

Lecture 1: Forces shaping genetic diversity:
mutation, drift, (migration) selection

Lecture 2: Introduction to coalescent theory
(practicals)

Anna-Sapfo Malaspinas (annasapfo.malaspinas@unil.ch)

Outline

Lecture 1

- Mutation
- Genetic drift and effective population size
- Selection

Lecture 2

- Coalescent theory and the site frequency spectrum

Typical population genetic questions

- What causes populations to change over time?
- What causes populations to diverge from each other?
- How much and what kind of variation exists in (natural) populations?
- What processes generate/preserve/destroy this variation?
- Is evolution largely governed by deterministic or stochastic effects?

What is evolution?

Change over time

What is evolution?

More specifically here : change in allele frequency

Four forces of evolution

- **Mutation** : changes in allele frequencies via the generation of (new) variants
- **Genetic drift** : changes in allele frequencies via stochastic fluctuations inherent to finite populations
- **Migration** : changes in allele frequencies via the introduction of migrant alleles
- **Selection** : changes in allele frequencies via the fitness effect (i.e., more or fewer offspring) of mutations

Four forces of evolution

- **Mutation** : changes in allele frequencies via the generation of (new) variants
- **Genetic drift** : changes in allele frequencies via stochastic fluctuations inherent to finite populations
- [**Migration** : changes in allele frequencies via the introduction of migrant alleles]
- **Selection** : changes in allele frequencies via the fitness effect (i.e., more or fewer offspring) of mutations

Mutation as a source of variation

- “Ultimate” source of variation
- Genetic drift and selection usually act to reduce variation
- Without mutation, there is no molecular evolution
- Without mutation, there is no variation on which selection may act

N.B: You may say migration can introduce new variants, but those were also generated via mutation

What types of mutation at the molecular level?

What types of mutation at the molecular level?

- Insertion
- deletion
- duplication
- inversion
- translocation
- **nucleotide substitution/point mutation**
 - **NB: alleles: A, C, G and T**

Mutational effects

- Deleterious mutation: reducing fitness
 - A deleterious mutation can have a selection coefficient $s < 0$
- Beneficial mutation: increasing fitness
 - A beneficial mutation can have a selection coefficient > 0
- Neutral mutation: no impact on fitness
 - A neutral mutation has a selection coefficient $s = 0$

How to measure μ , the mutation rate

A new mutation can arise at a given rate that we term μ .

- **Experimentally:** in viruses, yeast, fruit flies: inbred lines (all same genomes) and then count the number of mutations.
- **Pedigrees:** sequence parents and offspring and ask how many new mutations were not carried by either parent.
- **Evolutionarily/molecular clock:** if the time of separation is known, the number of differences between species is a direct estimate of mutation rate (i.e., the molecular clock).

Mutation-selection balance

- Mutation rate (and population size) alone does not dictate how much variation we observe in a population. We have to consider selective constraints.
- For example, lethal mutations are not fully sampled in a population. We under sample lethal mutations.
- There is an expected equilibrium of the input of mutations vs the elimination of deleterious variation by selection : «mutation-selection» balance.

Interplay of mutation and recombination

- While mutation is the source of variation, recombination (and reassortment) may shuffle this variation into different configurations
 - even though recombination is not creating new variation, it is creating new combinations of variation
- Recombination may uncouple beneficial and deleterious mutations, allowing selection to act more efficiently

Is recombination an evolutionary force?

Debatable. But here: remember that we defined evolution as the change in allele frequency.

Recombination does not change allele frequency; it re-arranges mutations.

How to measure the recombination rate ρ

- Experimentally: For example, genotype markers in parents and offspring, and count the number of crosses between those markers in one generation.
- Computationally: For example, utilizing observed patterns of linkage disequilibrium (that is, non-random associations between mutations in the genome) to infer how much recombination is likely occurring.

So, μ and ρ generate variation, but they act very slowly in terms of evolution

- One of the key take-aways: both mutation and recombination are critical in shaping the amount and distribution of variation in populations
- However, if only these two processes are operating, evolution moves slowly

Genetic drift

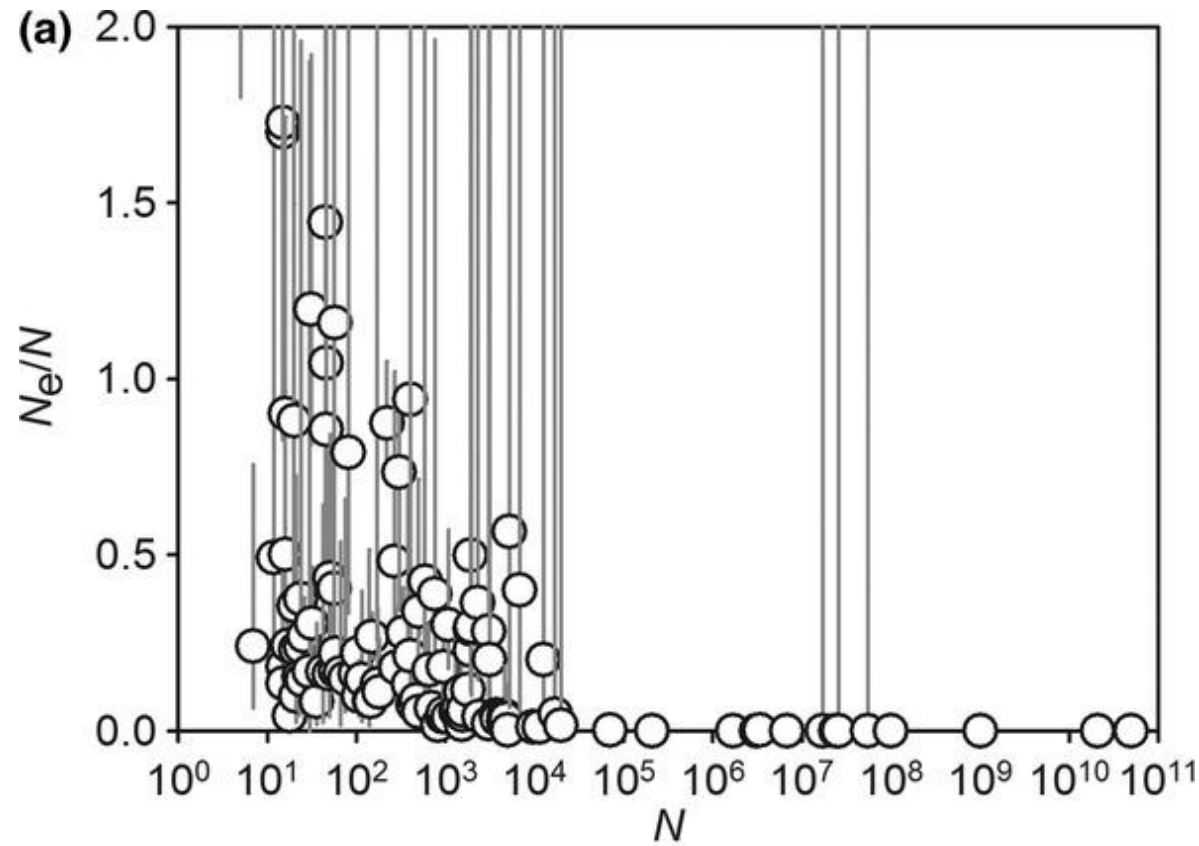
Genetic drift: random change in allele frequencies through time owing to the fact that populations are of finite size

- Often modelled as a binomial sampling process
 - Expectation of allele frequencies from one generation to the next remain the same
 - Variance can be large and thus deviations are common
 - Variance is larger in smaller populations
- With only genetic drift, the direction of change in allele frequency is random
- Drift will eventually eliminate or fix a mutation if no other forces are acting

Effect size and census size differ

- Effective population size: number of individuals that would result in the same level of genetic drift as is observed in the population
 - ⇒ You may think of it as the harmonic mean of the number of individuals contributing genetic material through time.
- The effective population size is equal to the census population size only in an idealized population:
 - Random mating
 - Constant population size
 - Discrete generations
 - Equal fitness in all individuals
 - Equal numbers of males and females (in sexual populations)

Effective and census size usually differ



Palestra and Fraser 2012

Why do the effective and census size usually differ

- One reason is that most populations are characterized by changing population sizes over time
 - In the case of humans, large population sizes are only observed in very recent history, where in deeper time populations in the dozens, hundreds or at most thousands were common
- However, there are other reasons - for example, highly skewed male/female ratios
 - In domestic cattle, where one male is often bred to produce almost the entirety of the next generation, the effective population size is greatly reduced compared to the total census size
- Selection (both positive and negative) also reduces effective population size, one of the reasons that N_e actually varies even across the genome of a population

Allele frequency change under genetic drift

- A key point to remember is that genetic drift is a more dominant force in small populations, and indeed in an infinitely large population, genetic drift can be neglected

Two nice results:

- Under genetic drift alone, the probability of fixation of a mutation is simply its current frequency.
 - If 10% of individuals in a population have a given mutation, that mutation has a 10% chance of ultimately fixing in the population
 - That also means that the probability of fixation of a newly arising neutral mutation is simply $1/2N_e$
- This underlies Kimura's molecular clock argument:
 - Rate of input * Prob of fixation = Rate of fixation
 $2N_e\mu * 1/2N_e = \mu$

The Wright-Fisher model

Assumptions:

- No migration
- Panmixia
- Here : constant N_e

Wright-Fisher model:

Model a population forward in time assuming discrete non overlapping generations

Individuals at every generation are obtained by sampling with replacement from the parents

NB. It is possible to incorporate selection and pop size changes in this model

time



The Wright-Fisher model

Assumptions:

- No migration
- Panmixia
- Here : constant N_e

Wright-Fisher model:

Model a population forward in time assuming discrete non overlapping generations

Individuals at every generation are obtained by sampling with replacement from the parents

NB. It is possible to incorporate selection and pop size changes in this model

time

Here: constant population of size 12

Generation 1



The Wright-Fisher model

Assumptions:

- No migration
- Panmixia
- Here : constant N_e

Wright-Fisher model:

Model a population forward in time assuming discrete non overlapping generations

Individuals at every generations are obtained by sampling with replacement from the parents

NB. It is possible to incorporate selection and pop size changes in this model

time

Here: constant population of size 12

Generation 1



A allele



a allele



$$p_1 = 1/12$$

The Wright-Fisher model

Assumptions:

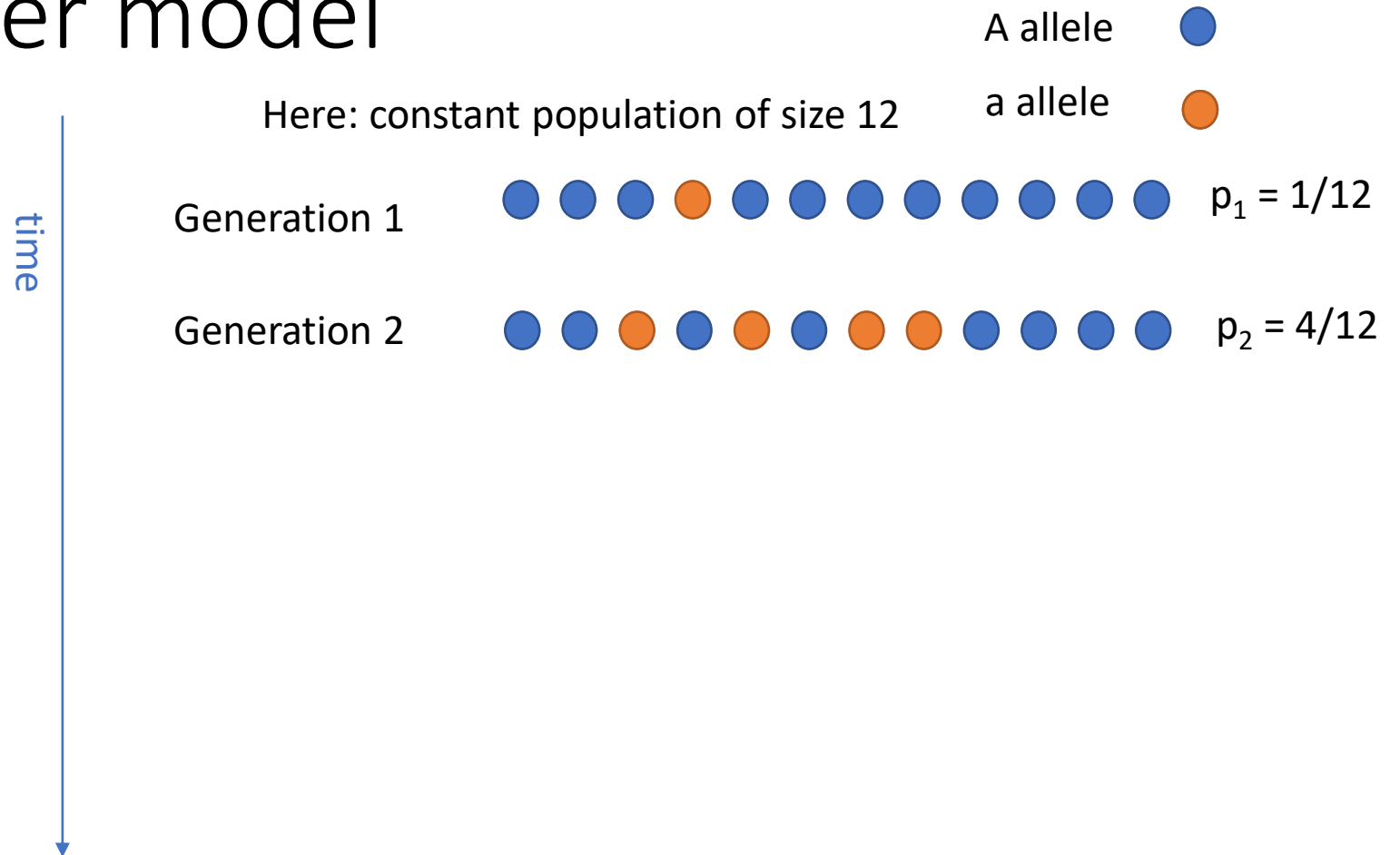
- No migration
- Panmixia
- Here : constant N_e

Wright-Fisher model:

Model a population forward in time assuming discrete non overlapping generations

Individuals at every generations are obtained by sampling with replacement from the parents

NB. It is possible to incorporate selection and pop size changes in this model



The Wright-Fisher model

Assumptions:

- No migration
- Panmixia
- Here : constant N_e

Wright-Fisher model:

Model a population forward in time assuming discrete non overlapping generations

Individuals at every generations are obtained by sampling with replacement from the parents

NB. It is possible to incorporate selection and pop size changes in this model

time
↓

Here: constant population of size 12

A allele ●

a allele ●

Generation 1



Generation 2



Generation 3



The Wright-Fisher model

Assumptions:

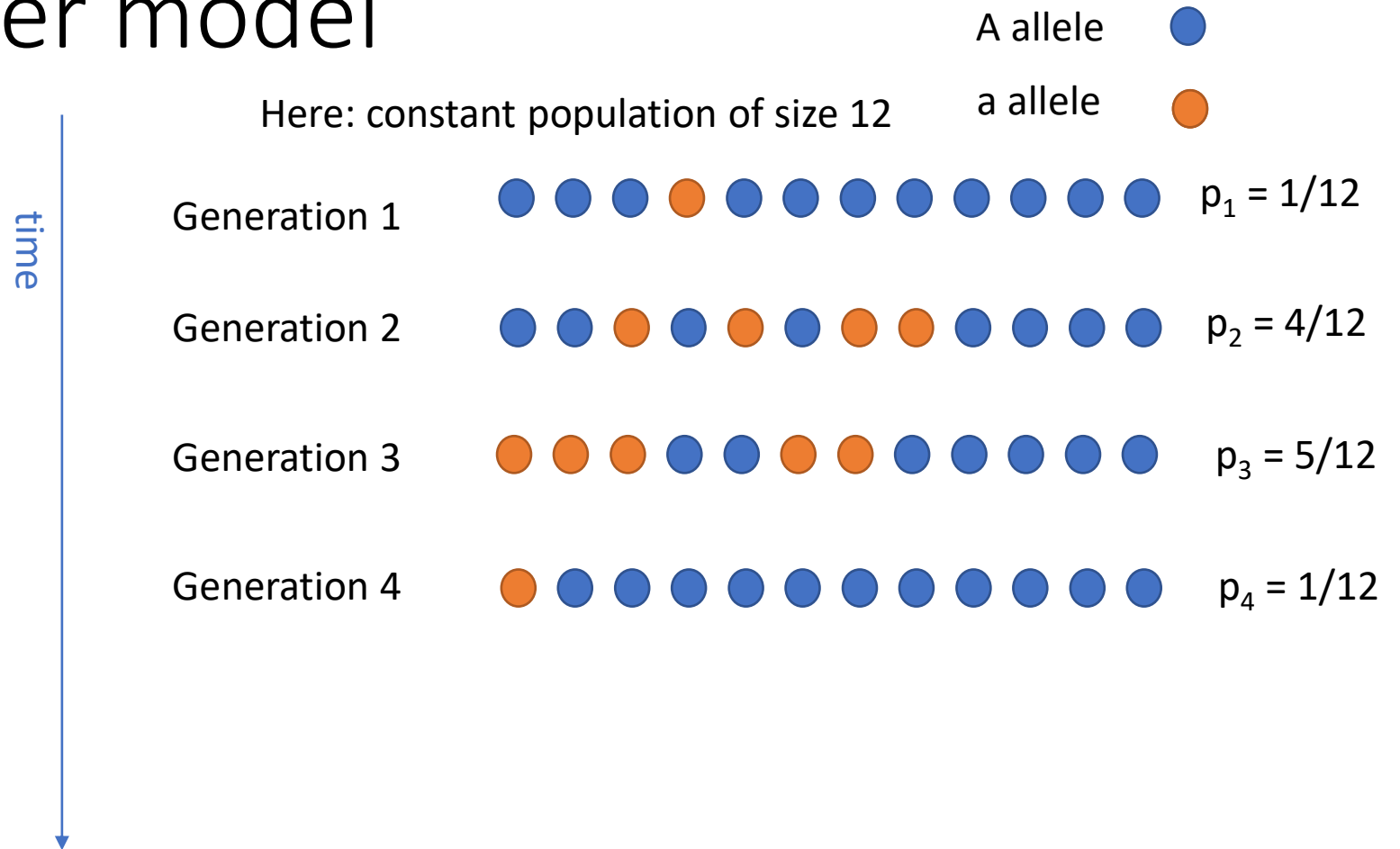
- No migration
- Panmixia
- Here : constant N_e

Wright-Fisher model:

Model a population forward in time assuming discrete non overlapping generations

Individuals at every generations are obtained by sampling with replacement from the parents

NB. It is possible to incorporate selection and pop size changes in this model



The Wright-Fisher model

Assumptions:

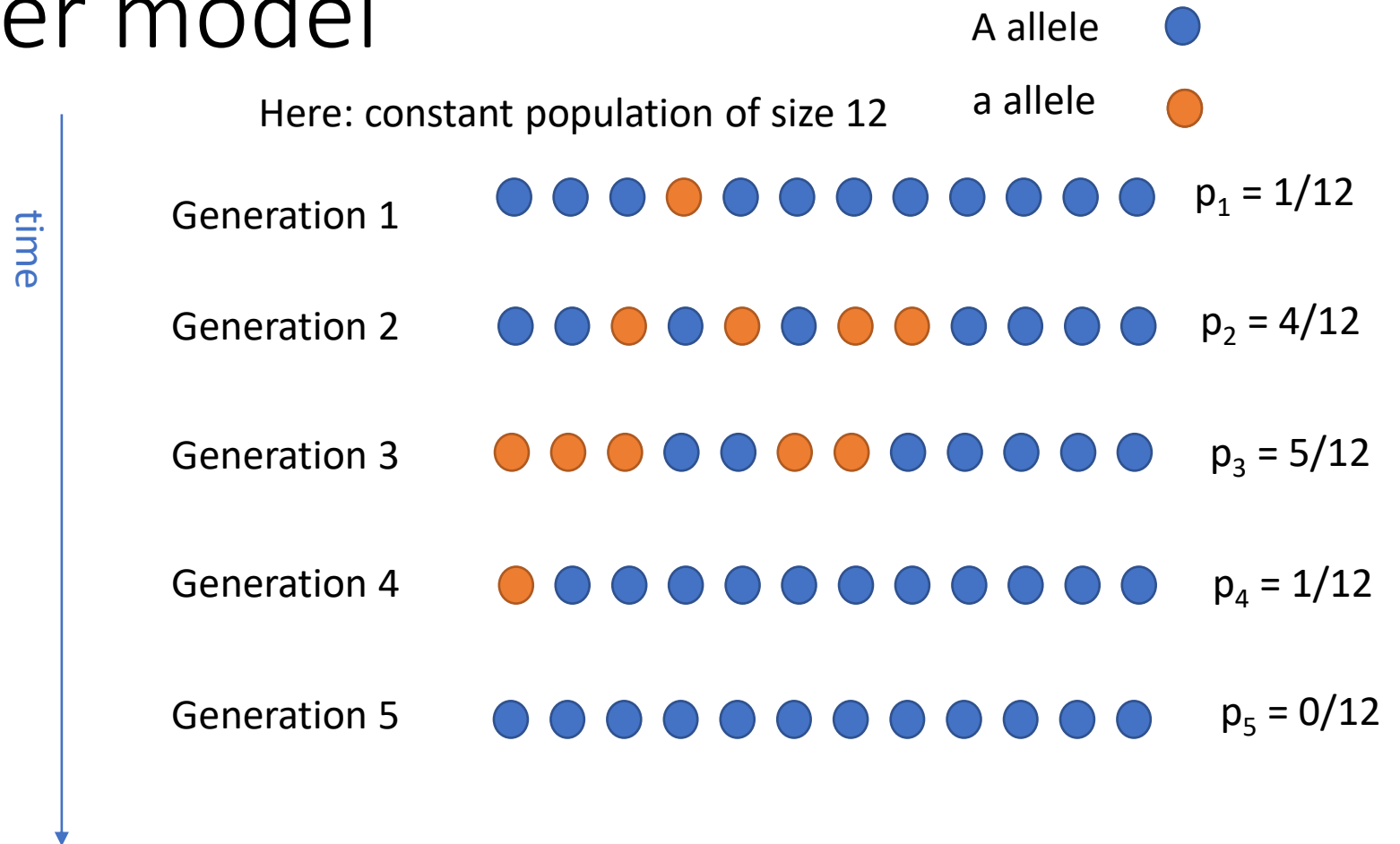
- No migration
- Panmixia
- Here : constant N_e

Wright-Fisher model:

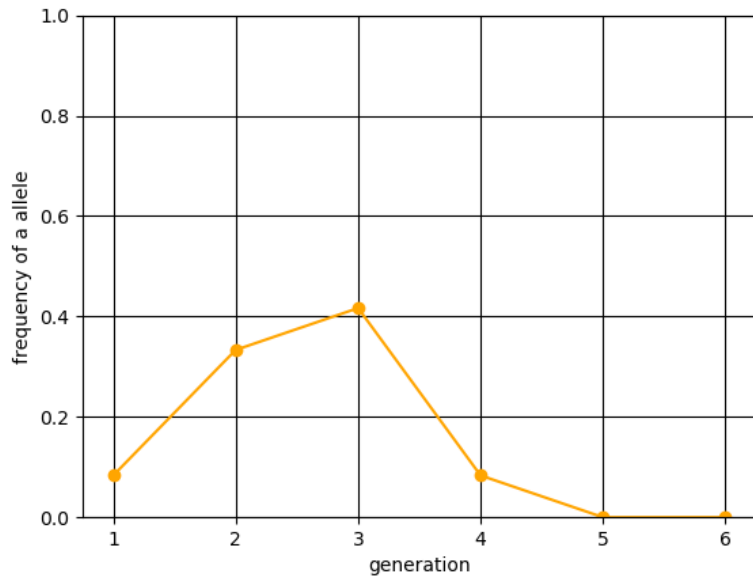
Model a population forward in time assuming discrete non overlapping generations

Individuals at every generations are obtained by sampling with replacement from the parents

NB. It is possible to incorporate selection and pop size changes in this model



The Wright-Fisher model



time











Here: constant population of size 12

A allele



a allele



Generation 1	 	$p_1 = 1/12$
Generation 2	 	$p_2 = 4/12$
Generation 3	 	$p_3 = 5/12$
Generation 4	 	$p_4 = 1/12$
Generation 5		$p_5 = 0/12$
Generation 6		$p_6 = 0/12$

We will discuss the math in the second part of the talk.

Simulating mutation and genetic drift

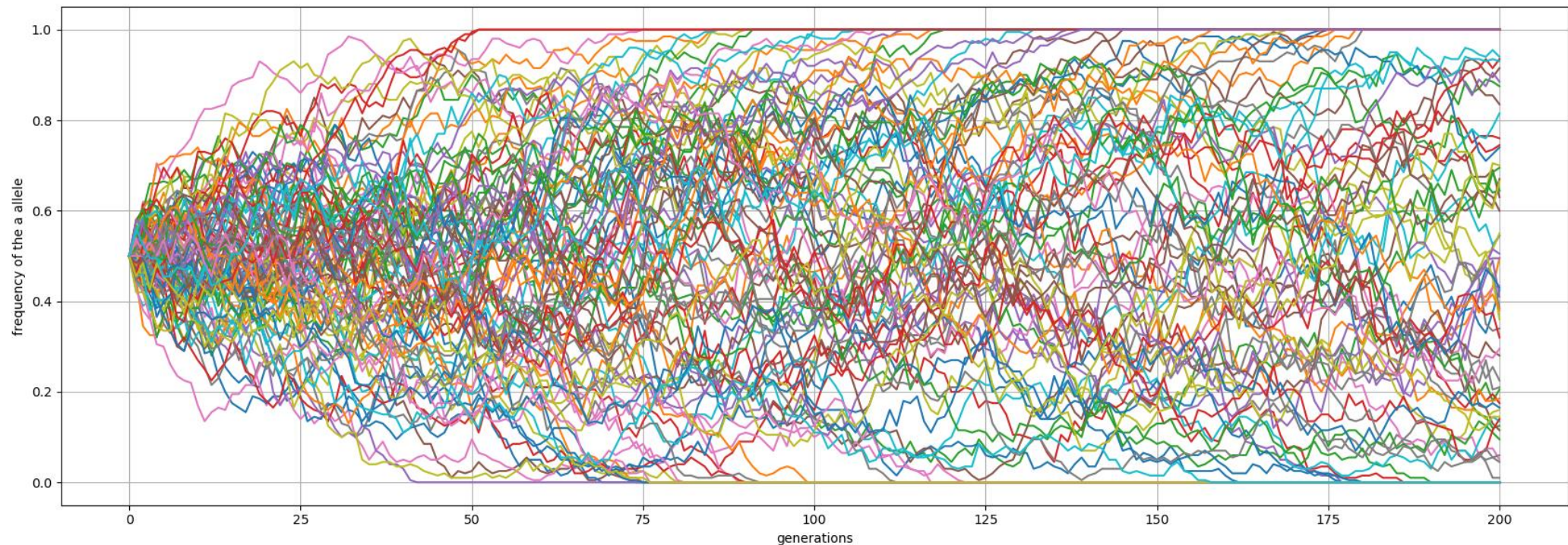
Theoretical result: Probability of fixation \sim starting frequency.

Simulations: alleles with a starting frequency of $p = 0.5$, generations = 200, replicates = 100, $N = 200$.

Simulating mutation and genetic drift

Alleles with a starting frequency of $p = 0.5$, generations = 200, replicates = 100, $N = 200$.

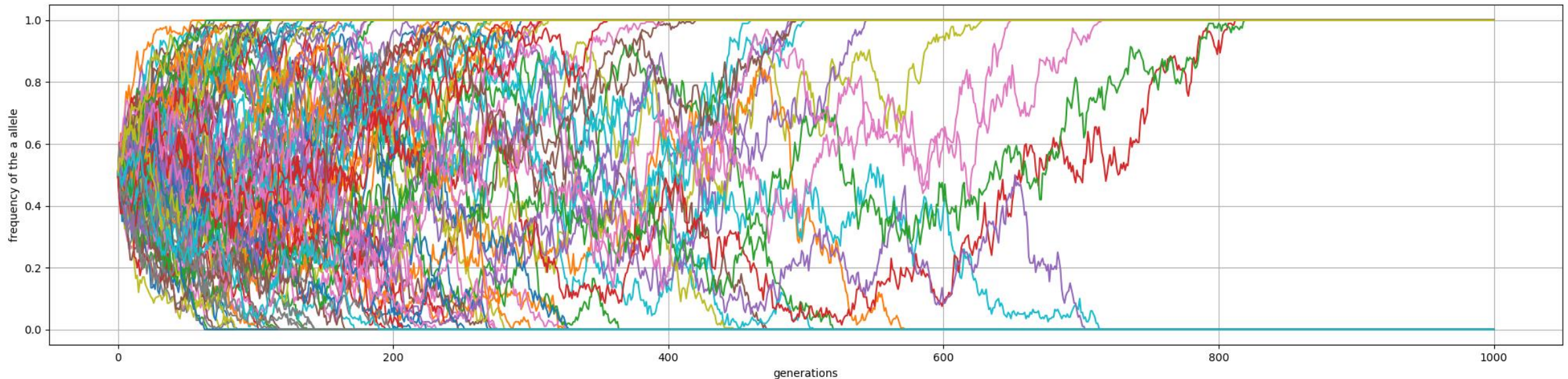
Fixation in 27 cases and losses in 13 cases. Probability of fixation \sim starting frequency.



Simulating mutation and genetic drift

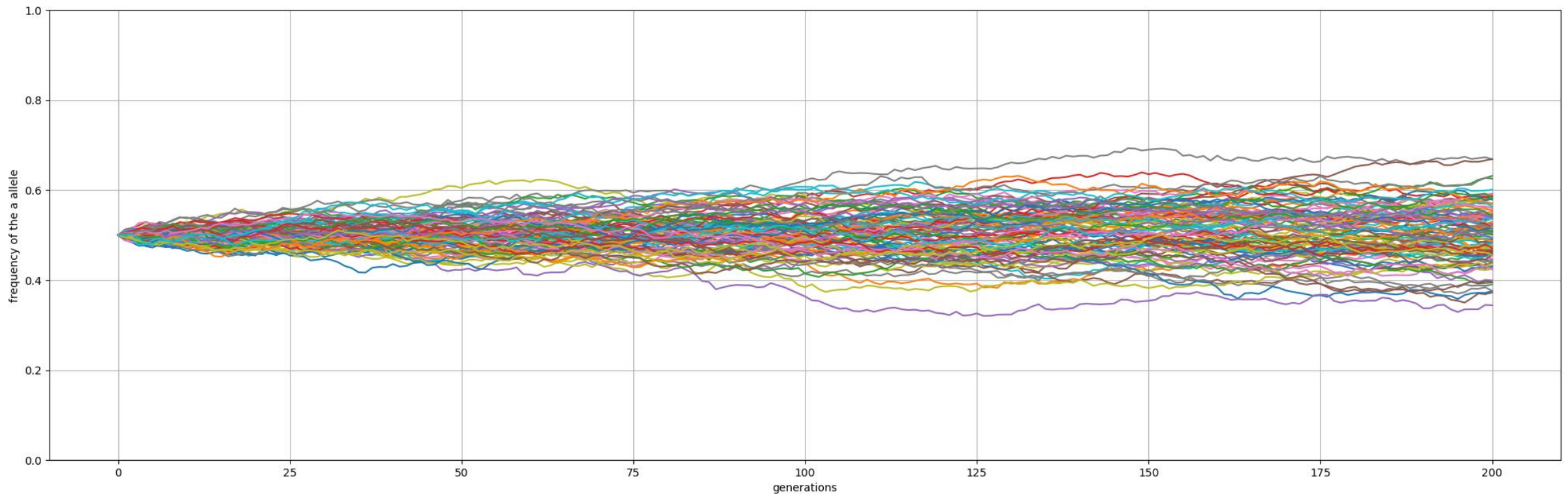
Alleles with a starting frequency of $p = 0.5$, generations = 1000, replicates = 100, $N = 200$.

Fixation in 54 cases and lost in 46 cases. Probability of fixation \sim starting frequency.



Simulating mutation and genetic drift

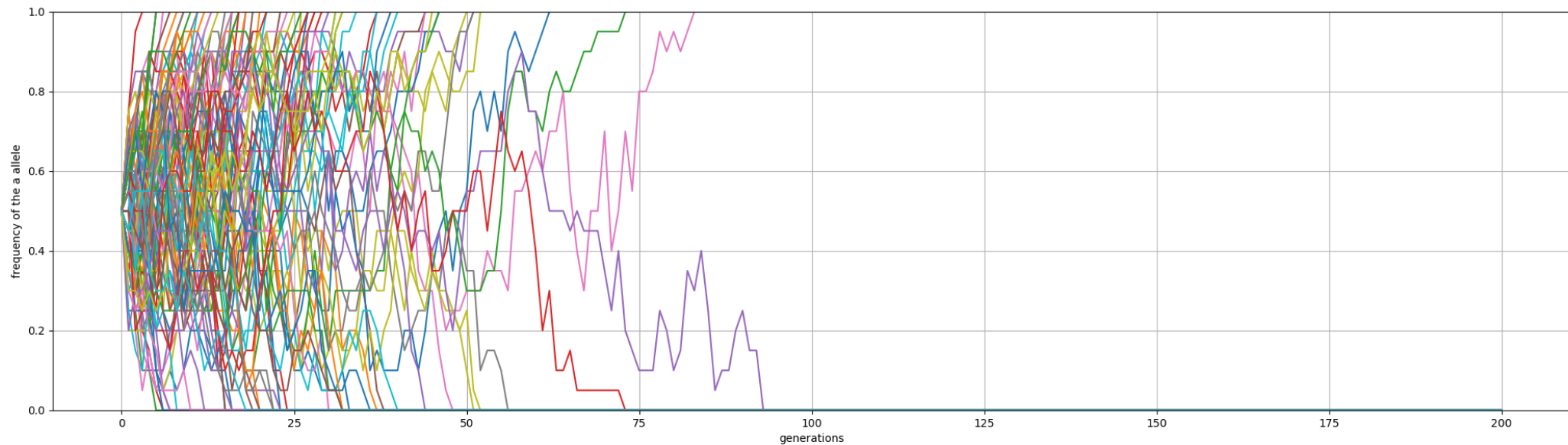
Larger $N = 10'000$ versus smaller $N = 200$ (effective) population size.



Simulating mutation and genetic drift

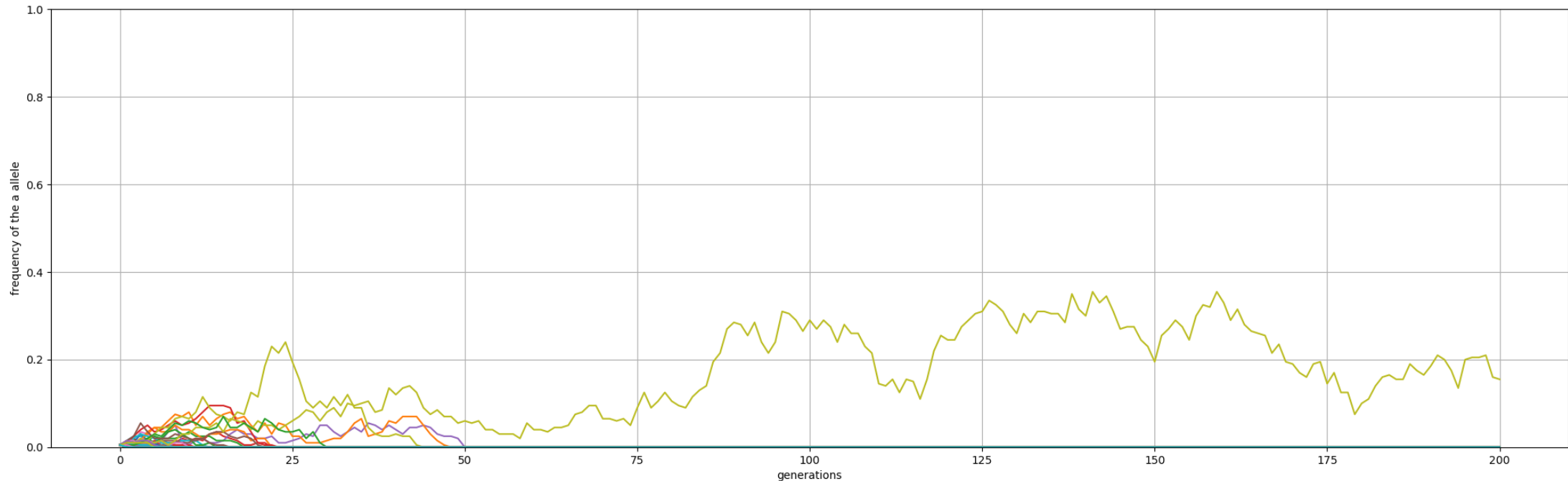
Larger $N = 10'000$ versus smaller $N = 200$ (effective) population size.

More variability with smaller effective size.



Simulating mutation and genetic drift

Alleles with a starting frequency of $p = 1/N = 0.005$, 200 generations, 50 replicates, $N = 100$. Fixation in 0 cases and lost in 49 cases.



How much variation (diversity) do we expect in a population?

$$\theta = 4N_e\mu$$

The diagram illustrates the components of the equation $\theta = 4N_e\mu$. Three arrows point from descriptive labels to the terms in the equation: an arrow from 'Watterson's Estimator' (in pink) points to the Greek letter θ ; an arrow from 'Effective population size' (in orange) points to N_e ; and an arrow from 'Mutation rate' (in blue) points to μ .

Watterson's Estimator

Effective population size

Mutation rate

And inversely, how might we use observed variation to estimate N_e ?

If one had a reliable measurement of mutation rate, it would be clearly feasible to estimate effective population size just based on observed levels of neutral variation

$$\theta = 4N_e\mu$$

Example: Human Cytomegalovirus (within host)

$$\theta^* = 0.0315$$

$$\mu = 0.0000002$$

$$\longrightarrow N_e = 78,750$$

Sources: Renzette et al. 2015 (PNAS),
Renzette et al. 2016 (Mol Ecol) *

Another way to think about N_e - per generation fluctuations in allele frequency

- Apart from levels of variation, genetic drift could be measured by tracking the frequencies of neutral mutations through time

⇒ In other words, the fluctuations induced by genetic drift in fact have a direct relationship with N_e for the reasons described

Effective population size and selection

The effective population size “determines” whether a mutation “is deleterious/beneficial or neutral”.

- While it is true that the selection coefficient is s (as defined earlier), the population selection coefficient ($N_e * s$) is what determines the strength of selection acting upon a mutation

Effective population size and selection

- The effective population size determines whether a mutation “is deleterious/beneficial or neutral”.
 - While it is true that the selection coefficient is s (as defined earlier), the population selection coefficient ($N_e * s$) is what determines the strength of selection acting upon a mutation.
- There is a relatively simple rule of thumb:
 - If $|N_e s| > 10$, the history of the mutation is dominated by selection
 - If $|N_e s| < 1$, the history of the mutation is dominated by genetic drift

Thinking about evolution in big vs. small populations

A few take-homes :

Population size is an absolutely key driver of many evolutionary processes.

$N_e \mu$ - dictates the total number of mutations entering a population, and thus how much variation is expected.

$N_e s$ - dictates the impact of selection (and genetic drift) on a mutation with a selective advantage or disadvantage

NB: the great majority of new mutations (even those that are beneficial) are stochastically lost from the population by genetic drift.

Fixing a deleterious mutation

There is a far greater likelihood of fixing mutations with negative selection coefficients in a small population.

⇒ relatedly, as inbreeding greatly reduces effective population size (i.e., the effective number of individuals contributing variation to the next generation), this is similarly associated with an increase in the number of deleterious mutations accumulating and fixing in the population.

Things you may know:

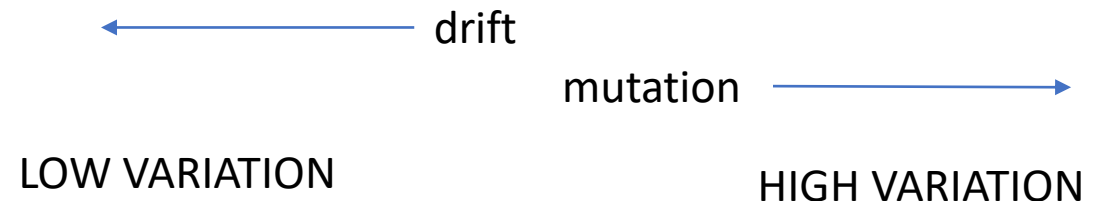
- inbreeding can be problematic
- it is critical to preserve maximal variation in endangered species (i.e., those with small population sizes)

“Mutation generates”, “Drift eliminates”

- In summary, mutation generates variation, and genetic drift serves to eliminate variation

⇒ The great majority of mutations being lost from the population are lost due to drift; but, drift can also drive mutations to fixation. These differing rates of loss and fixation result in the molecular clock observed for neutral mutations.

This also results in an equilibrium for neutral mutations - known as the **mutation-drift balance**



A word on the neutral theory

Mutation and genetic drift are important. What about selection?

Kimura's (1968) Neutral Theory:

What it is: Kimura proposed that the majority of new mutations were deleterious and eliminated by purifying selection. A very small fraction were advantageous and fixed so quickly that they would not be observed segregating in natural populations. Thus, the variation that is observed, is likely neutral with regards to fitness.

What it is not: It is not 'anti-selection' or 'counter-Darwinian' - as noted above, Kimura began on the principle that natural selection was of course acting (natural selection mainly meaning purifying selection).

Early molecular data

This was in response to the early observations that there was far more genetic variation in populations than had been expected.

⇒ By the early 1960's, most had the view that when a mutation would arise, it would either be beneficial or deleterious, and selection would act quickly to either remove it or fix it.

⇒ However, empirical results began accumulating suggesting that individuals carried a wide variety of both unique and shared mutations - including non-synonymous, synonymous, intronic, and non-coding mutations alike

Kimura's idea : perhaps this is just neutral variation that has no impact either way on fitness, and thus it simply exists in the population governed by the simple rules of binomial sampling (i.e., genetic drift).

The nearly-neutral theory

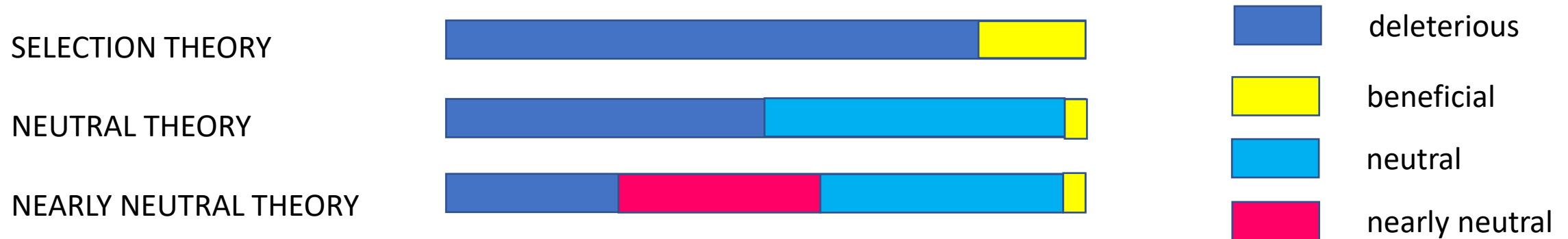
Issues with the neutral theory. E.g. Crow (1970): The number of genes subject to natural selection was overestimated in Kimura's calculations

This led to a revision of the Neutral Theory

Namely, Ohta (1971) proposed that these inconsistencies disappear if segregating variation, rather than being strictly neutral (i.e., $s = 0$), as Kimura had proposed - was 'nearly neutral (i.e., $|N_e s| < 1$).

The nearly-neutral theory

Namely, Ohta (1971) proposed that these inconsistencies disappear if segregating variation, rather than being strictly neutral (i.e., $s = 0$), as Kimura had proposed - was 'nearly neutral (i.e., $|Nes| < 1$).



In this way, mutations would sometimes “behave neutrally” (drift dominates) when population sizes were small, but also sometimes be governed by weak selection when population sizes were larger.

NB. This nicely explains for example the widespread observation that variation does not in fact increase indefinitely as population size grows