Introduction to Bioinformatics for Computer Scientists

Lecture 13

Plan for next lectures

- Today: Introduction to population genetics
- Alexis: Review of some knowledge check questions
- Alexis: Course review

Outline for today

- What is biological evolution?
- Units & Types of Evolution
- Good old G. Mendel (phenotypes)
- Alleles & SNPs (genotypes)
- Models of evolution for infinite populations (Hardy)
- Models of evolution of finite populations (Wright-Fisher)

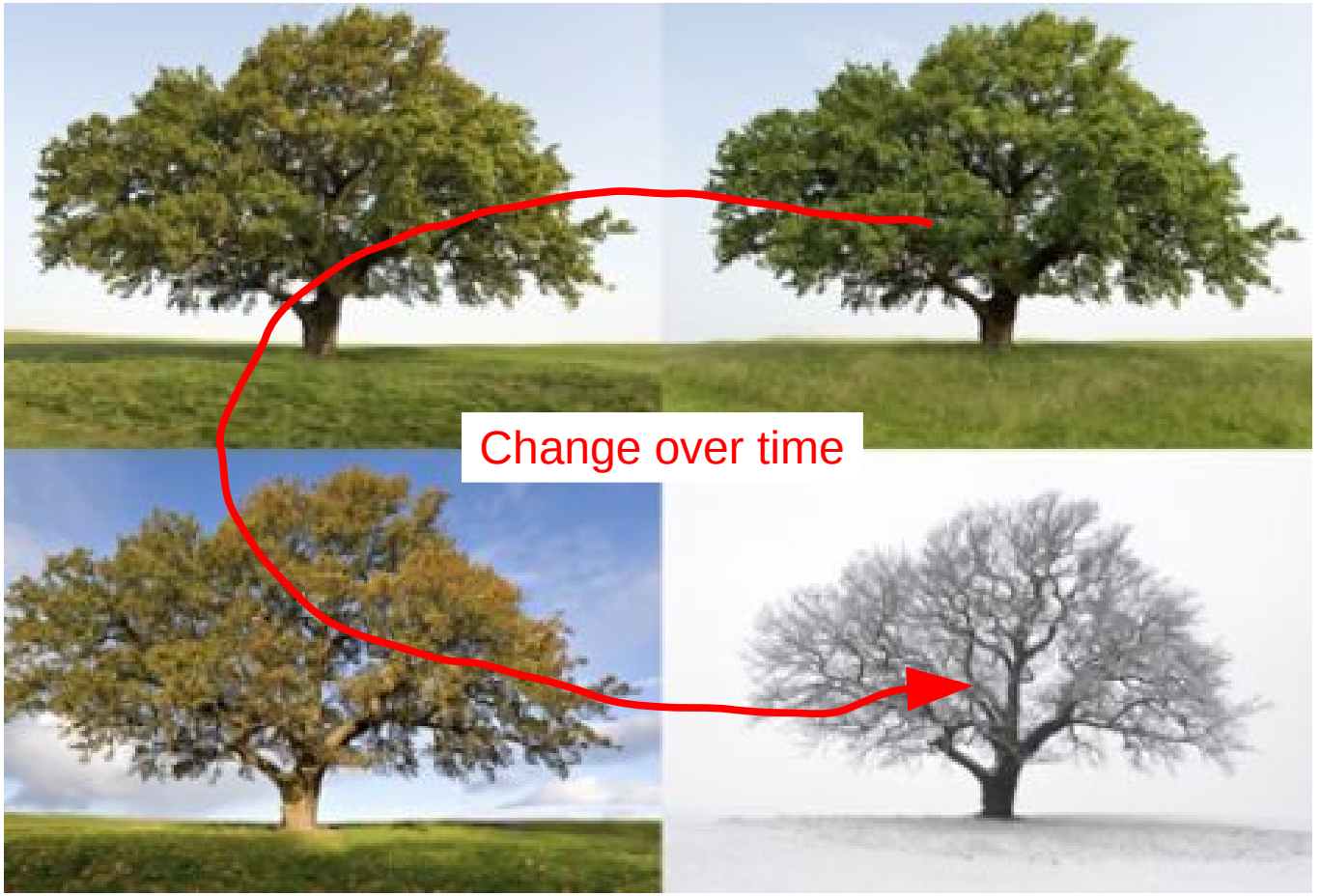
What is Evolution?

- Change over time
- Languages evolve → languages change
- Galaxies evolve → galaxies change
- Political systems change → political systems evolve

Biological Evolution

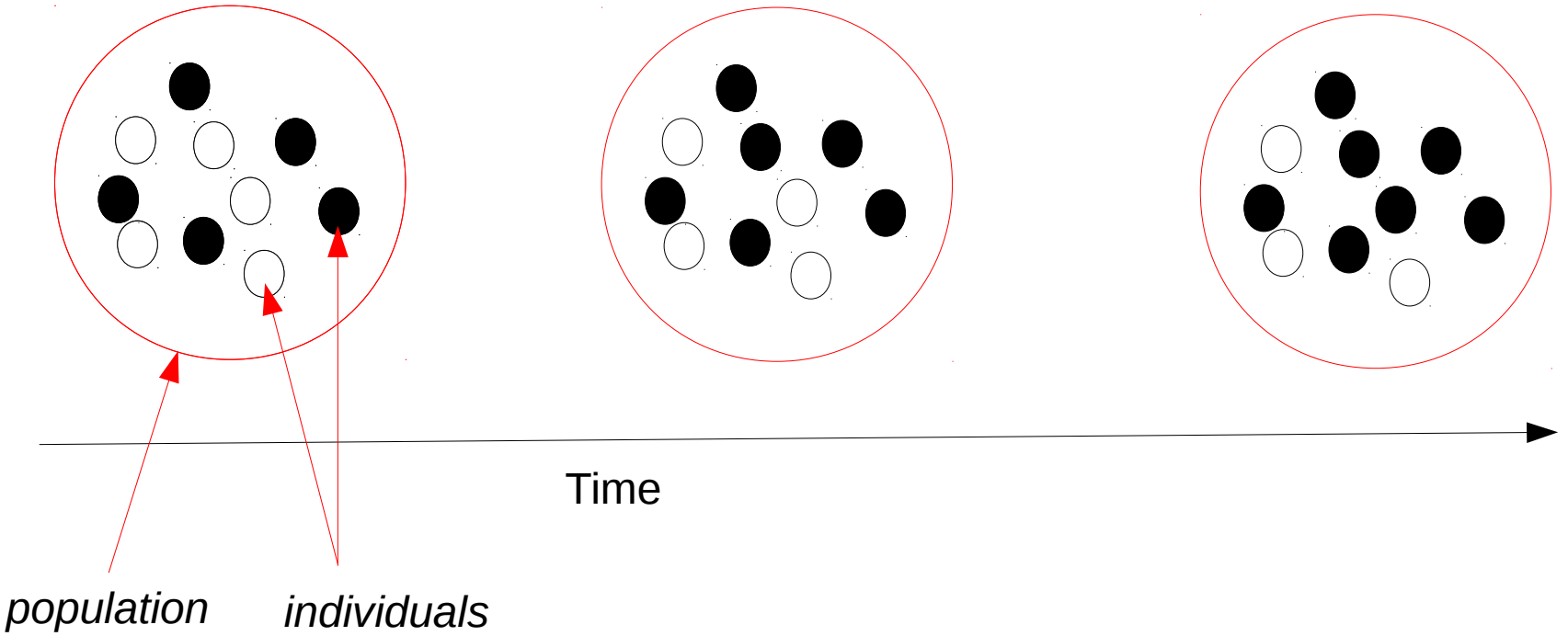
- In Biology one more condition, except for change, is required to characterize evolution
- Do you know which one?

Biological Evolution

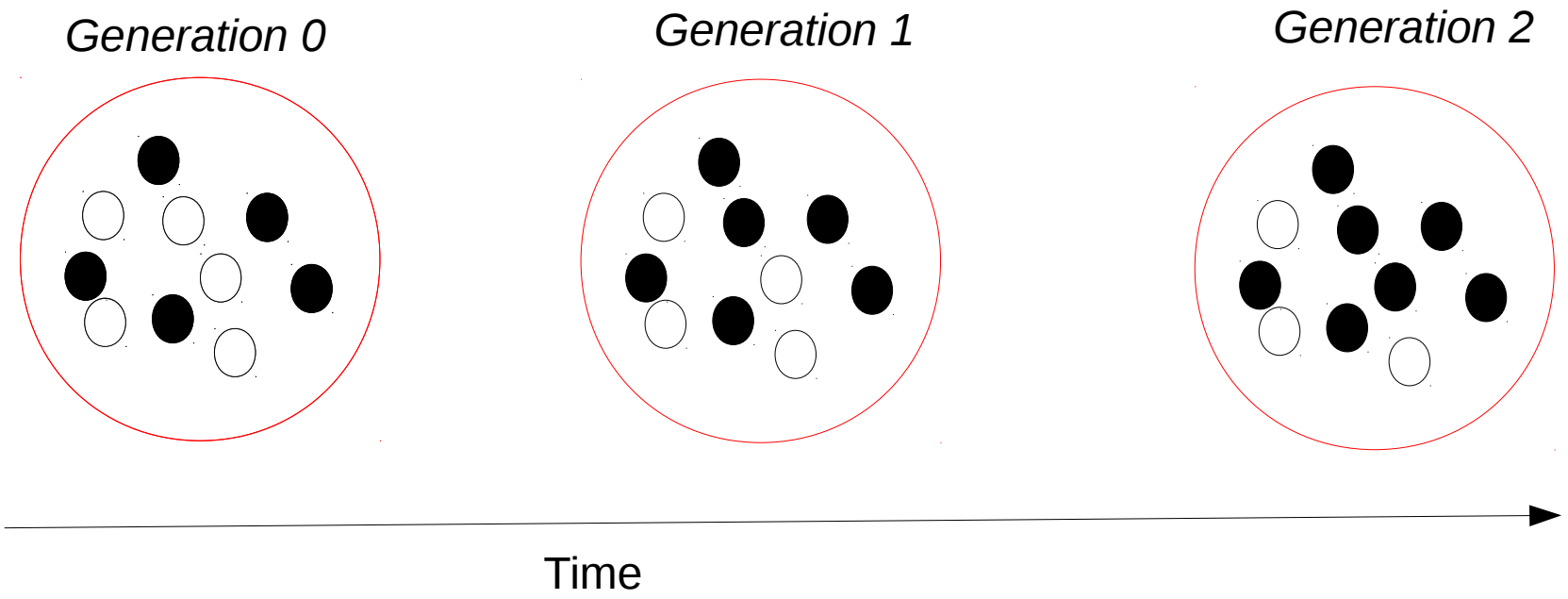


Is this evolution?

Biological Evolution



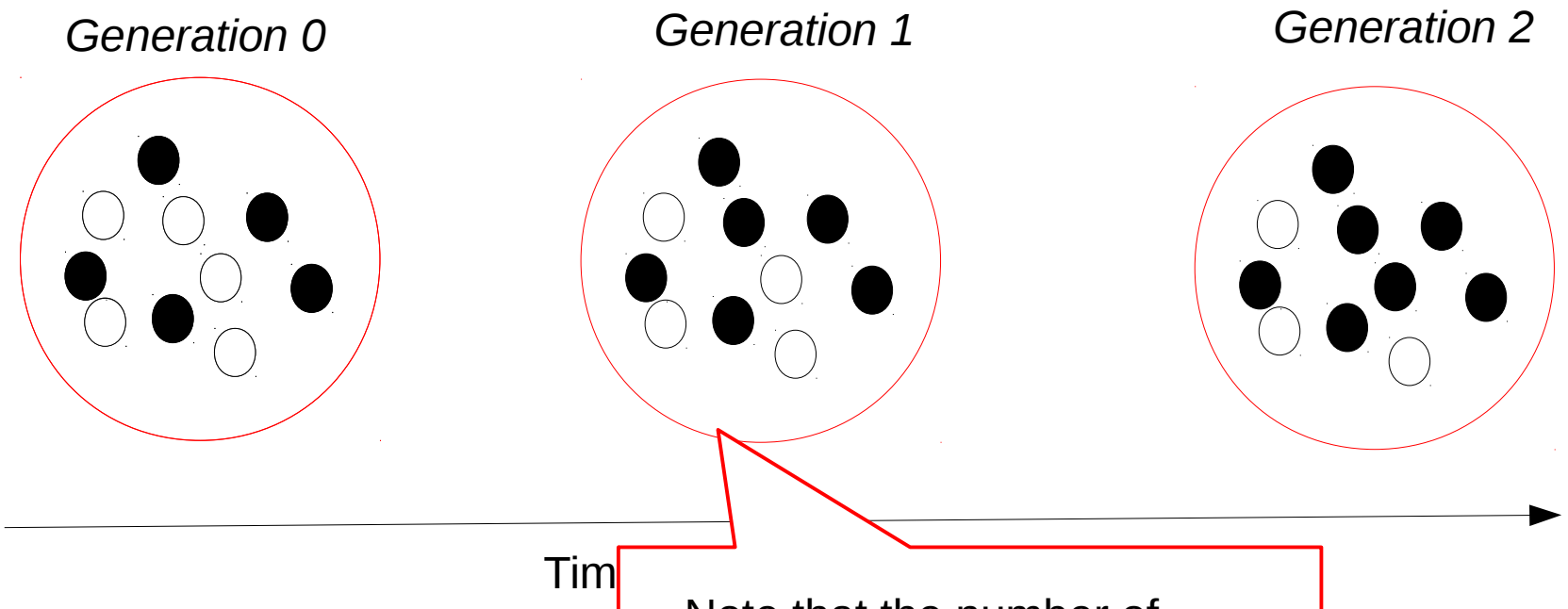
Biological Evolution



The frequency of black and white *individuals* in the *population* changes over time.

Is this evolution?

Biological Evolution

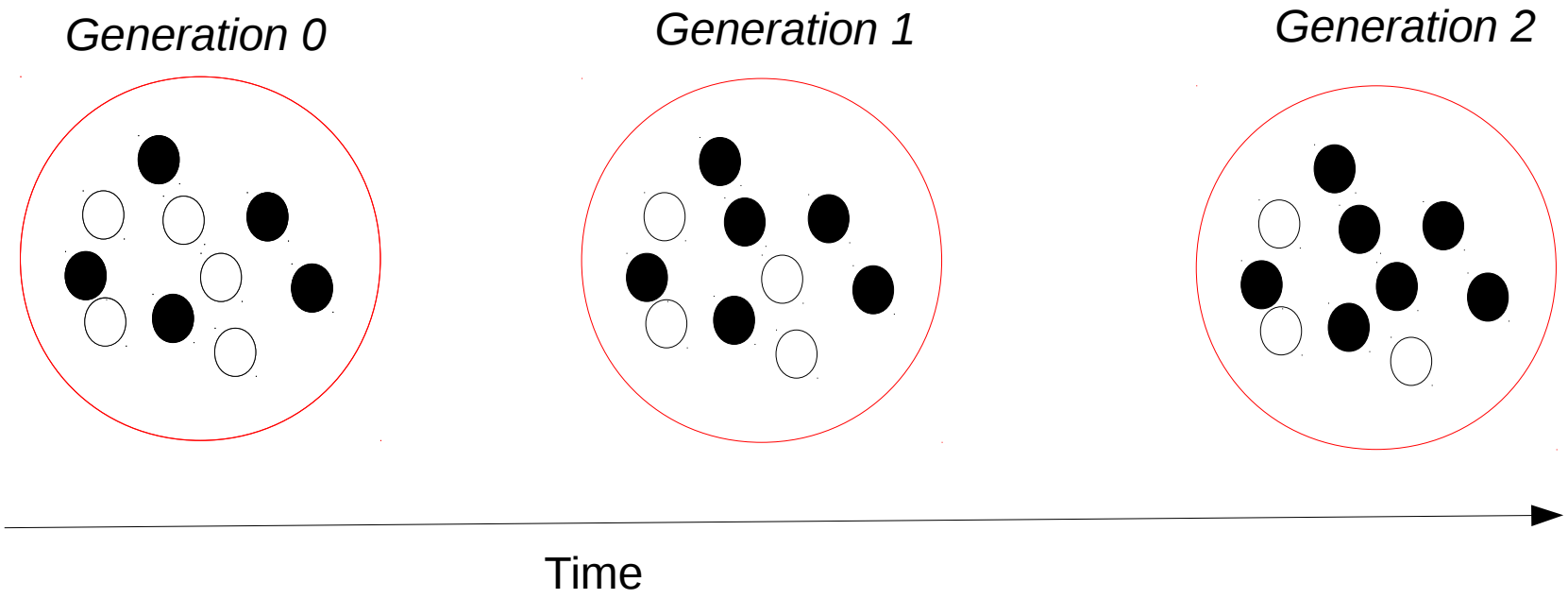


Note that the number of *individuals* remains constant here!
We refer to this as *constant population* size

The frequency of black a *on* changes over time.

Is this evolution?

Biological Evolution



In population genetics we are interested in how characteristics (e.g., ratio of black versus white individuals) of populations change over time.

Another Example

- Population of 5 white and 5 black individuals
  - *frequency(white) = 0.5*
  - *frequency(black) = 0.5*
- Suddenly 7 out of 10 individuals die → 2 white and 1 black left
  - *frequency(white) = 2/3*
  - *frequency(black) = 1/3*
- The population has changed!
- Is this evolution?

Yet Another Example

- Population of 5 white and 5 black individuals
  - *frequency(white) = 0.5*
  - *frequency(black) = 0.5*
- 3 individuals (2 white & 1 black) decide to leave and form a new colony
  - *frequency(white) = 2/3*
  - *frequency(black) = 1/3*
- The population of the new colony is different!
- Is this evolution?

Biological Evolution

- The phenomenon of change is not sufficient for defining biological change/evolution
- For talking about biological evolution, change needs to be inherited
- The reasons for the change are not important for the definition of biological evolution
- … but we are of course interested in them!

Biological Evolution

- Given these examples, by biological evolution we refer to
  - Change of the frequency of occurrence of features of individuals in the population
  - Features can be, for instance, resistance to antibiotics, color, etc.
- These features should be inherited from generation to generation
- Key question: What are the mechanisms of feature inheritance?
- We distinguish between *phenotype* and *genotype*!

The basic Unit of Biological Evolution

- Based on the previous examples, what is the biological unit of evolution:
  - An individual?
  - A population?
  - Something else?

Units of Evolution

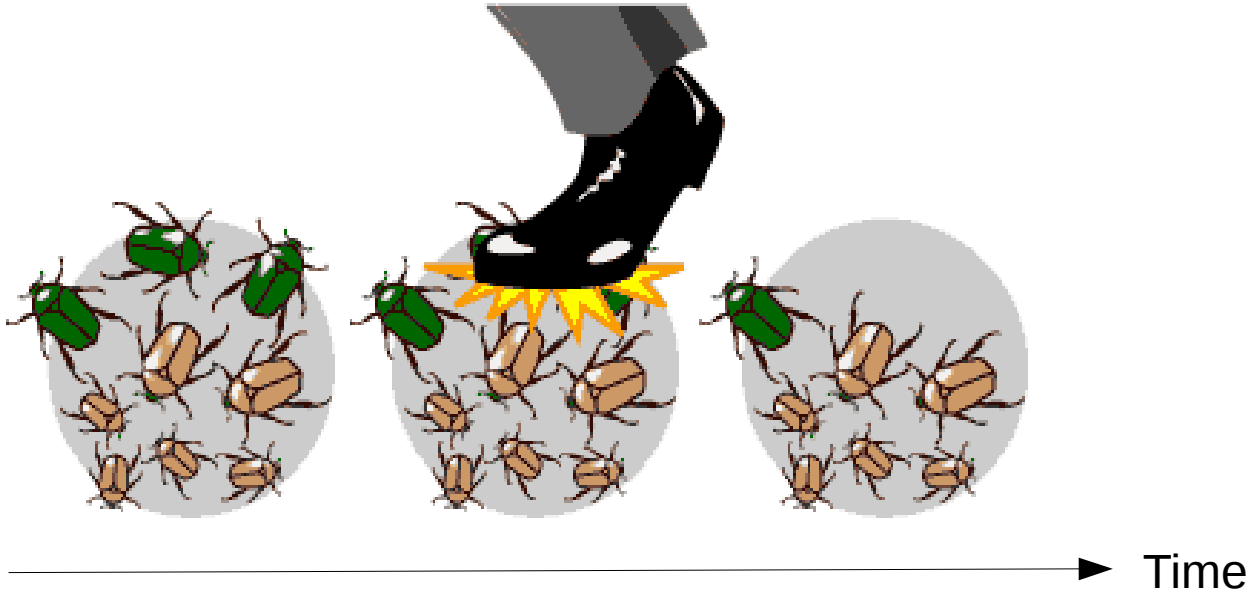
- The population
- A gene
- The genome of an individual
- One needs to define first at which level evolutionary forces act

  → what competes with what?

Units of Evolution: The Population

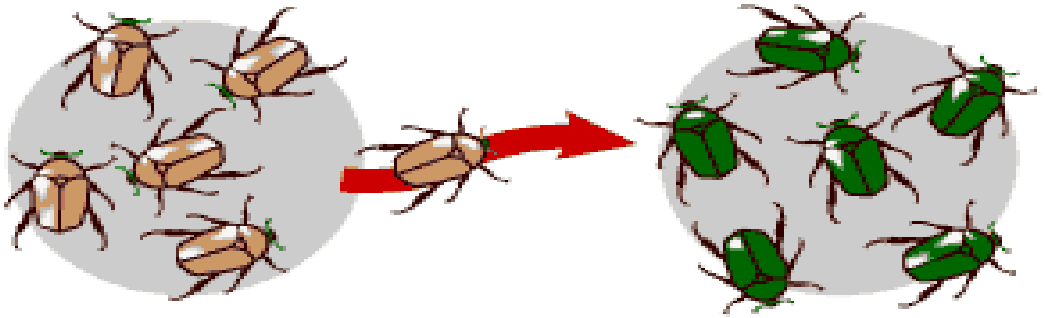
- A Population evolves because the *frequency* of the features of its individuals changes
- Features frequency can change due to
  - *Genetic Drift*: Chance (other than a random mutation)
  - *Migration*
  - *Mutation*
  - *Natural Selection: Response to some pressure (e.g., antibiotics, climate change)*
- Features can be:
  - *Genotype*
  - *Phenotype*

Genetic Drift



Composition of population changes by some random event

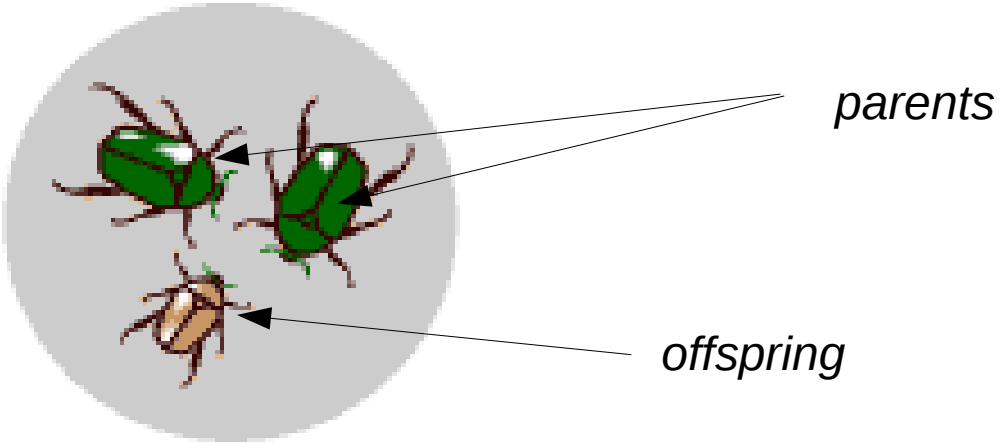
Migration



Population 1

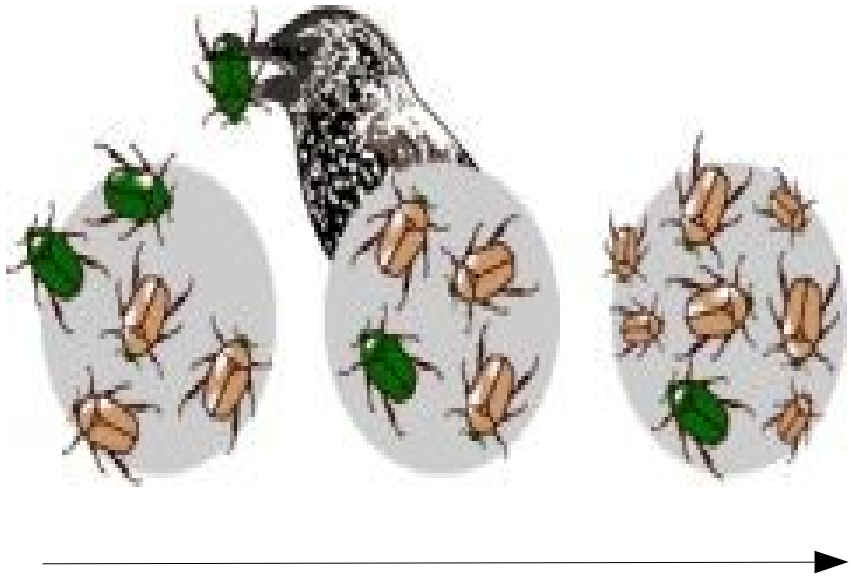
Population 2

Mutation



A random mutation may occur that changes the color of the offspring and hence the frequency of brown beetles in the population

Natural Selection



Time

Green Beetles may be easier to spot for birds → they will have less offsprings in the following generations

Units of Evolution: The Gene

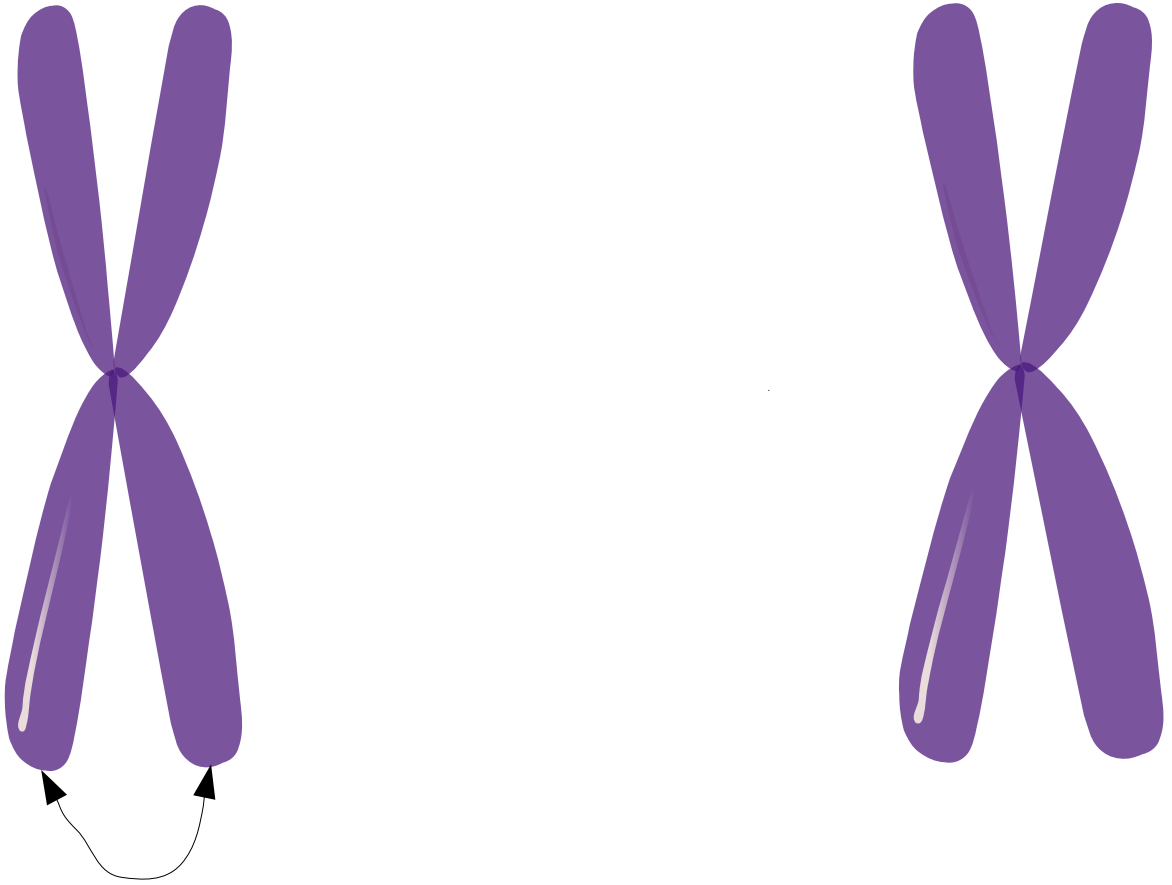
- Genes encode information
- Assume that gene **A** encodes eye color
  - In reality a total of about 15 genes encode eye color
- If **A** has the *form* A → color = blue
- If **A** has the *form* a → color = brown
- What does *form* mean?

Units of Evolution: The Gene

- Genes are inherited from generation → generation
- Inheritance take places via *Alleles*
- An *Allele* is a specific form (slightly different DNA sequence): a or A of gene **A**
- Most multi-cellular organisms are *diploid* → they have two sets of corresponding chromosomes that are called *homologous*
- Diploid organisms have one copy of each gene/allele in each of the homologous chromosome pairs
- If the Allele sequences in the two chromosomes are identical: *homozygous*
- If the Allele sequences in the two chromosomes are different: *heterozygous*

Units of Evolution: The Gene

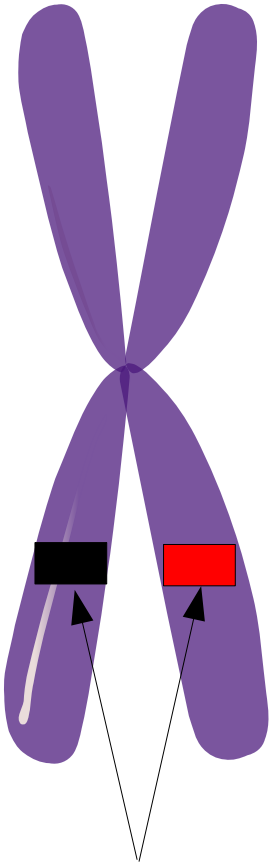
Diploid Chromosome



Homologous pair of chromosomes

Units of Evolution: The Gene

Homozygous → identical DNA



Heterozygous → different DNA sequence

Fraction of heterozygous Alleles

Ancient DNA Siberia
http://en.wikipedia.org/wiki/Denisova_Cave

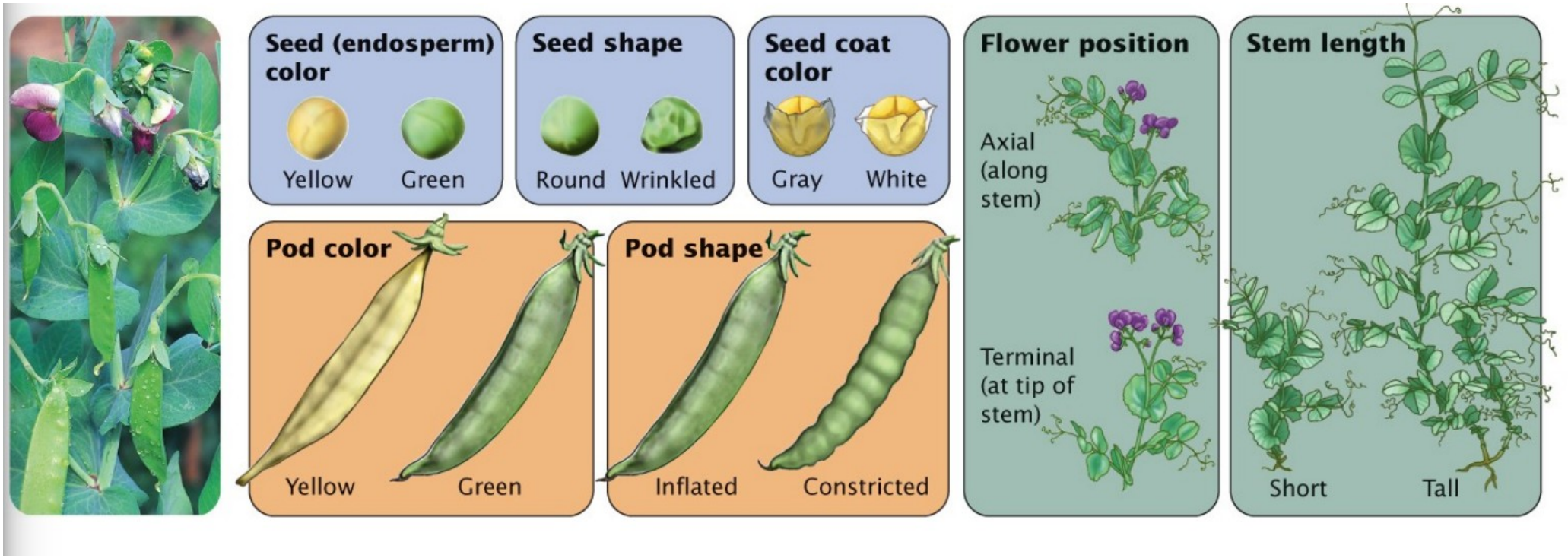
Individual	Heterozygosity estimate (%)
Denisova	0.0165
San	0.0721
Mandenka	0.0686
Yoruba	0.0649
Mbuti	0.0657
Dinka	0.0635
Sardinian	0.0490
French	0.0473
Dai	0.0465
Han	0.0454
Papuan	0.0386
Karitiana	0.0353

Table from:
http://genetics.med.harvard.edu/reich/Reich_Lab/Welcome_files/2013_Bryc_Genetics.pdf

Units of Evolution: The Gene

- Why are we interested in heterozygous versus homozygous Alleles?
- Inheritance →Humans inherit one allele from the father and one from the mother
- Some more terminology:
  - *Genotype* of a gene: the set of corresponding alleles in a diploid organism
  - *Phenotype* of a gene: observation for the trait/property that the gene controls (e.g. brown eye color) → in reality more complex genes interact on traits

Mendelian Inheritance



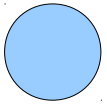
Pea plant traits (phenotype!) studied by G. Mendel

Dominance

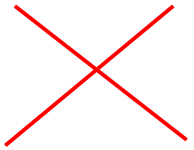
- In Mendel's experiment
  - An individual with the **R**ound-**W**rinkled *genotype* had the **R**ound *phenotype*, *i.e.*, ***RW*** → ***R***
  - We say that the round allele is *dominant* and the wrinkled allele is *recessive*
- What are the phenotypes of:
  - ***RR*** → ***?***
  - ***RW*** → ***?***
  - ***WR*** → ***?***
  - ***WW*** → ***?***
- If there is no dominance-recession relationship the phenotype is intermediate!

Mendel

Homozygous round seed: RR



cross

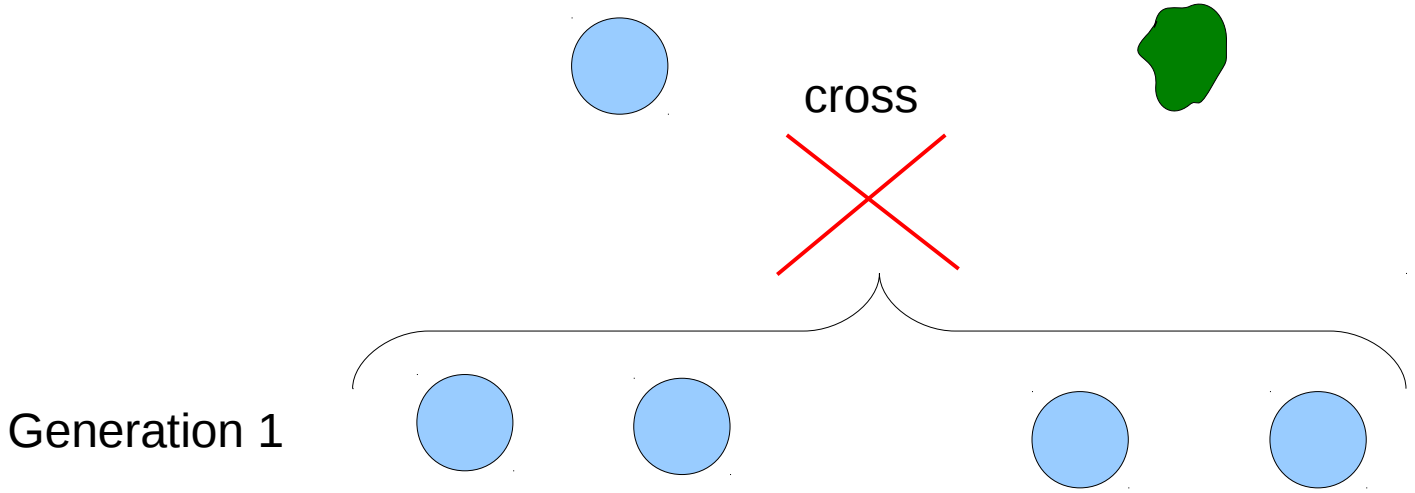


Homozygous wrinkled seed: WW

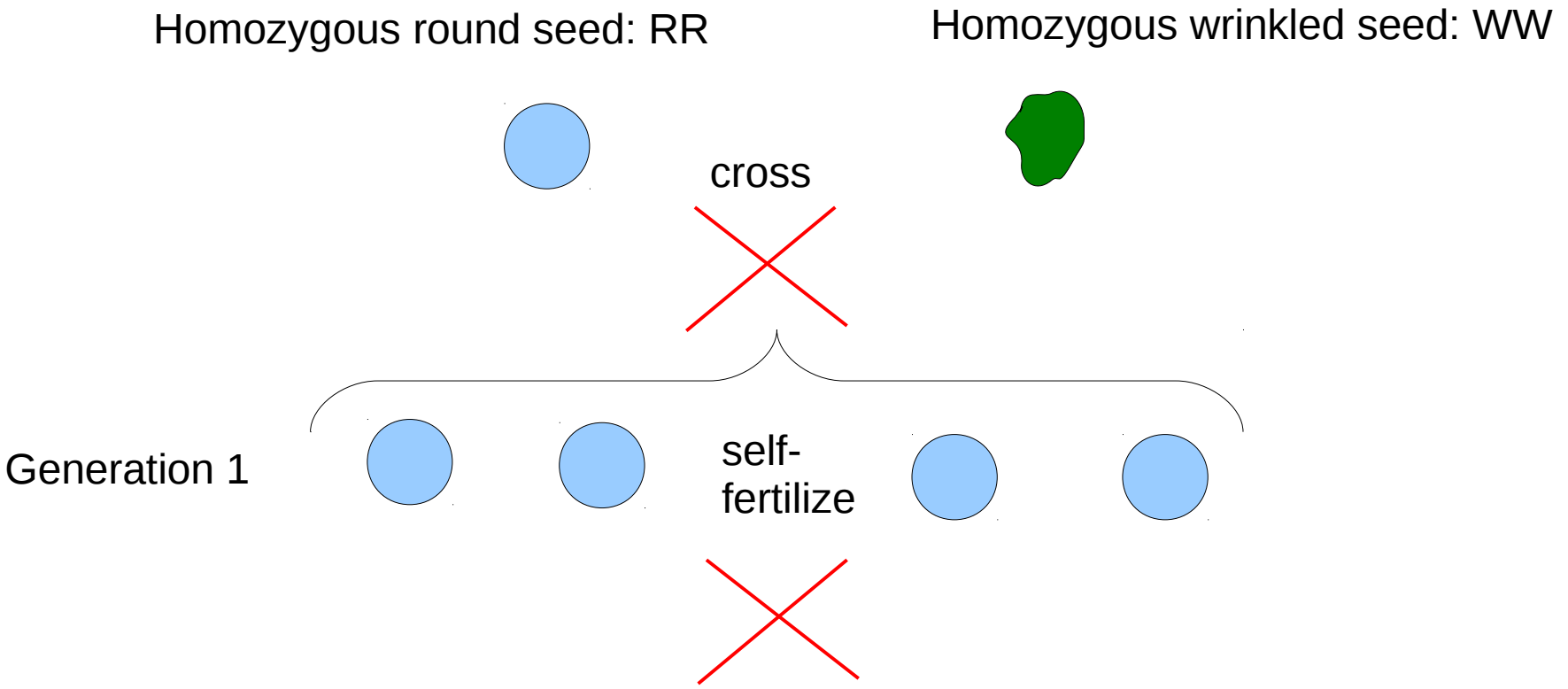
Mendel

Homozygous round seed: RR

Homozygous wrinkled seed: WW



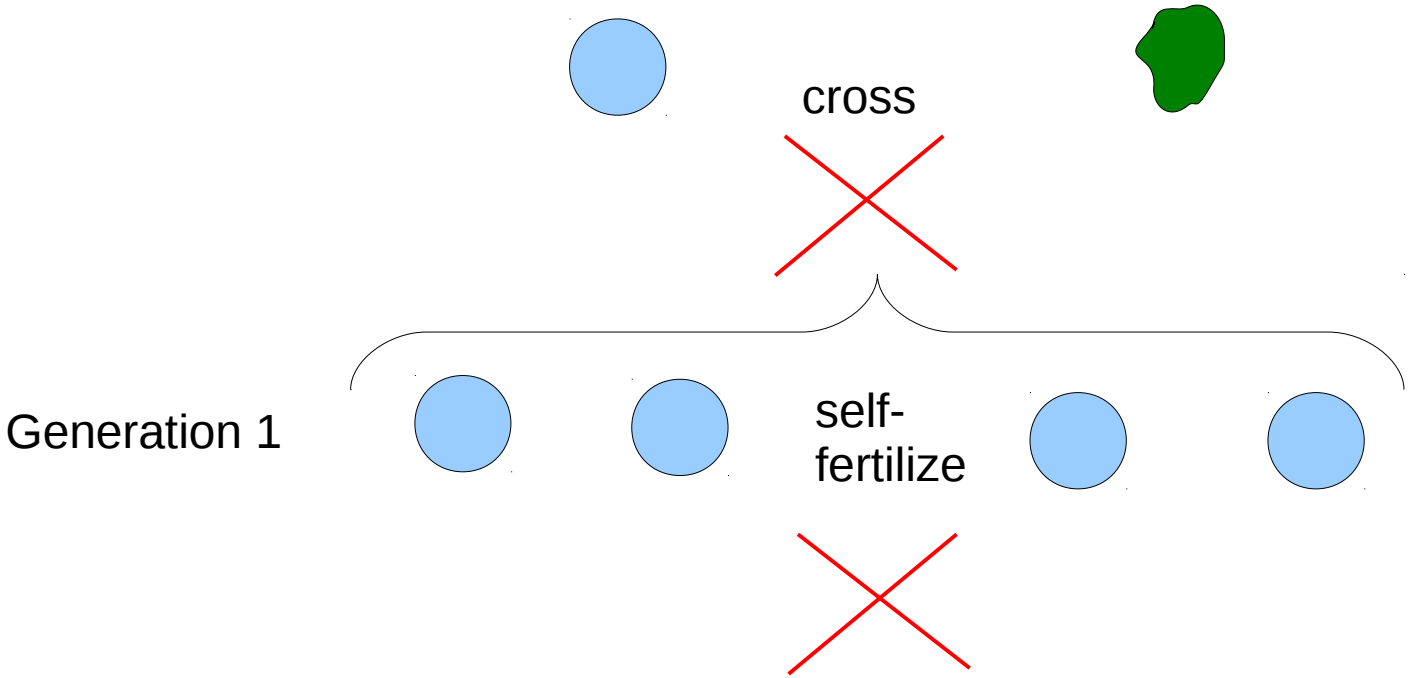
Mendel


Homozygous round seed: RR
Homozygous wrinkled seed: WW
cross
Generation 1
self-
fertilize

Mendel

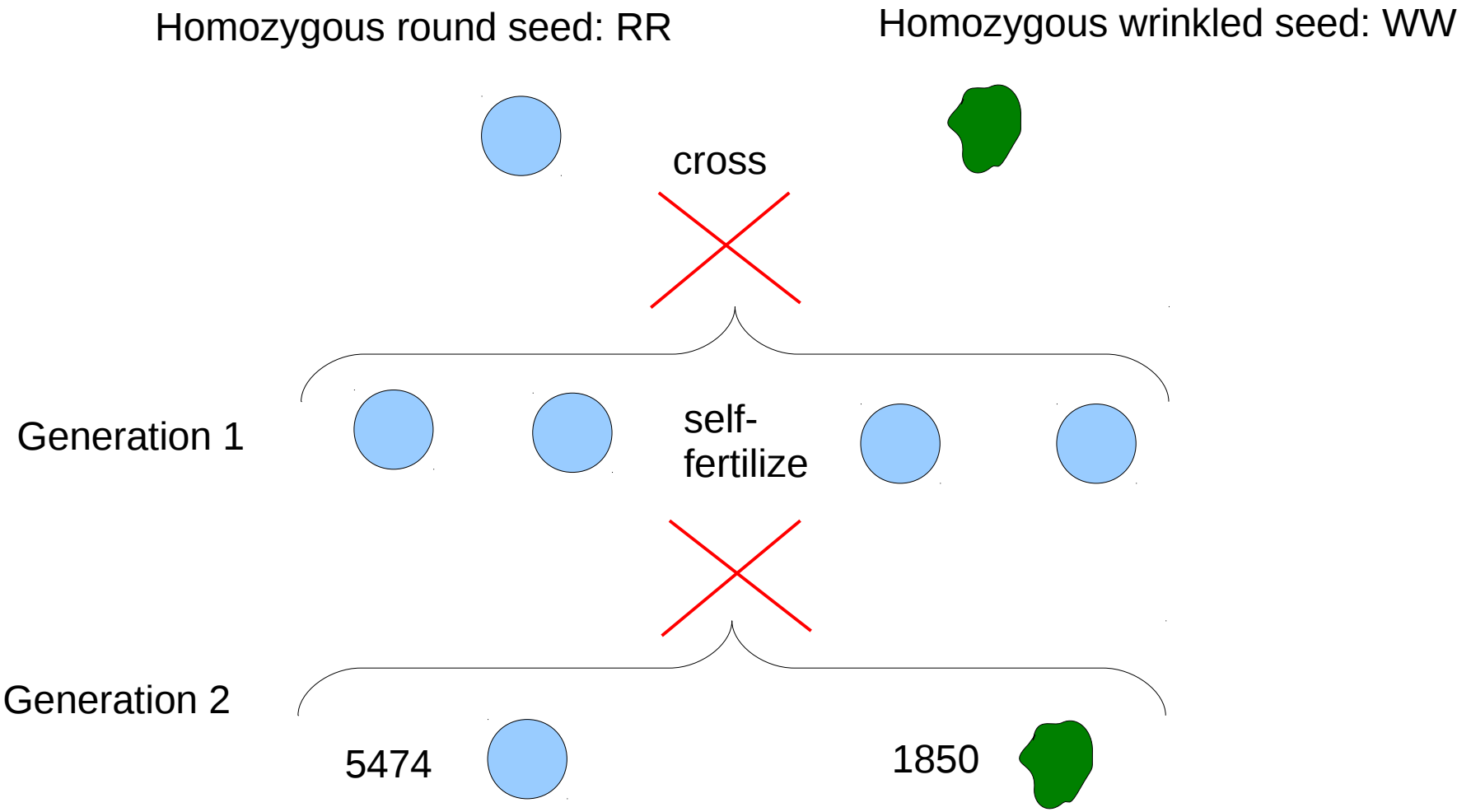
Homozygous round seed: RR

Homozygous wrinkled seed: WW

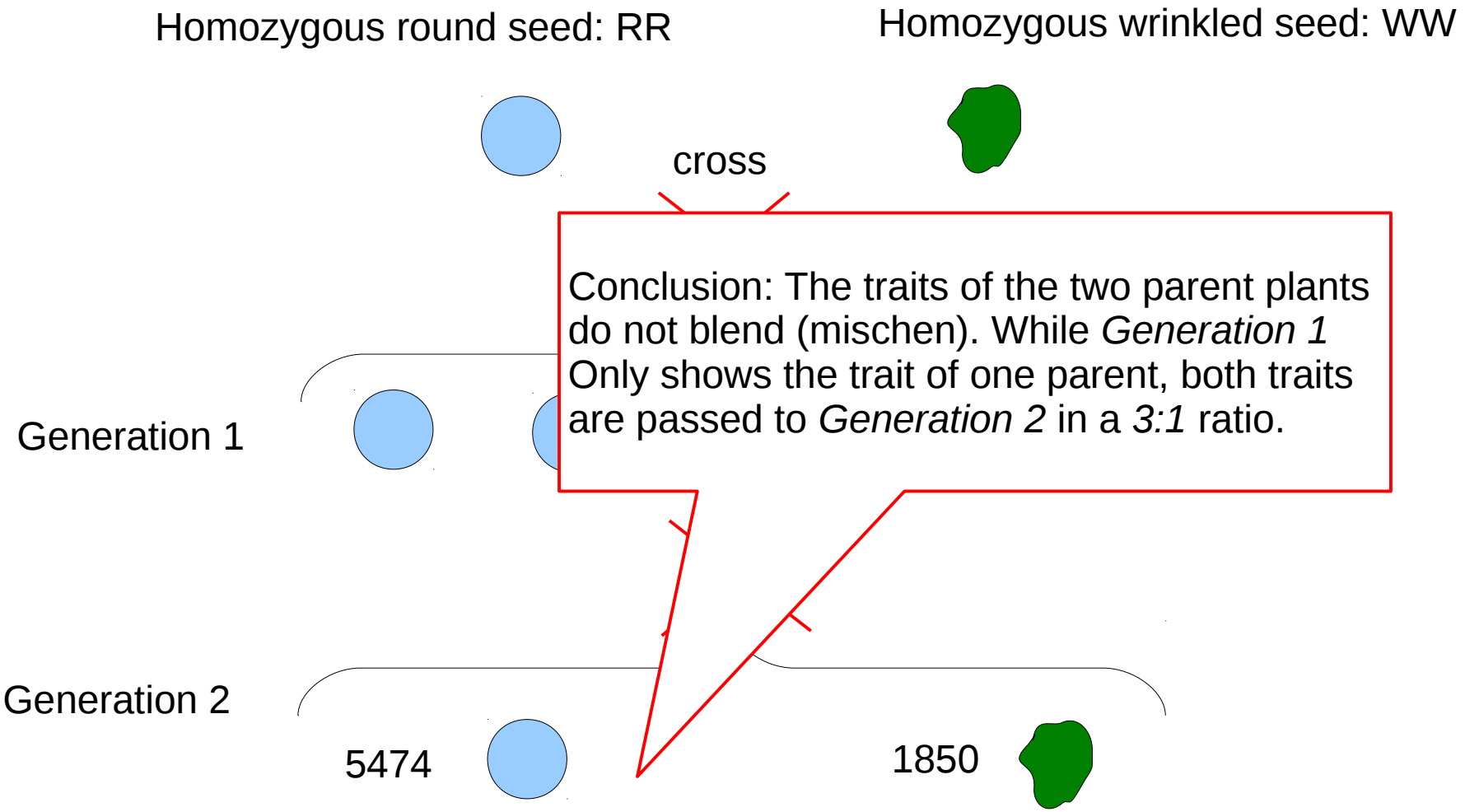


What do you expect?

Mendel


Homozygous round seed: RR
Homozygous wrinkled seed: WW
cross
Generation 1
self-
fertilize
Generation 2
5474
1850

Mendel


Homozygous round seed: RR
Homozygous wrinkled seed: WW
cross
Conclusion: The traits of the two parent plants do not blend (mischen). While *Generation 1* Only shows the trait of one parent, both traits are passed to *Generation 2* in a *3:1* ratio.
Generation 1
Generation 2
5474
1850

Mendel's 1st law
The principle of *Segregation*

Each physical trait of a diploid organism is determined by two factors (alleles). These two factors separate between the generations and re-unite in the next generation.

- **Observation**: the *2nd* generation shows all traits from the initial generation *0* even though the parents in generation *1* do not show all traits.
- **Conclusion**: Generation *1* must receive some information that causes this “hidden” trait to be revealed in generation *2*, in addition to the traits of generation *1*.

Allele Inheritance

- As we know, a diploid organism has *2* alleles per gene
- Alleles can either be *heterozygous* or *homozygous*
- One allele is inherited from the mother and one from the father

  → each parent will pass only one of his – possibly heterozygous – alleles to the offspring
- For a certain, single allele, there is a 50 % chance to have obtained it either from the mother or from the father

Allele Inheritance Terminology

- We denote a gene with the capital bold-font letter **A**
- We denote corresponding Alleles by *A* and *a* if two alleles exist or as A_1, A_2, A_3, … if more than two alleles exist
- *A* denotes both an allele and the corresponding gene which may sometimes lead to confusion
- I use bold font **A** to denote the gene and italic *A* to denote the corresponding Allele

Why do we care about Alleles?

- In population genetics we study the evolution of populations, that is:
  - How does the frequency of alleles change over time?
  - Why does the frequency change?
- As a consequence we are interested in the evolution of so-called *Polymorphisms*
- Polymorphism (Greek): many shapes

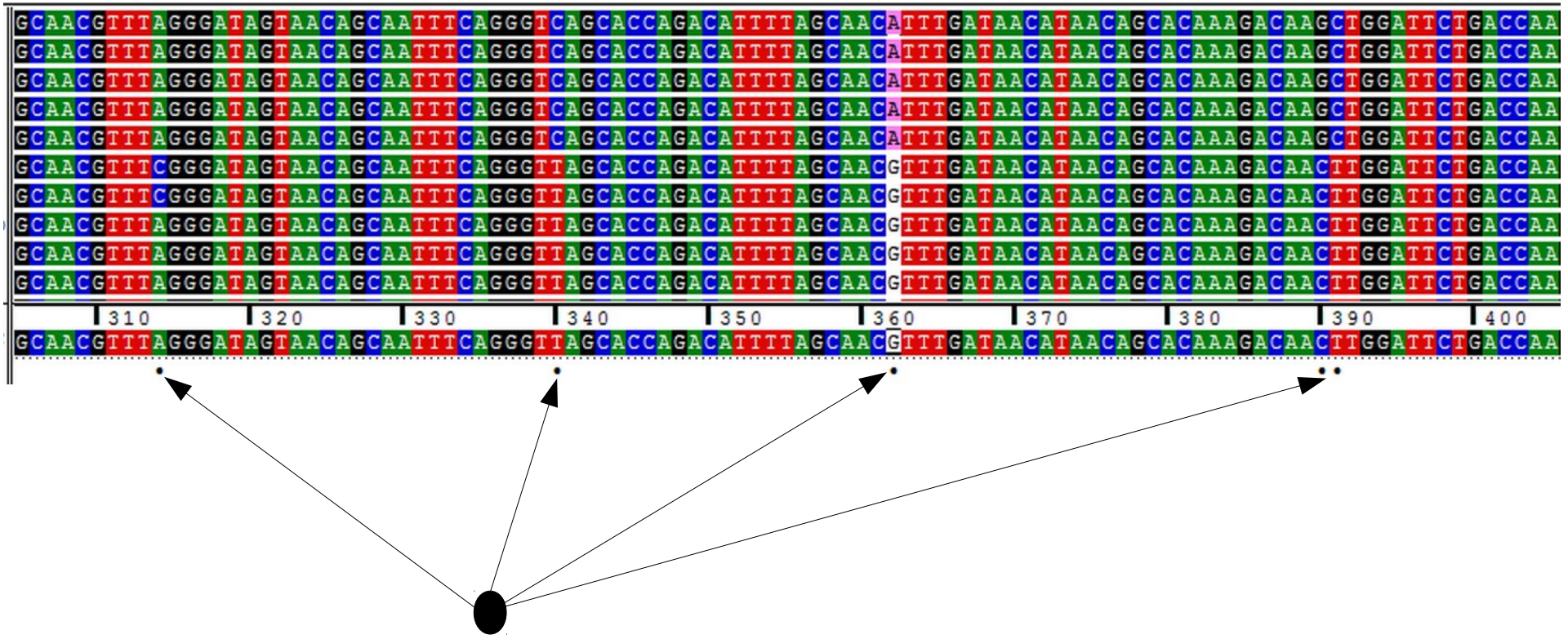
Polymorphism

- *Polymorphic* gene
  - A gene **A** is in the population is polymorphic when there exist multiple alleles (e.g. *A, a*)
- *Polymorphic* site
  - Today, we can sequence the entire DNA of several individuals of a population
  - After multiple sequence alignment we can observe sites in certain genes with more than one state
  - Such sites are called *polymorphic*!

Population genetics versus Phylogenetics

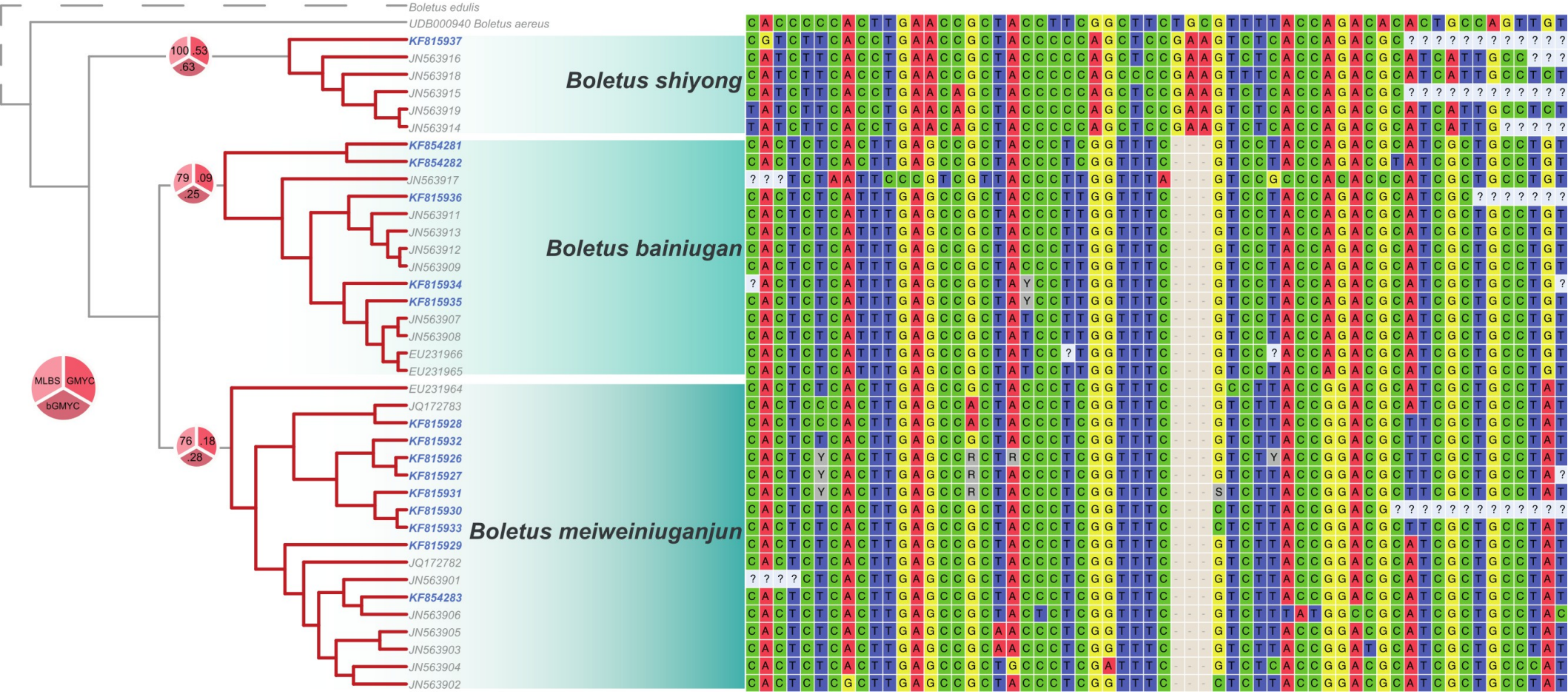
- Evolution at very different scales
- In an alignment of individuals of a single population (species) there will be far less mutations than in the phylogeny of mammals, for instance!
- Since in population genetics there are so few mutations and each mutation is much more important we need to absolutely get the alignment right!

An Alignment Of Individuals



Polymorphic sites

An Alignment of Species



Boletus is a Fungus

Polymorphic Sites – SNPs

- In the MSA of the individuals, we observe some sites, that have more than one nucleotide state
- Such sites are called *Polymorphic* sites or more commonly *SNPs = Single Nucleotide Polymorphisms*
- SNPs is pronounced: Snips
- Modern population genetic analyses mostly operate on SNPs

Modern Population Genetics

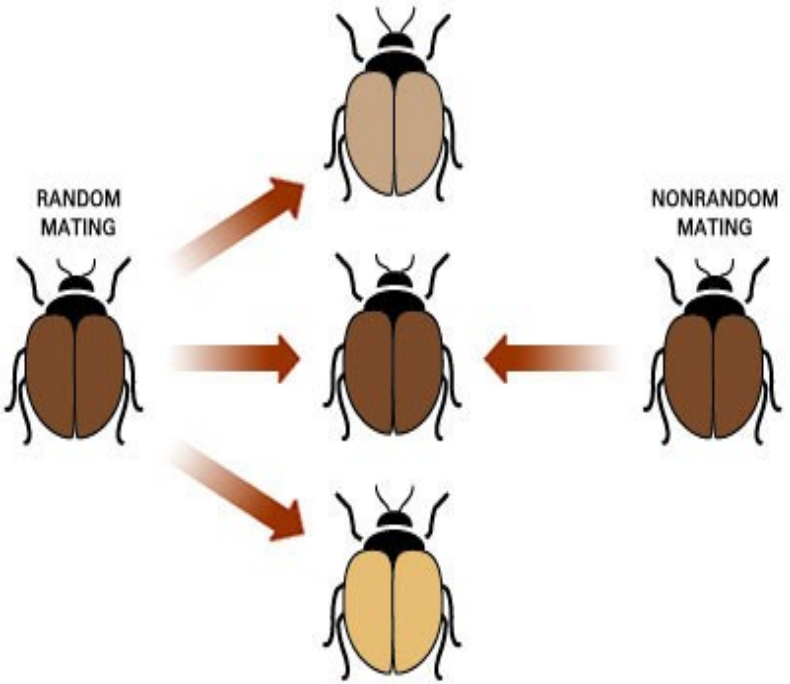
- Study of polymorphisms in a population
  - Which processes introduce polymorphisms into the population?
  - If a polymorphisms exists in a population will it be there for ever?
  - Is there some process that removes polymorphisms from the population?
  - Do polymorphisms exhibit patterns?
  - …

A simple Hypothesis & Model

- Question: Does dominance affect the frequency of alleles?
  - First tested by the famous mathematician G. Hardy at the beginning of the 20th century
  - Assume
    - infinite *population size*
    - *random mating (see next slide)*
    - a *diploid population*
    - A gene **A** with 2 alleles: *A* and *a*
    - Current frequencies (at *generation 0*) of *allele* pairs defining the *genotype*
      - $f_0(AA)$: x
      - $f_0(Aa)$: $2y$
      - $f_0(aa)$: z
    - Current frequencies of gametes (single Alleles) at *generation 0*
      - $f_0(A) = x+y$
      - $f_0(a) = z+y$
    - Sexes have same distribution of the three genotypes *AA, Aa, aa*
  - Does the frequency of occurrence of *A* change over generations?

    → note that we are not asking about the phenotype here!

Random Mating



See Whiteboard

- For the equations

Hardy-Weinberg Equilibrium

- What happens to the frequencies of two alleles at a single gene when the four evolutionary forces (*Natural selection, mutation, migration, genetic drift*) are not acting on a population, and where mating is random?
- If allele frequencies are the same between a parental and offspring generation → no evolution has occurred at that gene
- Serves as null hypothesis in evolutionary biology & population genetics

Effects of finite Population Size Random Genetic Drift

- Populations are of finite size!
  - Does this affect the evolution of allele frequencies over generations?
  - Assume:
    - there are *N* individuals in a diploid population → *2N* chromosomes
    - Frequency of *A* allele is *p*
- What will be the frequency of *A* in the next generation?

Random Genetic Drift

- Definition:

  *Genetic drift is a random process that causes changes in allele frequencies from one generation to the next. Some alleles will be passed on to the next generation disproportionally without being advantageous or harmful. Especially in **small** populations genetic drift is strong due to sampling errors. Alleles can be fixed or get lost by chance.*

The Wright-Fisher Model for finite populations

- Assume a diploid population:
  - *Population size*: *N* (*2N* chromosomes)
  - *Random mating*
  - *Non-overlapping generations* → something like discrete time steps from generation to generation (e.g., annual plants)
  - No *natural selection*
  - *Equal distribution of sexes*
- The *Wright-Fisher* model is the simplest model of evolution for a population of **finite** size

Wright-Fisher Rules/Simulation Example

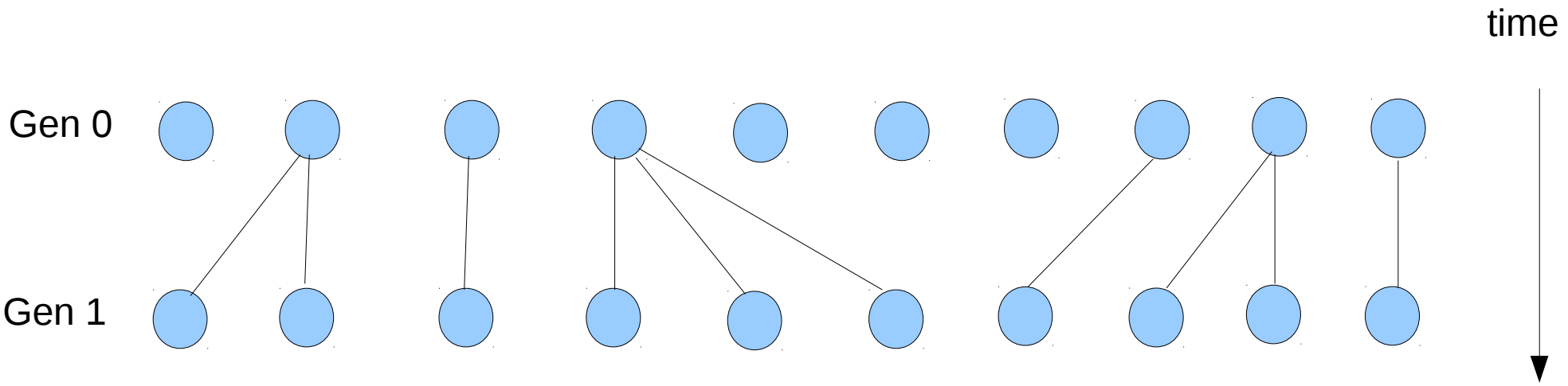
- We assume a constant population → say 10 individuals (or 5 diploid individuals) per generation
- Each individual from the offspring generation picks a parent at random from the previous generation

  → all parents have equal probability to be picked

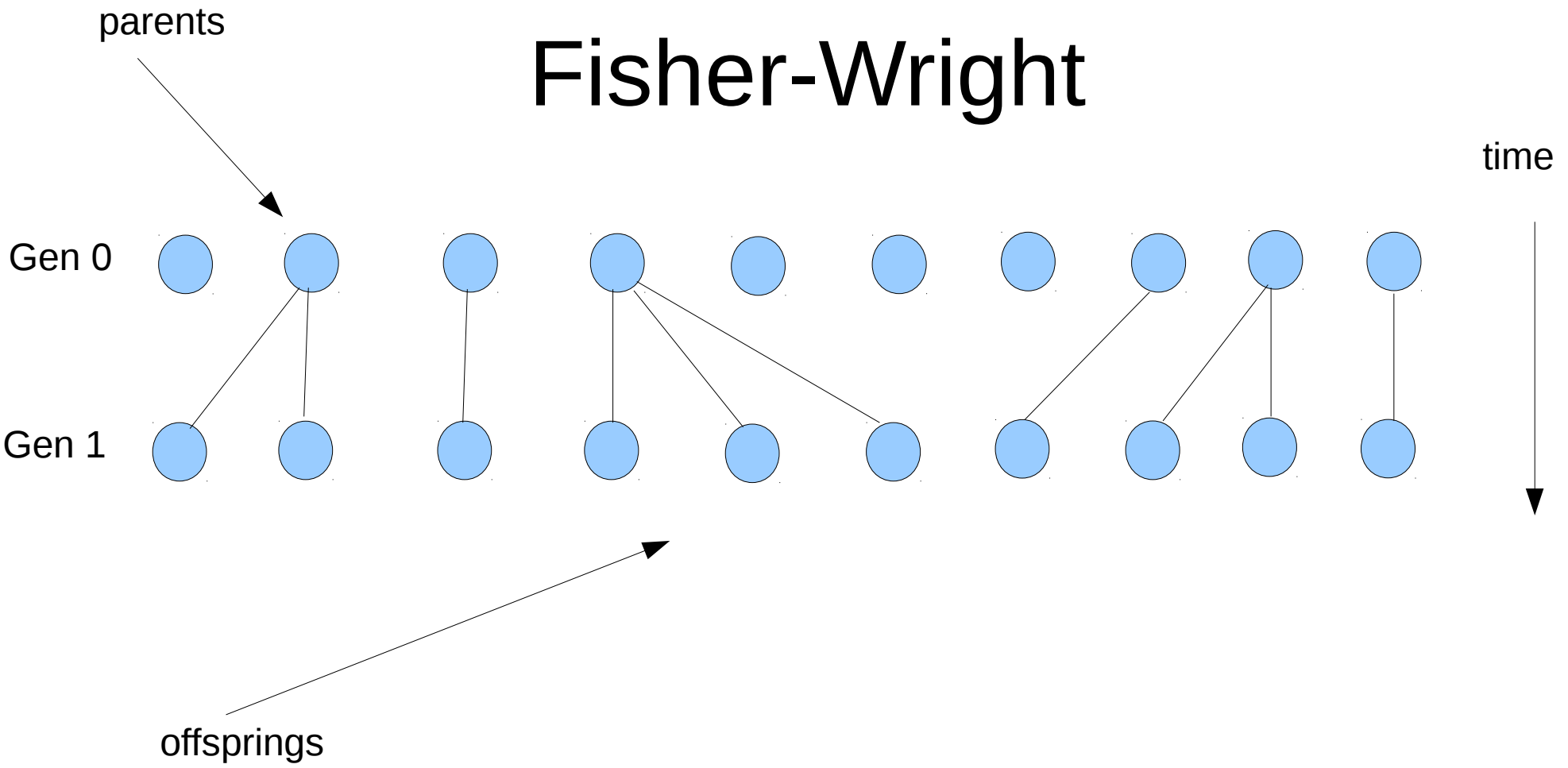
  → a parent can be picked more than once
- Each offspring inherits the genetic information of the parent
- The process and maths are easier to understand if we forget about alleles for a second and just think about individuals

Fisher-Wright

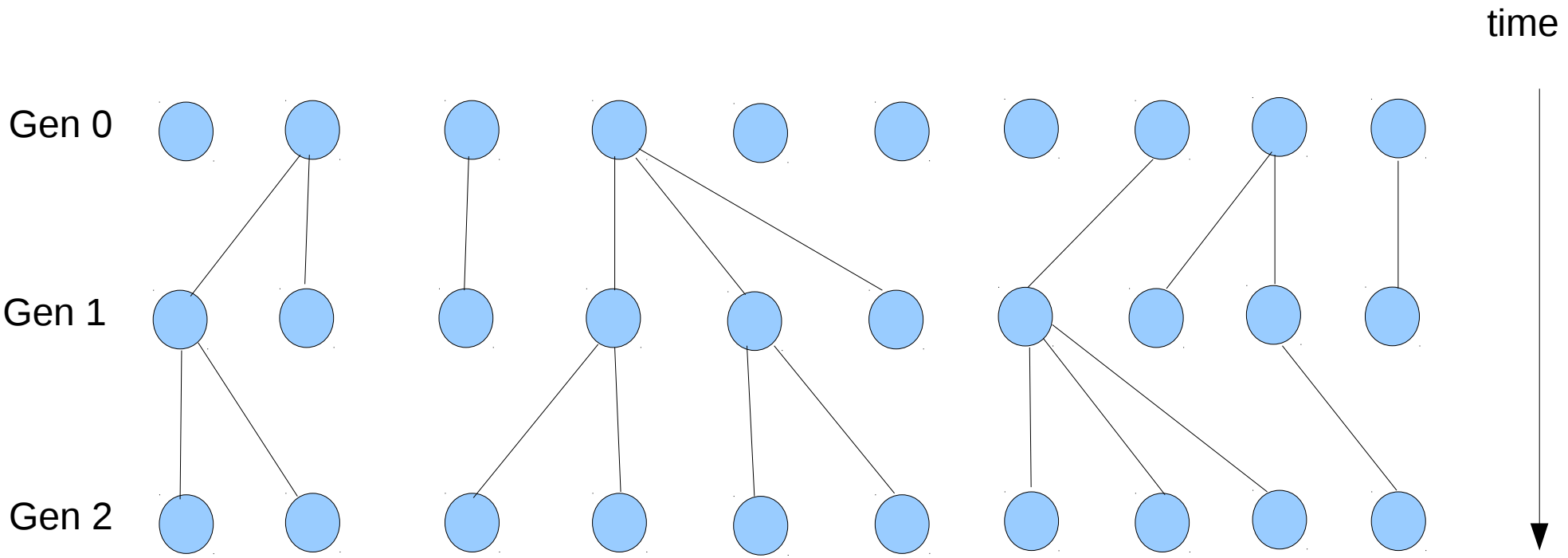
Fisher-Wright



Fisher-Wright



Fisher-Wright



Fisher-Wright
Binomial Random Sampling

- The probability to pick an individual *X* as ancestor of an individual in the next generation is *p = 1/2N*
- If the population remains constant then you have to sample *2N* (*2N = 10* in our example) times from the current generation to construct the next generation with *2N* offsprings
- For every sample, the probability to pick *X* remains constant at *p* → by definition of our model
- The number of offsprings for *X* follows a binomial distribution, thus the probability to pick *X* as an ancestor *k* times is

$$\mathbf{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

- Where *p := 1/2N* and *n := 2N*

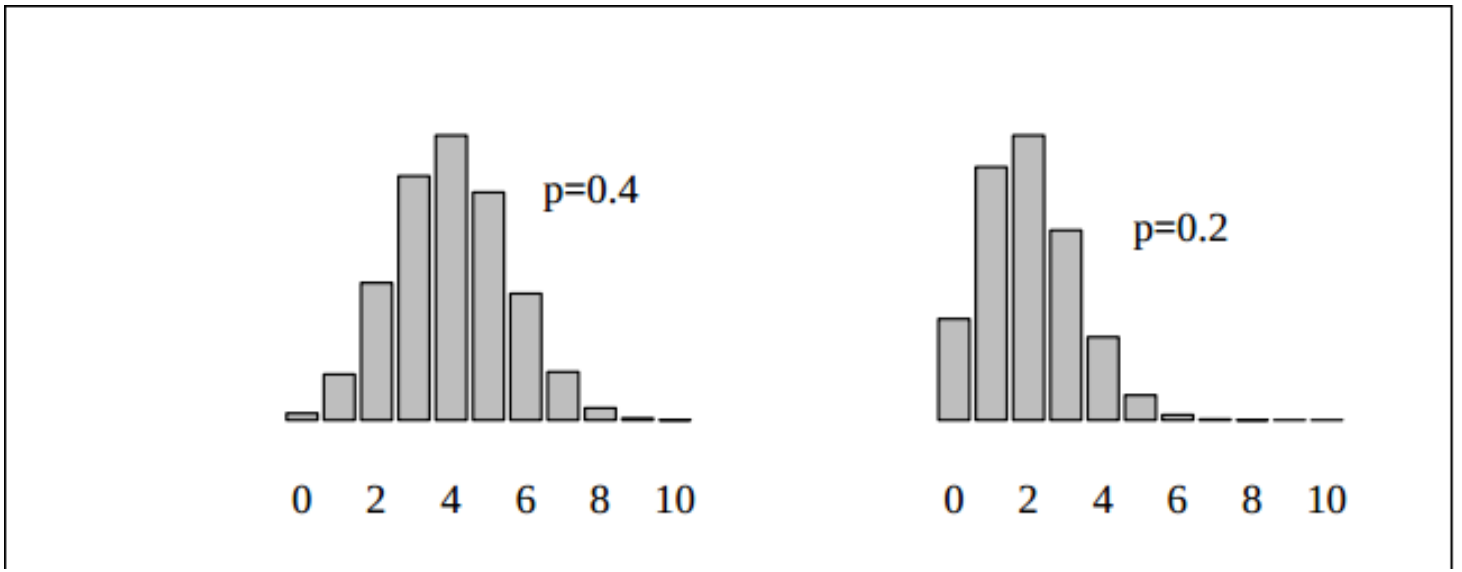
Binomial Random Sampling

- The probability to pick an allele *A* as ancestor of an individual in the next generation is *p = #A/2N*
- If the population remains constant then you have to sample *2N* (*2N = 10* in our example) times from the current generation to construct the next generation with *2N* offsprings
- For every sample, the probability to pick *A* remains constant at *p* → by definition of our model
- The number of offsprings for *A* follows a binomial distribution, thus the probability to pick A as an ancestor *k* times is

$$\mathbf{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

- Where *p := #A/2N* and *n := 2N*

Binomial Sampling of Alleles



Binomial distributions for frequency allele *A* in the next generation for *p=f(A)=0.4* and *p=f(A)=0.2* and a population size of *2N = 10*

Mean and Variance of Allelic Frequency due to drift

- From the properties of the binomial distribution we obtain
  - $E(\#A) = 2N * p$
  - $Var(\#A) = 2N * p * (1 - p)$

The evolution of the frequency of *A* as a Markov Chain

- The evolution of the frequency of *A* over generations is a stochastic process!
- Even if we know everything about the population we cannot predict the state at the next generation with certainty
- One important property of the process: the next state depends only on the current state

  → The process can be modeled as a Markov Chain

Transition Probabilities Fisher-Wright

Frequency in next generation *t+1*

Frequency in current generation *t*

Probability of changing from *i alleles* in generation *t* to *j* alleles in generation *t+1*

$$\text{Prob}\{X(t+1) = j \mid X(t) = i\}$$

$$= p_{ij} = \binom{2N}{j}\left\{\frac{i}{2N}\right\}^{j}\left\{1-\left(\frac{i}{2N}\right)\right\}^{2N-j}$$

$$i, j = 0, 1, 2, \ldots, 2N$$

Population size *2N* haploid or *N* diploid organisms

Example

- Prob of change from *i = 4 → j = 8* Alleles of same type for a population of size *2N := 10* from one generation to the next

$$p_{4,8} = \binom{10}{8}(\frac{4}{10})^8(1-(\frac{4}{10}))^{10-8} = 0.0106168$$

Wright-Fisher Model

- A state of a Markov process is called *absorbing* when the probability to exit this state once we have entered it is 0.
- Are there absorbing states in the Wright-Fisher model?

Probability to enter an absorbing state

- Useful to study the evolution in a Wright-Fisher model as a Markov Chain because you can answer a lot of questions via standard Markov Chain theory.
- For instance: What is the probability that the population will end up (after how many generations?) in the absorbing state where *f(A)=1?*

  → this is also called *fixation*
- Given that the frequency of *A* is *#A/2N* the probability that *A* will become fixed is *#A/2N*
- For details, see: http://people.sc.fsu.edu/~pbeerli/isc5317-notes/pdfs/01-populationmodels.pdf

Random genetic Drift

- The change in allele frequencies over generations in finite populations due to stochasticity (re-sampling) is called *random genetic drift*
- What is the effect of random genetic drift on the polymorphism level?
- Since our human population is finite, why do we still observe polymorphisms?

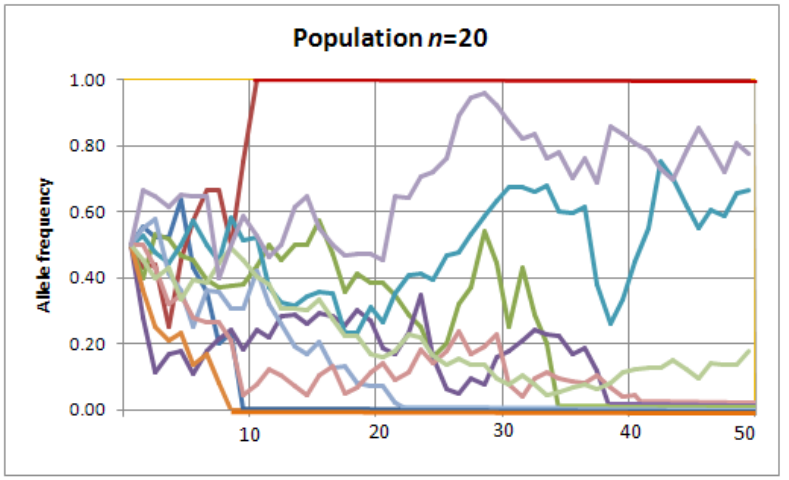
Heterozygosity and Genetic Drift

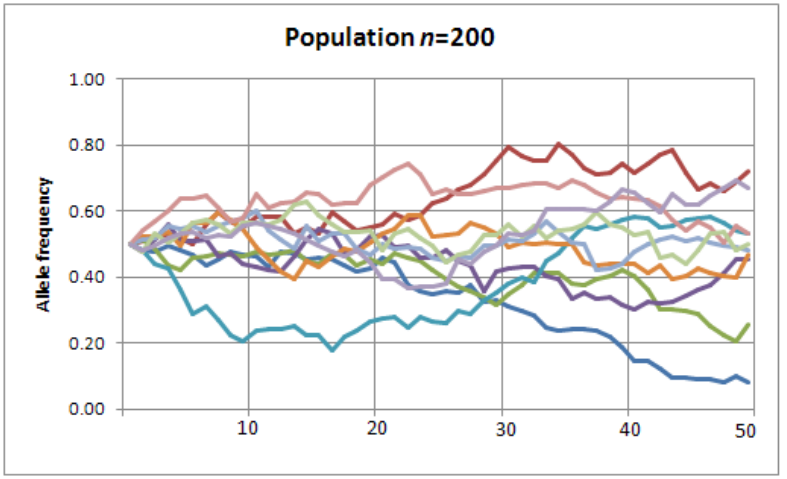
- Reduction of polymorphism is quantified by the degree of homozygosity → The probability that two alleles are identical

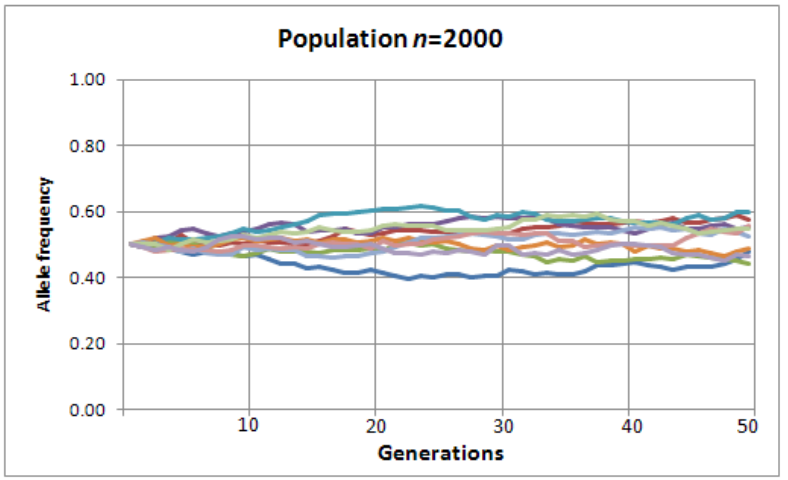
  → *heterozygosity (1 – homozigosity)* at generation *t* is defined as: Het_t

- Assume a population of size *2N*
- We can define the heterozygosity recursively as $Het_t = Het_{(t-1)} (1 – 1/2N)$
- Thereby we obtain: $Het_t = Het_0 (1 - 1/2N)^t$

Probability that two randomly
Alleles are different







Initial allele frequency
f(A) = 0.5

Mutation-Drift Balance

- Genetic drift removes polymorphisms (SNPs) from the population
- Mutations introduce polymorphism (SNPs) into the population
- Is there some balance?

Heterozygosity at mutation – drift balance

- *Define:*
  - *Het*: heterozygosity
  - *-1/2N * Het*: Loss of heterozygosity **per generation** due to genetic drift
  - *μ:* mutation rate **per gene** (remember two alleles per gene!) and **per generation**
  - *2μ(1 – Het)*: gain of heterozygosity due to mutation
- Pick two alleles
- Consider transition from generation $t \rightarrow t + 1$
- The probability that they are identical is: *(1-Het)*
- If they are identical, the probability that one out of the two will mutate is *2μ*

  → *2μ(1 – Het) gain in* heterozygosity due to mutation
- Overall: $Het_{t+1} = Het_t - 1/2N * Het_t + 2\mu (1-Het_t)$

  $\Delta Het = -1/2N * Het_t + 2\mu(1-Het_t)$
- $\Delta Het = 0 \rightarrow Het = (4\mu N) / (1 + 4\mu N)$

Rate of Evolution by mutation and genetic drift

- Rate of Evolution = The probability of a new mutation to arise in the population and to eventually become fixed
- Assume
  - μ is the probability of mutation per generation and per individual
  - $2N$ individuals → $2N\mu$ mutations per generation
- The probability that a particular mutation will be fixed is $1/2N$
- Thus, the rate at which a mutation will arise and fix in the population is $1/2N * 2N\mu = \mu$
- Why is this result remarkable?

Natural Selection

- So far, we have assumed that the probabilities of *fitness* and *reproduction* are the same for each individual, independently of its genotype
- Consequently, a random individual at generation $t+1$ descends from any individual in generation t, with the same probability
- We denote the ability of an individual “to survive and reproduce” as *fitness*
- We assume that *fitness* depends on the genotype

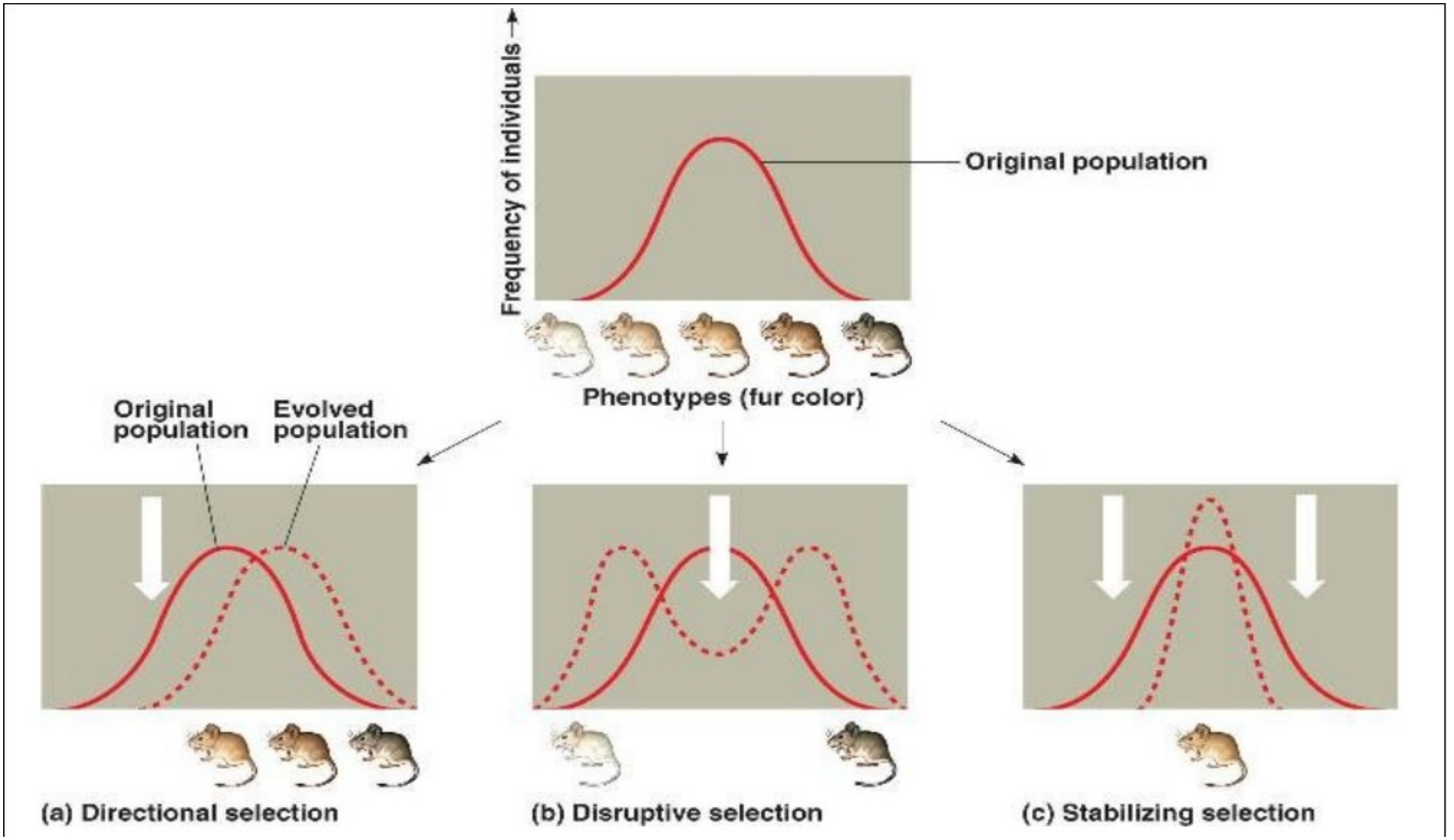
Natural Selection

- The term *selection* means that a genotype reproduces more frequently than others
- If a certain genotype, e.g., *AA* has better/higher fitness

  → it will fix in the population after several generations

  → consequently, the allele *A* will also fix
- We say: *Natural selection* has favored allele *A*
- In this case, the *natural selection* on *A* is termed *Positive Selection*

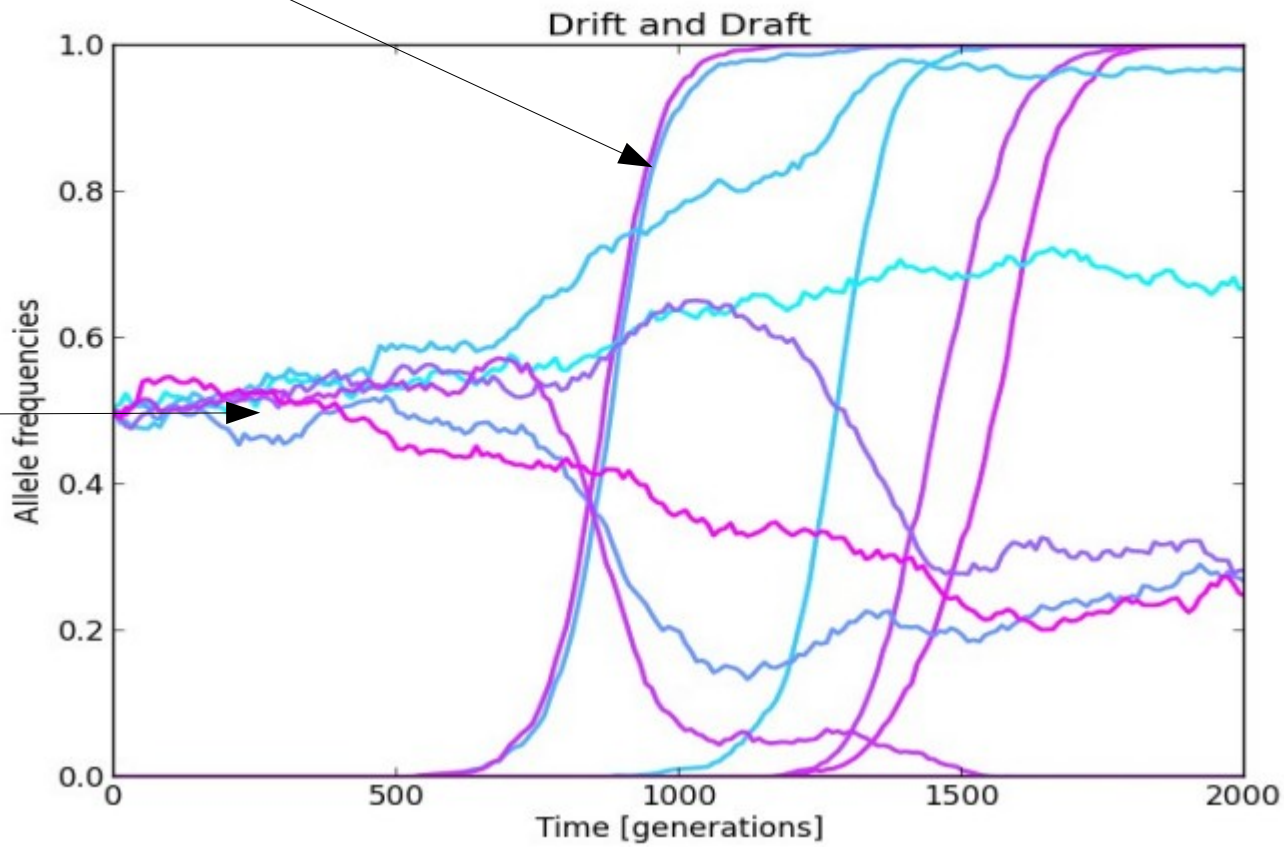
Different Modes of Selection



The Frequency Evolution of *A* under Positive Selection

Positive selection

Random genetic drift



Summary Statistics

- Summary statistics provide a summarized description of the dataset, e.g., the number of polymorphic sites
- Summary statistics are important because:
  - They allow to estimate parameters of the population
  - They help us to assess if positive selection occurred
- Differences to phylogenetics
  - Given the data (MSA of individuals)
  - We don't reconstruct a population tree for the individuals
  - We simulate evolution under different scenarios (including more complex models with changing population sizes etc)
  - Then we compare if one of the scenarios fits the summary statistics (e.g. # SNPs) of our empirical dataset