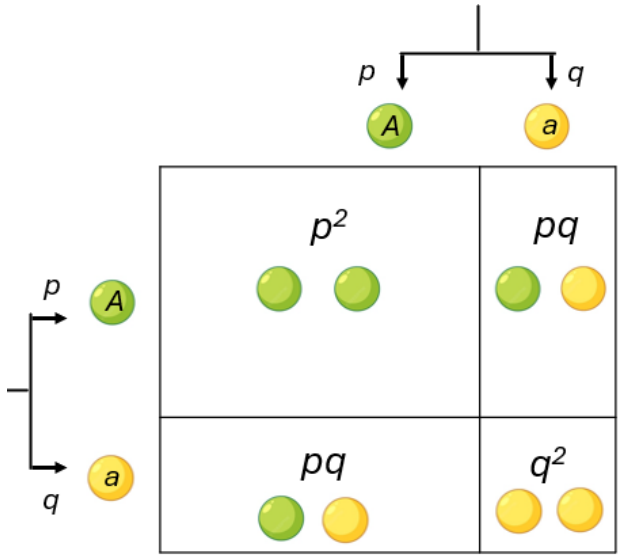
We can build a model of evolution using our Hardy-Weinberg's equilibrium..


G.H. Hardy
(1877-1947)

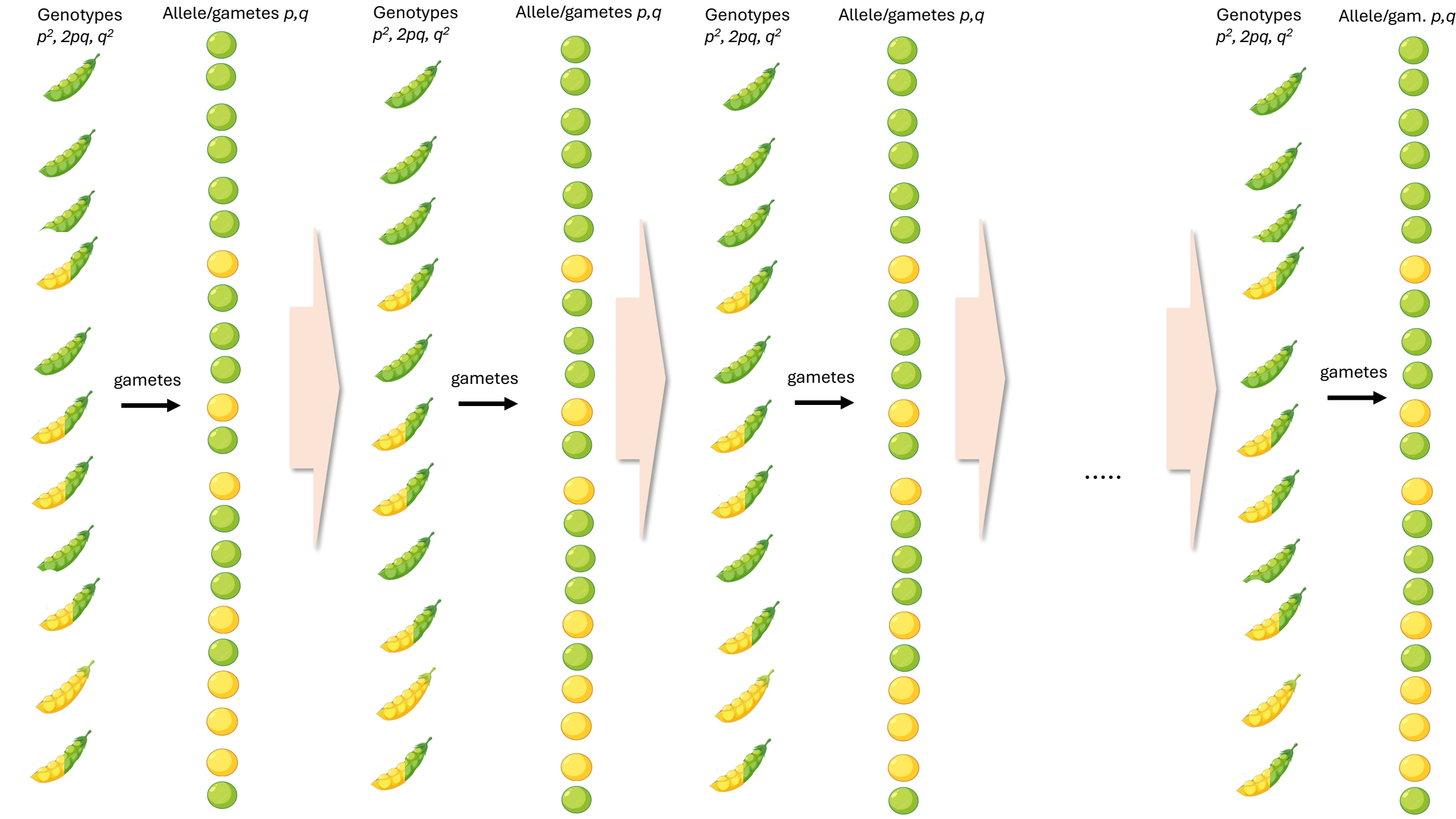

Wilhelm Weinberg
(1862-1937)

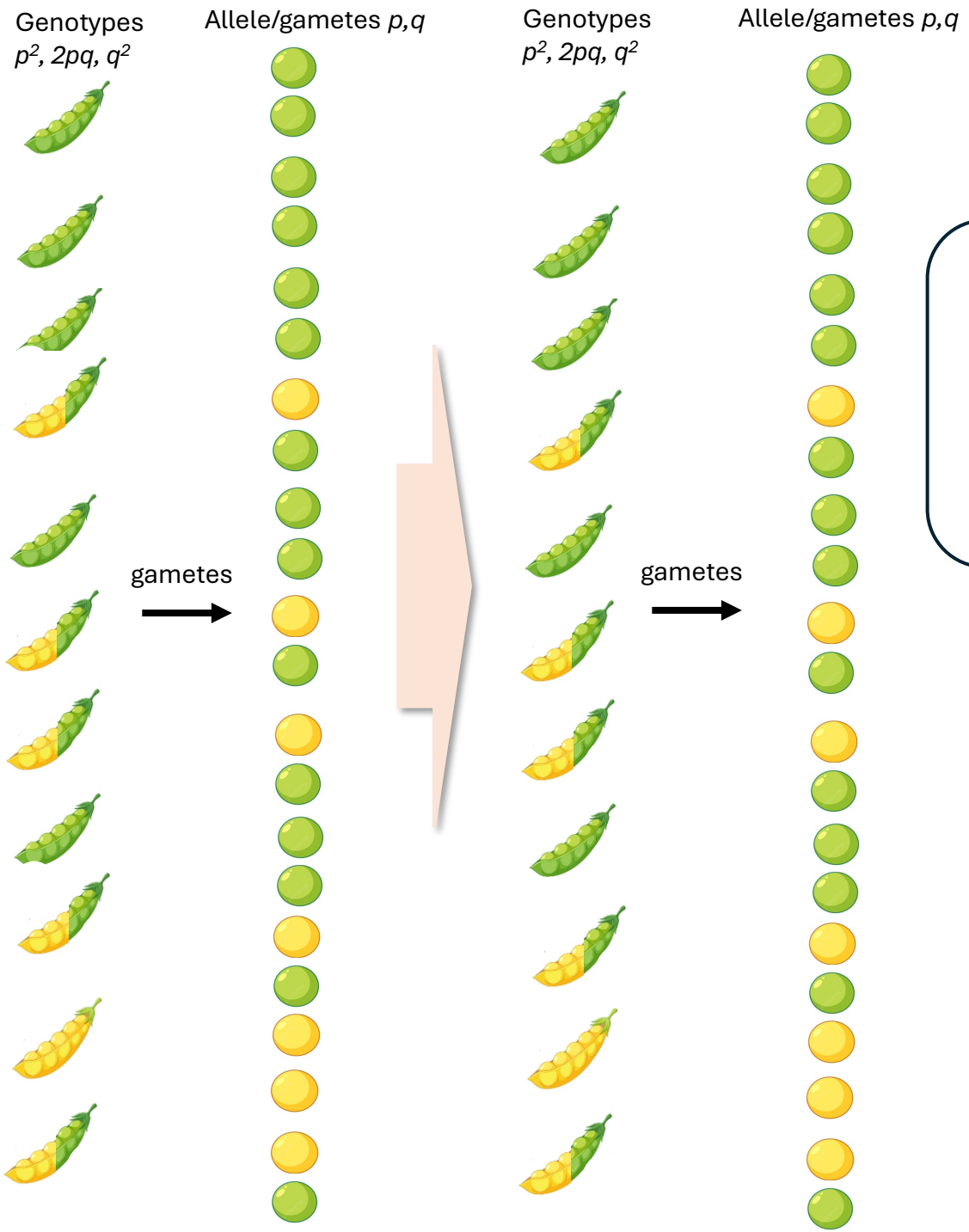

Genotypes


Genotypes

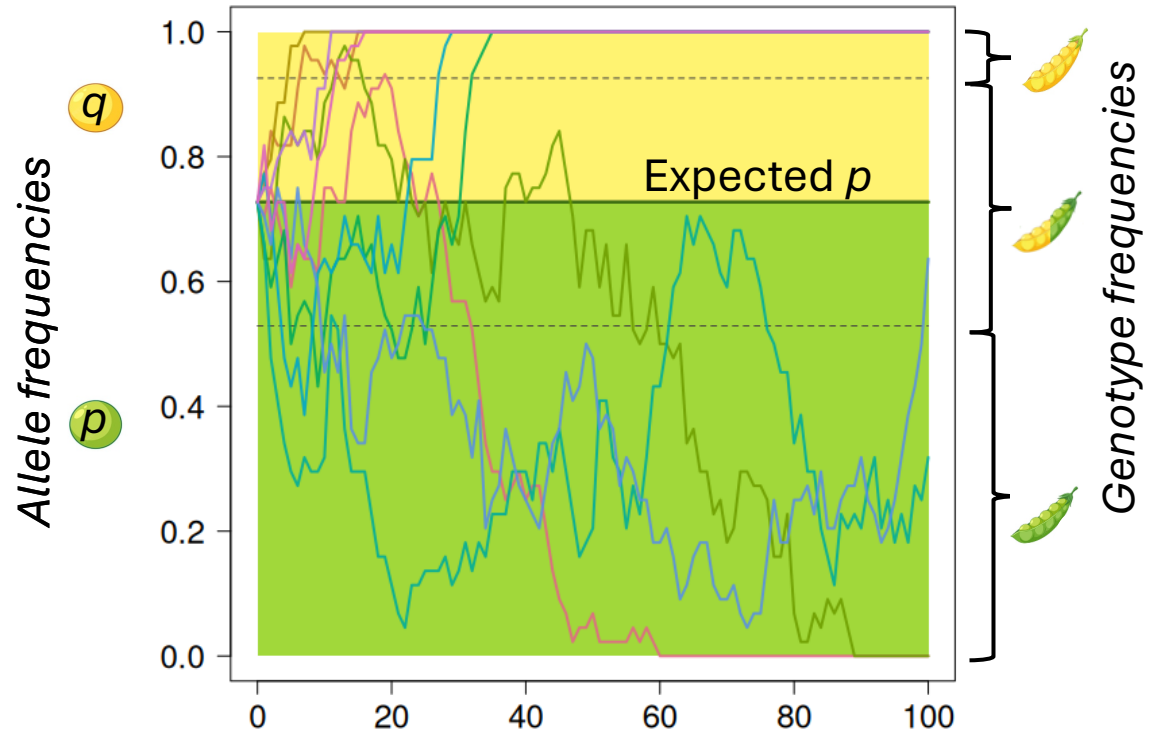


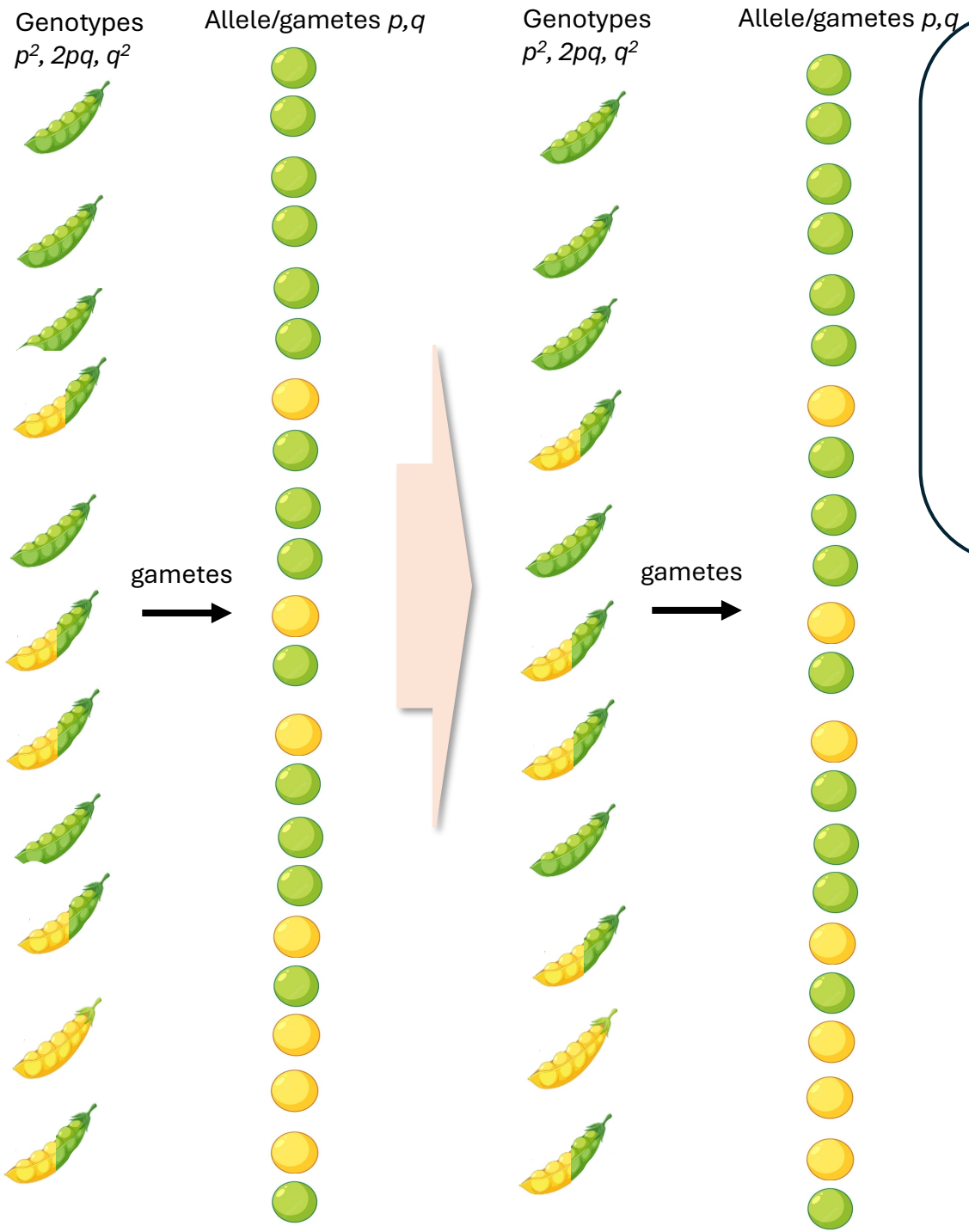
Frequency of (future) genotypes


Genotypes
p^2, 2pq, q^2
Allele/gametes p,q
gametes
Genotypes
p^2, 2pq, q^2
Allele/gametes p,q
gametes
Genotypes
p^2, 2pq, q^2
Allele/gametes p,q
gametes
.....
Genotypes
p^2, 2pq, q^2
Allele/gam. p,q
gametes



..but it ignores «sampling noise».

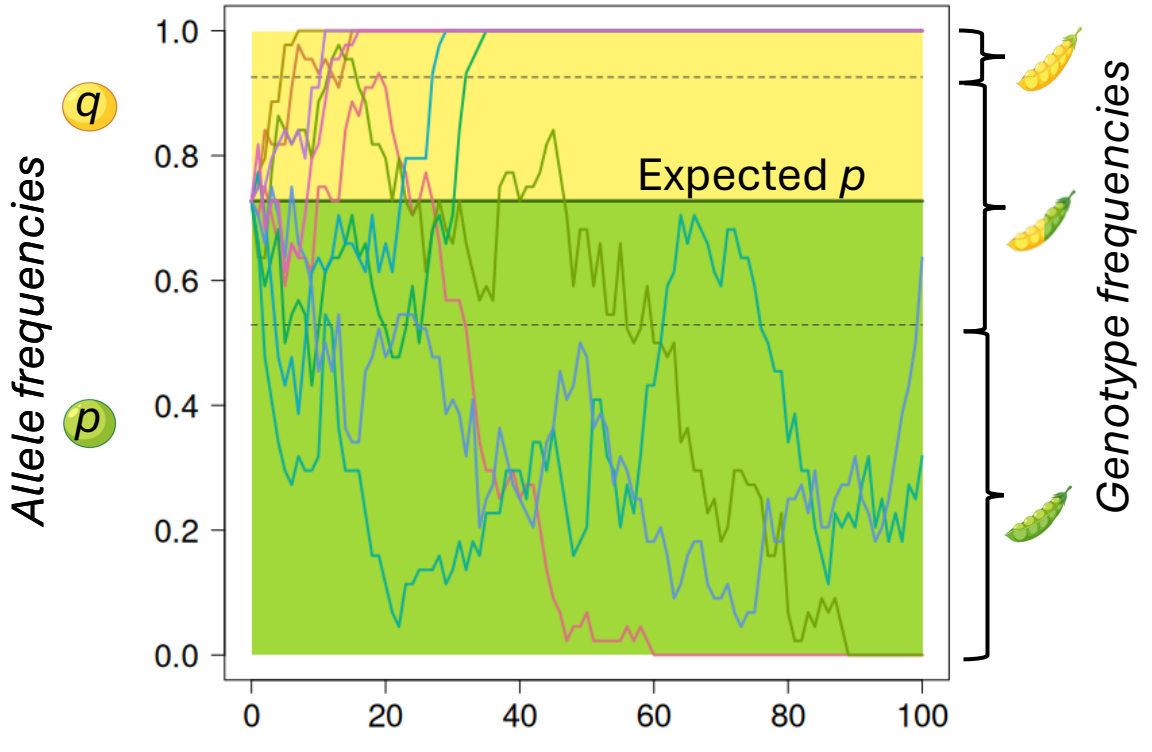




..but it ignores «sampling noise».

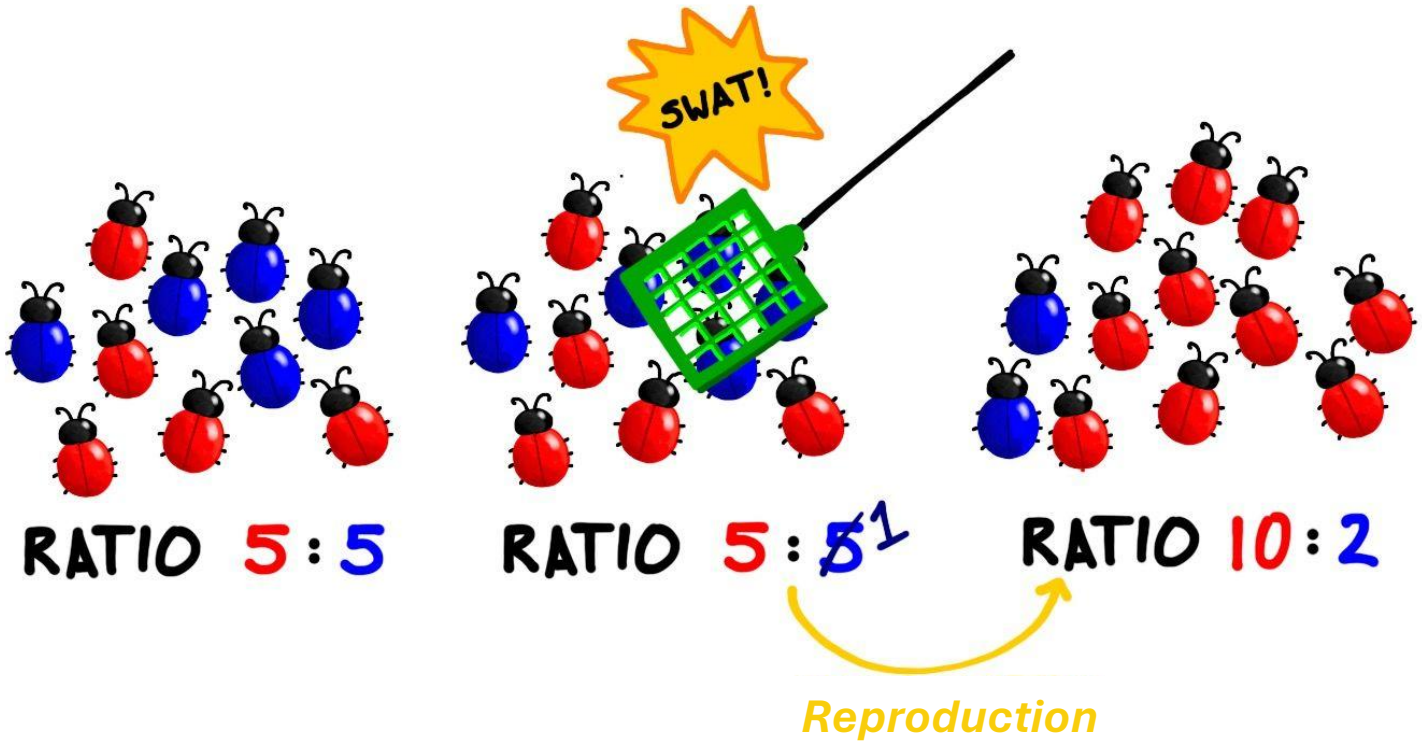
Each single population/evolutionary trajectory can be quite different.

Alleles can eventually go extinct (frequency 0) or go to fixation (frequency 1). These are *absorbing states*.



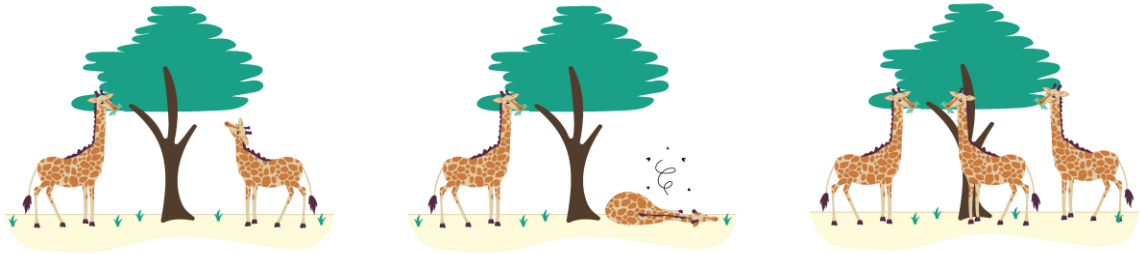
Genetic drift

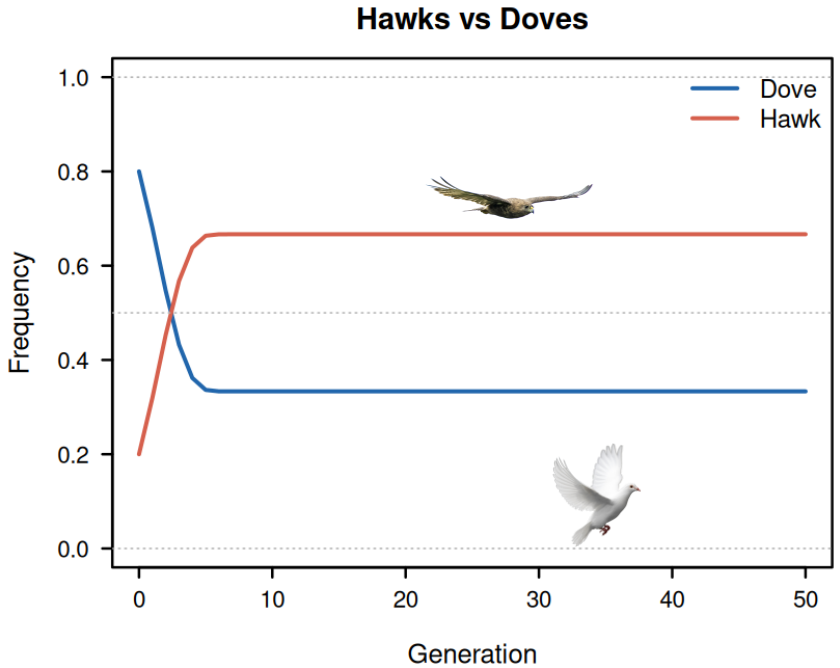
a change in allele frequencies due to random events



modified from https://i.pinimg.com

To build realistic models of evolution we need to include «***random drift***»





So far we have been neglecting it, *de facto* assuming «**infinite populations**» and thus *deterministic dynamics*.

The Wright-Fisher model

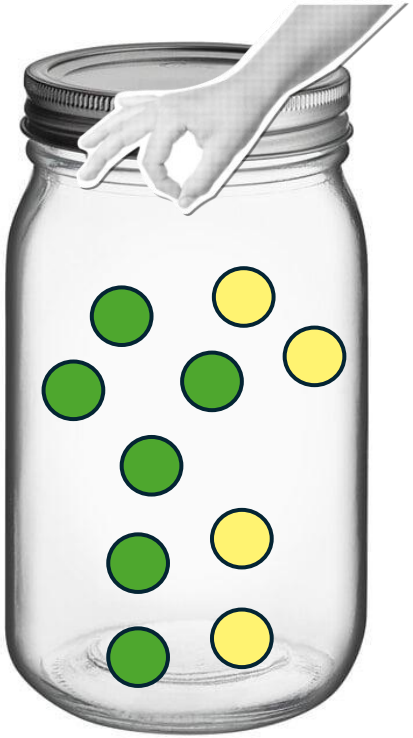
Let’s randomly sample gametes from a population of A a alleles

```
s = sample(c("A","A","A","a","A","a","A","A","a","a","A","a","A","A","A","A","A","a","A","A","A","A"), replace=T)
```

The Wright-Fisher model

Let's randomly sample gametes from a population of A a alleles

```
s = sample(c("A","A","A","a","A","a","A","A","a","a","A","a","A","A","A","A","A","a","A","A","A","A"), replace=T)
```



sampling with replacement: like sampling from an infinite jar (or placing back the marbles in the jar after having sampled them)

This is equivalent to
`sample(c("A","a"),prob=6/22,replace=T)`

The Wright-Fisher model

Generation 1

Let's randomly sample gametes from a population of A a alleles

16 6

```
s = sample(c("A","A","A","a","A","a","A","A","a","a","A","a","A","A","A","A","A","a","A","A","A","A"), replace=T)
```

Generation 2

19 3

```
s = sample(c("A","A","A","a","A","a","A","A","A","a","A","A","A","A","A","A","A","A","A","A","A","A"), replace=T)
```

The Wright-Fisher model

Generation 1

Let's randomly sample gametes from a population of A a alleles

16 6

```
s = sample(c("A","A","A","a","A","a","A","A","a","a","A","a","A","A","A","A","A","a","A","A","A","A"), replace=T)
```

Generation 2

19 3

```
s = sample(c("A","A","A","a","A","a","A","A","A","a","A","A","A","A","A","A","A","A","A","A","A"), replace=T)
```

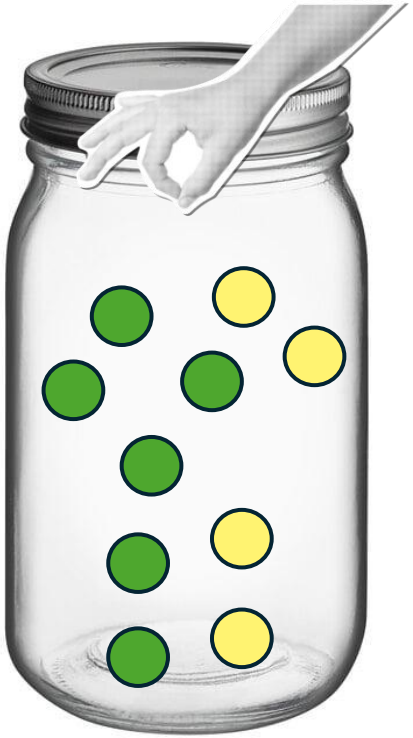
Generation 3

21 1

The Wright-Fisher model

Let's randomly sample gametes from a population of A a alleles

```
s = sample(c("A","A","A","a","A","a","A","A","a","a","A","a","A","A","A","A","A","a","A","A","A","A"), replace=T)
```

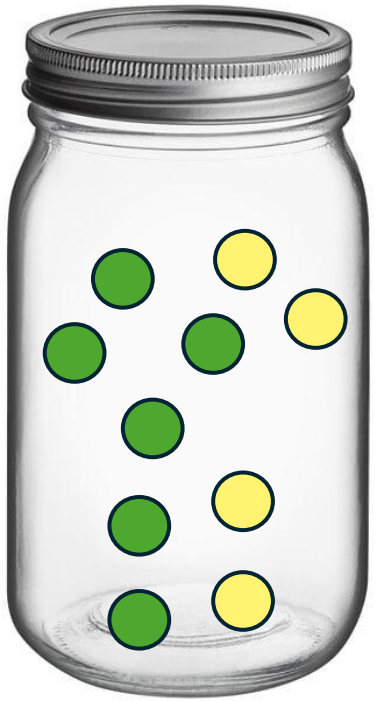


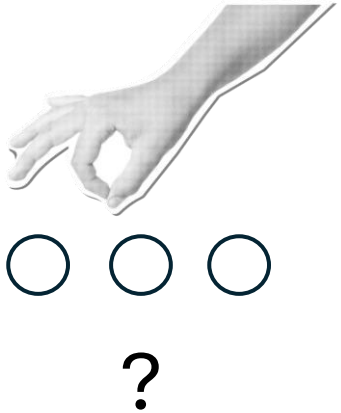
sampling with replacement: like sampling from an infinite jar (or placing back the marbles in the jar after having sampled them)

This is equivalent to
`sample(c("A","a"),prob=6/22,replace=T)`

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?





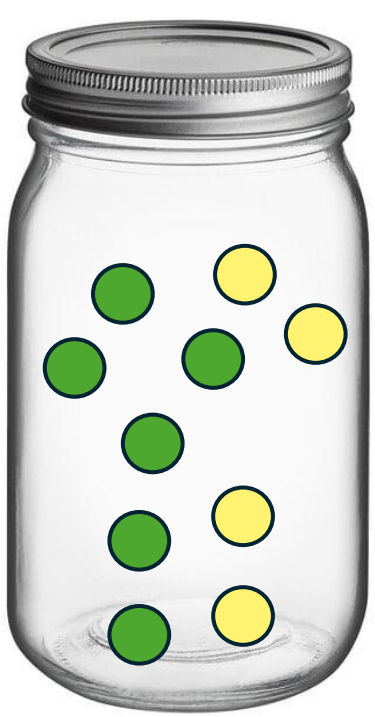
?

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

What are these two terms?

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?



Order 1

Order 2

Order 3

There can be more than one order to sample the colored marbles!

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

The $\binom{n}{k}=n!/(k!(n-k)!)$ is called «binomial coefficient» and describes the number of possible combinations of sampled marbles.

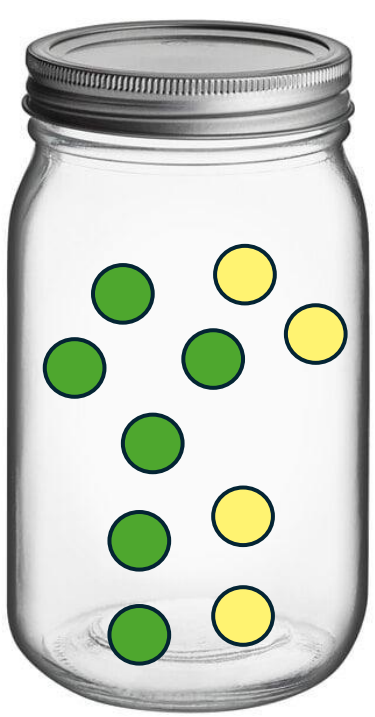
The factorial $n!=n(n-1)(n-2)....1$

e.g.

$\binom{3}{2} = 3!/(2!1!)=3*2/2=3$

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?



Order 1

Order 2

Order 3

Note that all order have the same proportions of green, so they have all the same probability!

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

The $\binom{n}{k}=n!/(k!(n-k)!)$ is called «binomial coefficient» and describes the number of possible combinations of sampled marbles.

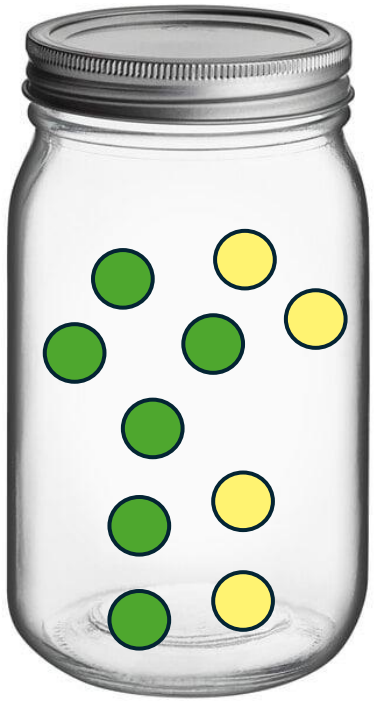
The factorial $n!=n(n-1)(n-2)....1$

e.g.

$\binom{3}{2} = 3!/(2!1!)=3*2/2=3$

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?



Order 1 0.6*0.6*0.4

Order 2 0.6*0.4*0.6

Order 3 0.4*0.6*0.6

Note that all order have the same proportions of green, so they have all the same probability!

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

Potential order of sampled marbles

Probability of each order

$$Pr(2)=3*0.6^2\ 0.4=0.432$$

The Wright-Fisher model

Generation 1

Let's randomly sample gametes from a population of A a alleles

16 6

```
s = sample(c("A","A","A","a","A","a","A","A","a","a","A","a","A","A","A","A","A","a","A","A","A","A"), replace=T)
```

$$Pr(3)=\binom{22}{3} 6/22^3 (1-6/22)^{19}$$

Generation 2

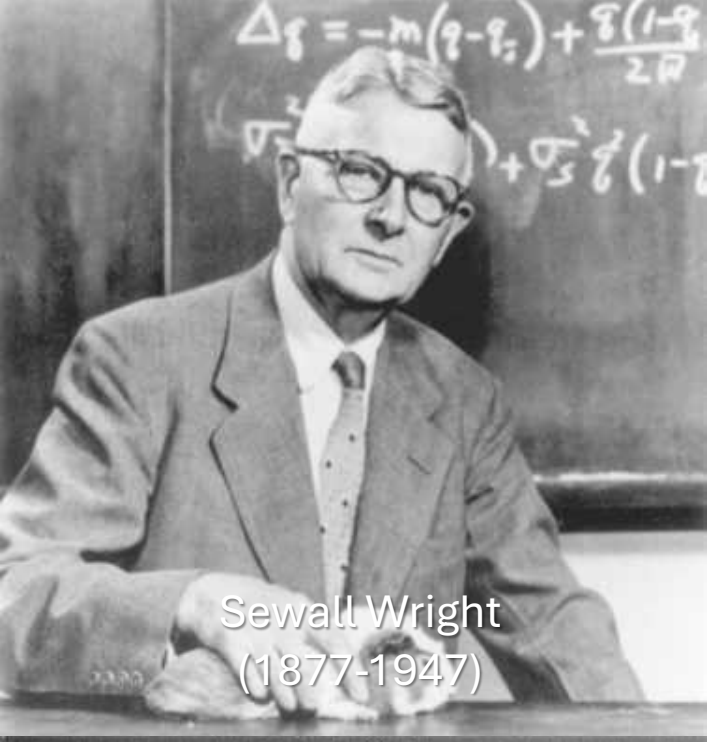
19 3

```
s = sample(c("A","A","A","a","A","a","A","A","A","a","A","A","A","A","A","A","A","A","A","A","A","A"), replace=T)
```

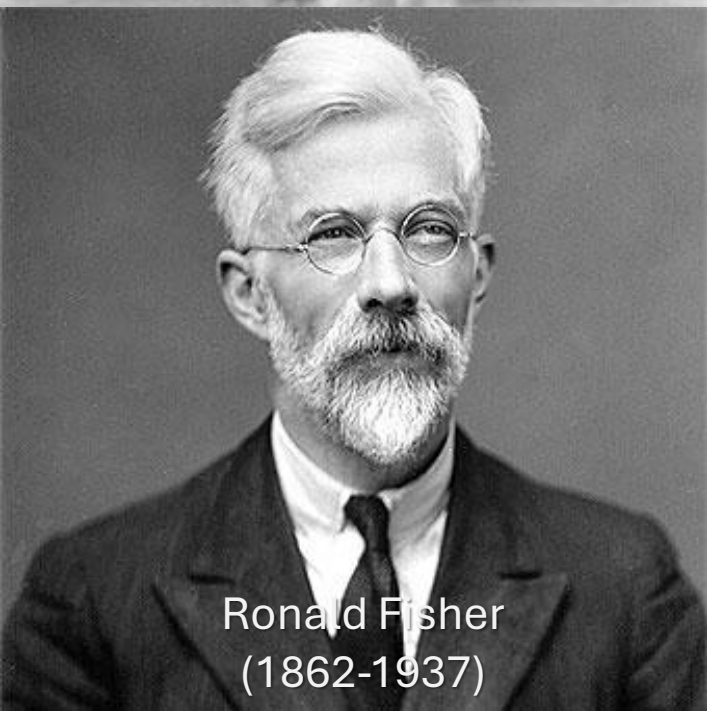
$$Pr(1)=\binom{22}{1} p(1-p)^{21}$$

Generation 3

21 1

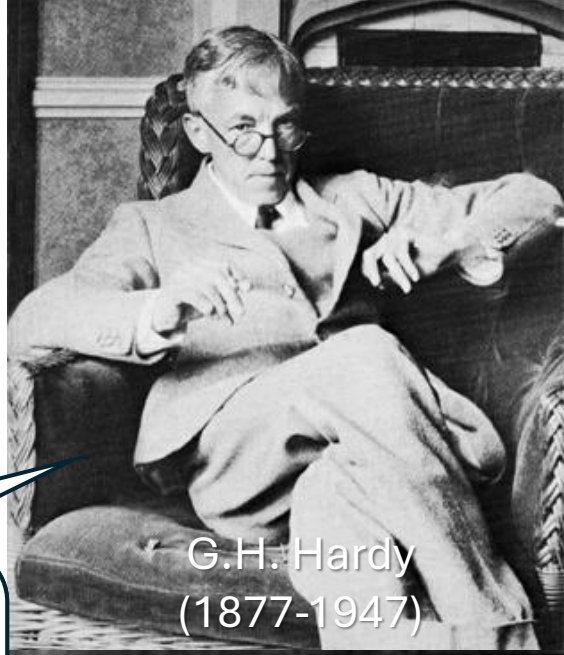


Sewall Wright
(1877-1947)



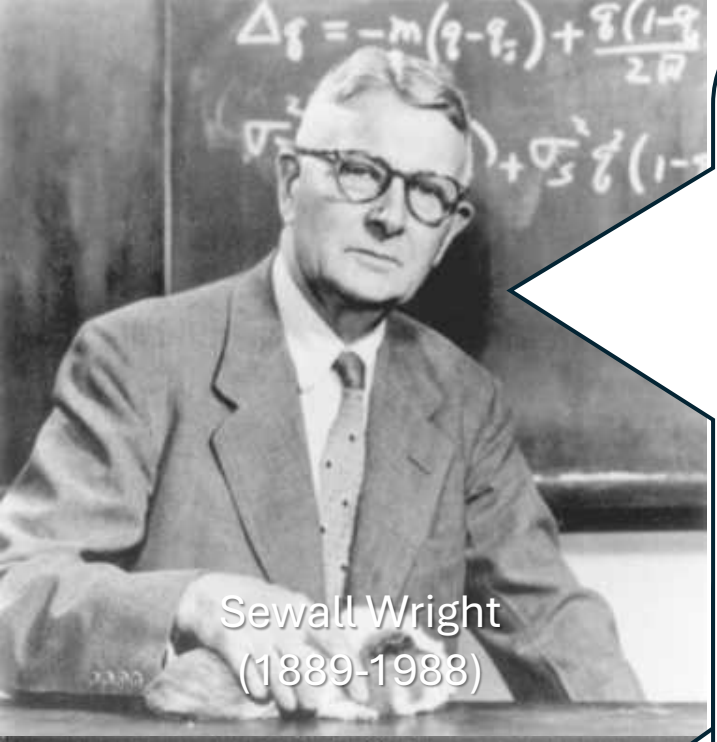
Ronald Fisher
(1862-1937)

Infinite populations remain in Hardy-Weinberg equilibrium...

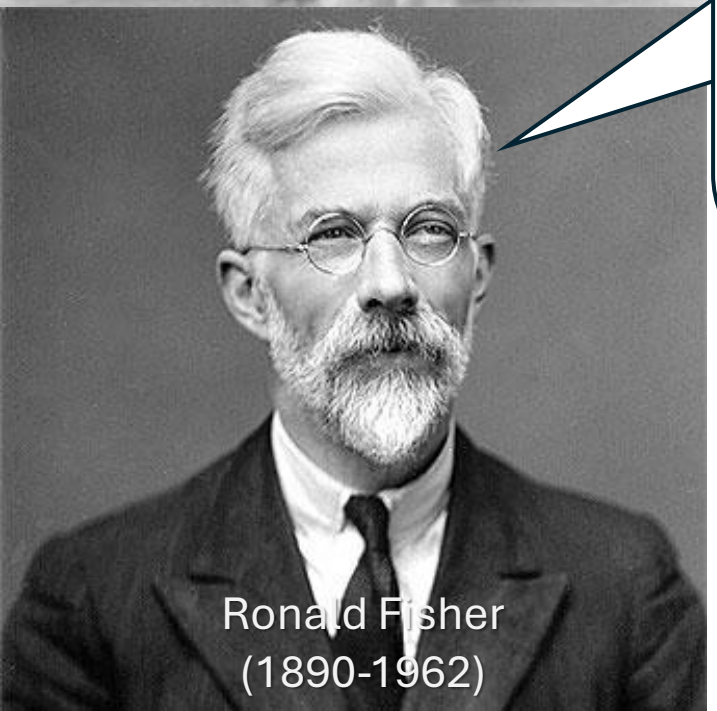


G.H. Hardy
(1877-1947)

Wilhelm Weinberg
(1862-1937)



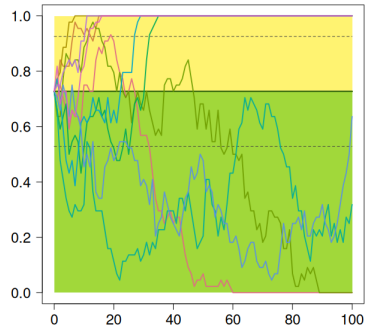
Sewall Wright
(1889-1988)



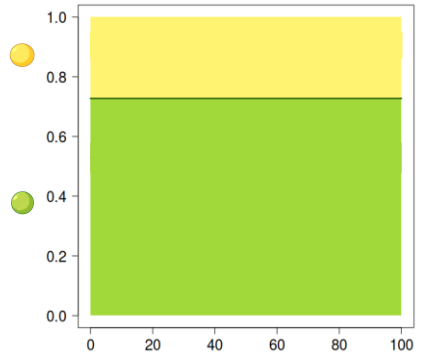
Ronald Fisher
(1890-1962)

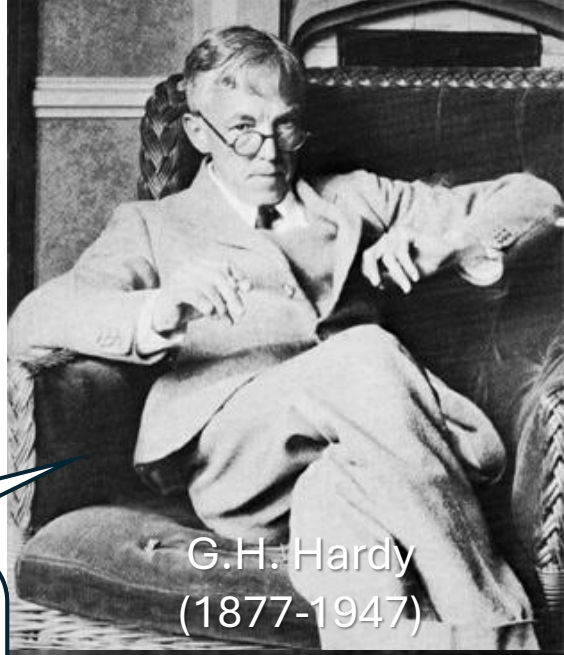
But in real populations allele frequencies **fluctuates because of random genetic drift**.

And now you can **simulate them** (just sample with replacement) and **calculate the probabilities** that they evolve along certain trajectories!



Infinite populations remain in Hardy-Weinberg equilibrium...





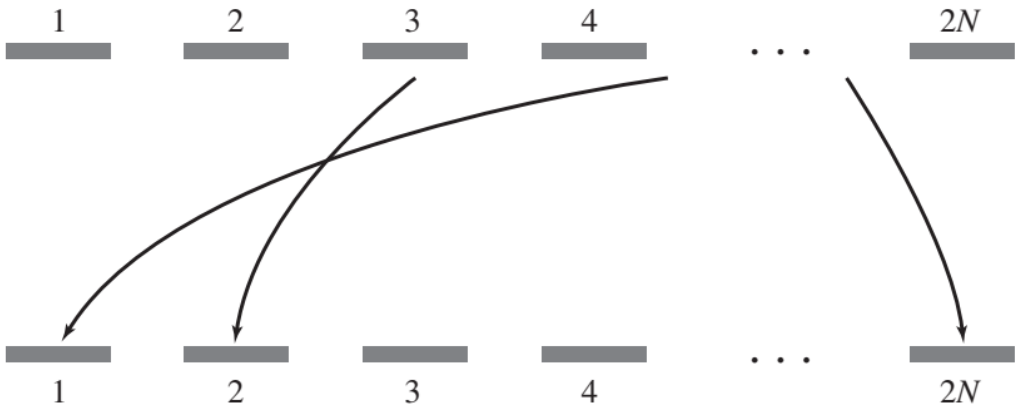
G.H. Hardy
(1877-1947)

Wilhelm Weinberg
(1862-1937)

Let’s first explore a Wright-Fisher model with simple assumptions:

- Constant population size 2N (either a haploid population with 2N individuals or a diploid with N individuals and 2N chromosomes – and even sex-ratio)
- No natural selection/alleles are all the same
- One single panmictic population

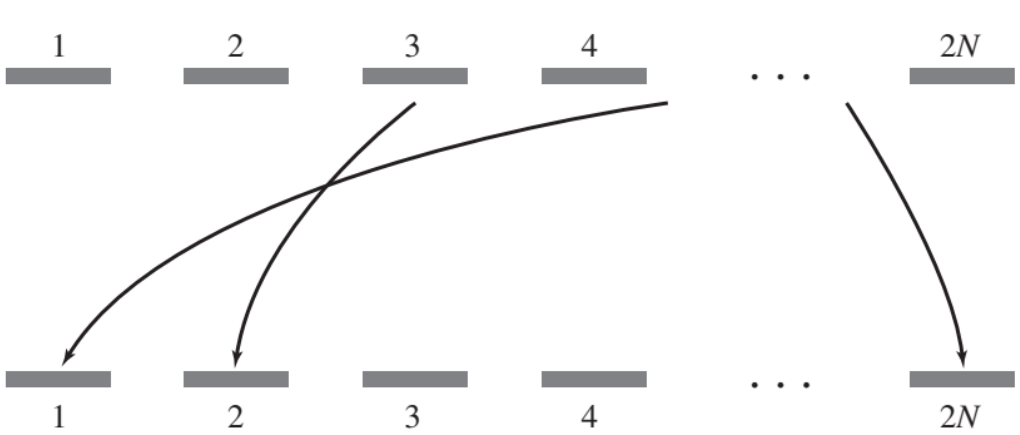
Haploid Wright-Fisher

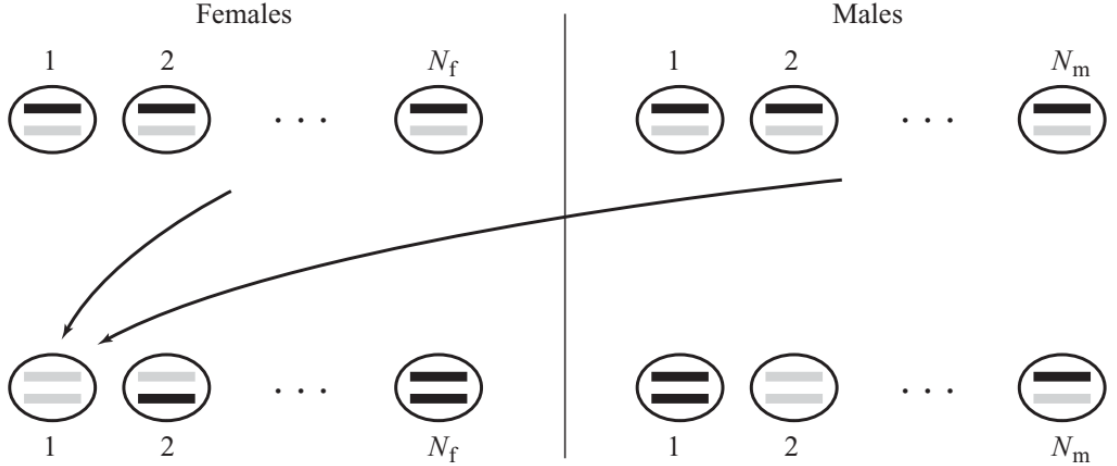


Let’s first explore a Wright-Fisher model with simple assumptions:

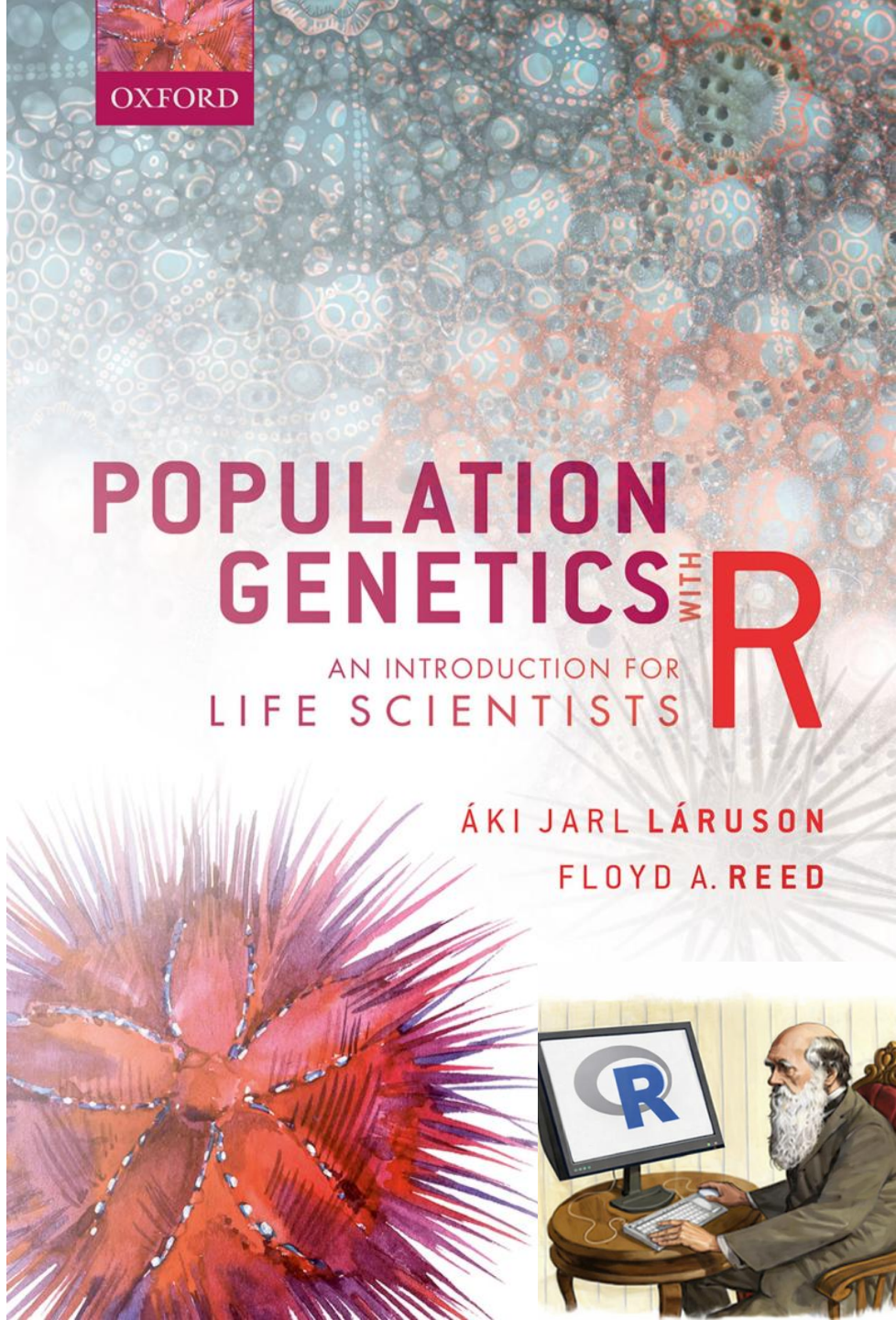
- Constant population size 2N (either a haploid population with 2N individuals or a diploid with N individuals and 2N chromosomes – and even sex-ratio)
- No natural selection/alleles are all the same
- One single panmictic population

Haploid Wright-Fisher ← when Nf=Nm they concide → Diploid Wright-Fisher





If you want to play interactively with random genetic drift visit: https://chrisnajman.github.io/genetic-drift/



```
> init_p <- 0.25 #Initial allele frequency
> gen <- 100 #Number of generations
> reps <- 500 #Lots of replicates to run
> colors <- rainbow(reps) #Grab some colors for our reps
> N <- 100 #Population size

> #Initialize a plot
> plot(x=NULL, y=NULL, xlim=c(1, gen), ylim=c(0,1),
    xlab="Generations", ylab="Allele frequency")

> Freq <- NULL #Create an object to save each replicates output

> #Iterate through the replicates
> for(i in 1:reps){
    p <- init_p
    for(j in 1:(gen-1)){
        a <- rbinom(n=1,size=2*N,prob=p[j])
        f <- a/(2*N)
        p <- c(p,f)
        }
    Freq <- rbind(Freq, p) #Save p
    lines(x=1:gen, y=p, lwd=2, col=colors[i])
}

> #Add the mean of all the replicates to the plot
> lines(1:gen, colMeans(Freq), lwd=2, col="black")
```

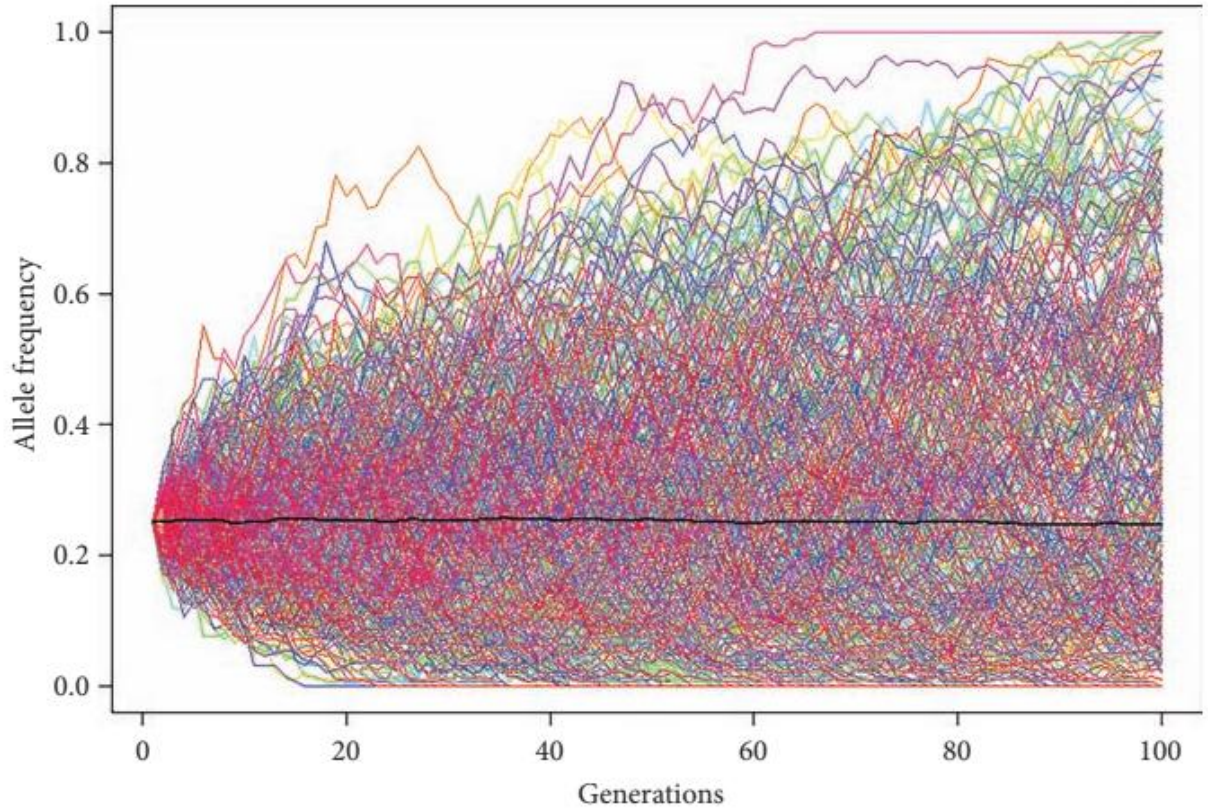


Figure 6.7 Five hundred replicates of random drift in a population of 100 individuals, starting with an allele frequency of 0.25. The average allele frequency across all replicates is plotted in black and stays close to the initial allele frequency.

```
> init_p <- 0.25 #Initial allele frequency
> gen <- 100 #Number of generations
> reps <- 500 #Lots of replicates to run
> colors <- rainbow(reps) #Grab some colors for our reps
> N <- 100 #Population size

> #Initialize a plot
> plot(x=NULL, y=NULL, xlim=c(1, gen), ylim=c(0,1),
    xlab="Generations", ylab="Allele frequency")

> Freq <- NULL #Create an object to save each replicates output

> #Iterate through the replicates
> for(i in 1:reps){
    p <- init_p
    for(j in 1:(gen-1)){
        a <- rbinom(n=1,size=2*N,prob=p[j])
        f <- a/(2*N)
        p <- c(p,f)
        }
    Freq <- rbind(Freq, p) #Save p
    lines(x=1:gen, y=p, lwd=2, col=colors[i])
}

> #Add the mean of all the replicates to the plot
> lines(1:gen, colMeans(Freq), lwd=2, col="black")
```

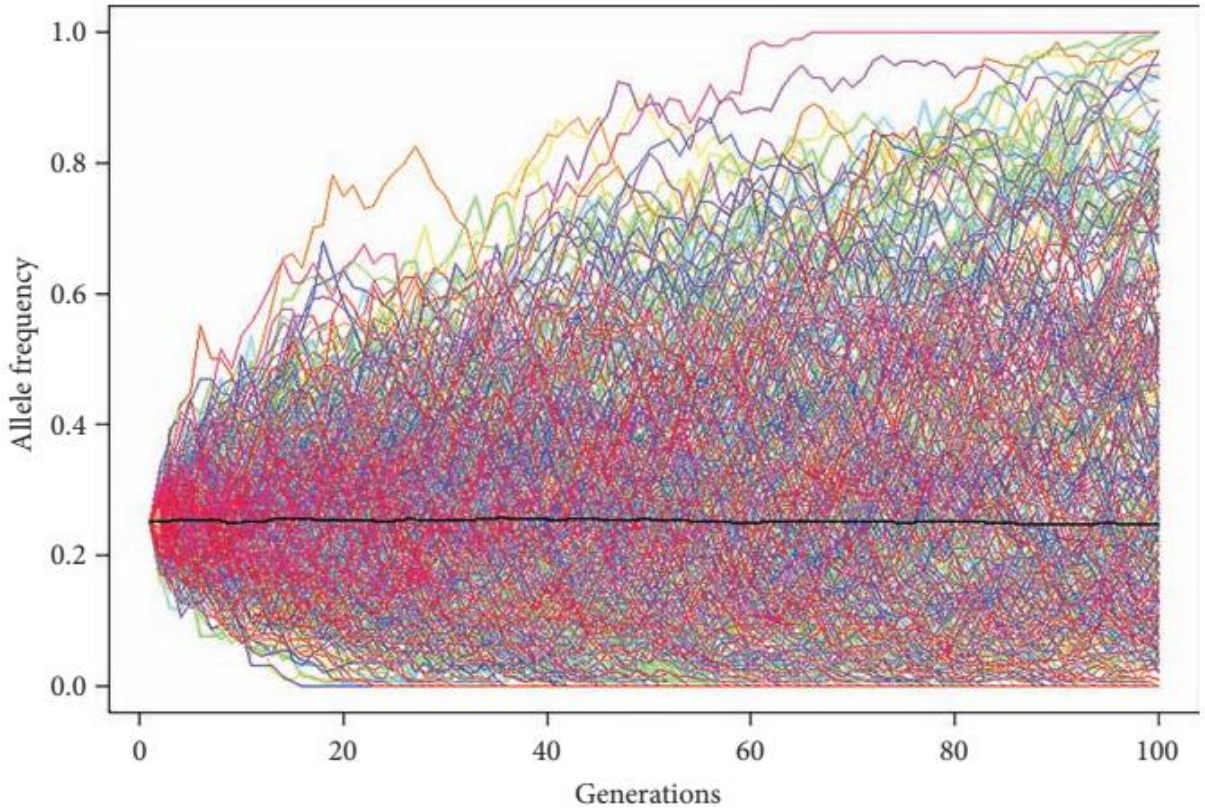
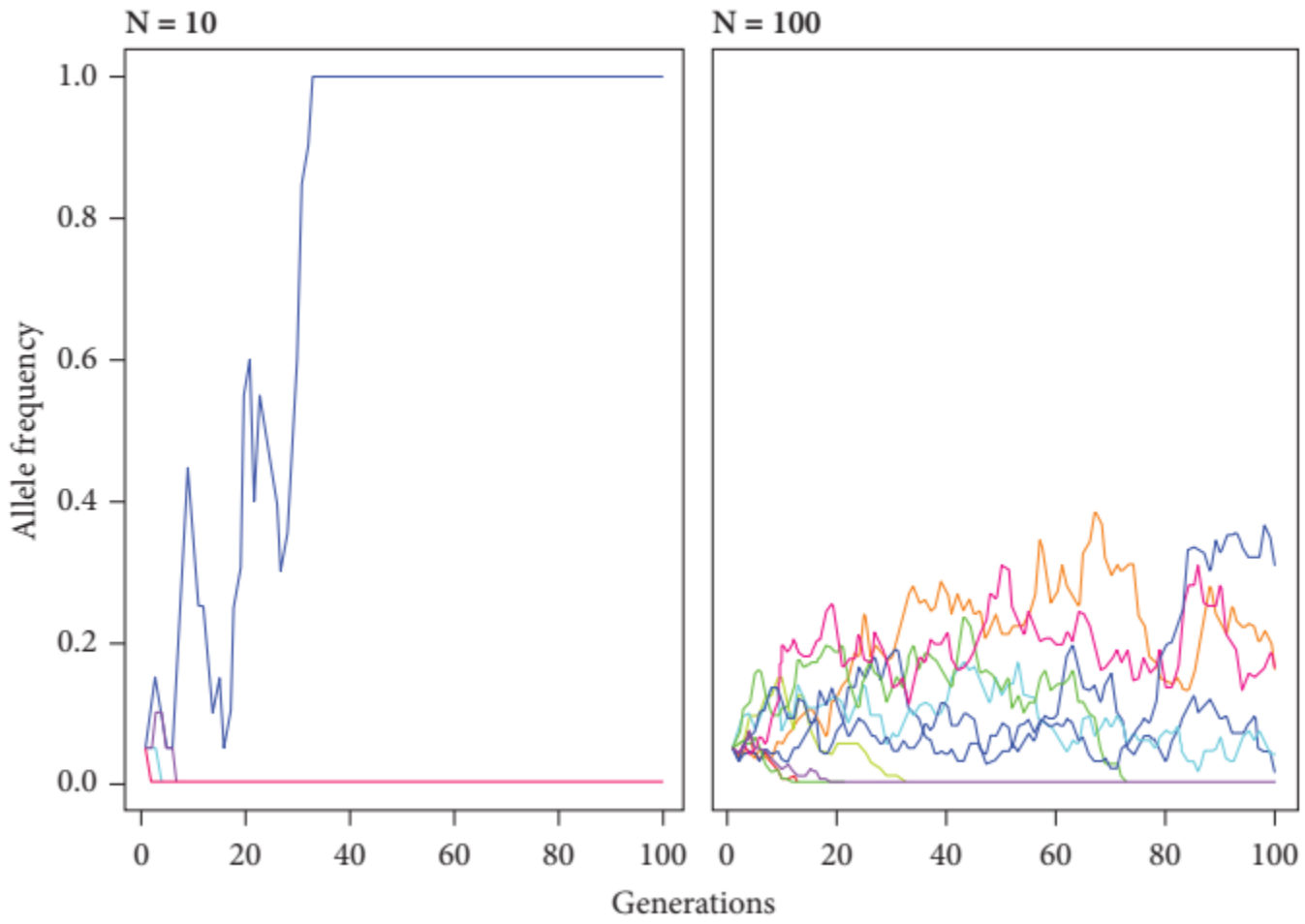


Figure 6.7 Five hundred replicates of random drift in a population of 100 individuals, starting with an allele frequency of 0.25. The average allele frequency across all replicates is plotted in black and stays close to the initial allele frequency.

Do you think that the average frequency will remain about 0.25 forever?

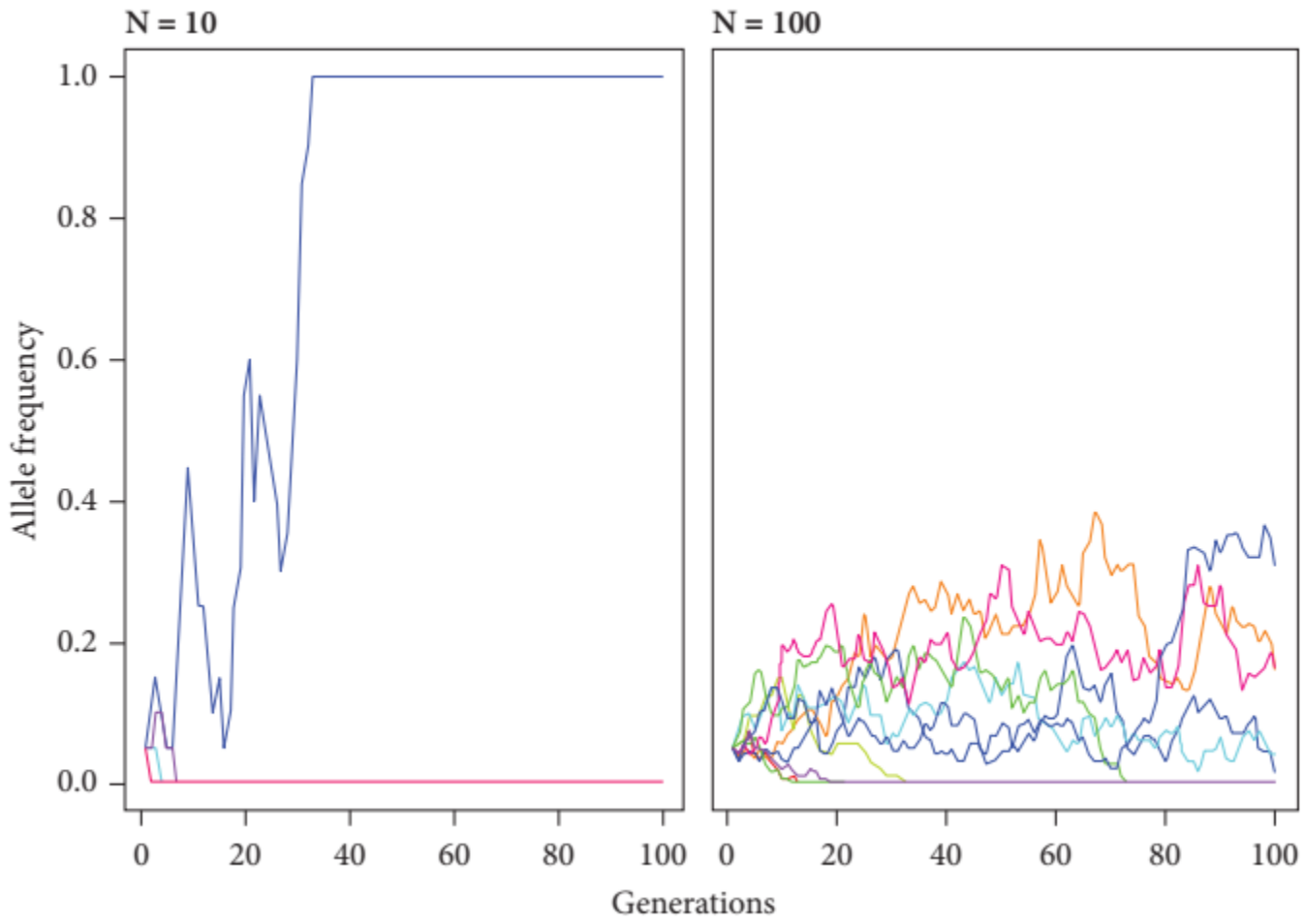
What happens to alleles on the long-term?

```
> init_p <- 0.05 #Initial allele frequency

> gen <- 100 #Number of generations

> reps <- 10 #How many replicates to run
> colors <- rainbow(reps) #Grab some colors for our reps.

> N <- 10 #Set the population size

> #Initialize a plot that we can add lines to later
> plot(x=NULL, y=NULL, xlim=c(1, gen), ylim=c(0,1),
    xlab="Generations", ylab="Allele frequency")

> #For each replicate: draw new copy numbers of the allele
    per generation, then re-set the allele frequency 'p'
    and use that frequency for the next random draw
> for(i in 1:reps){
    p <- init_p
    for(j in 1:(gen-1)){
        a <- rbinom(n=1,size=2*N,prob=p[j])
        f <- a/(2*N)
        p <- c(p,f)
        }
    lines(x=1:gen, y=p, lwd=2, col=colors[i])
}
```

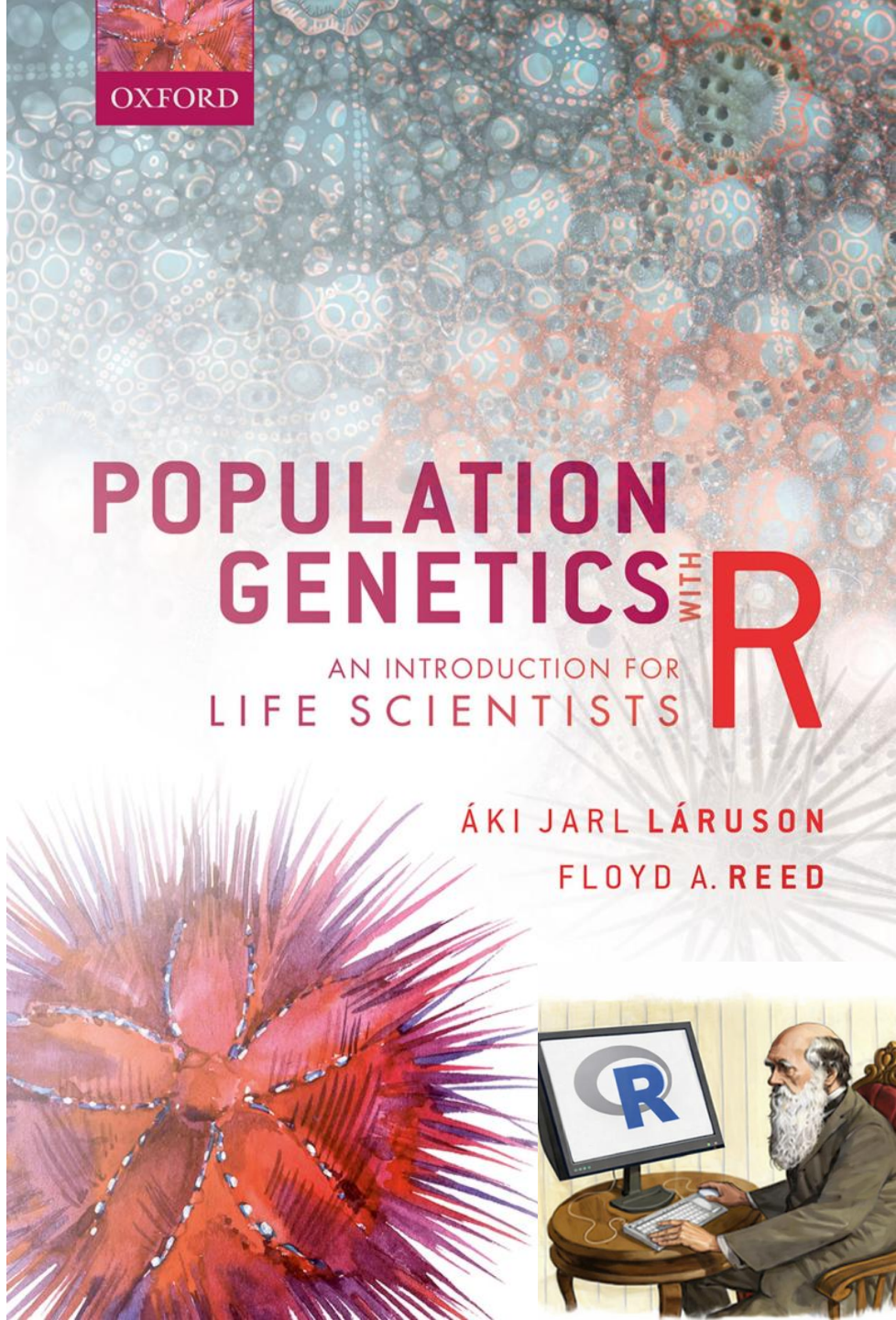


Alleles can reach **extinction** (frequency 0%) or **fixation** (frequency 100%) just by chance

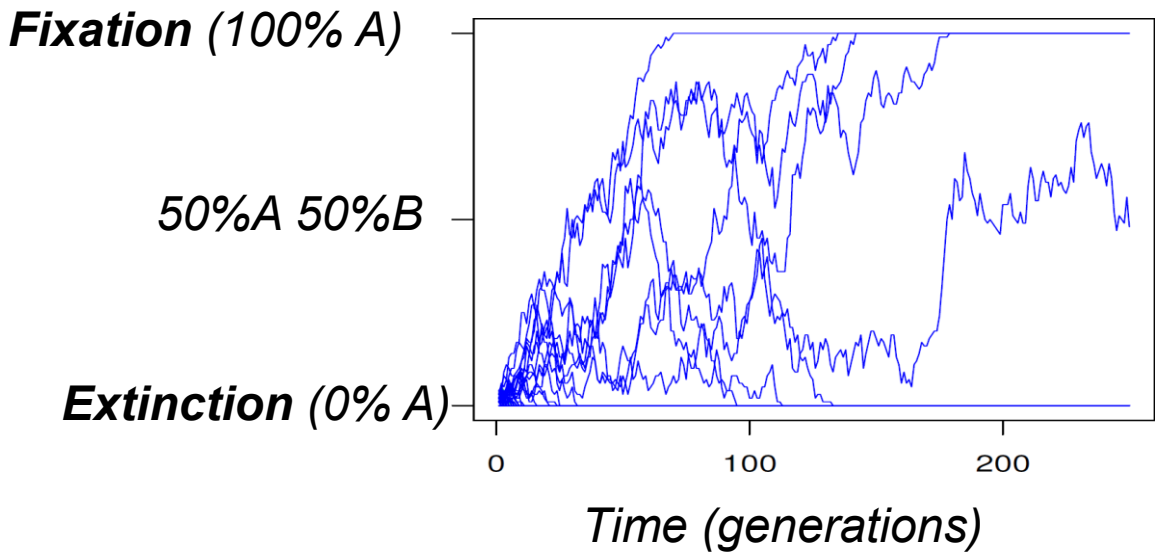
```
> init_p <- 0.05 #Initial allele frequency

> gen <- 100 #Number of generations

> reps <- 10 #How many replicates to run
> colors <- rainbow(reps) #Grab some colors for our reps.

> N <- 10 #Set the population size

> #Initialize a plot that we can add lines to later
> plot(x=NULL, y=NULL, xlim=c(1, gen), ylim=c(0,1),
    xlab="Generations", ylab="Allele frequency")

> #For each replicate: draw new copy numbers of the allele
    per generation, then re-set the allele frequency 'p'
    and use that frequency for the next random draw
> for(i in 1:reps){
    p <- init_p
    for(j in 1:(gen-1)){
        a <- rbinom(n=1,size=2*N,prob=p[j])
        f <- a/(2*N)
        p <- c(p,f)
        }
    lines(x=1:gen, y=p, lwd=2, col=colors[i])
}
```



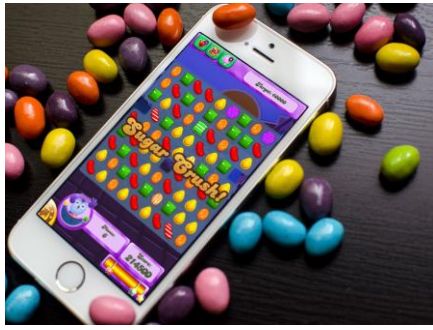
If you want to play interactively with random genetic drift visit: https://chrisnajman.github.io/genetic-drift/



Alleles can reach **extinction** (frequency 0%) or **fixation** (frequency 100%) just by chance

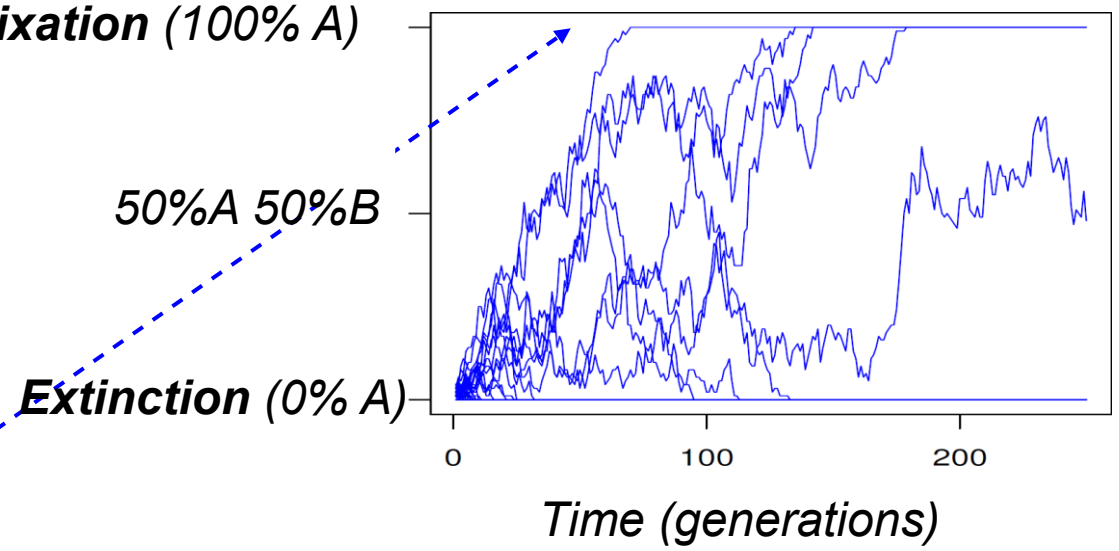


If you want to play interactively with random genetic drift visit: https://chrisnajman.github.io/genetic-drift/



Alleles can reach **extinction** (frequency 0%) or **fixation** (frequency 100%) just by chance

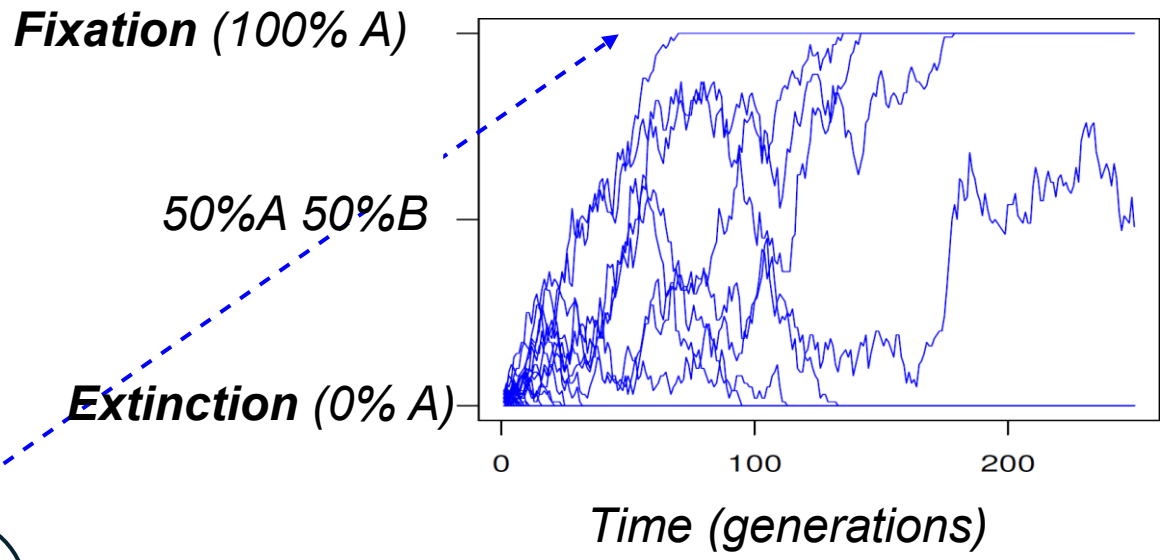
What is the fixation probability of a new mutation (initial frequency 1/2N) in a diploid popolazione of N individuals (or an haploid population of 2N)?



- ☐ *1/2N*
- ☐ *1/N*
- ☐ *1%*

The **fixation probability** of a new neutral mutation is **1/2N**

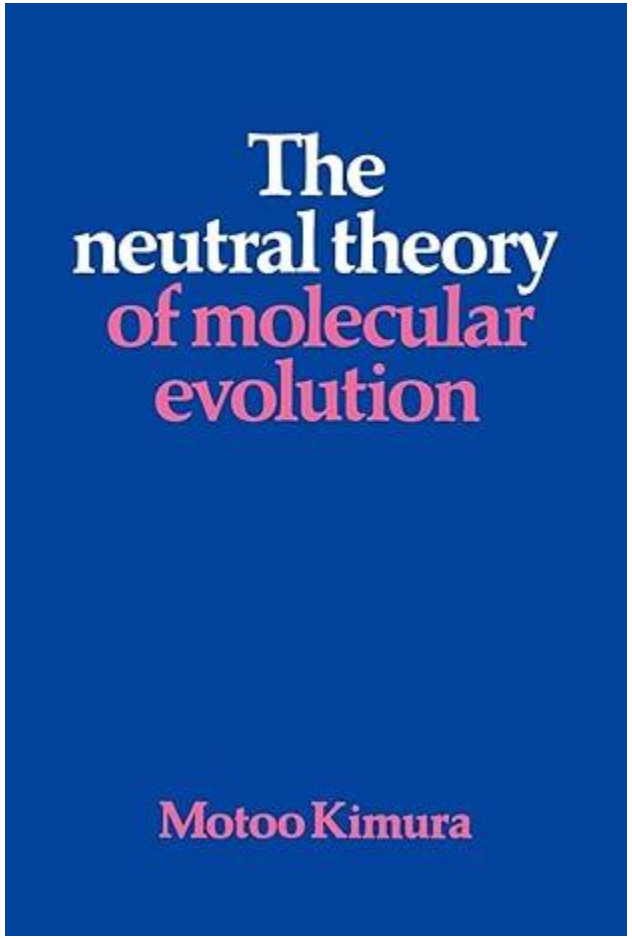
Every one of the 2N alleles has the same chance to reach fixation!



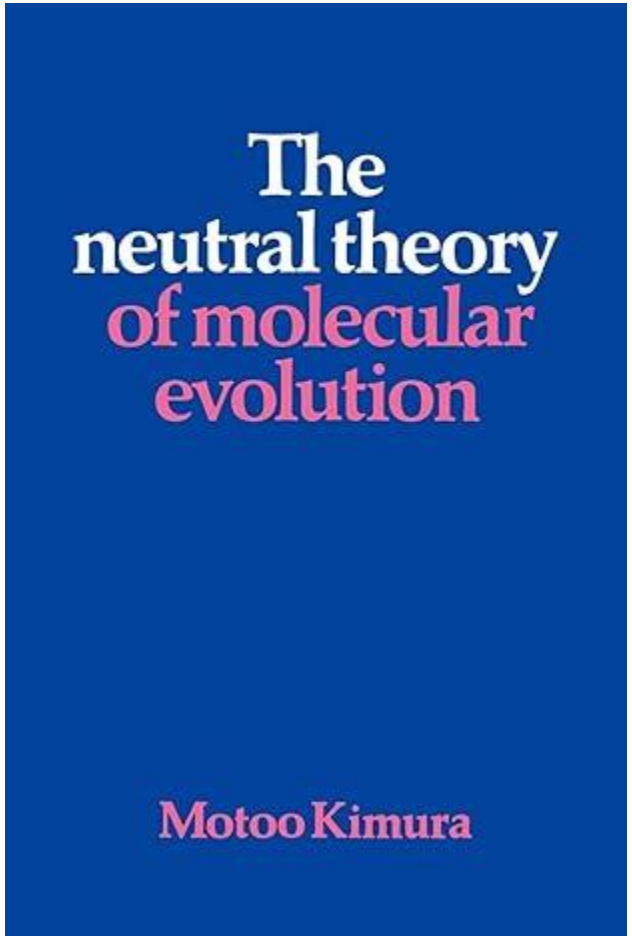
- [x] *1/2N*
- [ ] *1/N*
- [ ] *1%*

The Neutral Theory

- Most mutation in the genome are neutral



The Neutral Theory



•Most mutation in the genome are neutral

•Since the probability to reach fixation for all of them is 1/2N and new mutations arise with 2Nμ:

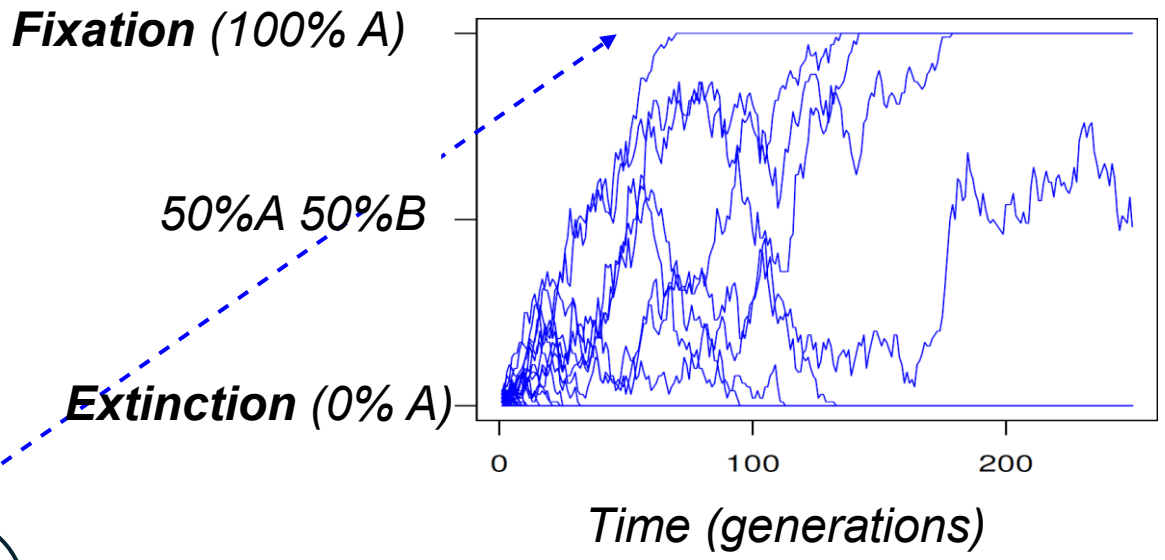
Divergence between two species = 2Nμ***1/2N*** = μ

This is the basic of phylogenetic and the molecular clock!

μ=mutation rate

Extinction probability

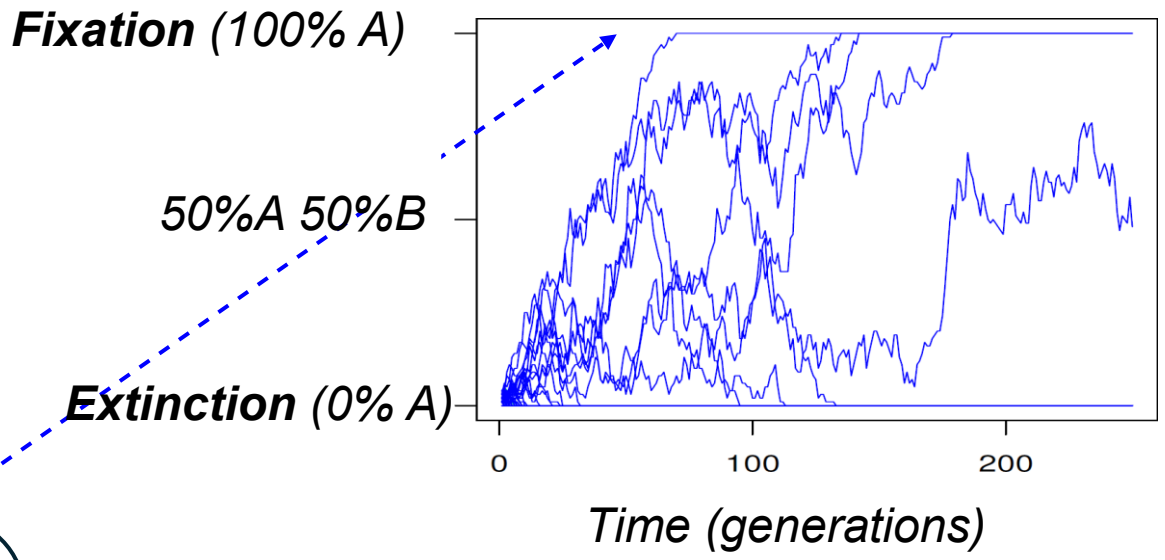
What is then the extinction probability of new mutation?



- ☐ *1/2N*
- ☐ *50%*
- ☐ *1-1/2N*

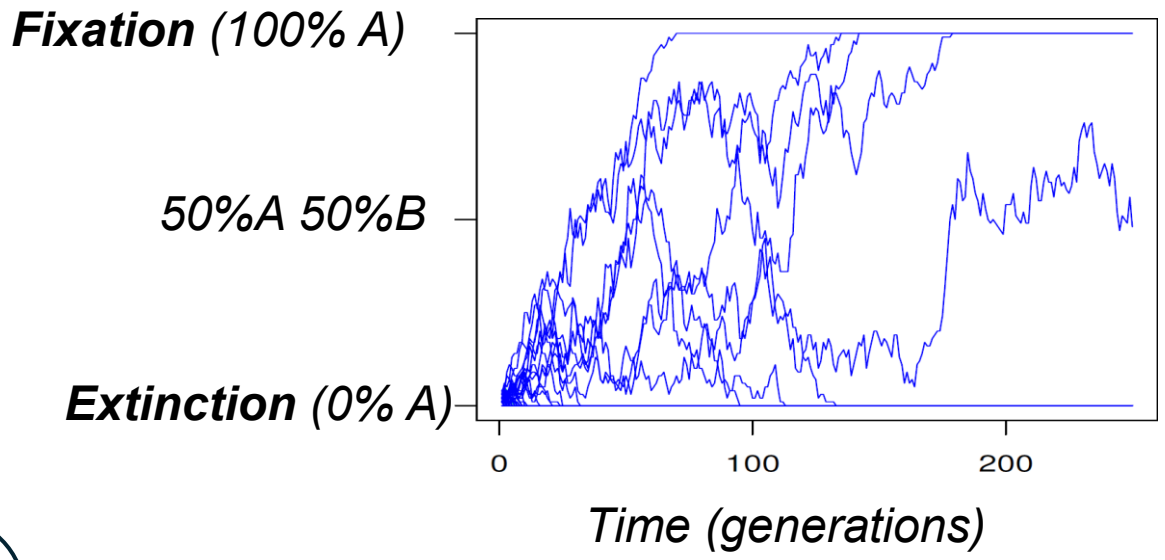
Most new mutations are eventually lost!

If the allele does not reach fixation (probability 1/2N) eventually it will be lost.



- ☐ *1/2N*
- ☐ *50%*
- ☑ *1-1/2N*

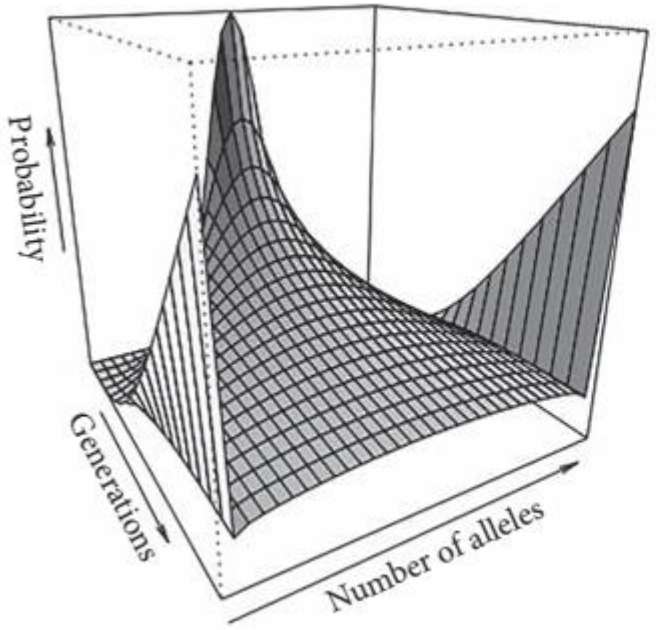
And what is the fixation probability for an allele with 50% initial frequency?



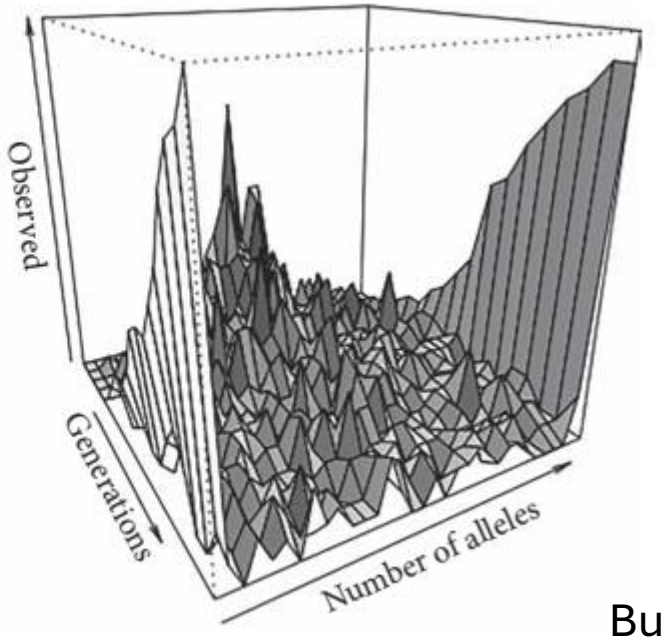
- ☐ *1/2N*
- ☐ *50%*
- ☐ *1-1/2N*

And what is the fixation probability for an allele with 50% initial frequency?

Wright-Fisher expectation



Real *Drosophila* population (*bw*^{75} allele)

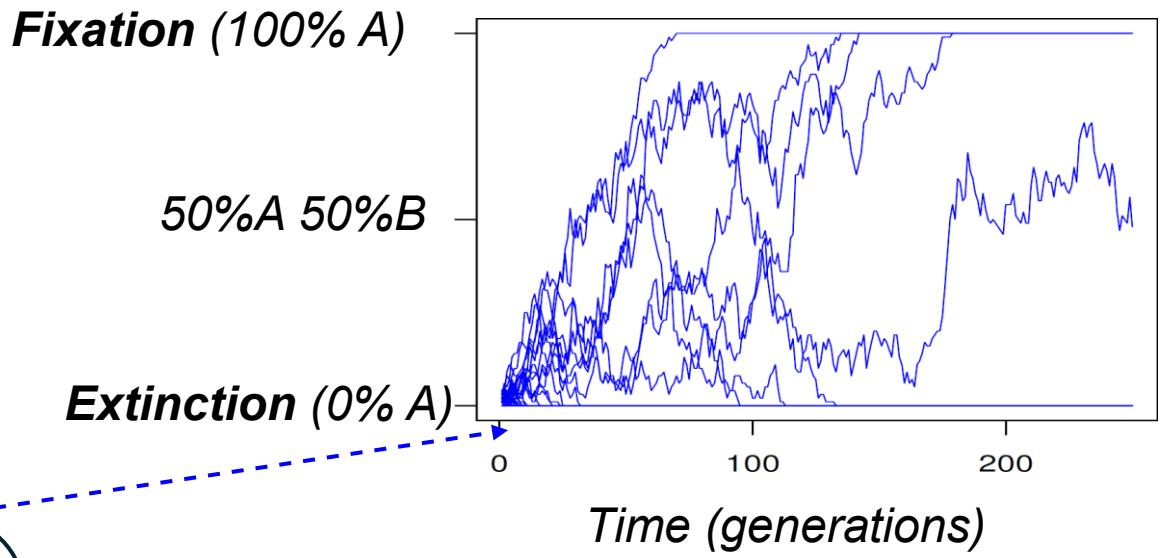


Buri et al.,1956

- ☐ *1/2N*
- ☑ *50%*
- ☐ *1-1/2N*

Extinction probability in a single generation

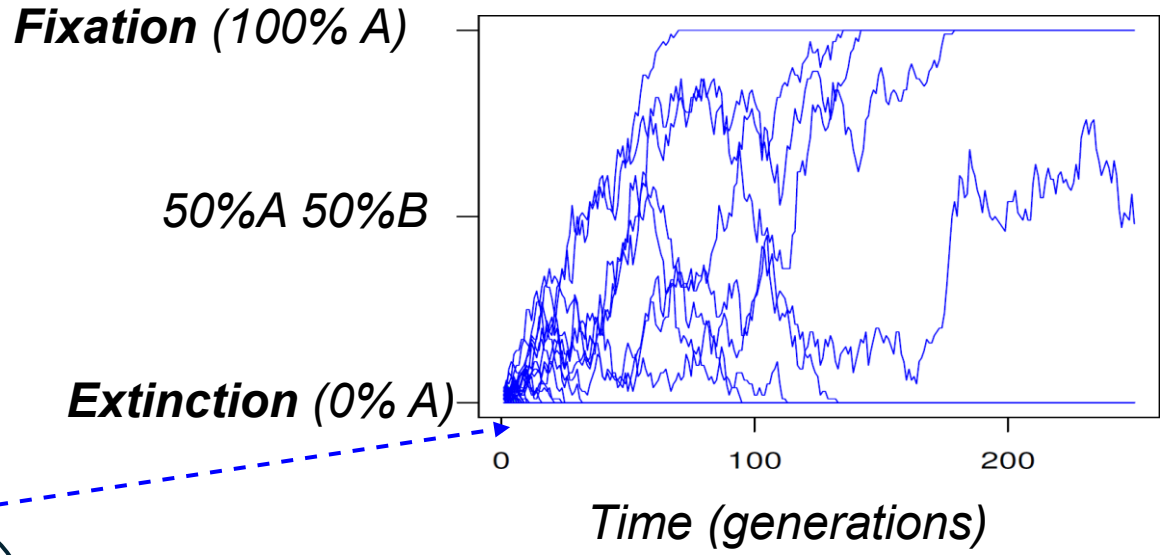
What is the probability that an allele get lost in a single generation in a large population?



- ☐ *It is inversely proportional to 1/2N*
- ☐ *About 37%*
- ☐ *Almost 1-1/2N*

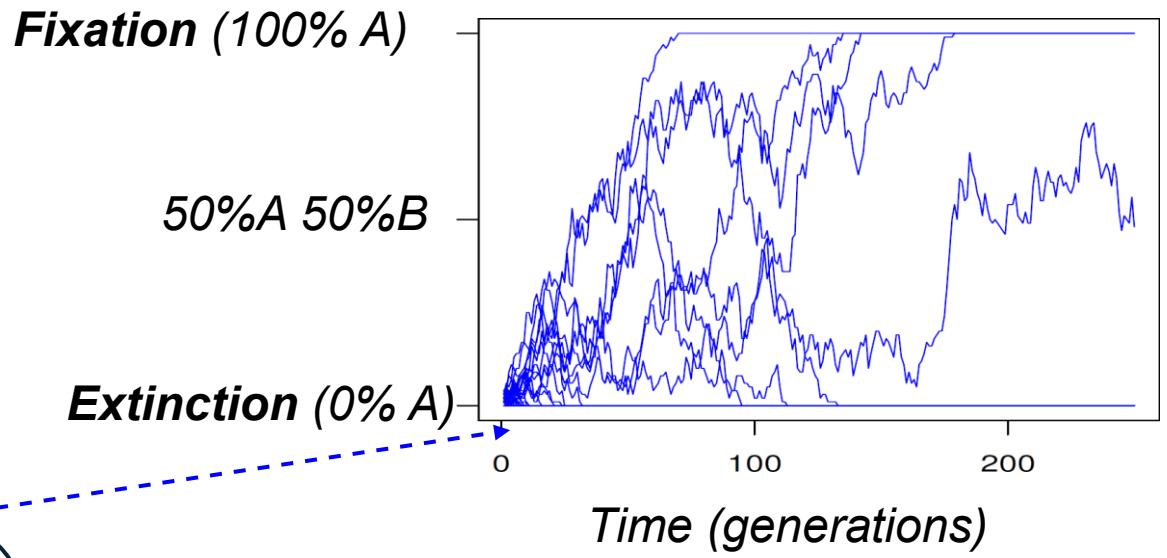
New mutations disappear very quickly in large populations

This can be calculated with the Binomial Probability Binomial(k=0,n=2N,p=1/2N),



- ☐ *It is inversely proportional to 1/2N*
- ☑ *About 37%*
- ☐ *Almost 1-1/2N*

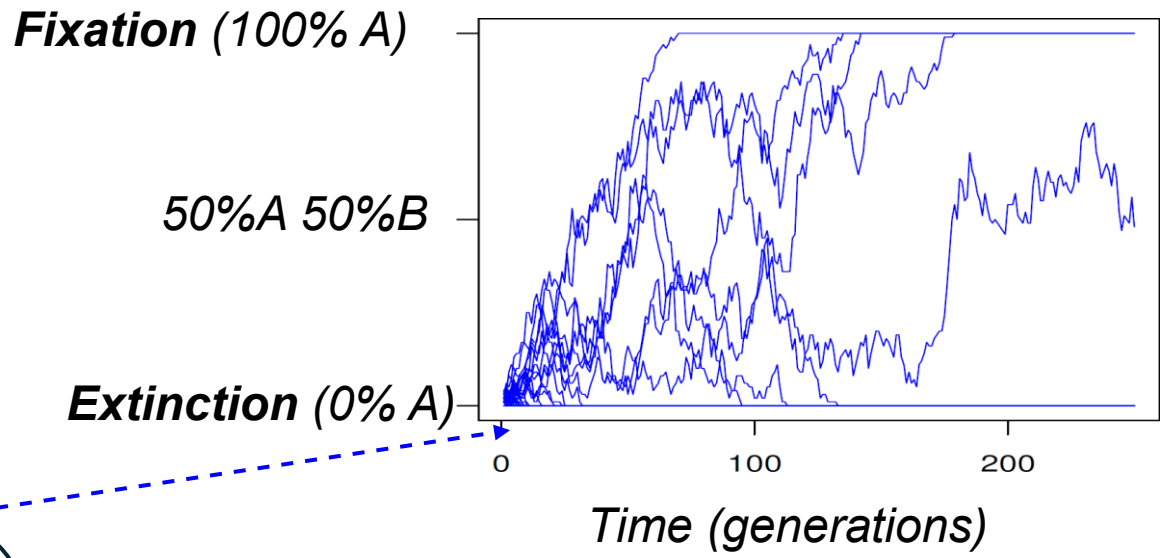
New mutations disappear very quickly in large populations



This can be calculated with the Binomial Probability Binomial(k=0,n=2N,p=1/2N), or more easily $(1-1/2N)^{2N}$

- ☐ *It is inversely proportional to 1/2N*
- ☑ *About 37%*
- ☐ *Almost 1-1/2N*

New mutation disappear very quickly in large populations



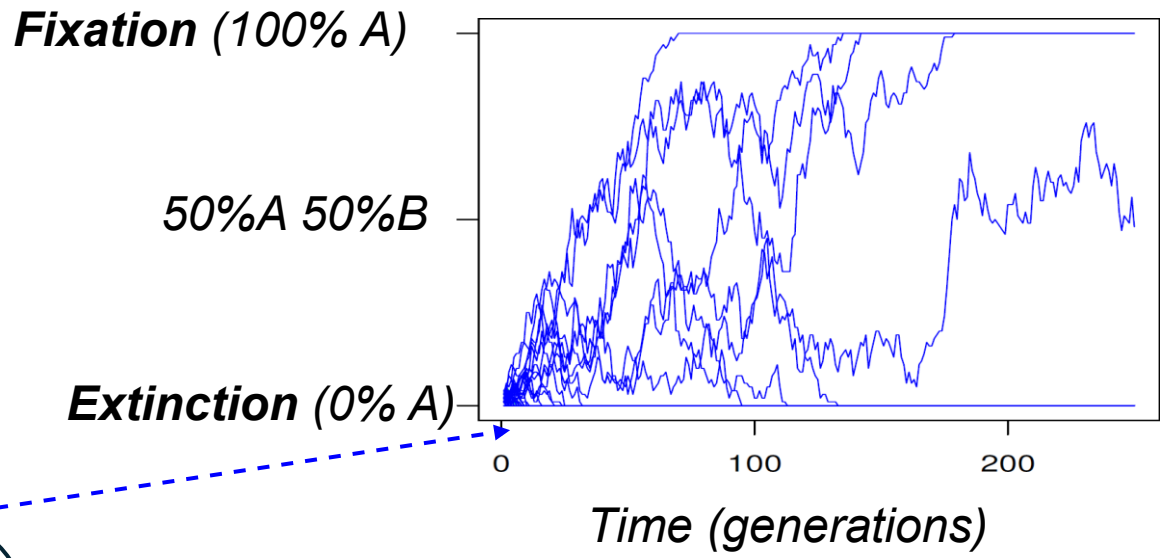
This can be calculated with the Binomial Probability Binomial(k=0,n=2N,p=1/2N), or more easily $(1-1/2N)^{2N}$

And you might remember from school that $\lim_{N\to\infty} (1-1/N)^N = e^{-1} = 0.37$

- ☐ *It is inversely proportional to 1/2N*
- ☑ *About 37%*
- ☐ *Almost 1-1/2N*

New mutation disappear very quickly in large populations

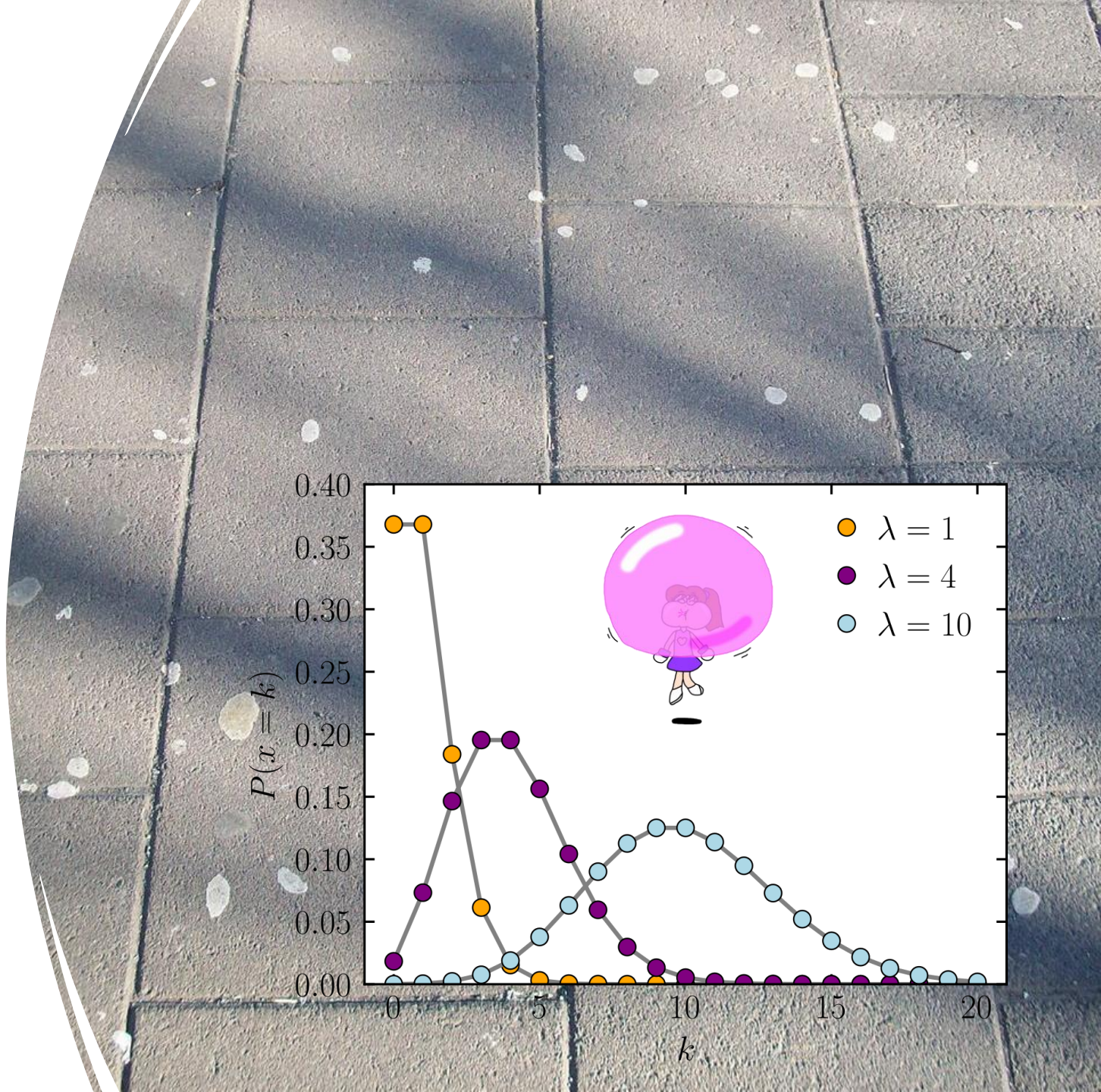
Another way to see it is that in large populations, all alleles leave the a (pretty independent) number of children with the same mean. What distribution does this remind you?



- ☐ *It is inversely proportional to 1/2N*
- ☑ *About 37%*
- ☐ *Almost 1-1/2N*

The Poisson distribution describes the probability of seeing **k** chewingums in a tile, given an average of **λ** chewingums per tile

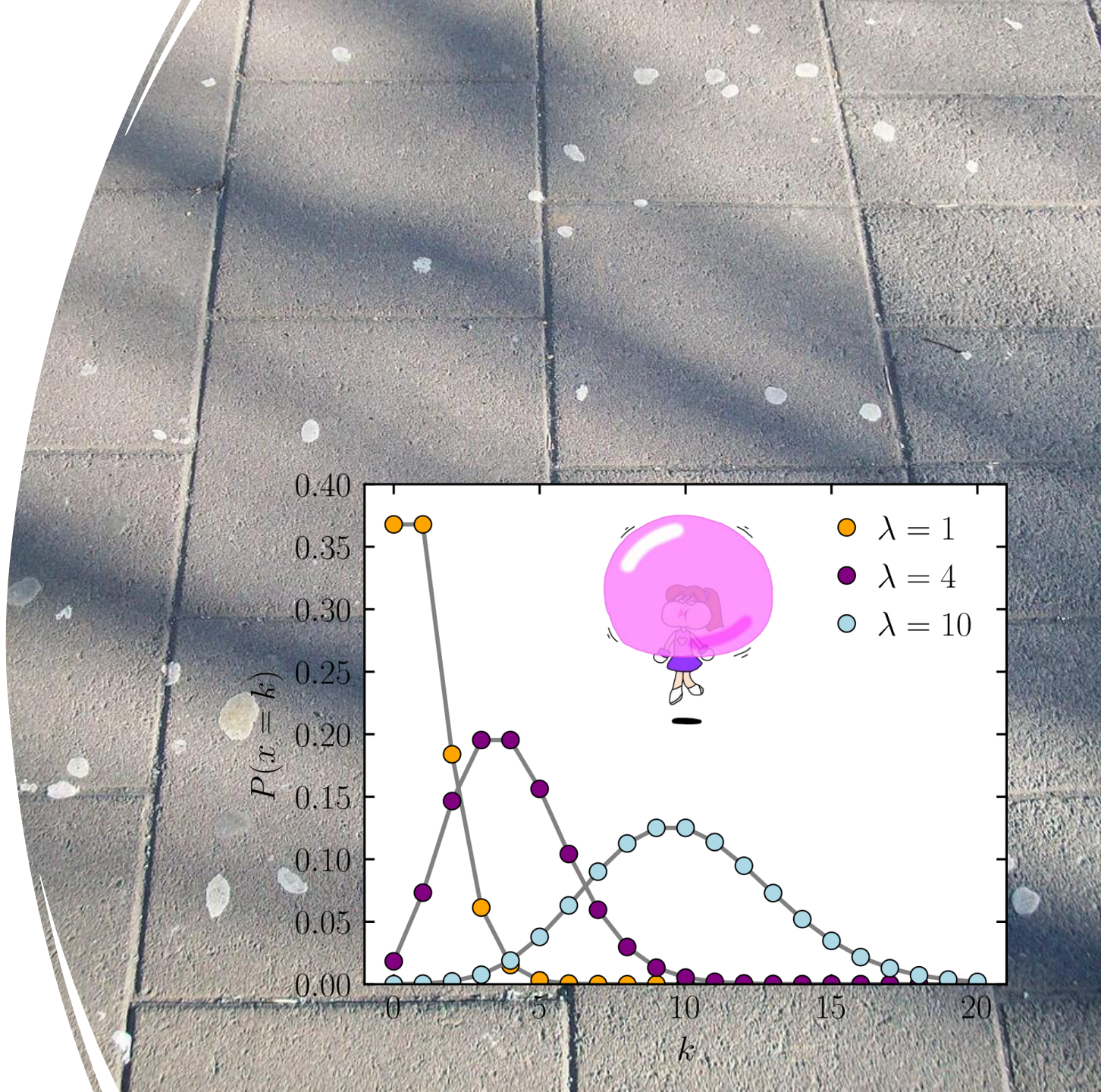
$$\Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



The Poisson distribution describes the probability of seeing **k** chewingums in a tile, given an average of **λ** chewingums per tile

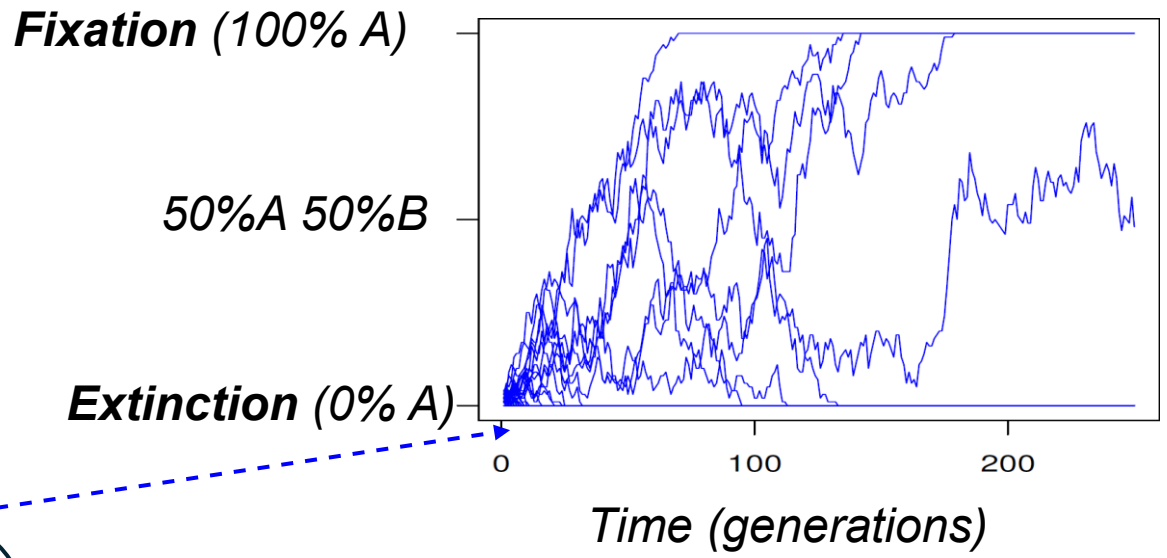
$$\Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

In our case we want to observe k=0 alleles (k!=1), and since all alleles are the same, the all expect on average a single allele: **λ=1.**
Thus: $\Pr(X=0)=e^{-1}=37\%$

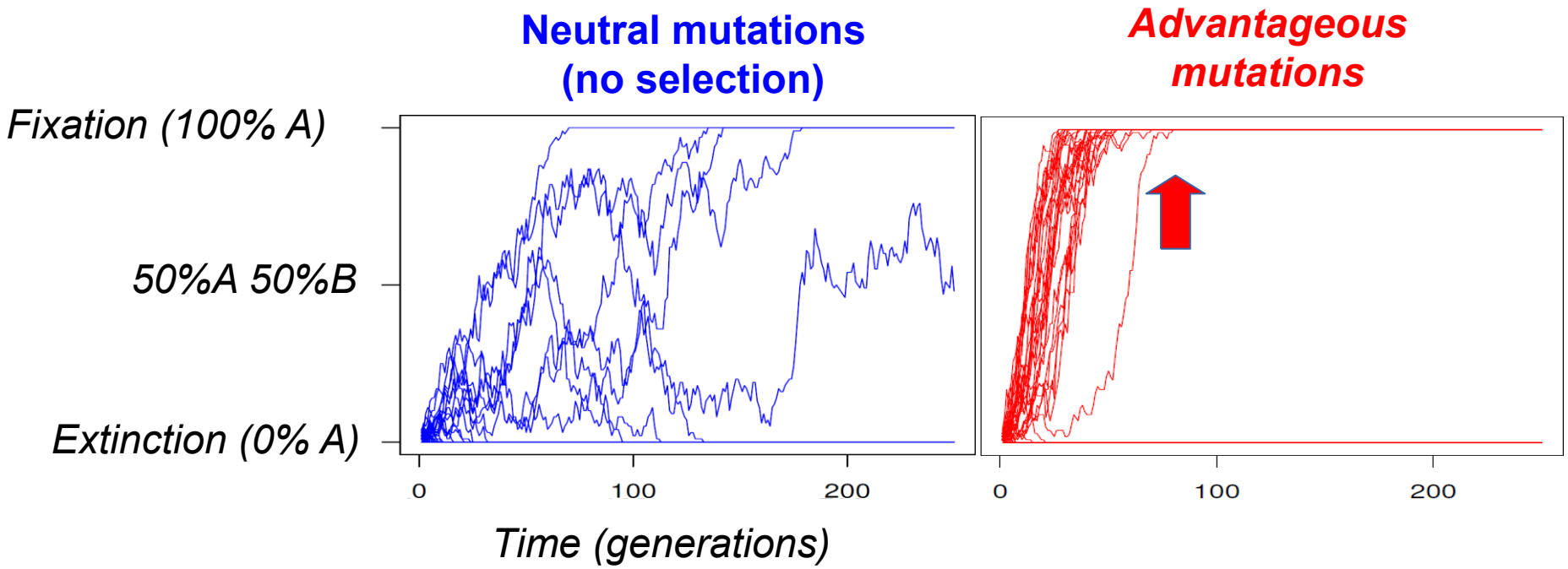


New mutation disappear very quickly in large populations

Mutations often disappears so quickly that natural selection cannot even act on them to select them!

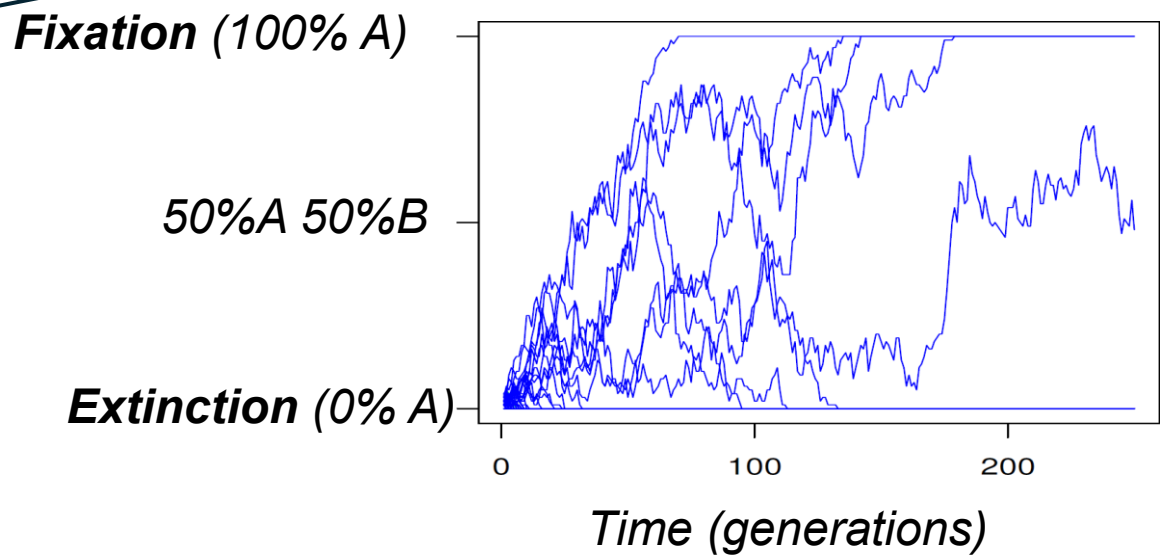


- ☐ *It is inversely proportional to 1/2N*
- ☑ *About 37%*
- ☐ *Almost 1-1/2N*

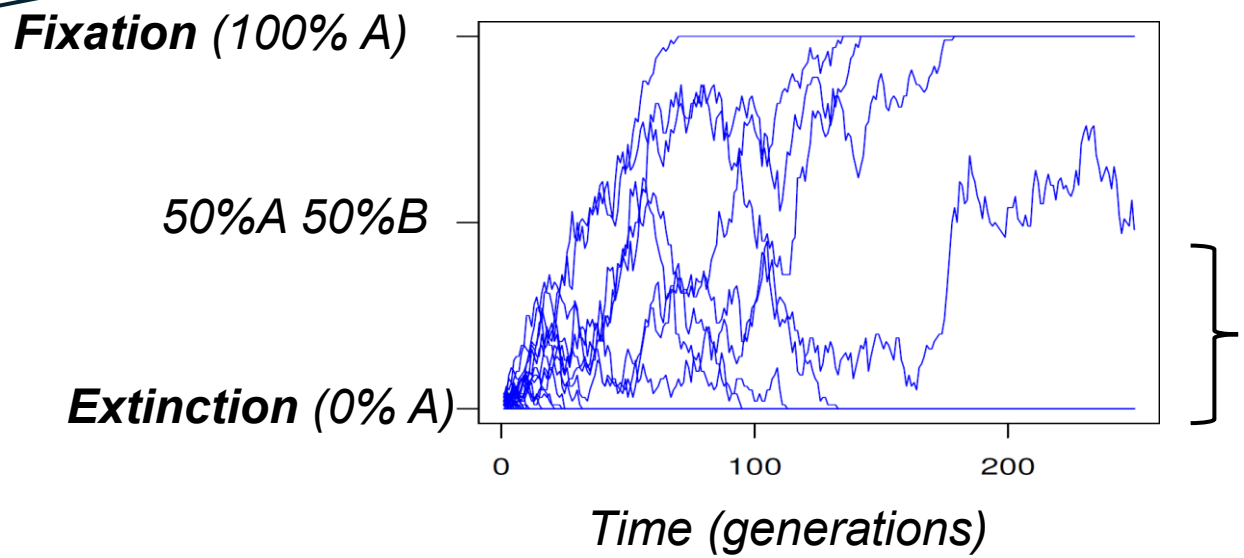


Higher probability of fixation, but still many rapidly disappear in finite populations!

Back to our case, neutral mutations. What do you notice?



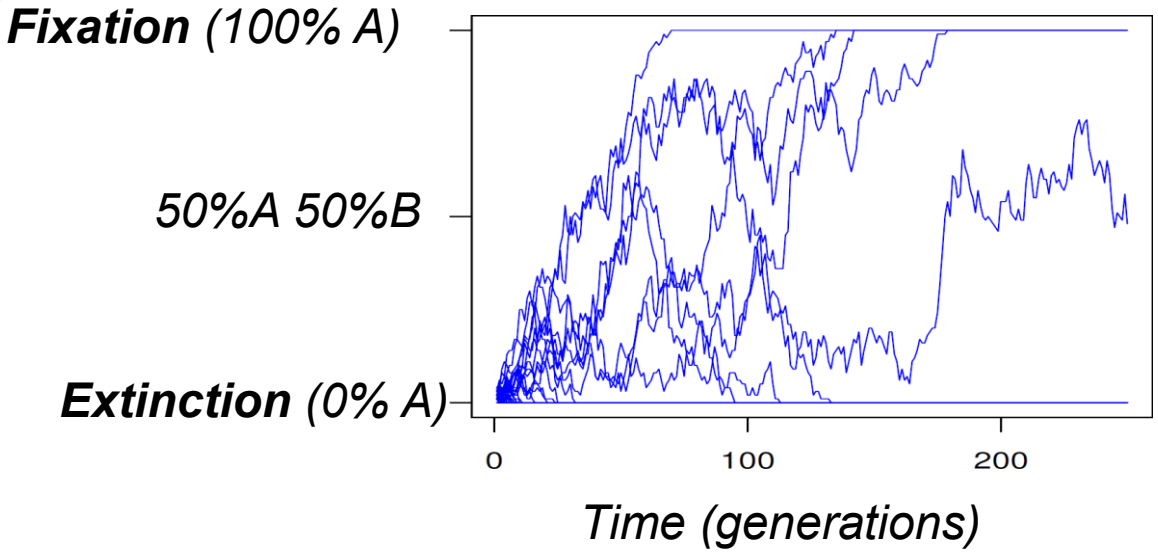
Most alleles should be at low frequencies



If one samples alleles in a population, most alleles would be at low frequencies

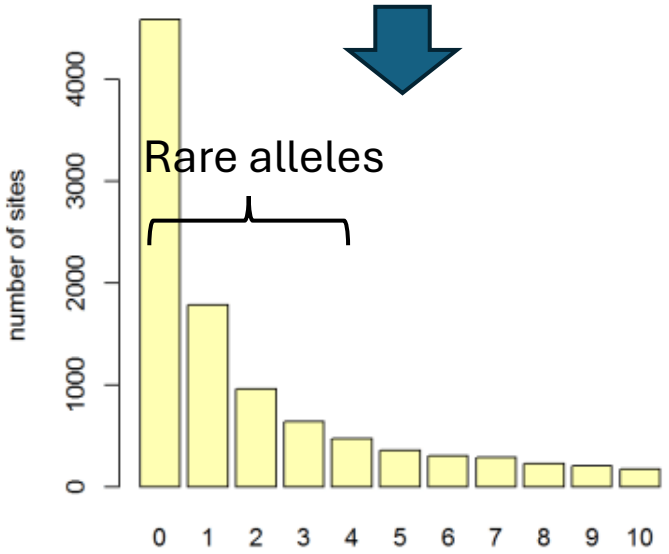
Number of alleles at a given count/freq.

Most alleles should be at low frequencies: the **SFS**



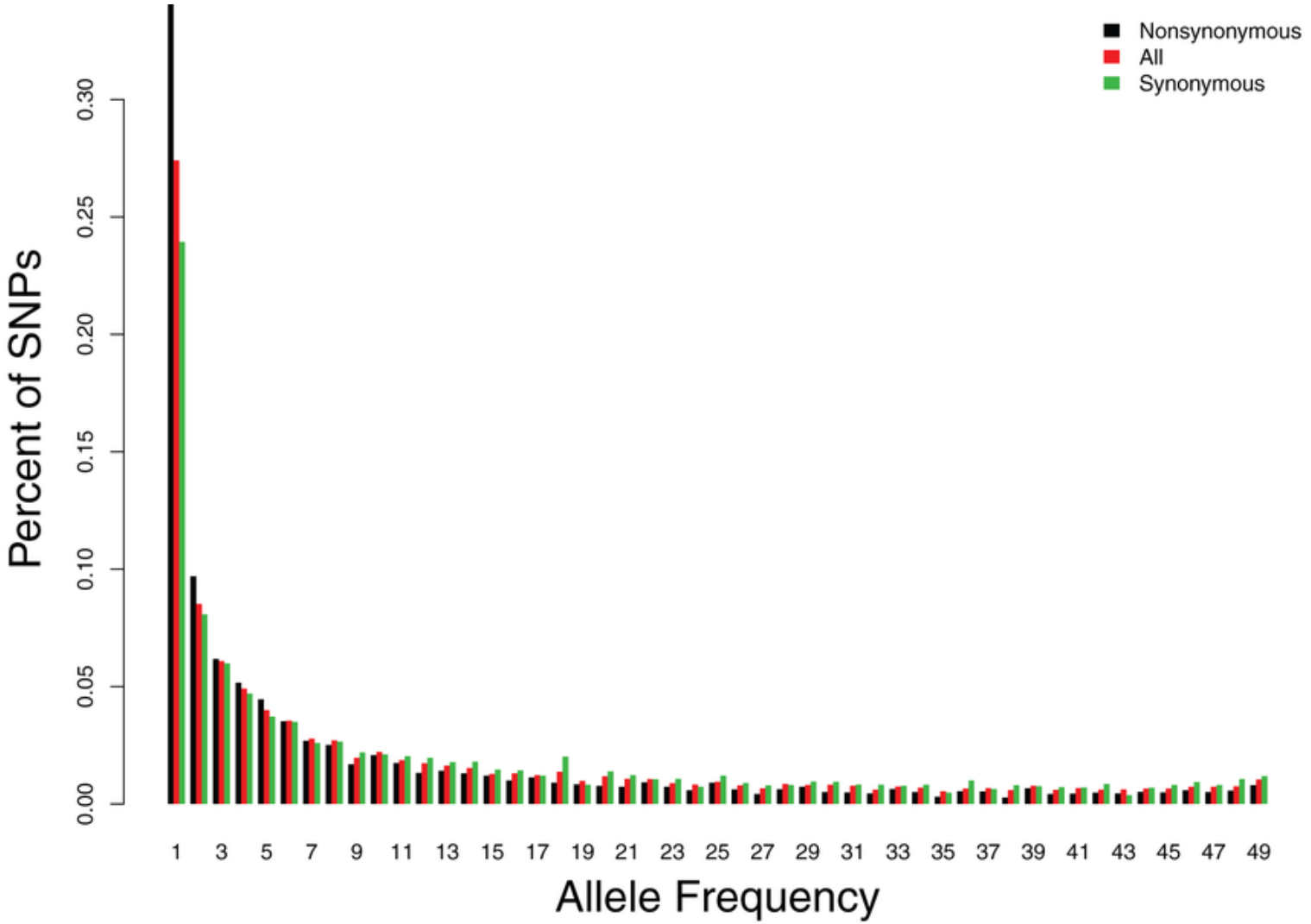
If one samples alleles in a population, most alleles would be at low frequencies

Site Frequency Spectrum (SFS)
(Histogram of allele frequencies)

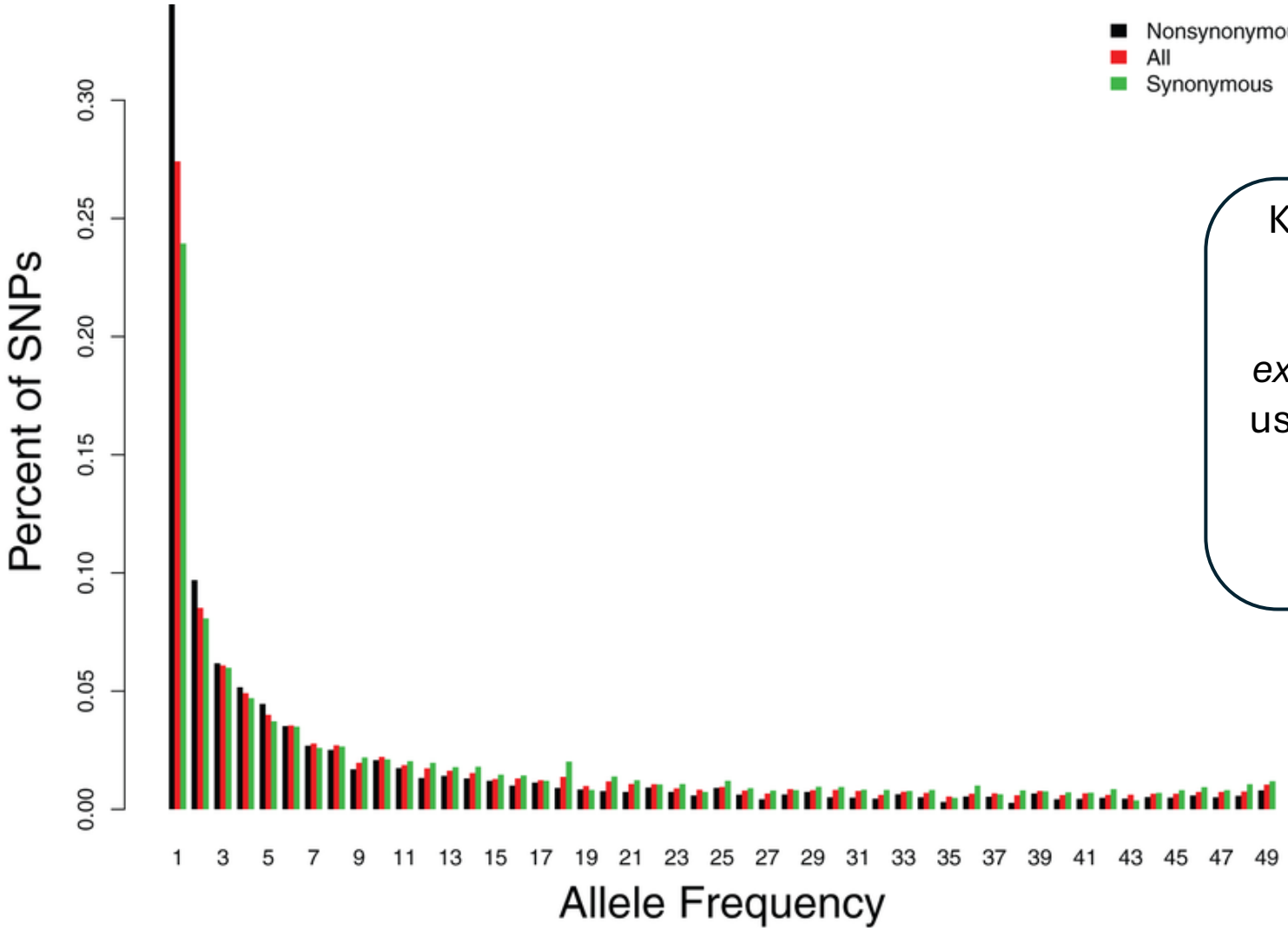


Number of alleles at a given count/freq.

The Site Frequency Spectrum of 25 Danish individuals



The Site Frequency Spectrum of 25 Danish individuals



Knowing what to expect when no selection is present (*neutral expectations*) will provide us a null model to identify targets of **natural selection** and other processes

We will explore this more in future lectures

We saw that we can model populations forward-in-time using Wrigh-Fisher

```
> init_p <- 0.25 #Initial allele frequency
> gen <- 100 #Number of generations
> reps <- 500 #Lots of replicates to run
> colors <- rainbow(reps) #Grab some colors for our reps
> N <- 100 #Population size

> #Initialize a plot
> plot(x=NULL, y=NULL, xlim=c(1, gen), ylim=c(0,1),
    xlab="Generations", ylab="Allele frequency")

> Freq <- NULL #Create an object to save each replicates output

> #Iterate through the replicates
> for(i in 1:reps){
    p <- init_p
    for(j in 1:(gen-1)){
        a <- rbinom(n=1,size=2*N,prob=p[j])
        f <- a/(2*N)
        p <- c(p,f)
        }
    Freq <- rbind(Freq, p) #Save p
    lines(x=1:gen, y=p, lwd=2, col=colors[i])
}

> #Add the mean of all the replicates to the plot
> lines(1:gen, colMeans(Freq), lwd=2, col="black")
```

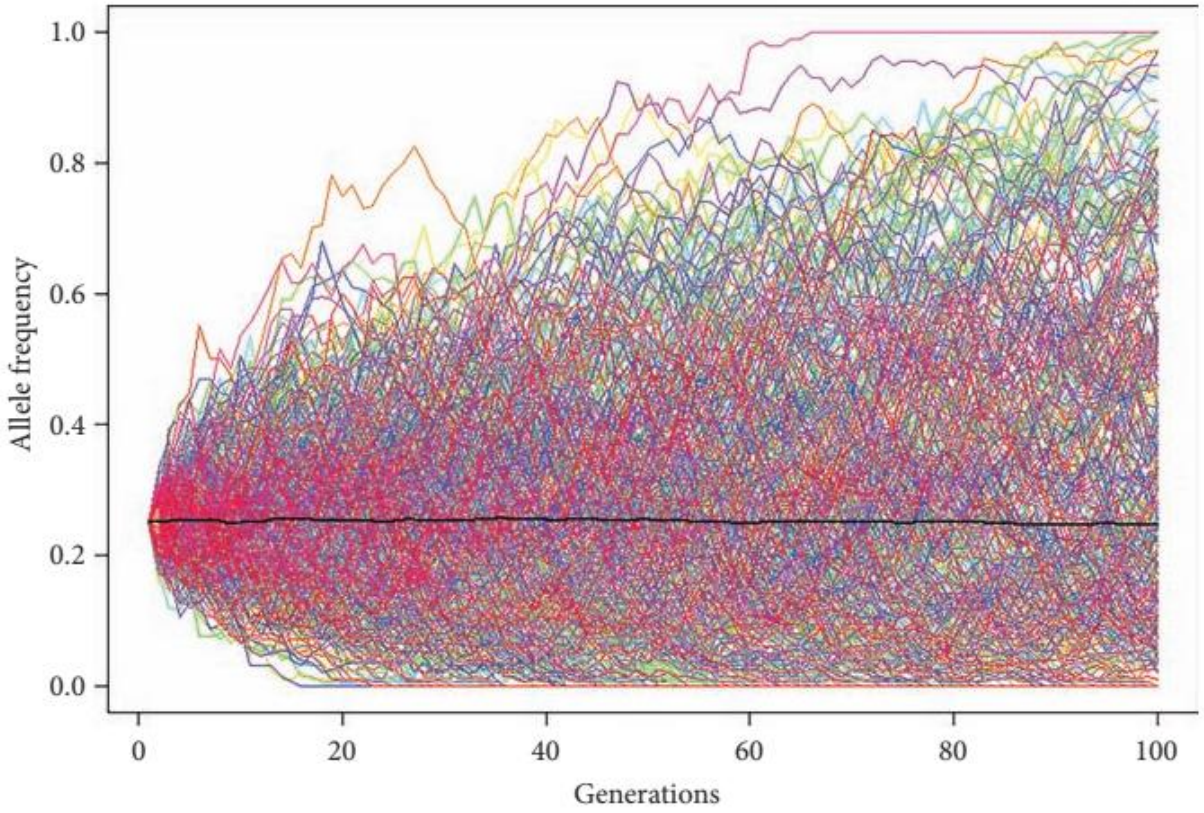


Figure 6.7 Five hundred replicates of random drift in a population of 100 individuals, starting with an allele frequency of 0.25. The average allele frequency across all replicates is plotted in black and stays close to the initial allele frequency.

Since Wright-Fisher is just a binomial, we could even calculate exactly the probability of each step

```
> N <- 10 #Ten diploid individual
> possible <- 0:(2*N) #Number of possible copies of an allele

> P <- NULL #Vector to hold our probabilities
> for(i in possible){
    P <- c(P,dbinom(possible, size=2*N, prob=i/(2*N)))
}

> #Our transition matrix should be 21 rows by 21 columns
> Q <- matrix(P, ncol=2*N+1, byrow=T)

> #Create our starting state matrix
> x <- matrix(c(rep(0,2*N+1)), ncol=2*N+1, byrow=T)

> x[2] <- 1 #Set the prob. of starting with one copy to 100%

> R <- x%*%Q #Get our first gen. transition probabilities

> #Change our color and shape parameters to include
    all the entries between our first "Extinct" state,
    and our final "Fixed" state
> color <- c("brown", rep("blue",ncol(R)-2),"grey")
> shape <- c(15, rep(19,ncol(R)-2), 17)

> #Start with generation 1 for all states
> g <- rep(1,ncol(R))

> #Let's increase our x-axis range to 100
> plot(points(x=NULL, xlim=c(1,100), ylim=c(0,1)
    ylab="Probability",
    xlab="Generations")

> #The unique() function avoids replicates for color & shape
> legend("bottomleft",
    legend=c("Extinct","Intermediate","Fixed"),
    col=unique(color), pch=unique(shape),
    inset=c(0,1), xpd=TRUE, bty="n")

> while(g[1]<100){
    R <- R%*%Q
    g <- g+1
    points(g, R, col=color, pch=shape)
}

> #Finally let's add two horizontal lines to our plot:
> #The starting allele frequency of A
> abline(h=1/(2*N), lwd=2)
> #And the starting allele frequency of a
> abline(h=1-(1/(2*N)), lwd=2, col="orange", lty=2)
```

$$P(k|p,n) = \binom{n}{k} p^k (1-p)^{n-k}.$$

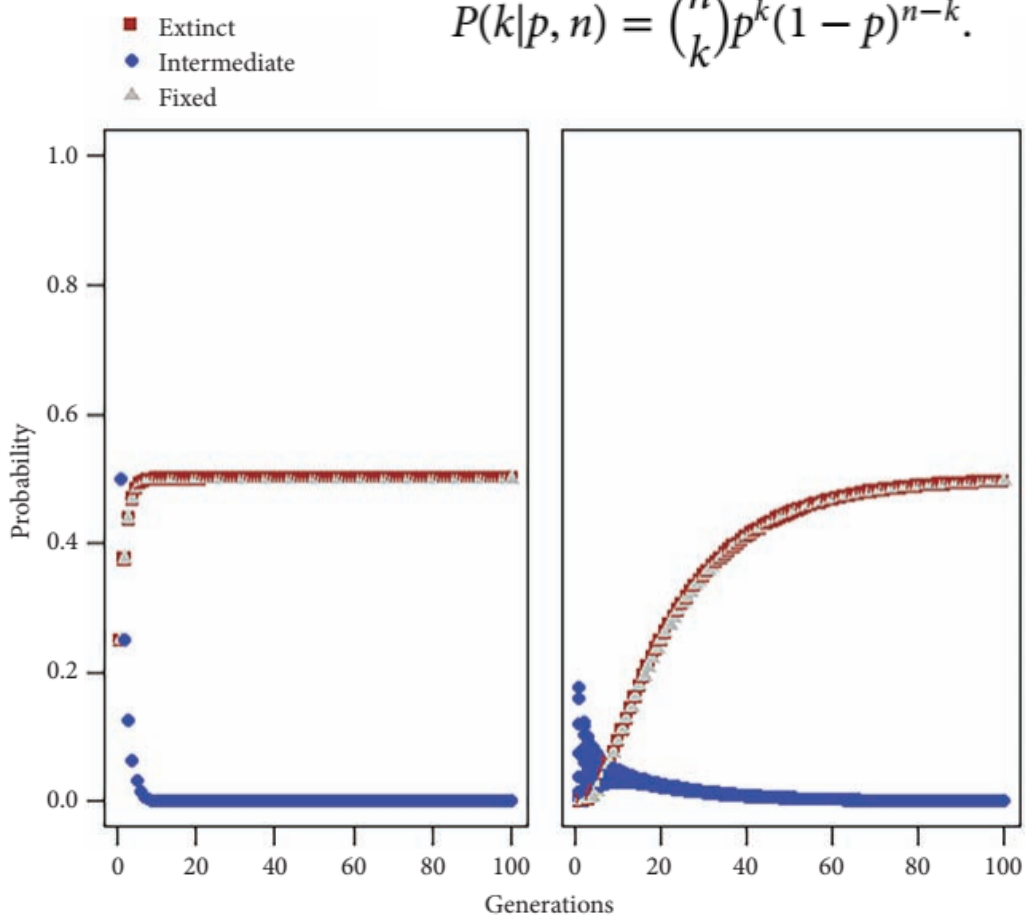
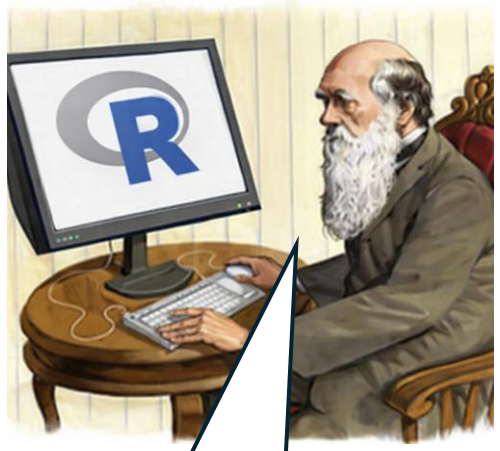
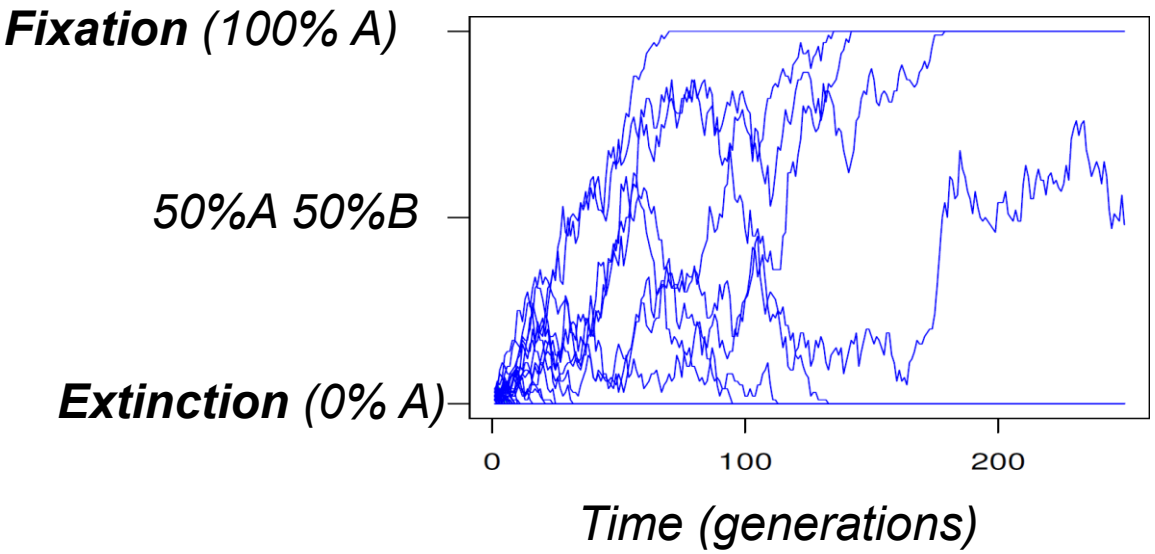


Figure 6.4 Probability of an allele state across generations when starting allele frequency is equal to 50%. Comparing single-diploid individual (left) with ten-diploid individuals (right).

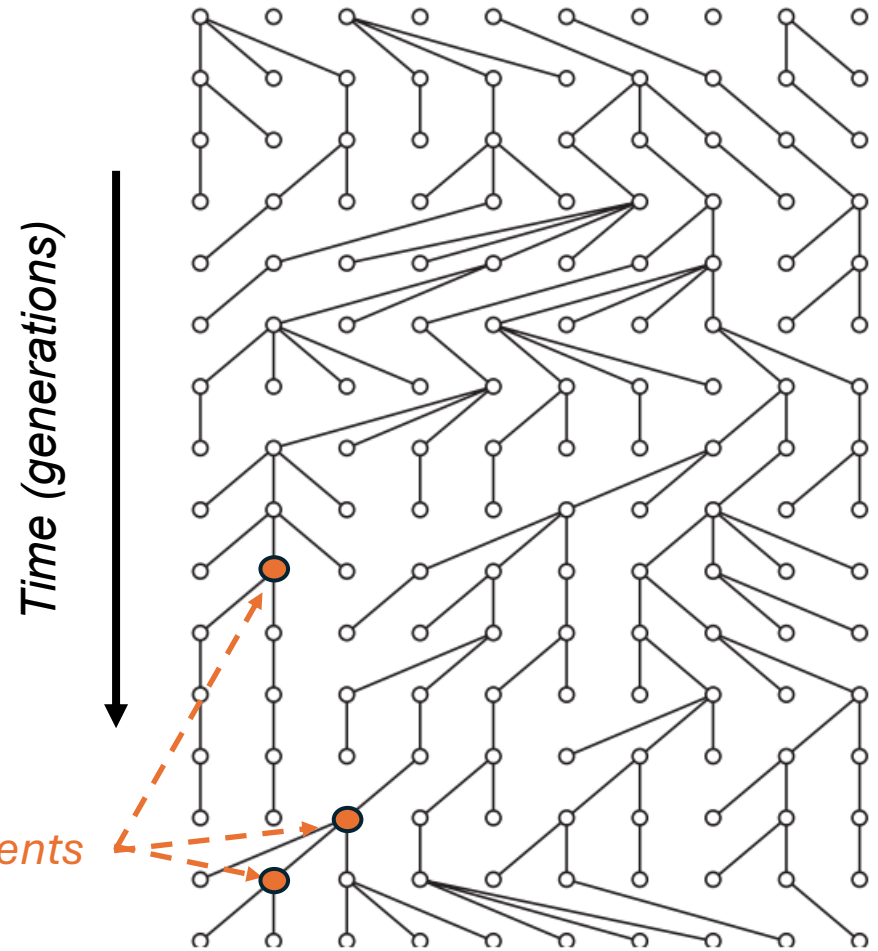


But long, computationally intensive and quite boring

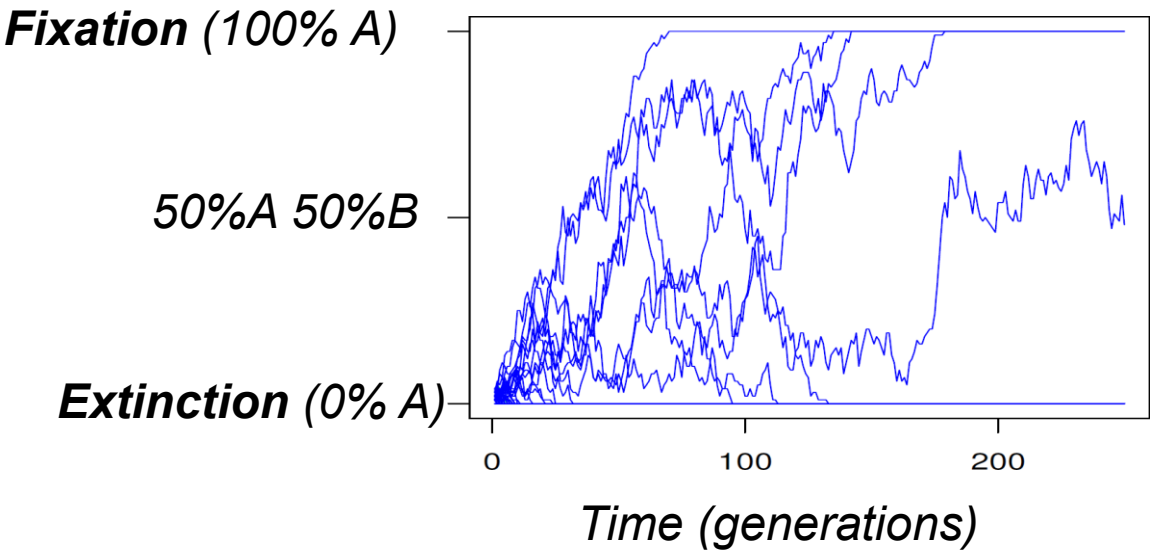
Coalescent theory



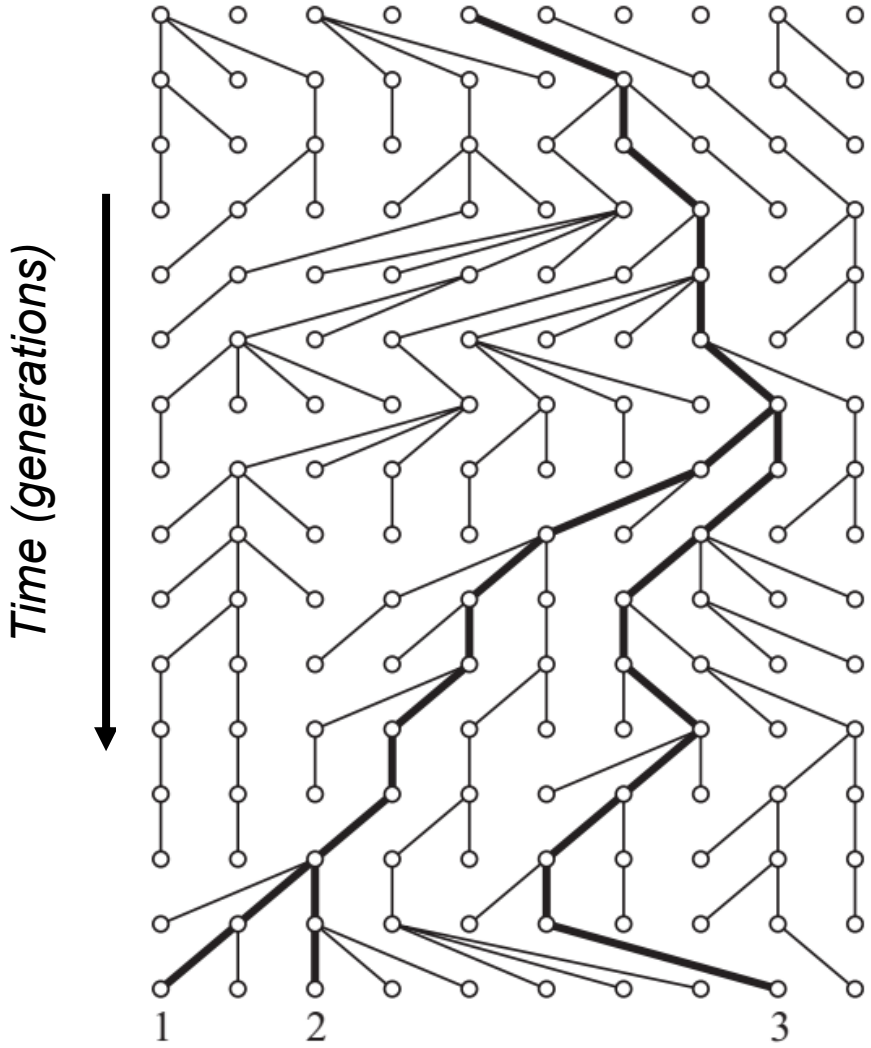
We can represent our evolving population as genealogies



Coalescent theory



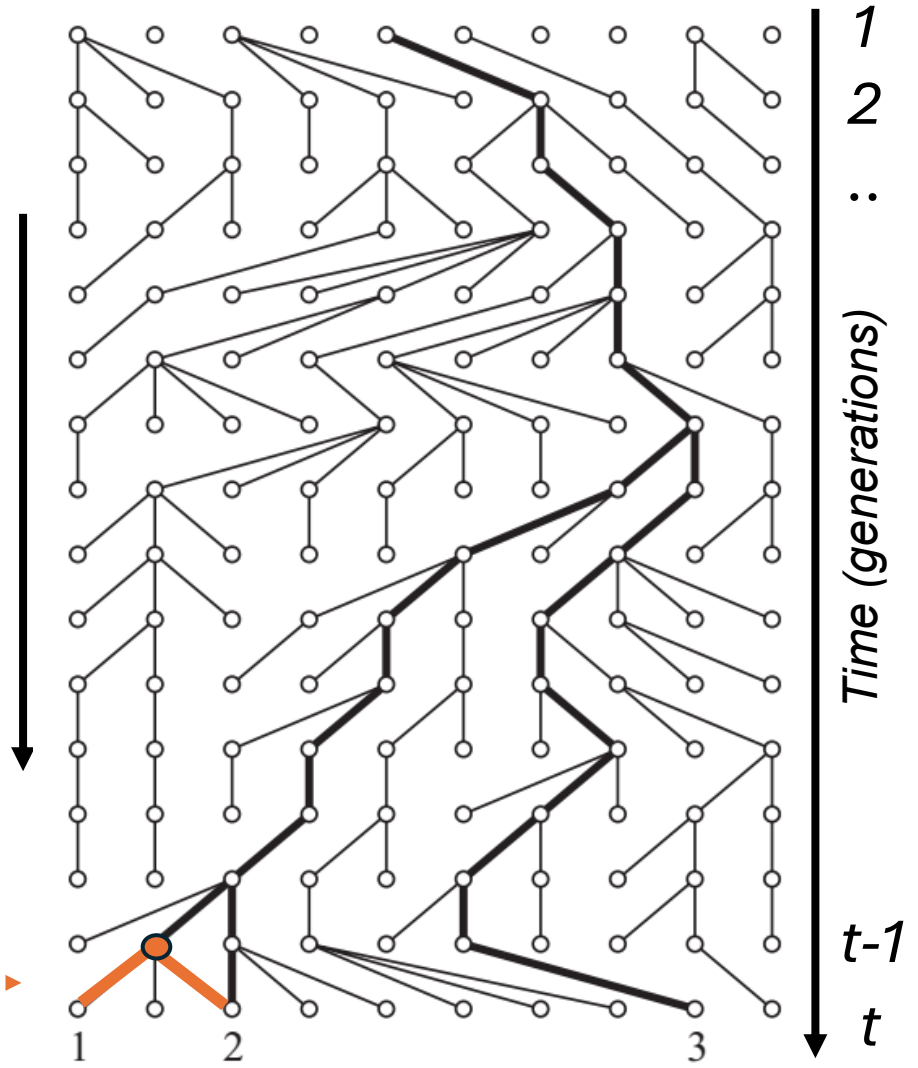
We can represent our evolving population as genealogies



Where do lineages 1, 2 and 3 coalesce?

Coalescent theory

We can represent our evolving population as genealogies

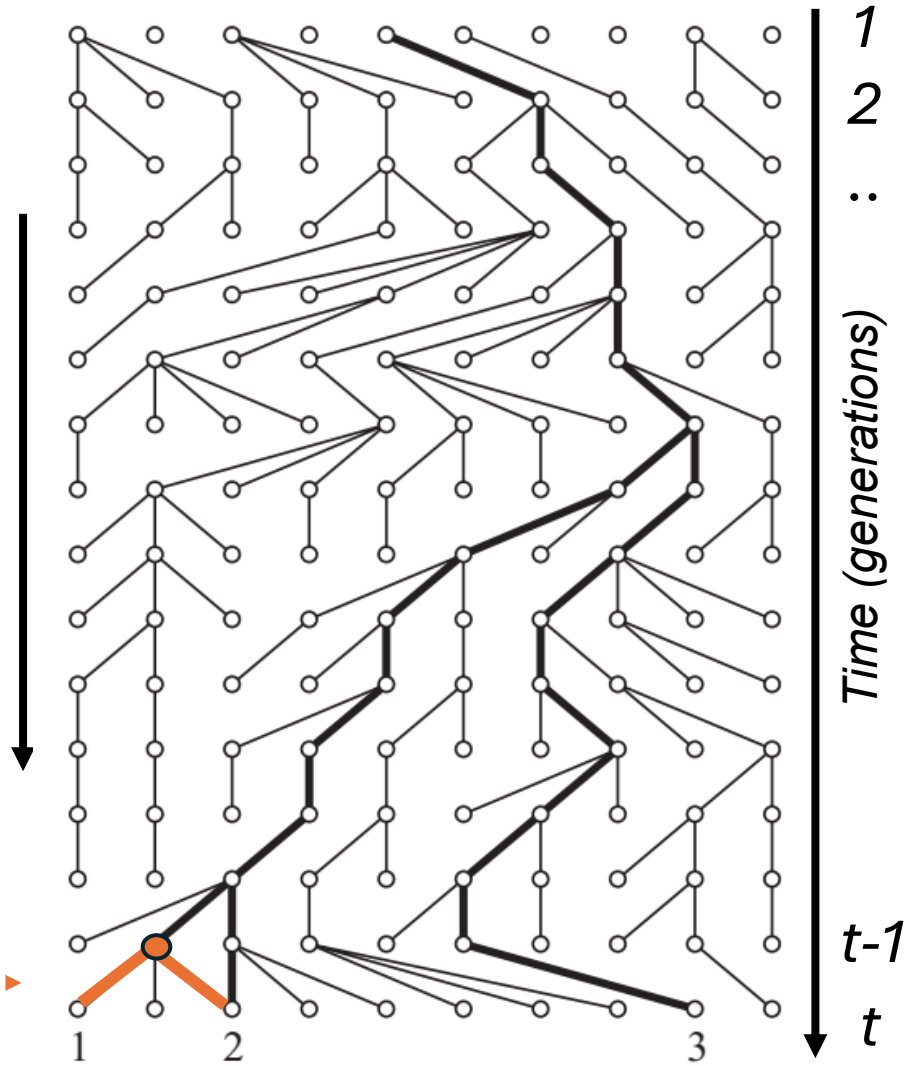


What is the probability that two lineages (e.g. 1 and 2) coalesce at time *t-1*?

Coalescent theory

- ☐ *1/2N*
- ☐ $1/(2N)^2$
- ☐ *1/N*

We can represent our evolving population as genealogies



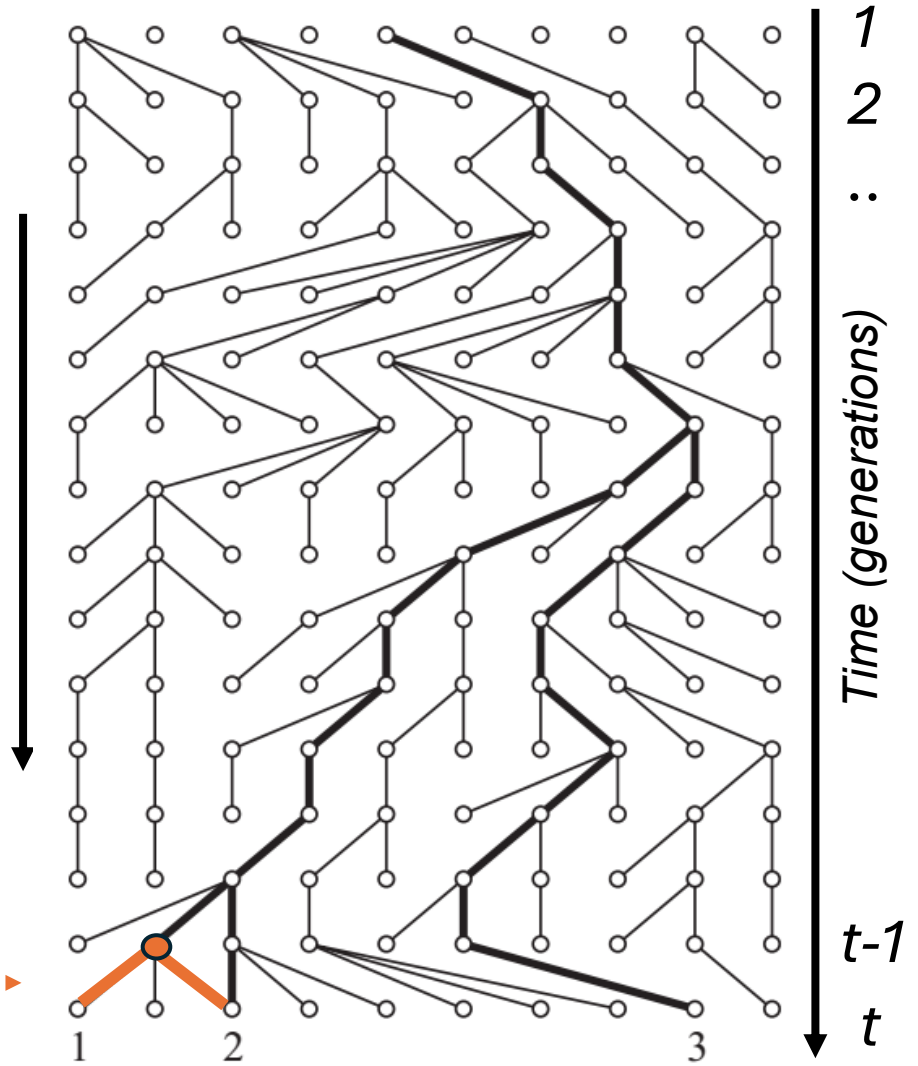
What is the probability that two lineages (e.g. 1 and 2) coalesce at time *t-1*?

Coalescent theory

Thin of it like this: first I choose the ancestor of *1*, which can be anyone. The chance that that of *2* is the same is then 1/2N

- [x] *1/2N*
- [ ] *1/(2N)²*
- [ ] *1/N*

We can represent our evolving population as genealogies

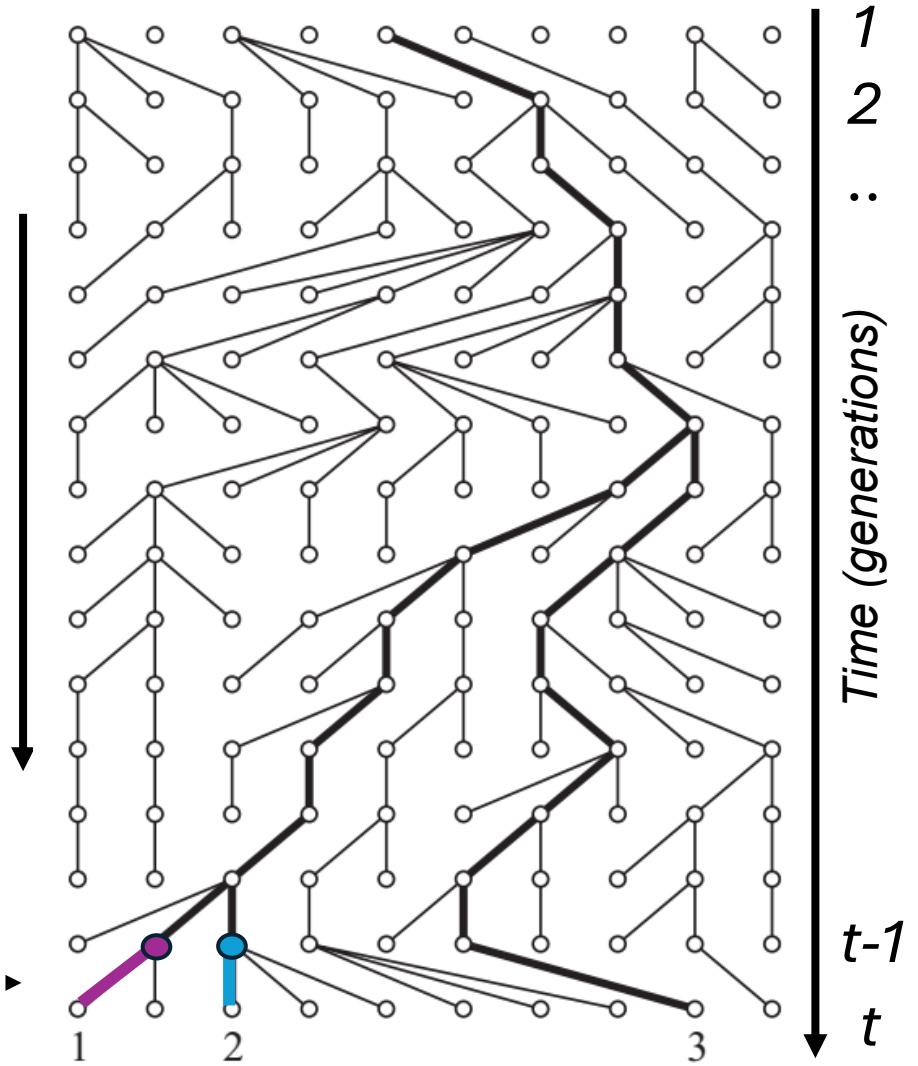


What is the probability that two lineages (e.g. 1 and 2) coalesce at time *t-1*?

Coalescent theory

- ☐ *1-1/2N*
- ☐ *1/2N*
- ☐ *1/N*

We can represent our evolving population as genealogies

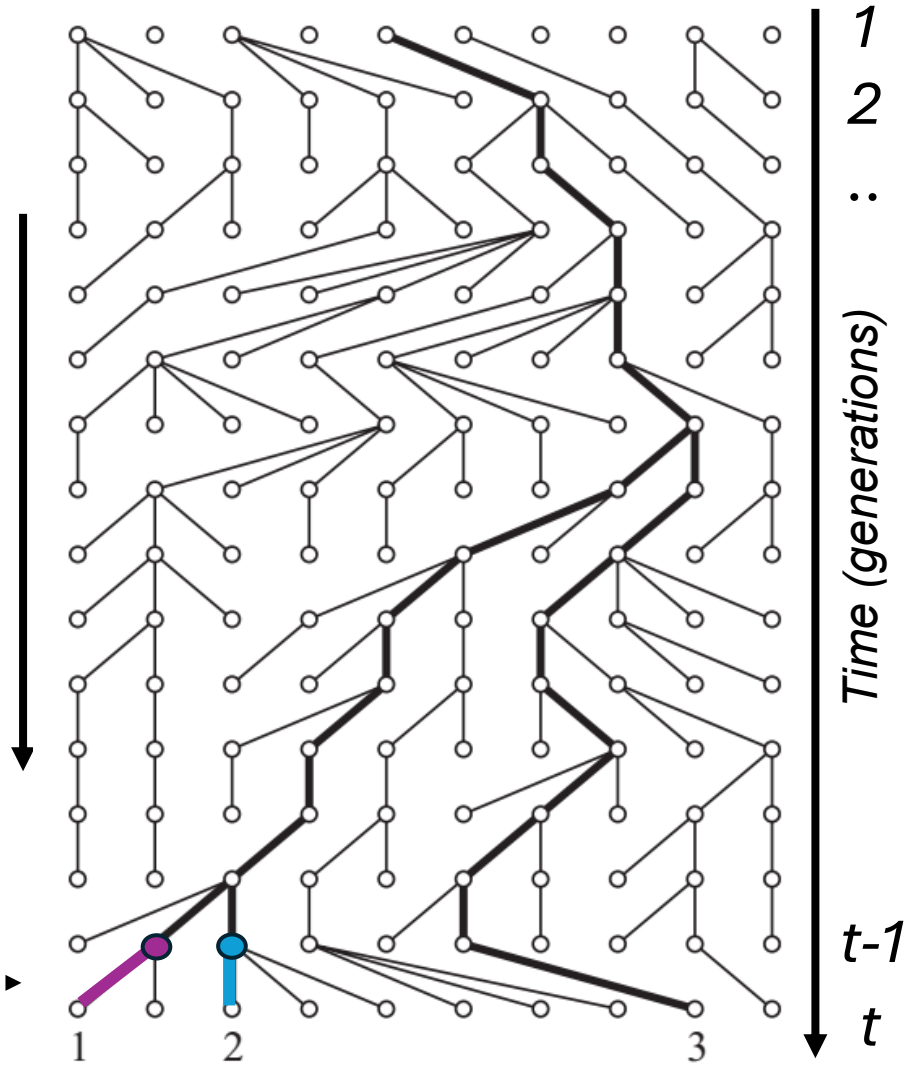


What is the probability then they have a different ancestor at time t-1?

Coalescent theory

- [x] *1-1/2N*
- [ ] *1/2N*
- [ ] *1/N*

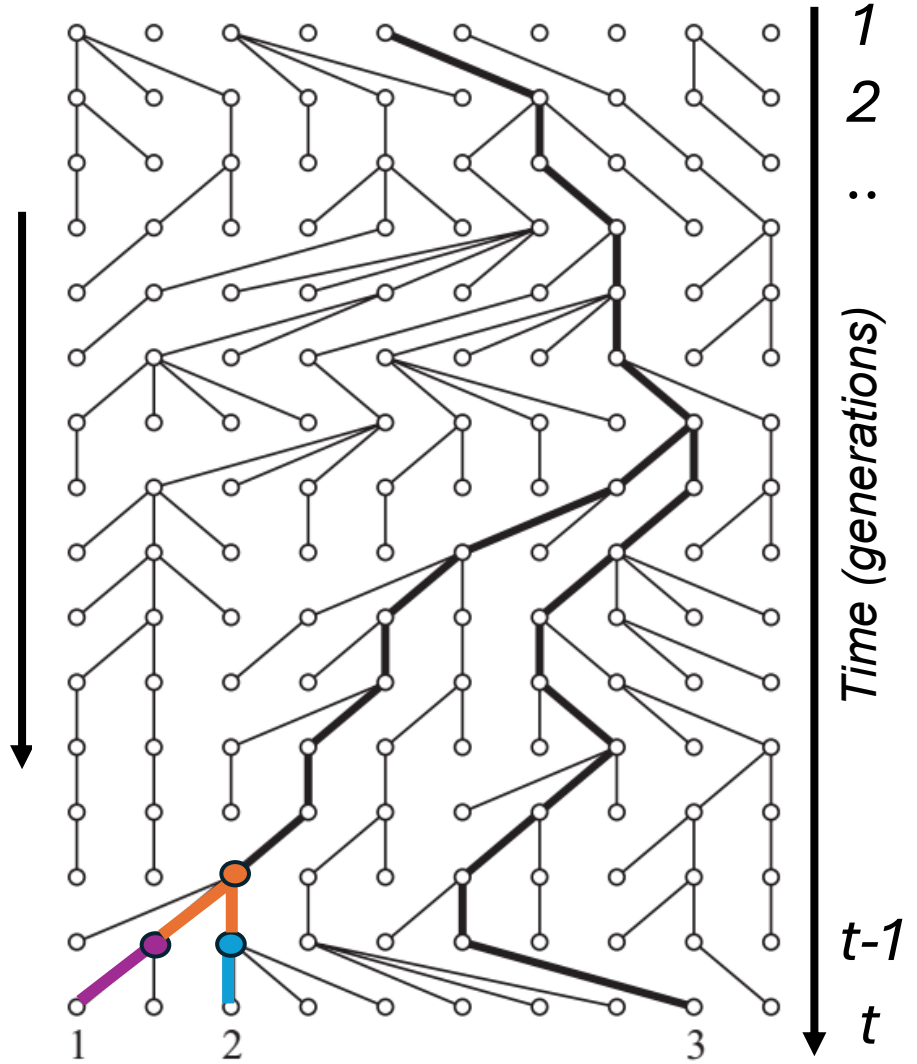
We can represent our evolving population as genealogies



What is the probability then they have a different ancestor at time t-1?

Using coalescent theory we can calculate the probability that two lineages will coalesce back in time without computing the WF

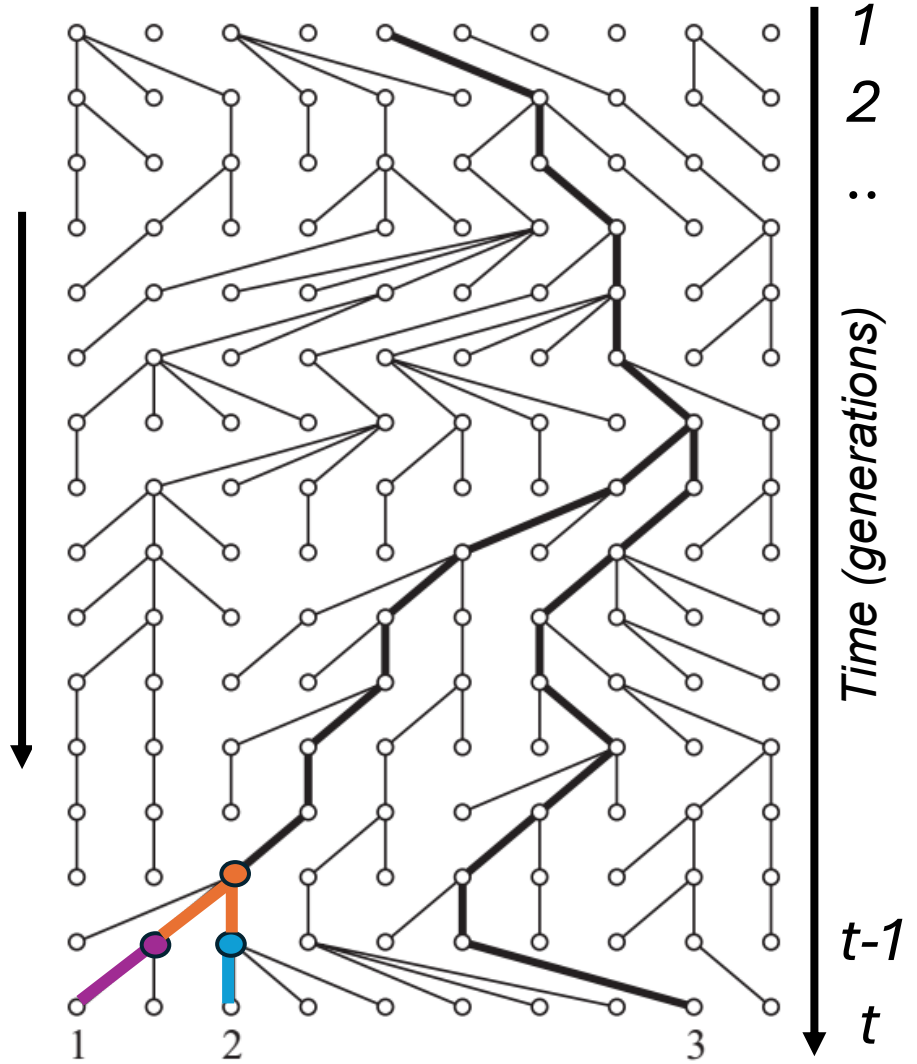
What is the probability then they have a different ancestor at time *t-2*?



Using coalescent theory we can calculate the probability that two lineages will coalesce back in time without computing the WF

- ☐ *(1-1/2N)*(1/2N)*
- ☐ *(1-1/2N)*(1-1/2N)*
- ☐ *1/2N*

What is the probability then they have a different ancestor at time *t-2*?

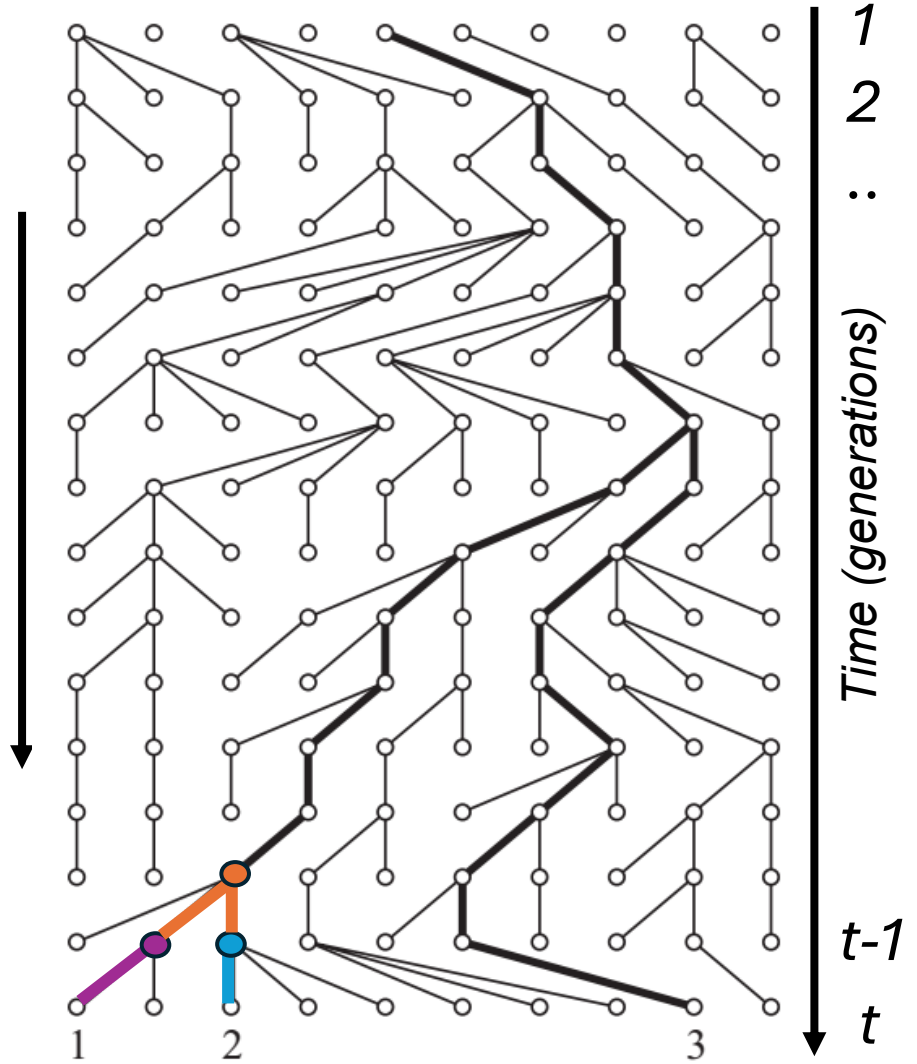


Using coalescent theory we can calculate the probability that two lineages will coalesce back in time without computing the WF

Non coalesce at t-1 | Coalesce at t-2

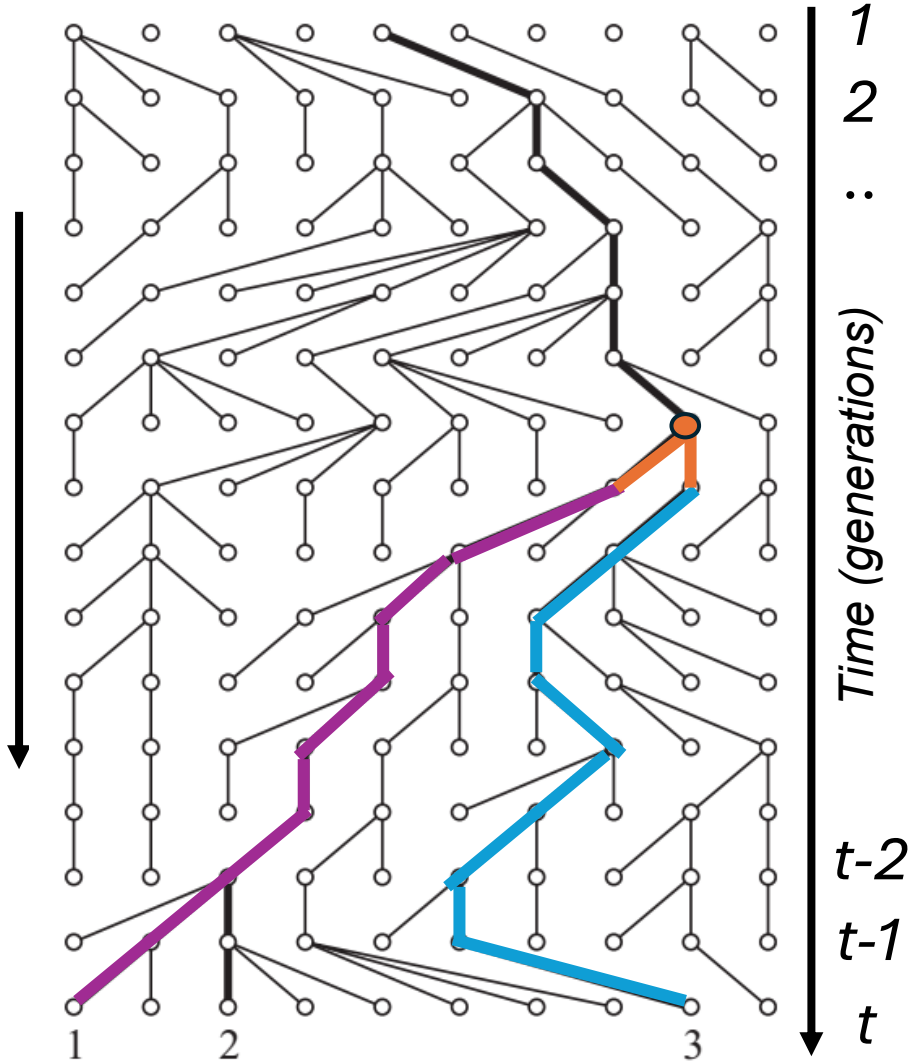
- [x] *(1-1/2N)*(1/2N)*
- [ ] *(1-1/2N)*(1-1/2N)*
- [ ] *1/2N*

What is the probability then they have a different ancestor at time *t-2*?



Using coalescent theory we can calculate the probability that two lineages will coalesce back in time without computing the WF

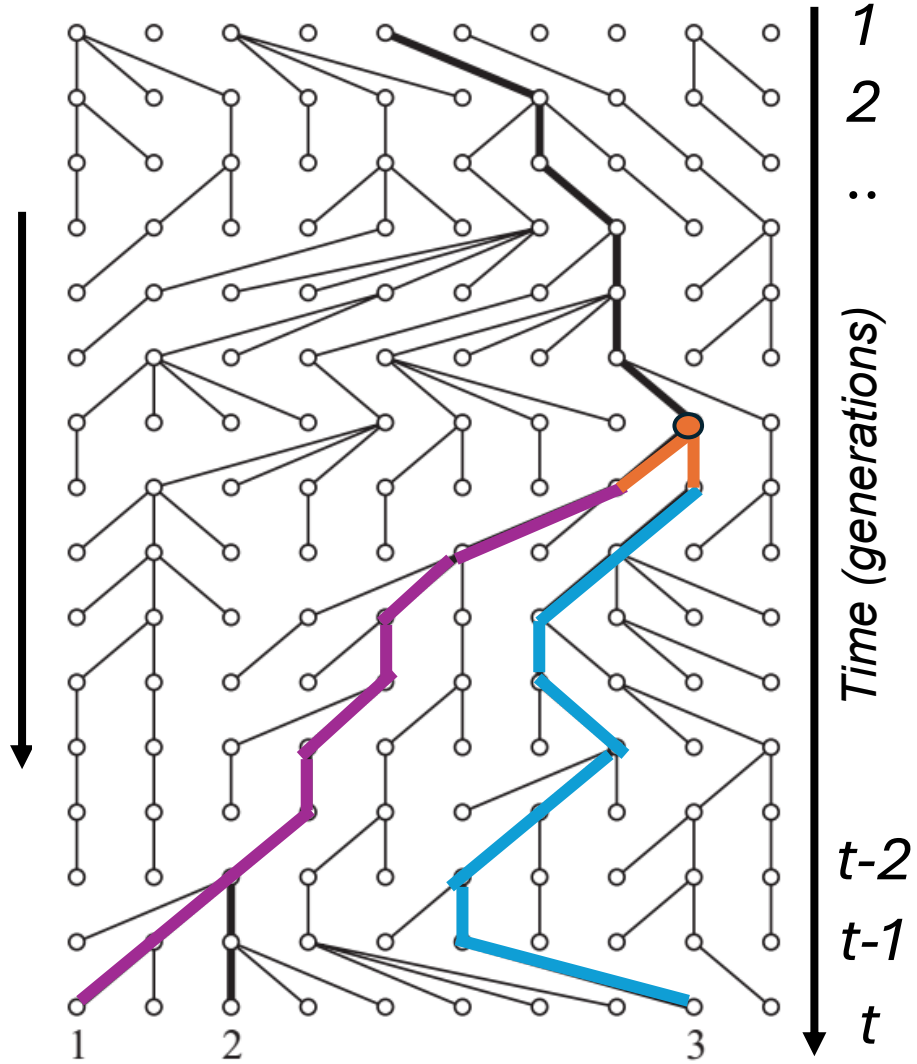
What is the probability that *1* and *3* have a different ancestor at time *t-9*?



Using coalescent theory we can calculate the probability that two lineages will coalesce back in time without computing the WF

- [x] $(1-1/2N)^8/2N$
- [ ] $(1-1/2N)^9$
- [ ] $(1/2N)^9$

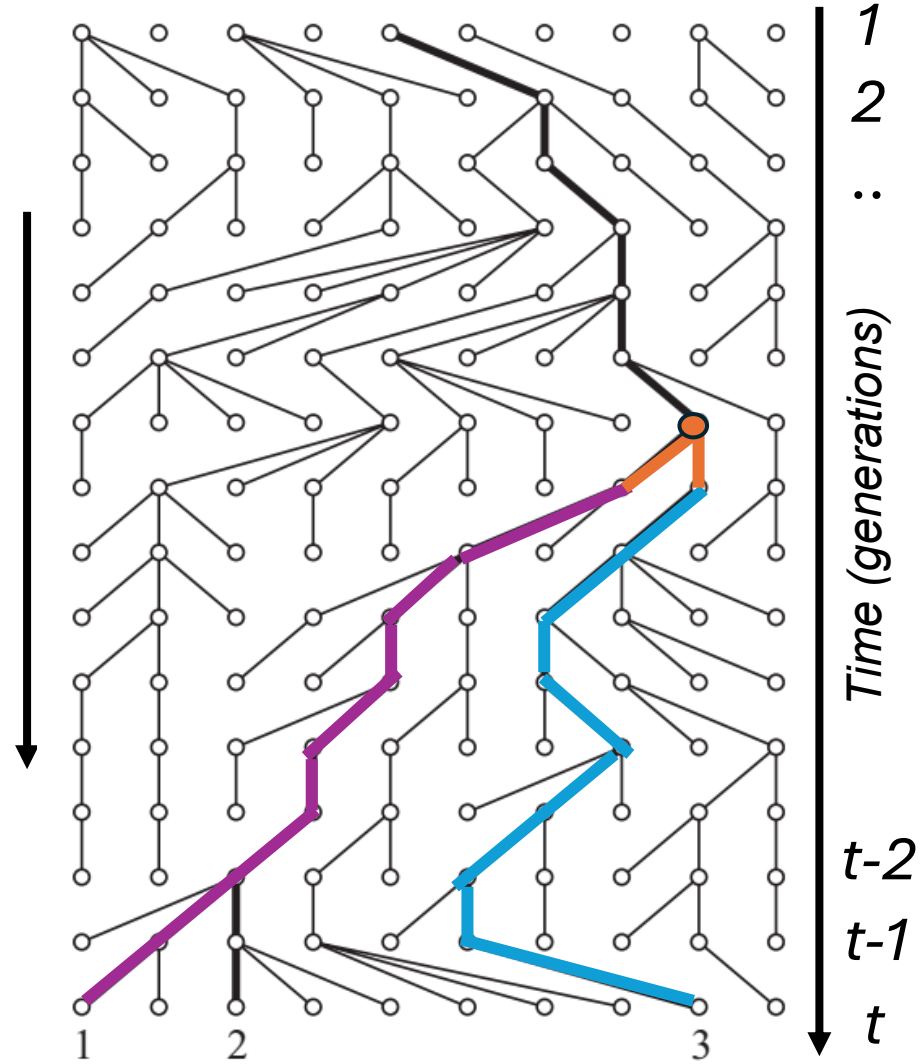
What is the probability that *1* and *3* have a different ancestor at time *t*-9?



Using coalescent theory we can calculate the probability that two lineages will coalesce back in time without computing the WF

- [x] $(1-1/2N)^8/2N$
- [ ] $(1-1/2N)^9$
- [ ] $(1/2N)^9$

What is the probability that *1* and *3* have a different ancestor at time *t-9*?



The time till a coalescent event for two lineages is geometrically distributed

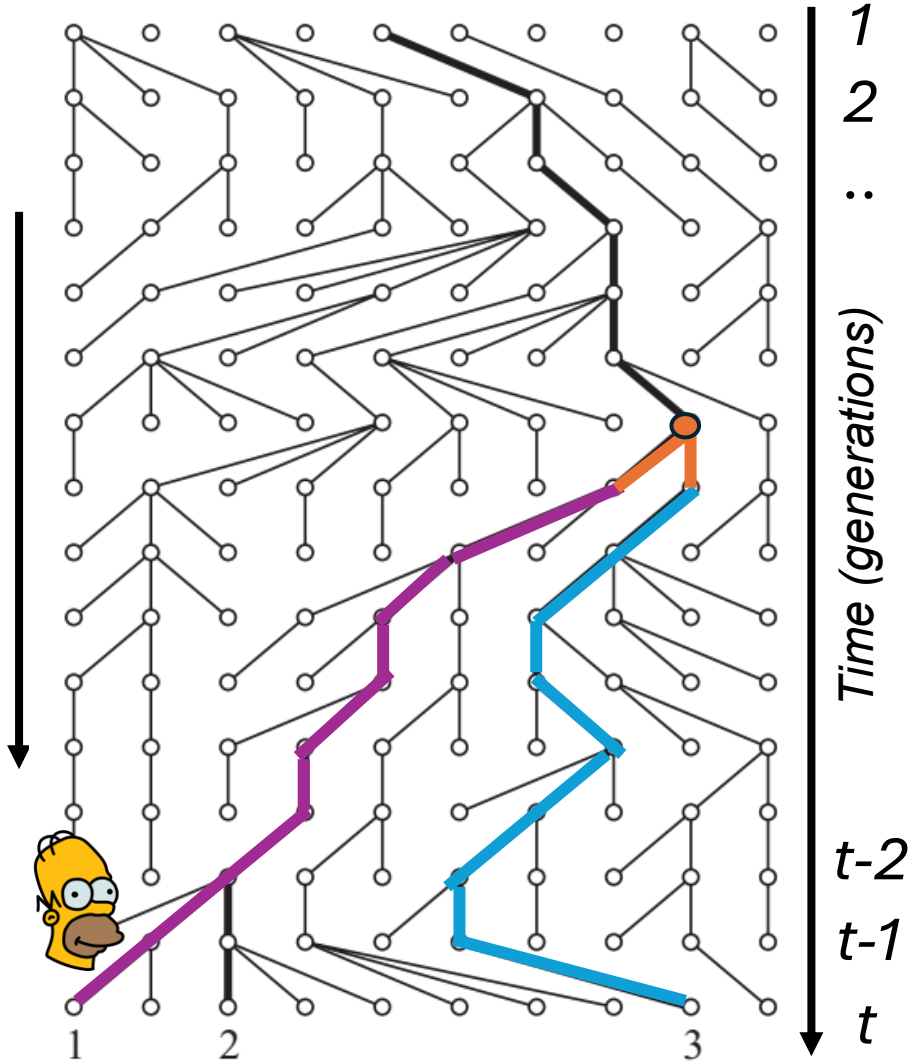
$$P(T_2 = j) = \left(1 - \frac{1}{2N}\right)^{j-1} \frac{1}{2N},$$

where *j* is the number of generations

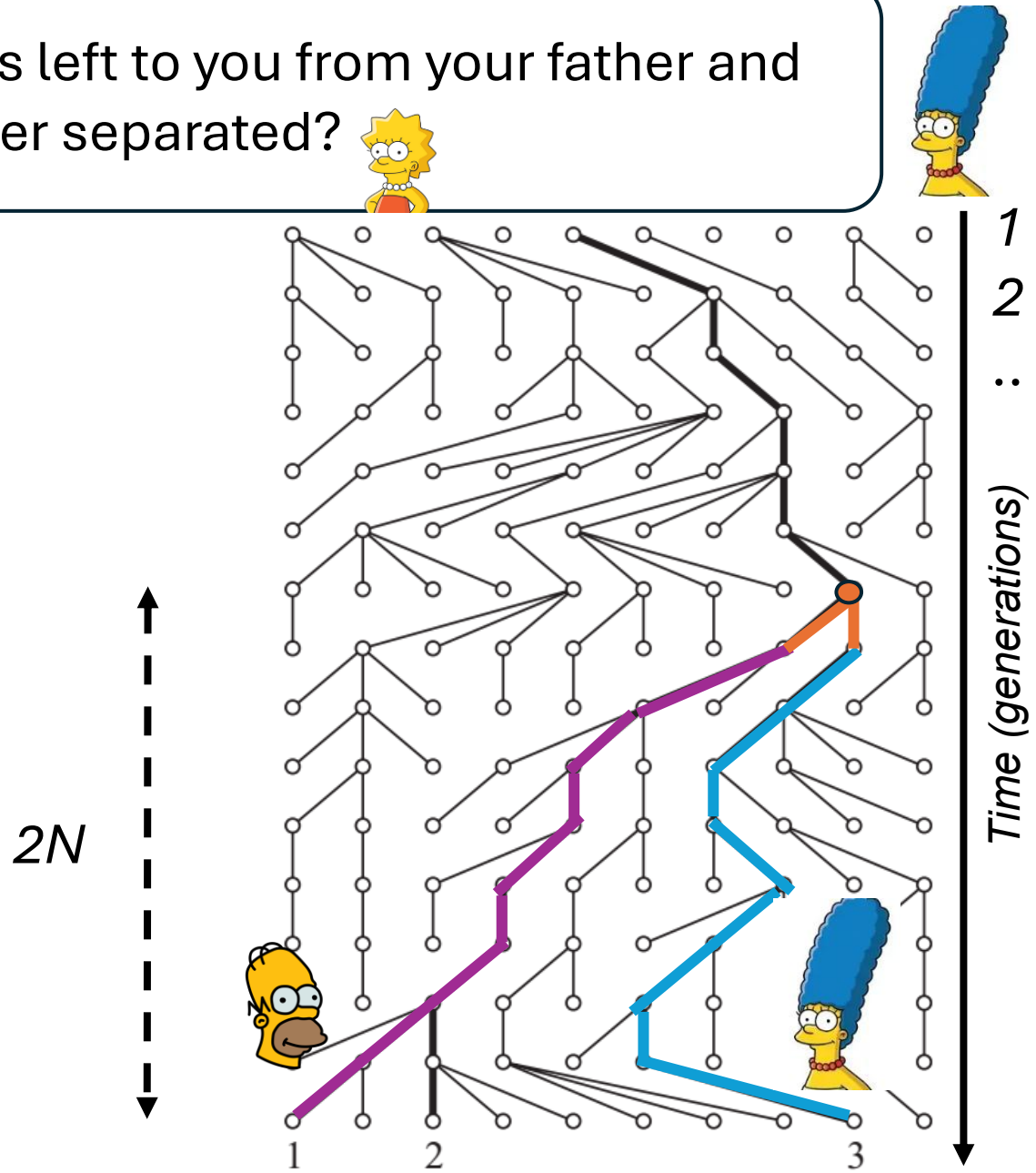
The time till a coalescent event for two lineages is geometrically distributed

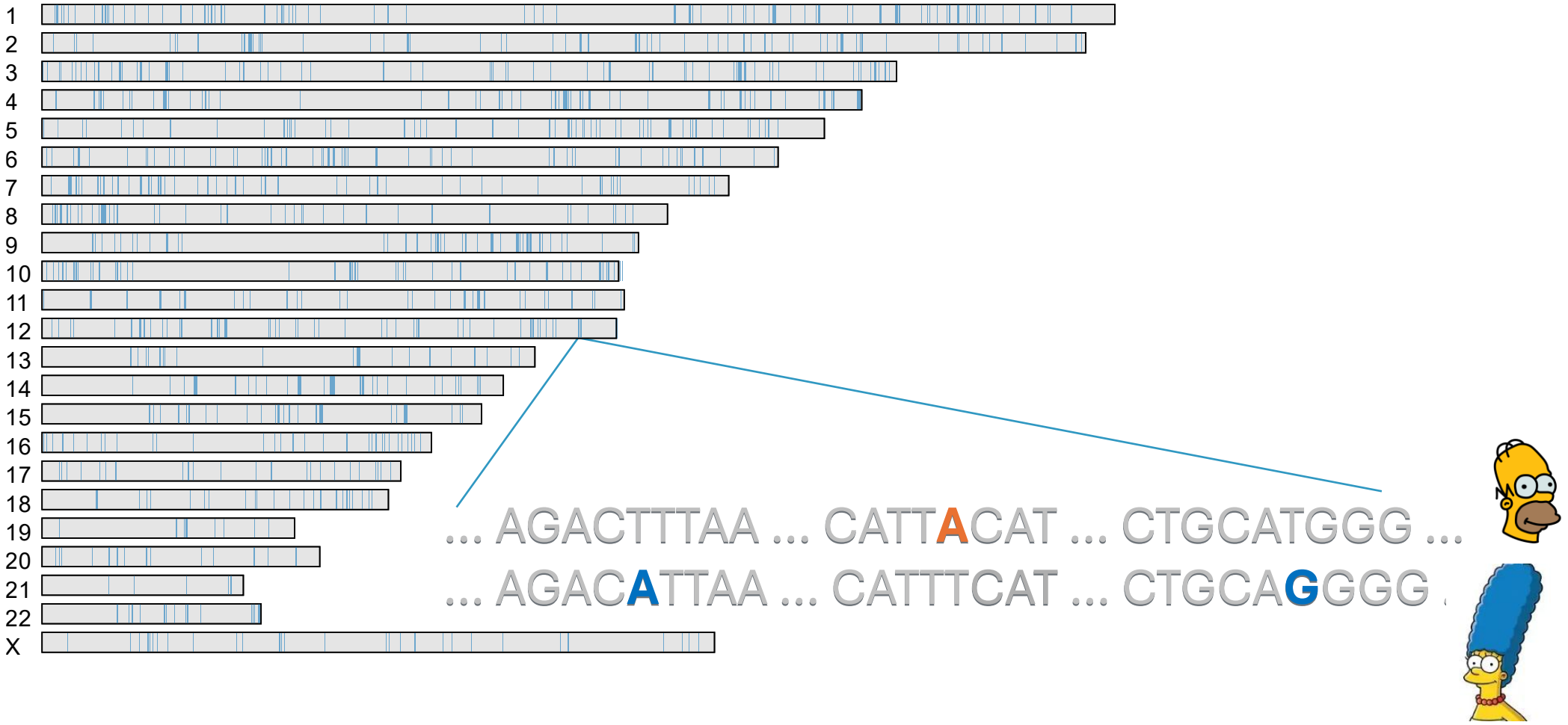
$$P(T_2 = j) = \left(1 - \frac{1}{2N}\right)^{j-1} \frac{1}{2N},$$

The geometric distribution has mean 2N, thus on average two «alleles» in a population coalesce 2N generations ago



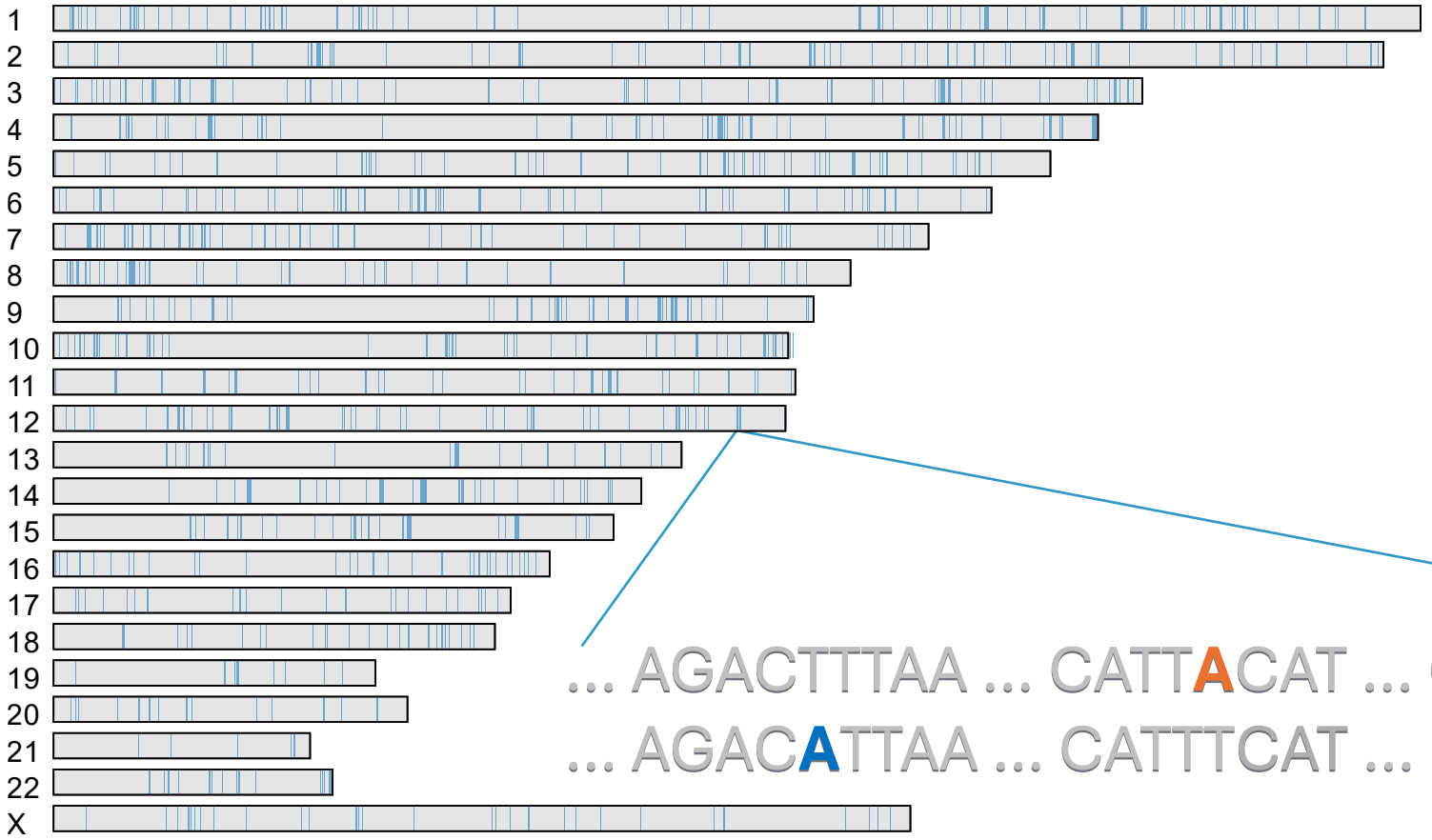
How long ago two alleles left to you from your father and mother separated?





If we take two of your alleles (the one given by your father and your mother) separated about 700.000 years ago

(2 N_{Humans}*years per generation = 2*12000*29years)



... AGACTTTAA ... CATTACAT ... CTGCATGGG ...

... AGACATTAA ... CATTTCAT ... CTGCAGGGG ...

Why is N so small, aren't we 9 billions?

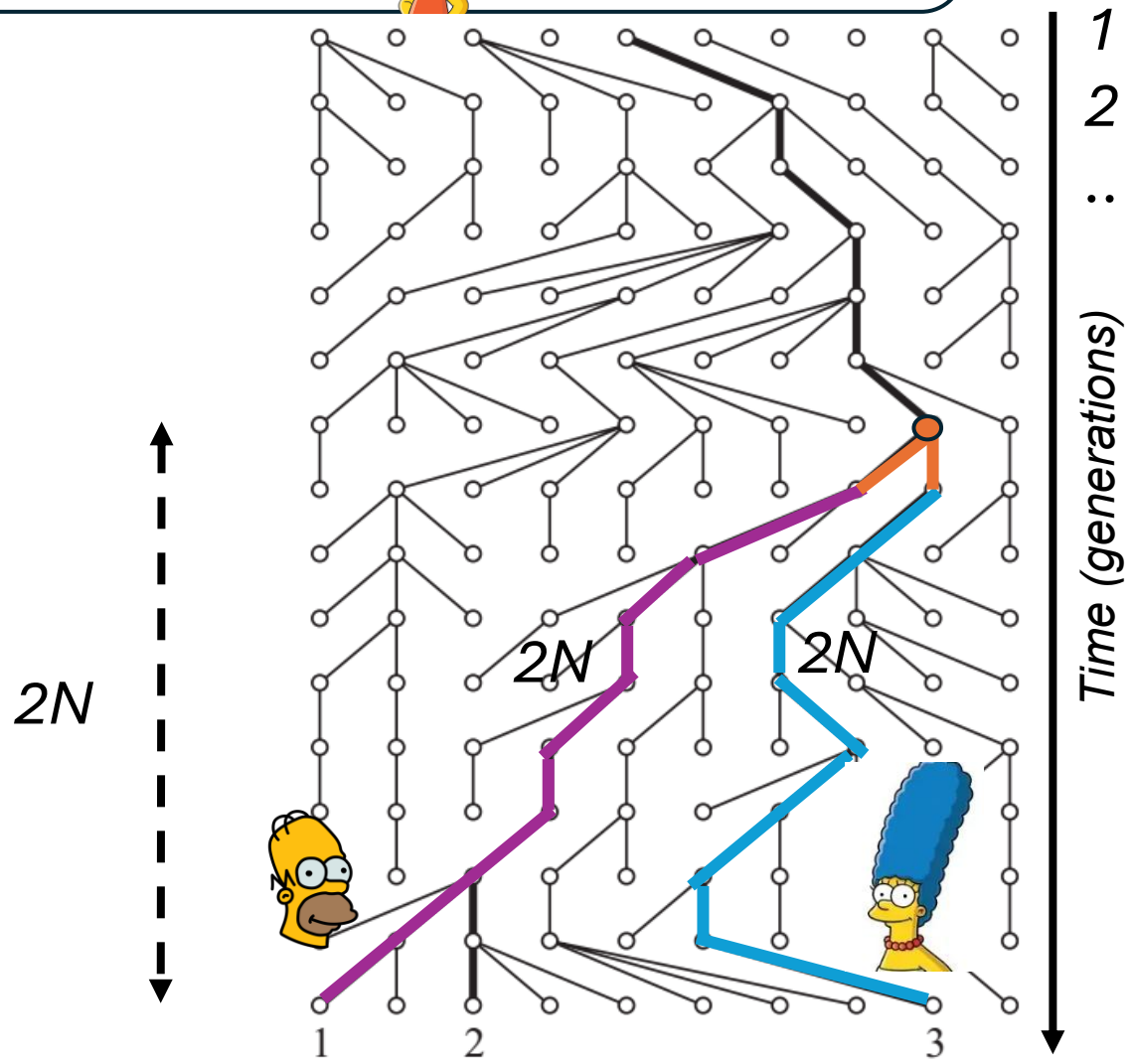
If we take two of your alleles (the one given by your father and your mother) separated about 700.000 years ago

(2 N_{Humans}*years per generation = 2*_**12000**_*29years)

How similar are the «chromosomes» left to you from your father and mother?

Time to accumulate mutations (diverge from each other):

$(2N+2N)$



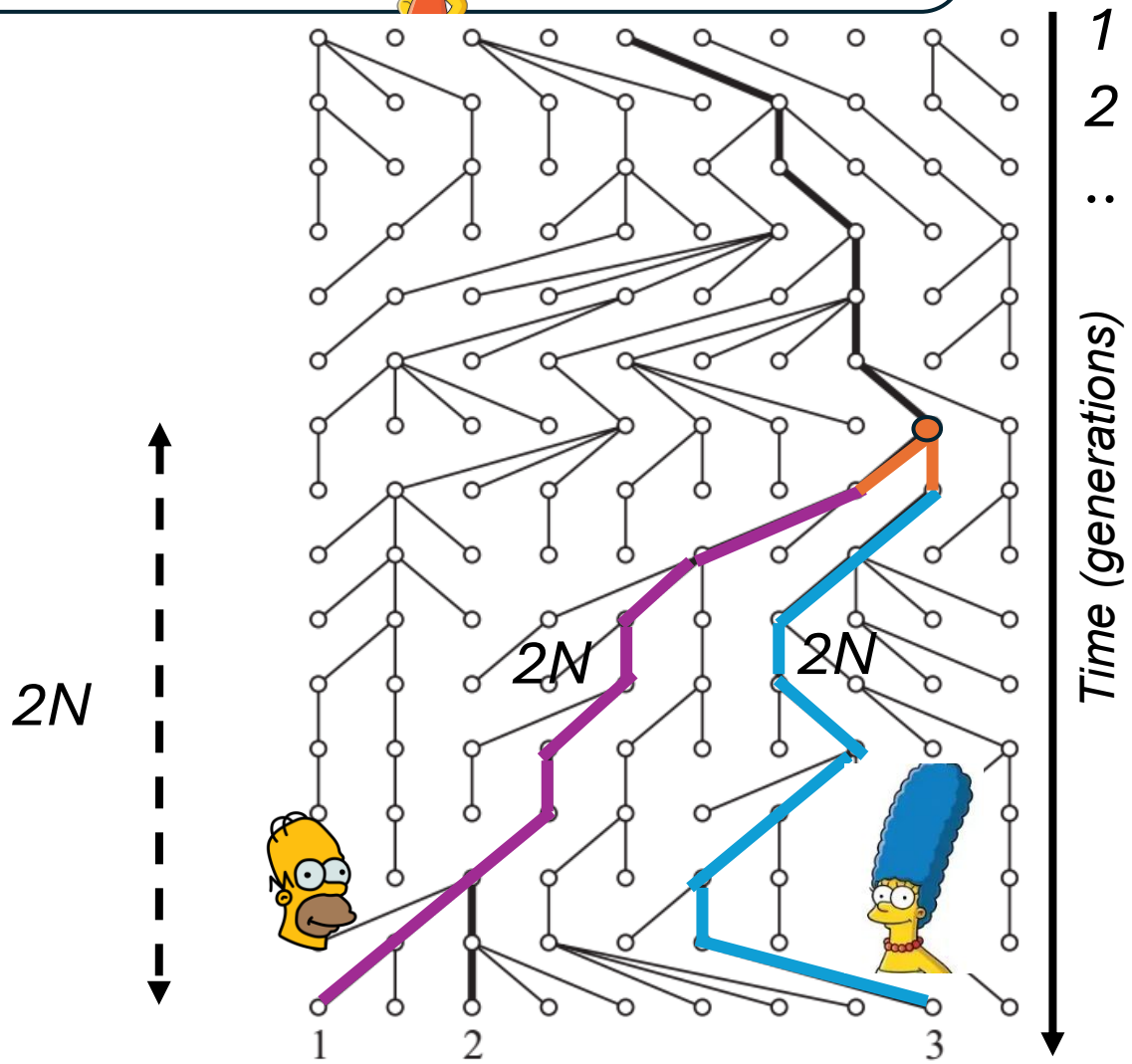
How similar are the «chromosomes» left to you from your father and mother?

Time to accumulate mutations (diverge from each other):

$(2N+2N)$

Mutation rate (n.mutations per site per generation)

$\mu = 4N\mu$



How similar are the «chromosomes» left to you from your father and mother?

Time to accumulate mutations (diverge from each other):

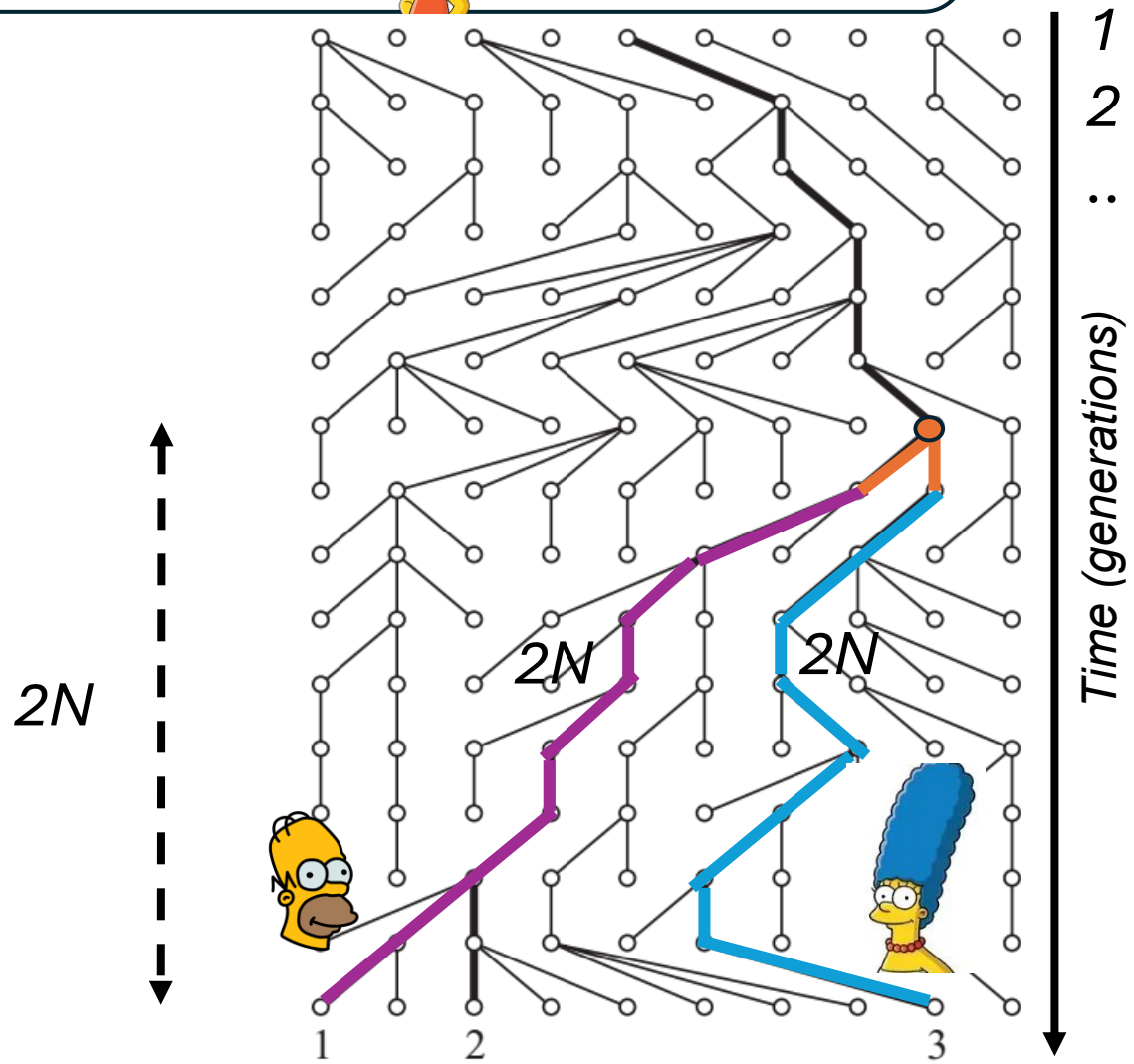
$(2N+2N)$

Mutation rate (n.mutations per site per generation)

$\mu = 4N\mu$

In population genetics, we refer to 4Nμ as theta:

$4N\mu=\theta$



How similar are the «chromosomes» left to you from your father and mother?

Tim o accumulate mutations (erge from each other):

2N+2N)

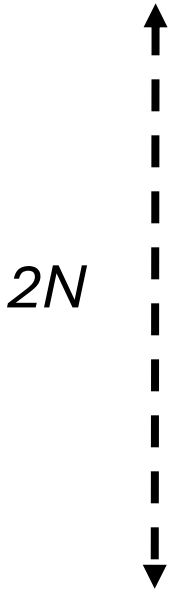
Mutation rate (n.mutations per site per generation)

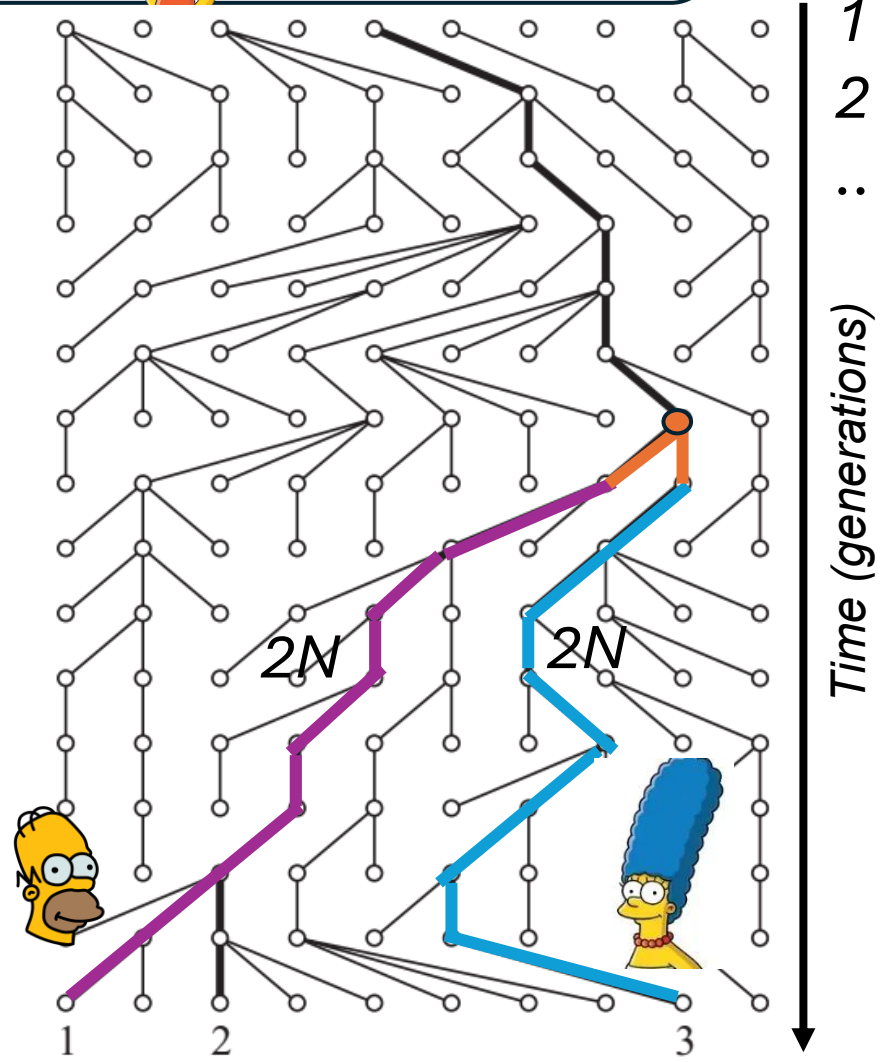
$\mu = 4N\mu$

Note that **θ** represents thus the ***expected heterozigosity***/difference between random alleles

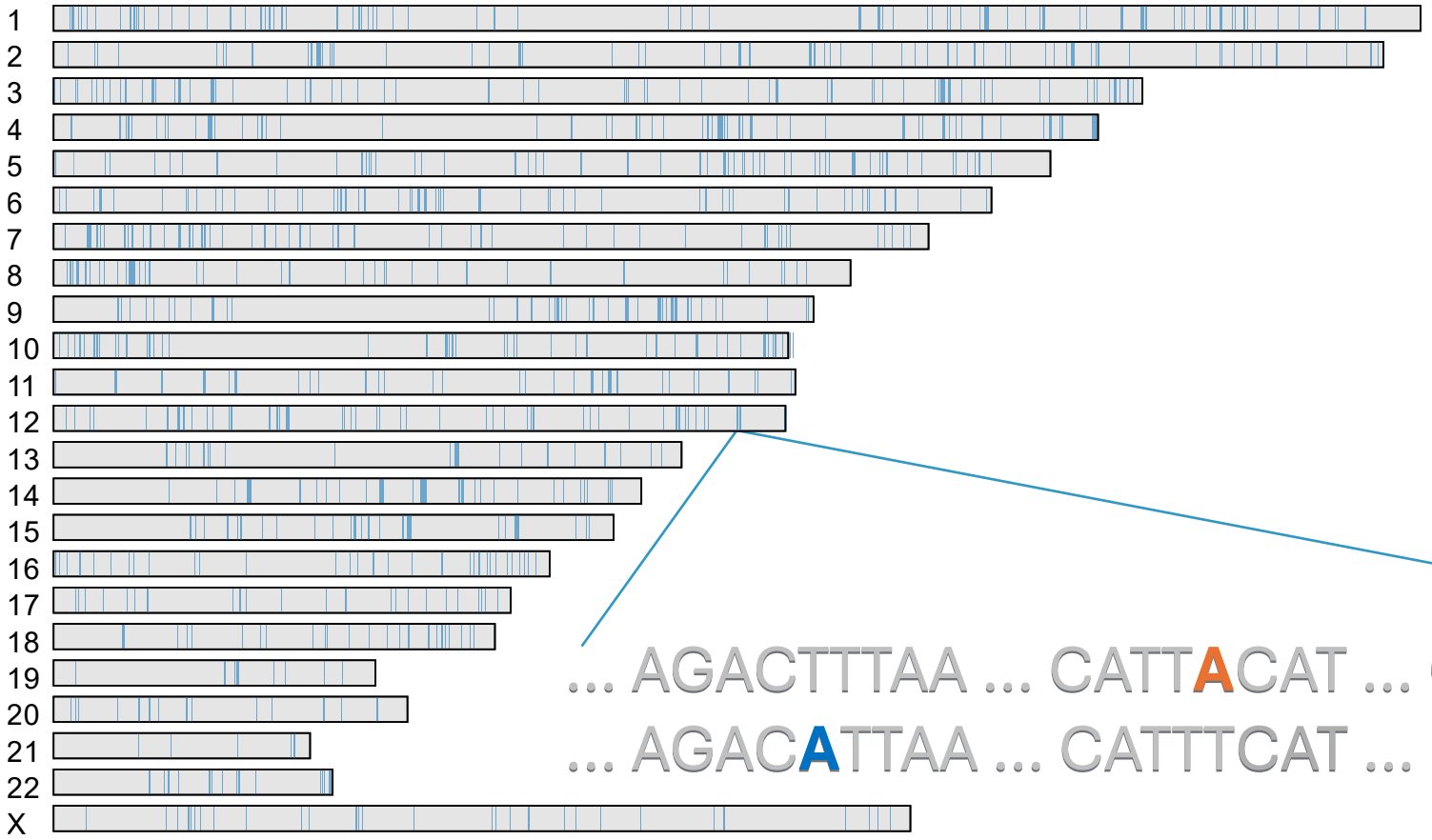
In population genetics, we refer to 4Nμ as theta:

$4N\mu=\theta$





Heterozigosity and *Ne* in humans



... AGACTTTAA ... CATTACAT ... CTGCATGGG ...

... AGACATTAA ... CATTTCAT ... CTGCAGGGG ...

Since we can measure θ and μ, we can calculate N solving N=θ/4 μ

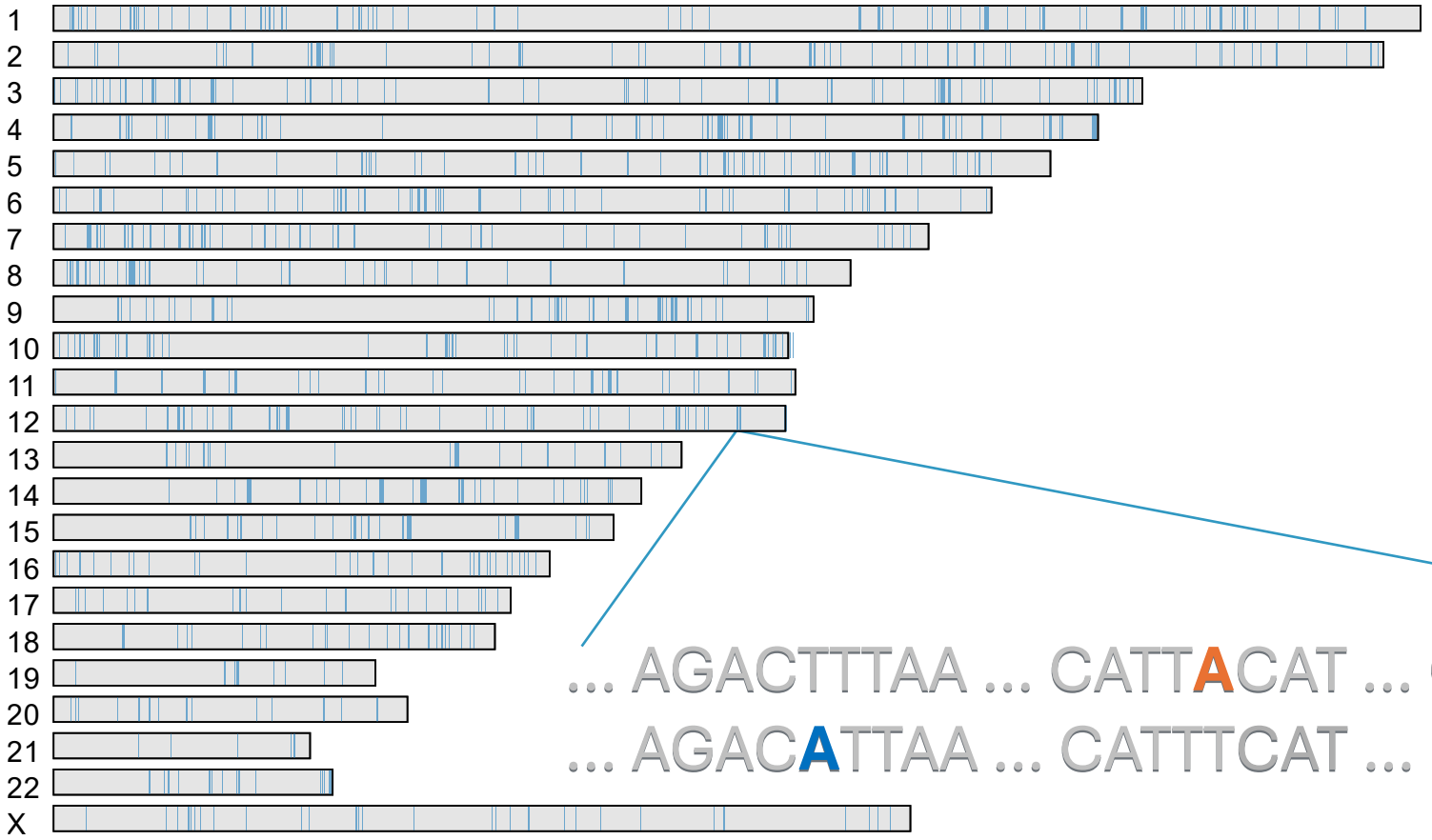
$$\theta = 4N\mu$$

In non-Africans, θ≈0.0006 and

$\mu = 1.2*10^{-8}$

$Ne = 0.0006/(4*1.2*10^{-8}) = 12500$

Heterozigosity and *Ne* in humans



... AGACTTTAA ... CATTACAT ... CTGCATGGG ...

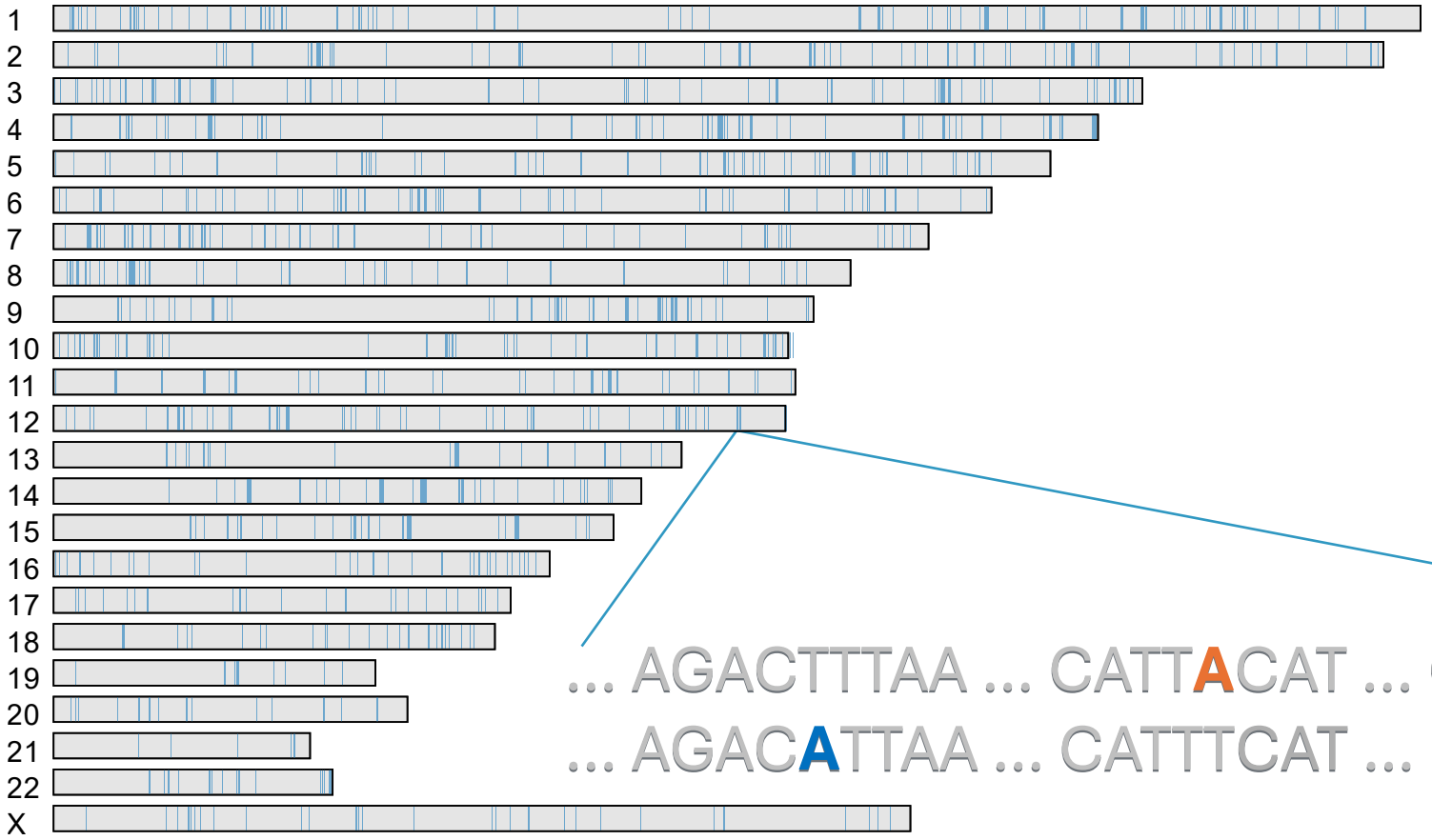
... AGACATTAA ... CATTTCAT ... CTGCAGGGG ...

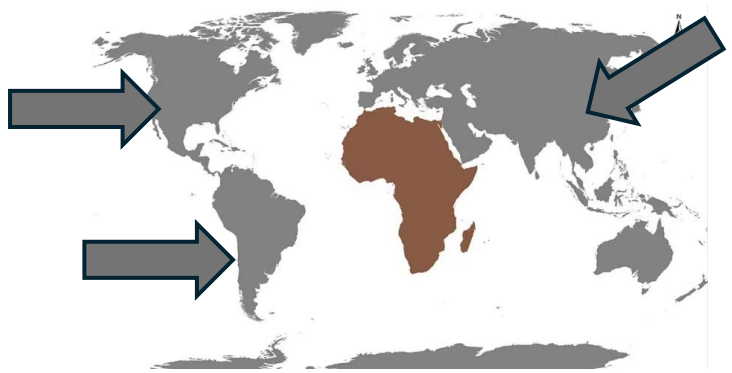
Why is N so small, aren't we 9 billions?

$$\theta = 4N\mu$$

In non-Africans, $\theta \approx 0.0006$ and $\mu = 1.2 * 10^{-8}$

$$Ne = 0.0006/(4 * 1.2 * 10^{-8}) = 12500$$





... AGACTTTAA ... CATTACAT ... CTGCATGGG ...

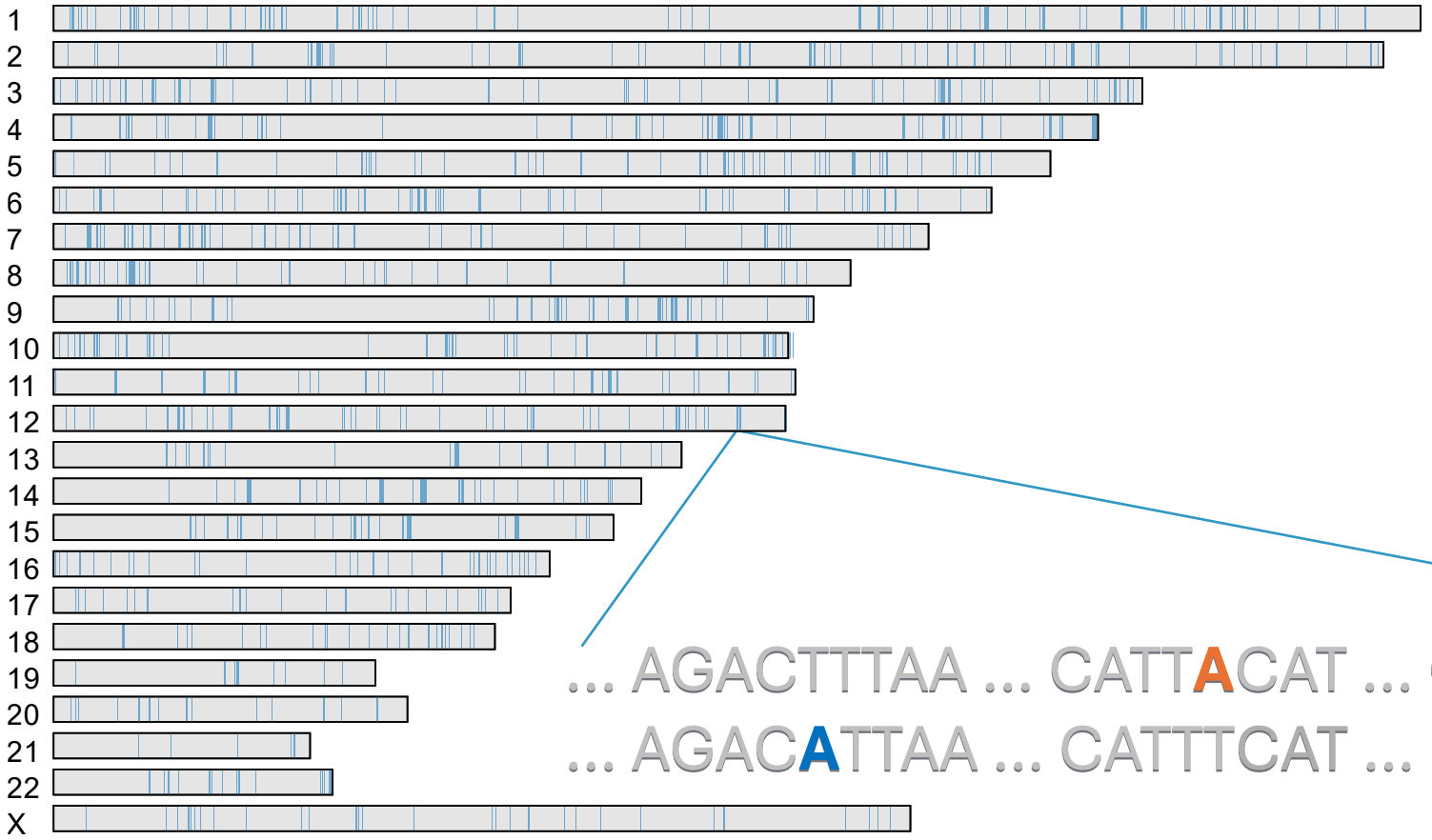
... AGACATTAA ... CATTTCAT ... CTGCAGGGG ...

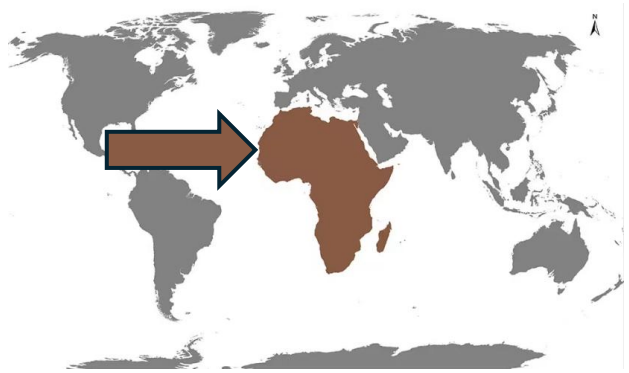
How come? There is more than 12.500 non-Africans!!

$$\theta = 4N\mu$$

In non-Africans, $\theta \approx 0.0006$ and $\mu = 1.2 * 10^{-8}$

$$Ne = 0.0006/(4 * 1.2 * 10^{-8}) = 12500$$





... AGACTTTAA ... CATTACAT ... CTGCATGGG ...

... AGACATTAA ... CATTTCAT ... CTGCAGGGG ...

How come? More Africans than everybody else? But also they are more!

$$\theta = 4N\mu$$

In Africans, $\theta \approx 0.001$ and $\mu = 1.2 * 10^{-8}$

$$Ne = 0.001/(4 * 1.2 * 10^{-8}) = 20833$$

Which factors cause a population to have high or low diversity/heterozygosity?

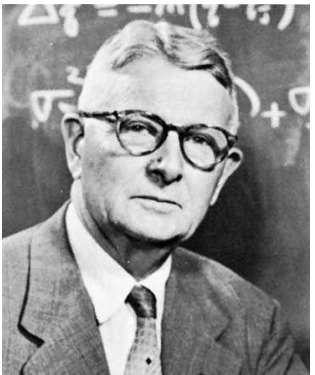
Which factors cause a population to have high or low diversity/heterozygosity?

- Genetic drift
- Size of the census population
- Natural selection
- Mutation rates
- Subdivision in subpopulations
- .....

The effective population size (or Ne)

The **effective population size of a population, *Ne,*** refers to the size of an idealized Wright-Fisher population (panmictic, no selection, etc.) that would show the same amount of genetic drift (or related effects) to the real population.

This quantity differs from the censun population (***N***, the real number of individuals) since real population do not usually fulfill all properties of idealized ones.



Sewall Wright

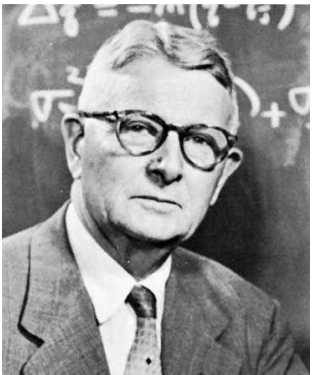
Motoo Kimura

The effective population size (or Ne)

The **effective population size of a population, *Ne,*** refers to the size of an idealized Wright-Fisher population (panmictic, no selection, etc.) that would show the same amount of genetic drift (or related effects) to the real population.

This quantity differs from the censun population (***N***, the real number of individuals) since real population do not usually fulfill all properties of idealized ones.

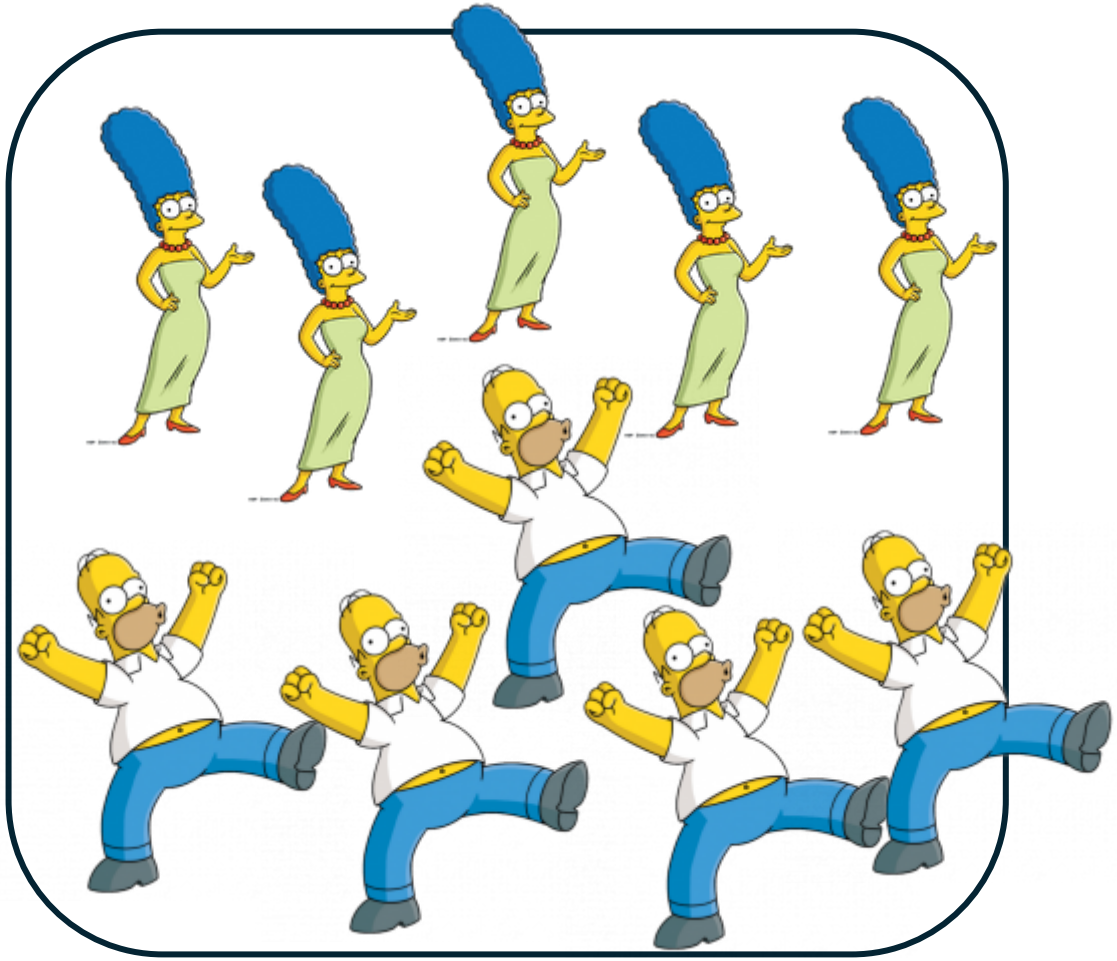
Note that we can quantify genetic drift in different ways (e.g. heterozigosity, coalescent times) and thus we can define Ne in slightly different ways for a population (which are the same only in idealized cases). However, they are usually very similar. Now I will refer to inbreeding/coalescent Ne unless stated differently.



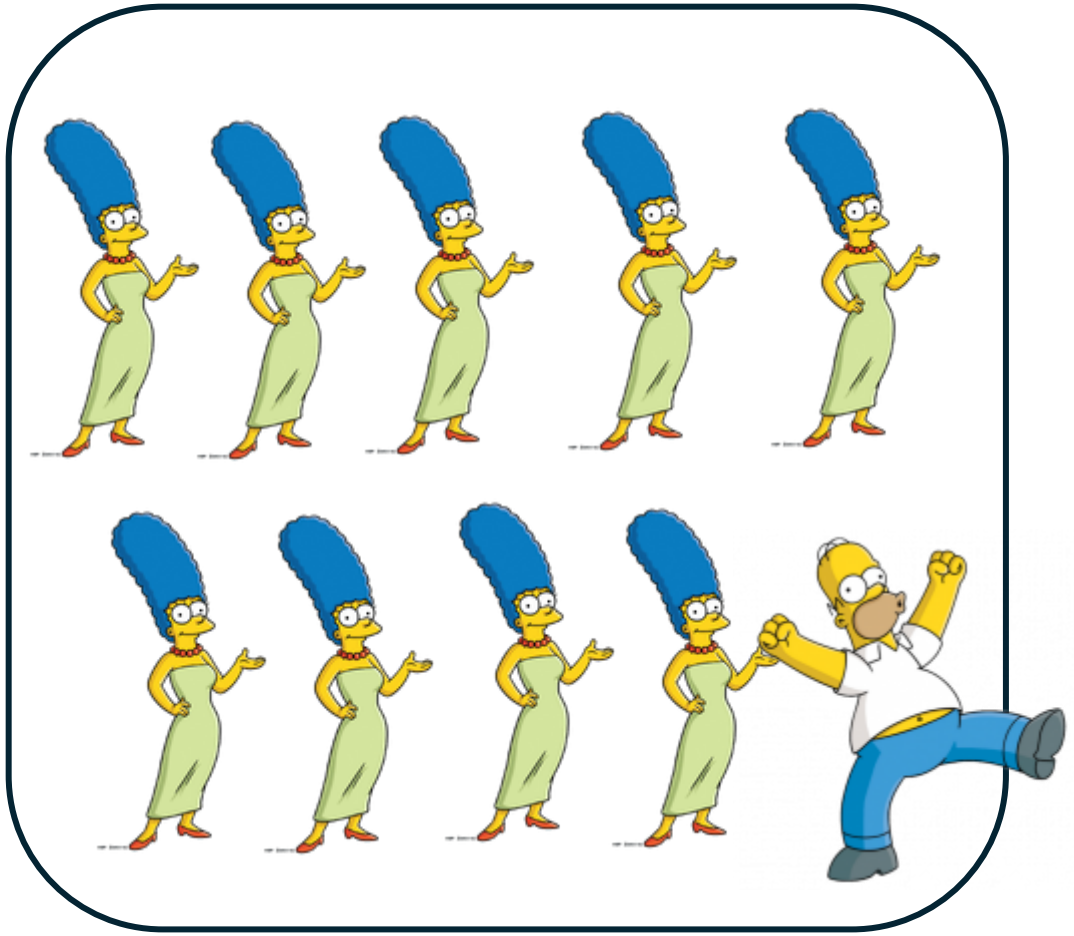
Sewall Wright

Motoo Kimura

An example showing how *N* and *Ne* differ: sex-ratio effect on *Ne*



A



B

Sex-ratio effect on Ne

$$Ne = 4 (N_{ef} \times N_{em})/(N_{ef} + N_{em})$$

Ne is higher when the number of males and females is balanced. The explanation is that when one of the two sexes is rarer, it becomes the limiting factor for the diversity of the population.

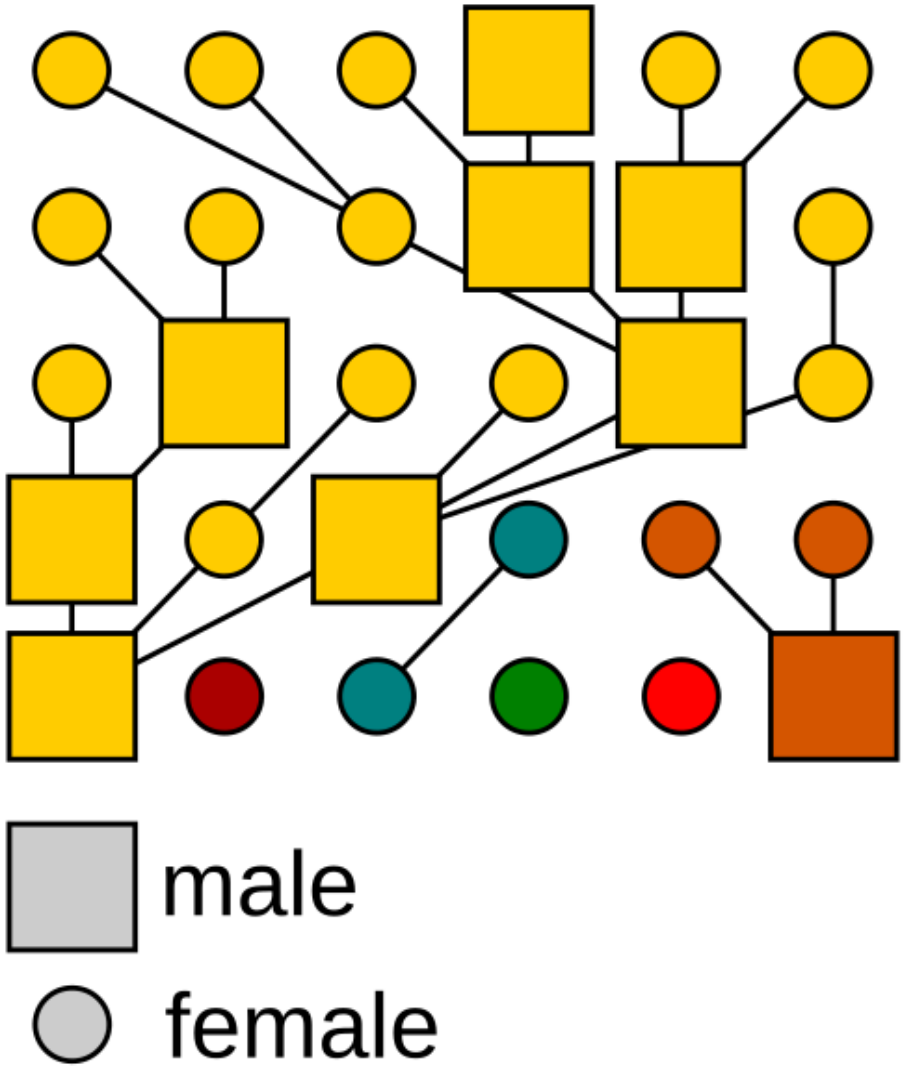
Mirounga angustirostris

$$N_e = \frac{4(105 \times 28)}{105 + 28} = 88$$

Effective population size (N_e) and *census* size (N)

Why do *family sizes* vary more than randomly?

- **Uneven sex ratio**
  - Members of the rare sex have more reproductive opportunities than members of the abundant sex



Variation in reproductive success and Ne

$$N_e = \frac{4N_c - 2}{\mathrm{VRS} + 2}$$

The variance in reproductive success reduces Ne since only a limited number of generations/individuals contribute disproportionally to the population

Geospiza fortis

$$N_e = \frac{4(500) - 2}{7.12 + 2} = 219$$

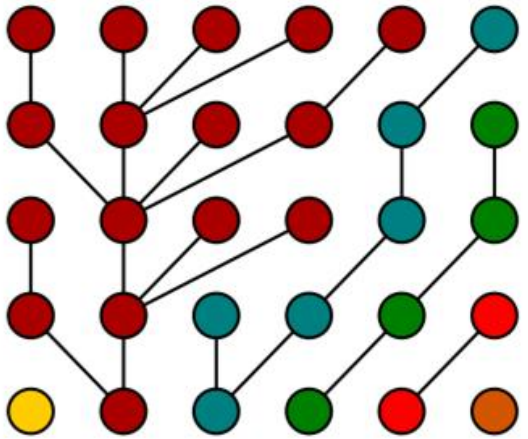
Effective population size (N_e) and *census* size (N)

N/N_e depends on
variation in *family size*

(number of children per parent)

$$N_e = N$$

Random reproduction



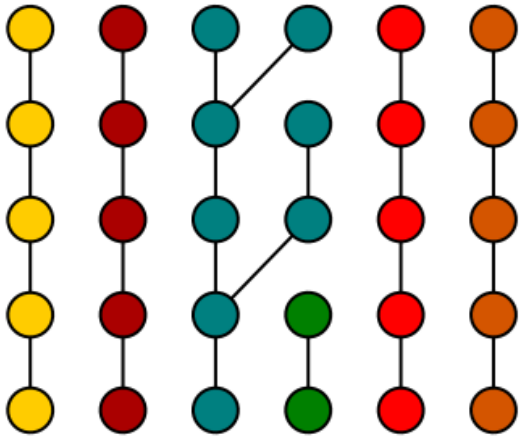
Effective population size (N_e) and *census* size (N)

N/N_e depends on
variation in *family size*

(number of children per parent)

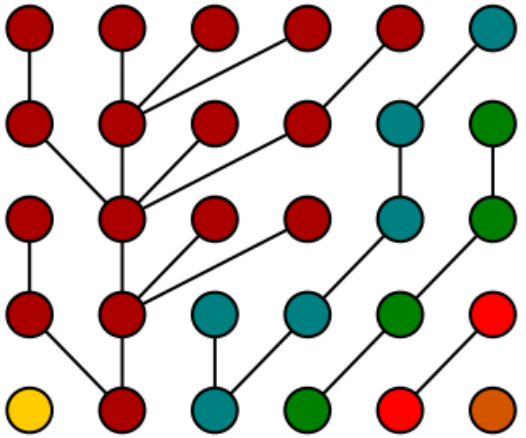
$N_e > N$

Uniform reproduction



$N_e = N$

Random reproduction



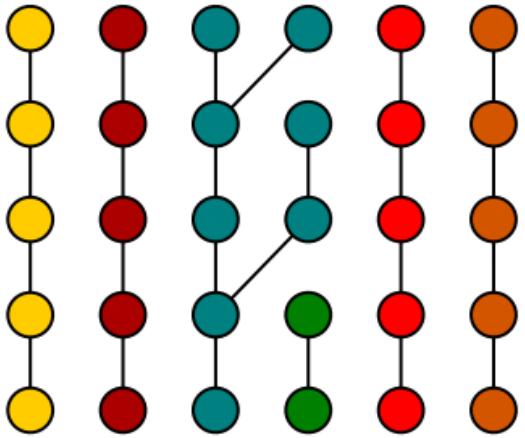
Effective population size (N_e) and *census* size (N)

N/N_e depends on
variation in *family size*

(number of children per parent)

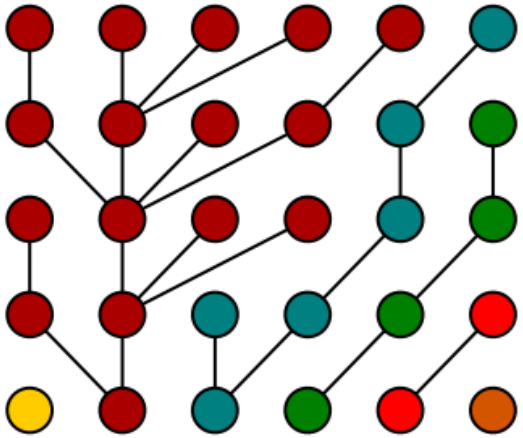
$N_e > N$

Uniform reproduction



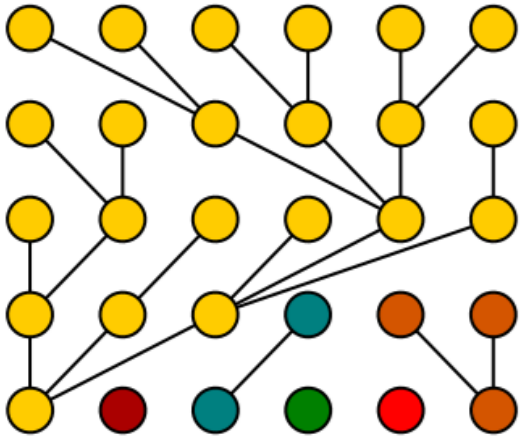
$N_e = N$

Random reproduction



$N_e < N$

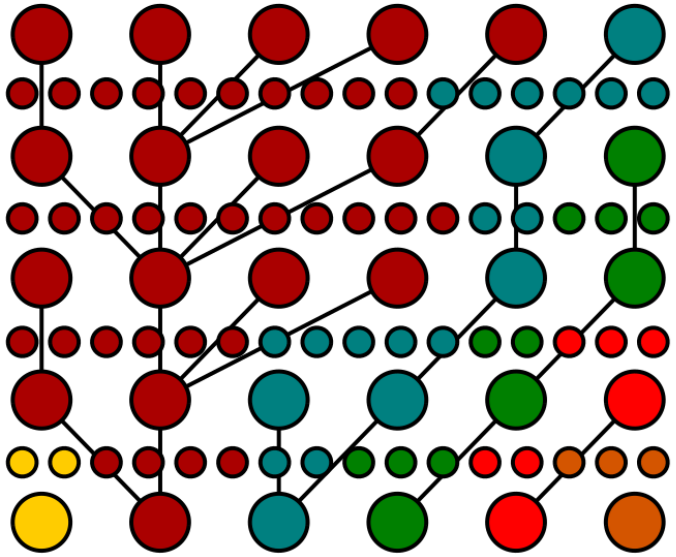
Uneven reproduction

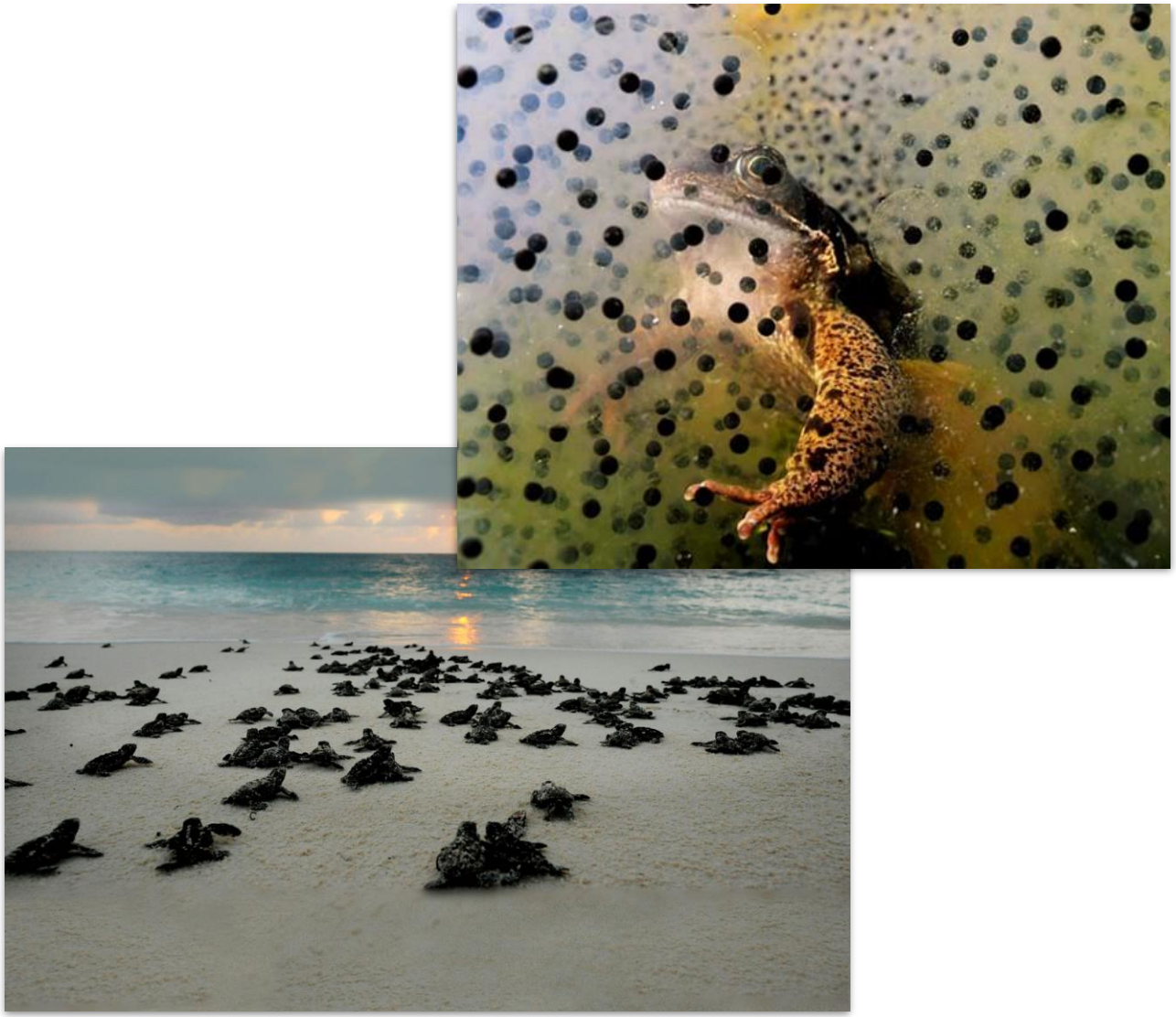


Effective population size (N_e) and *census* size (N)

Why do *family sizes* vary more than randomly?

- Pre-reproductive mortality

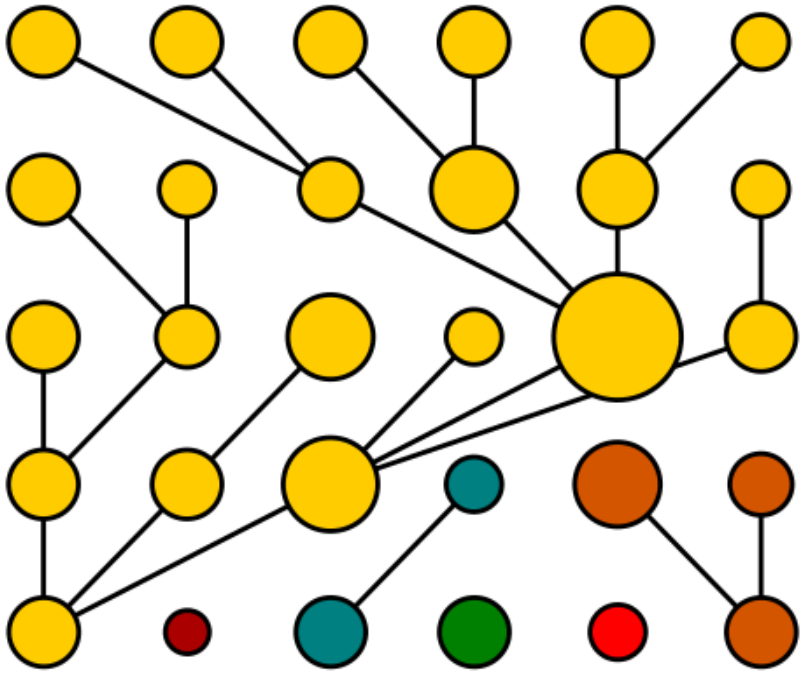




Effective population size (N_e) and *census* size (N)

Why do *family sizes* vary more than randomly?

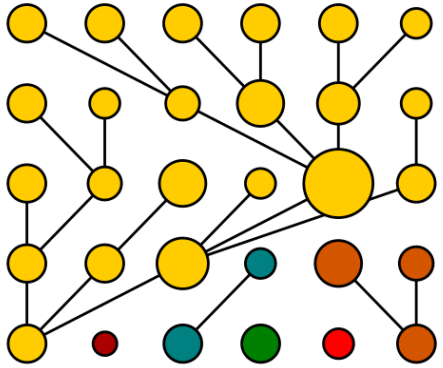
- Overlapping generations
  - Individuals that live through more reproductive “seasons” are expected to leave more children



Effective population size (N_e) and *census* size (N)

Why do *family sizes* vary more than randomly?

- **Reproductive skew**
  - In many animals, a few individuals (especially males) tend to monopolize access to the other sex and produce a large share of offspring .



Fluctuations in Ne

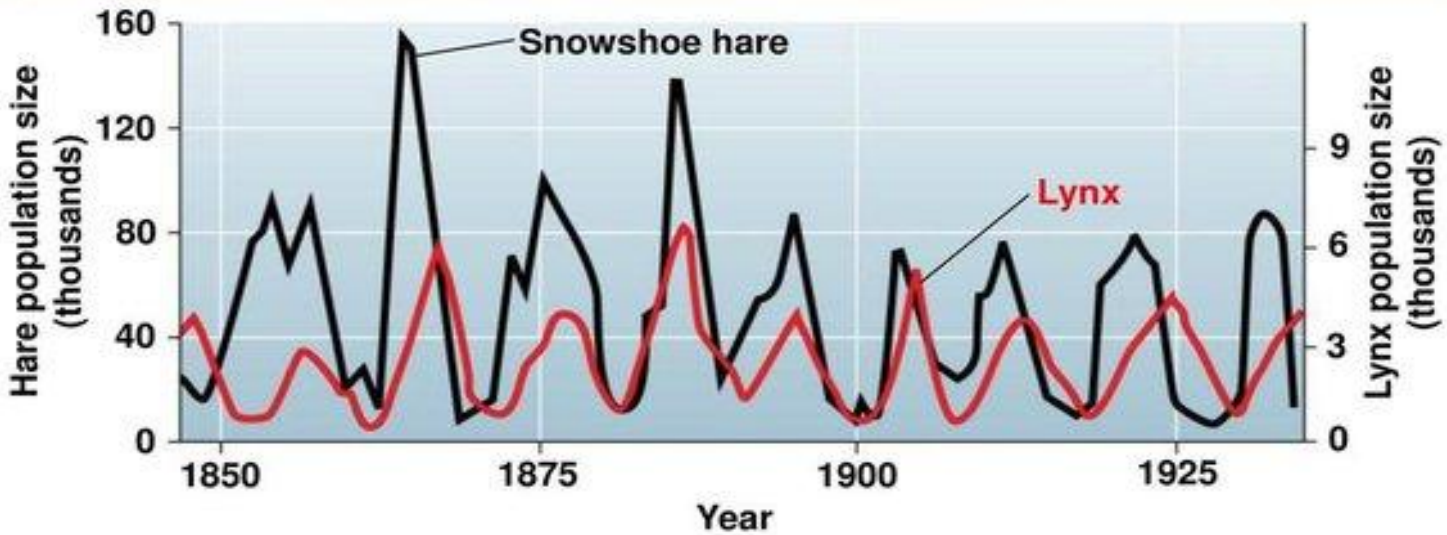
$$Ne = \frac{t}{1/N_{e1} + 1/N_{e2} + 1/N_{e3} + \ldots + 1/N_{et}}$$

Ne is about the harmonic mean of the effective population sizes over time.
In the harmonic mean the terms with lower population sizes matter the most.

$$N_e = \frac{4}{(1/220) + (1/70) + (1/40) + (1/200)}$$
$$= \text{approx. } 82$$

Fluctuations in Ne

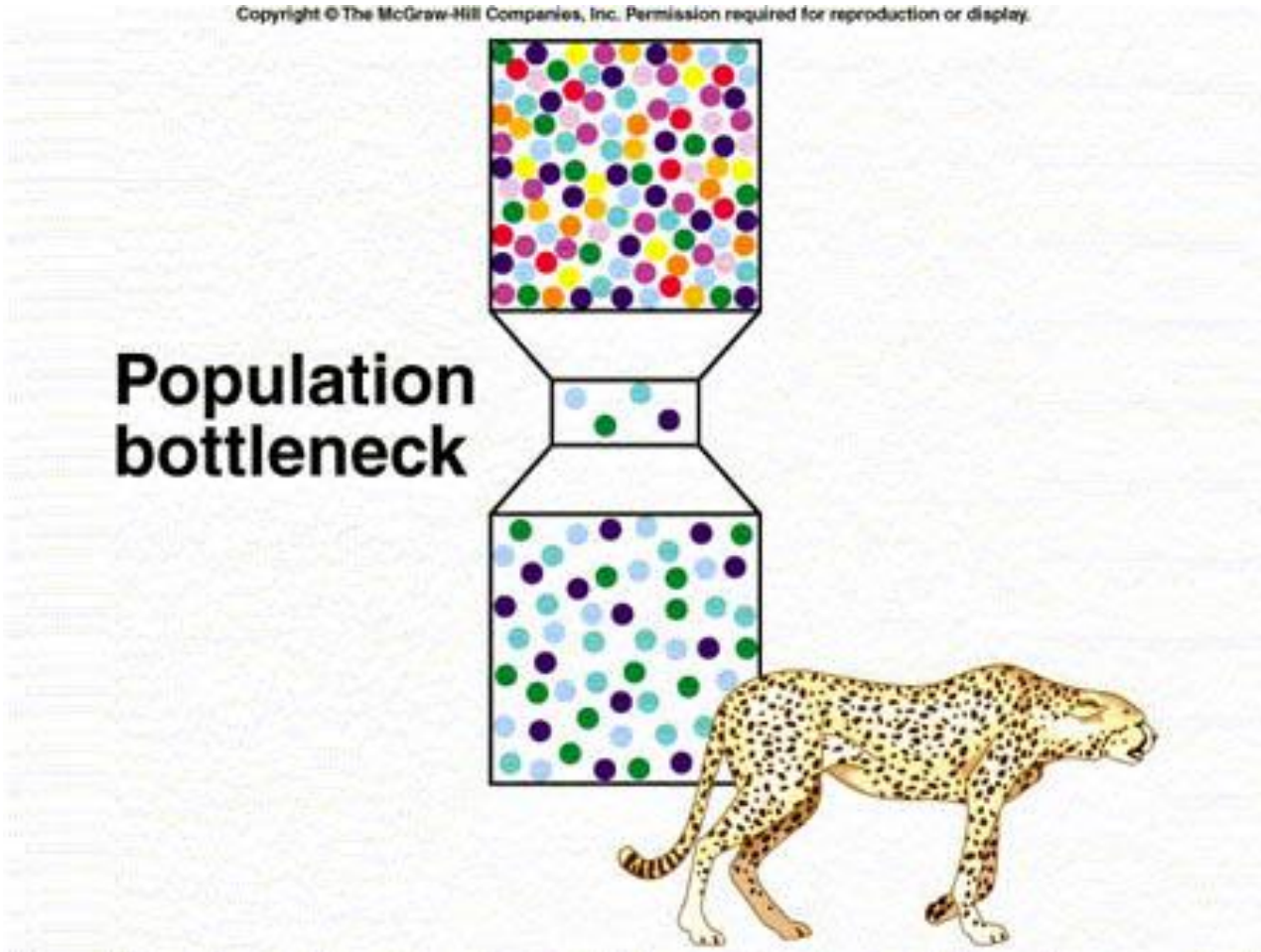
Figure 36.6-0



Data from C. Elton and M. Nicholson, The ten-year cycle in numbers of the lynx in Canada, *Journal of Animal Ecology* 11 : 215–244 (1942).

In general Ne < N

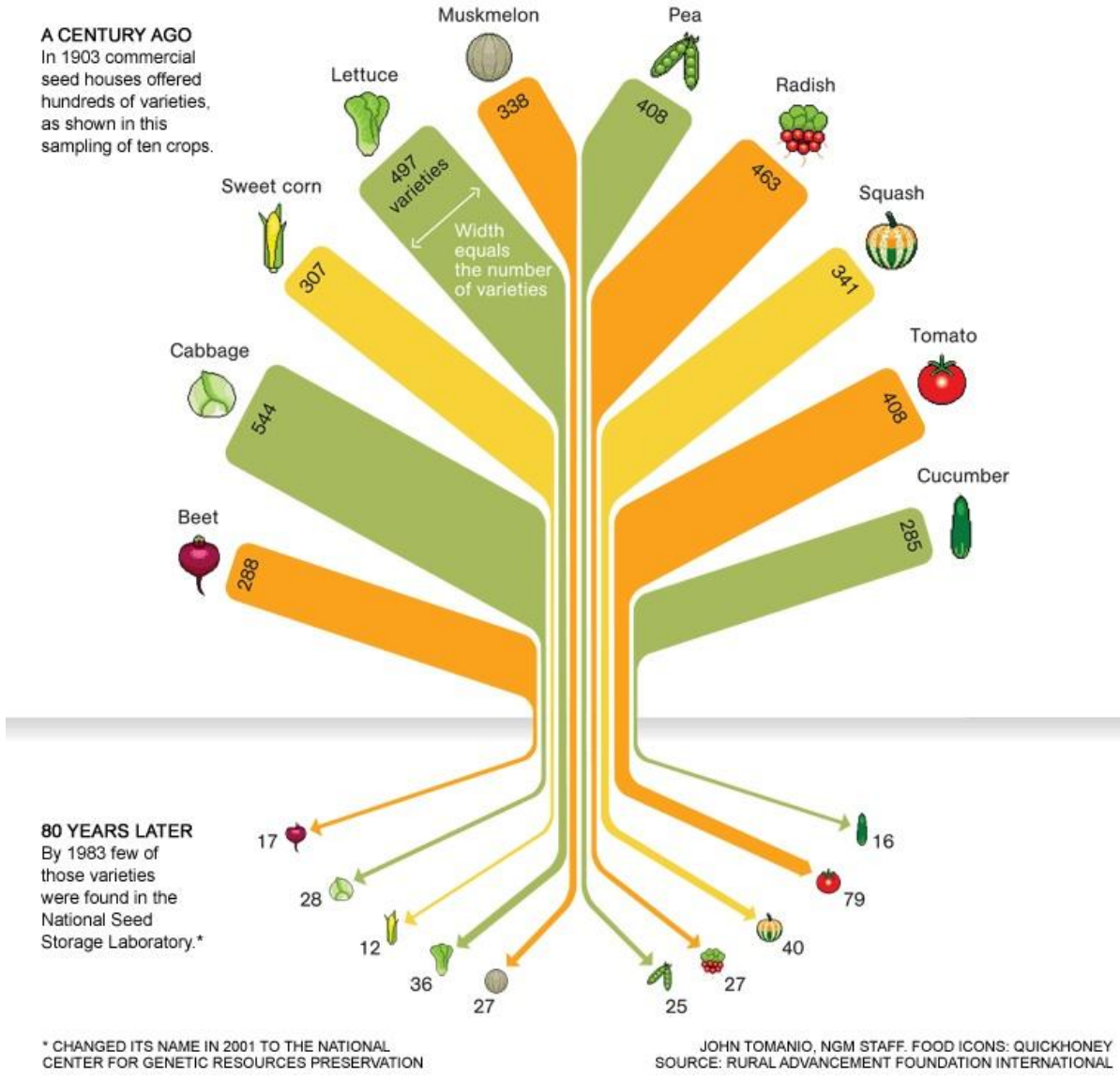
Demographic bottlenecks



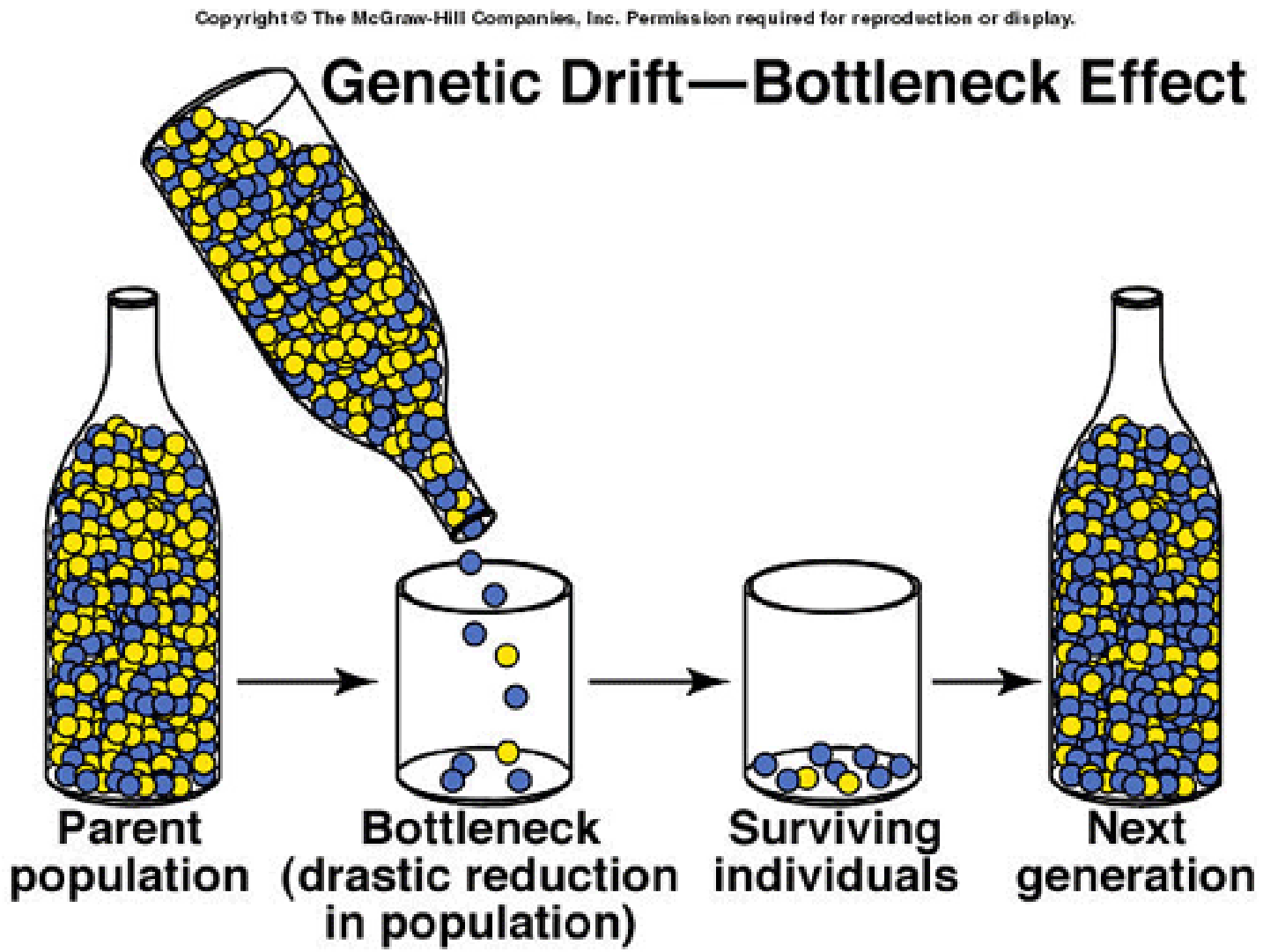
Founder effect

Tay-Sachs in Ashkenazi jews

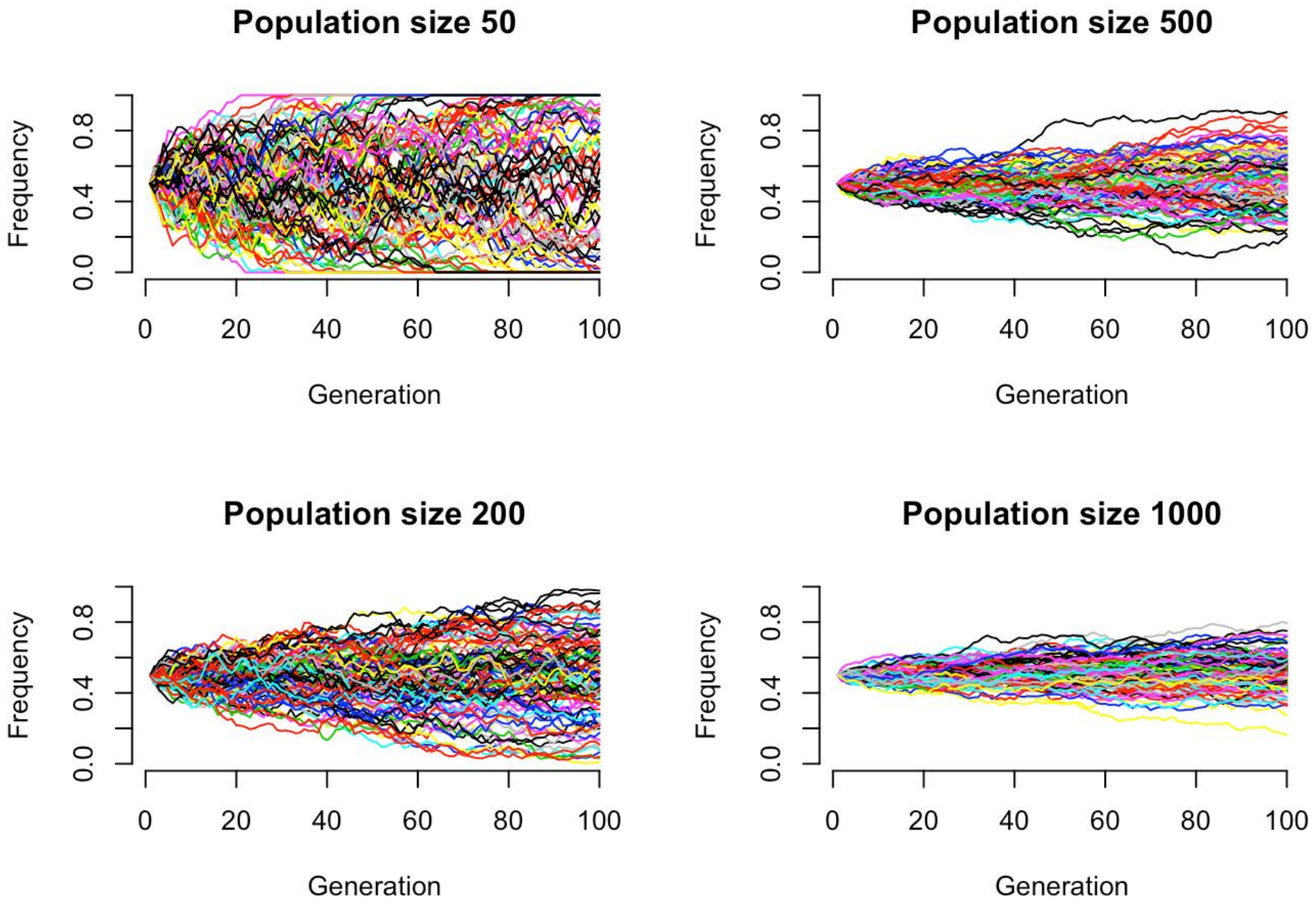
Loss of advantageous variation in domesticated species



Demographic bottlenecks



Genetic drift is stronger in small populations



Conclusions

- We can represent random genetic drift with the Wright-Fisher model, which capture ideal populations reproducing randomly
- Using the Wright-Fisher model we can infer the probability of fixation and extinction of new mutations
- Coalescence theory allows us to infer the time to the most recent common ancestor of two sequences
- The concept of effective population size captures deviations of real populations from the Wright-Fisher model
- Usually population bottlenecks, unequal sex ratios and variance in reproductive effects leads to $\boldsymbol{Ne} < N$