

# Introduction to population genetics & evolution

# Course Organization

- Exam dates:
  - Feb 19
  - March 1<sup>st</sup>
  - Has everybody registered?
  - Did you get the email with the exam schedule
- Summer seminar: “Hot topics in Bioinformatics”
  - On Thursdays 11:30 – 13:00 room -120

# Outline

- Evolution
- Genetics
- Population genetics

# Tutorial

- <http://www.stochastik.uni-freiburg.de/homepages/pfaffelh/teaching/2008S/S/Manuscript.pdf>

# What is evolution

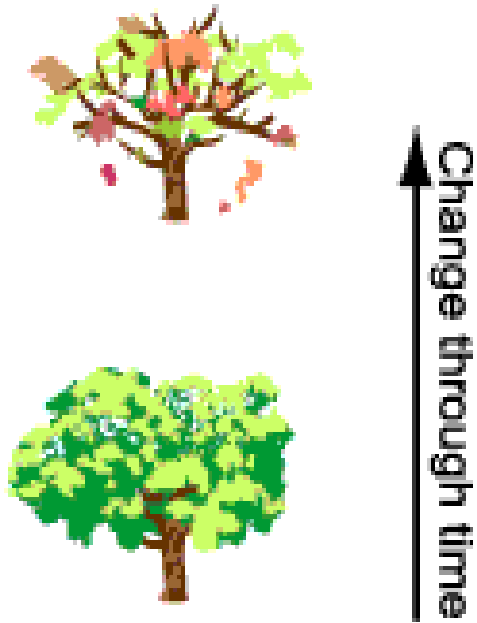
- In a broad sense evolution = change over time
- Languages evolve = Languages change
- Galaxies evolve = Galaxies change
- Political systems evolve = Political systems change

# Biological Evolution

- In biology one more condition is needed

?

# Biological Evolution



- Several trees drop their leaves during winter

- The color of leaves change over time

Is this evolution?

# Biological Evolution



The frequency of red bugs has changed in the population.

**Is this evolution?**



# Biological Evolution

- A population of insects consists of 5 white and 5 black individuals.
- This means: frequency of white = 50%, frequency of black = 50%
- Suddenly, out of 10 individuals, 7 die, and only 3 survive
- Let's say that these 3 are: 2 white, i.e. frequency of white = 66%, frequency of black = 33 %
- The population has changed
- Is this evolution?

# Biological Evolution

- A population of insects consists of 5 white and 5 black individuals.
- This means: frequency of white = 50%, frequency of black = 50%
- 3 random individuals decide to leave and establish a new colony ( 2 white + 1 black)
- The new colony consists now of 66 % white and 33 % black, which is different than the initial population.
- Is this evolution?

# Biological Evolution

- In biological evolution “the change” is not a sufficient condition for evolution
- Change should be inherited
- The reason of change is not important

# Biological Evolution and the nature of change

- It is evident from previous examples that when we talk about biological evolution we mean:
- Change of frequencies of features from generation to generation (e.g. resistance to antibiotics, color etc)
- This feature should be inherited from generation to generation → inheritance through ????

# Examples of Biological evolution2

- Resistance to drugs / soap, vaccines, antibiotics

# The basic unit of biological evolution

- From the previous examples, what do you think is the basic unit of biological evolution? i.e. what is really evolving?
  - An individual?
  - The population?
  - The species?
  - Something else?

# Units of Evolution

- The population
- The Gene
- The species
- The genome of an individual
- You must define “at which level do evolutionary forces act” and “what competes with what” in order to define the unit of evolution

# The population as the basic unit of evolution

- Population evolves because the frequency of its features changes.
- Frequency of features (e.g. resistance to antibiotics) changes because of
  - Chance
  - Response to some pressure (antibiotics for microbes, climate change)
  - Migration
  - Mutation

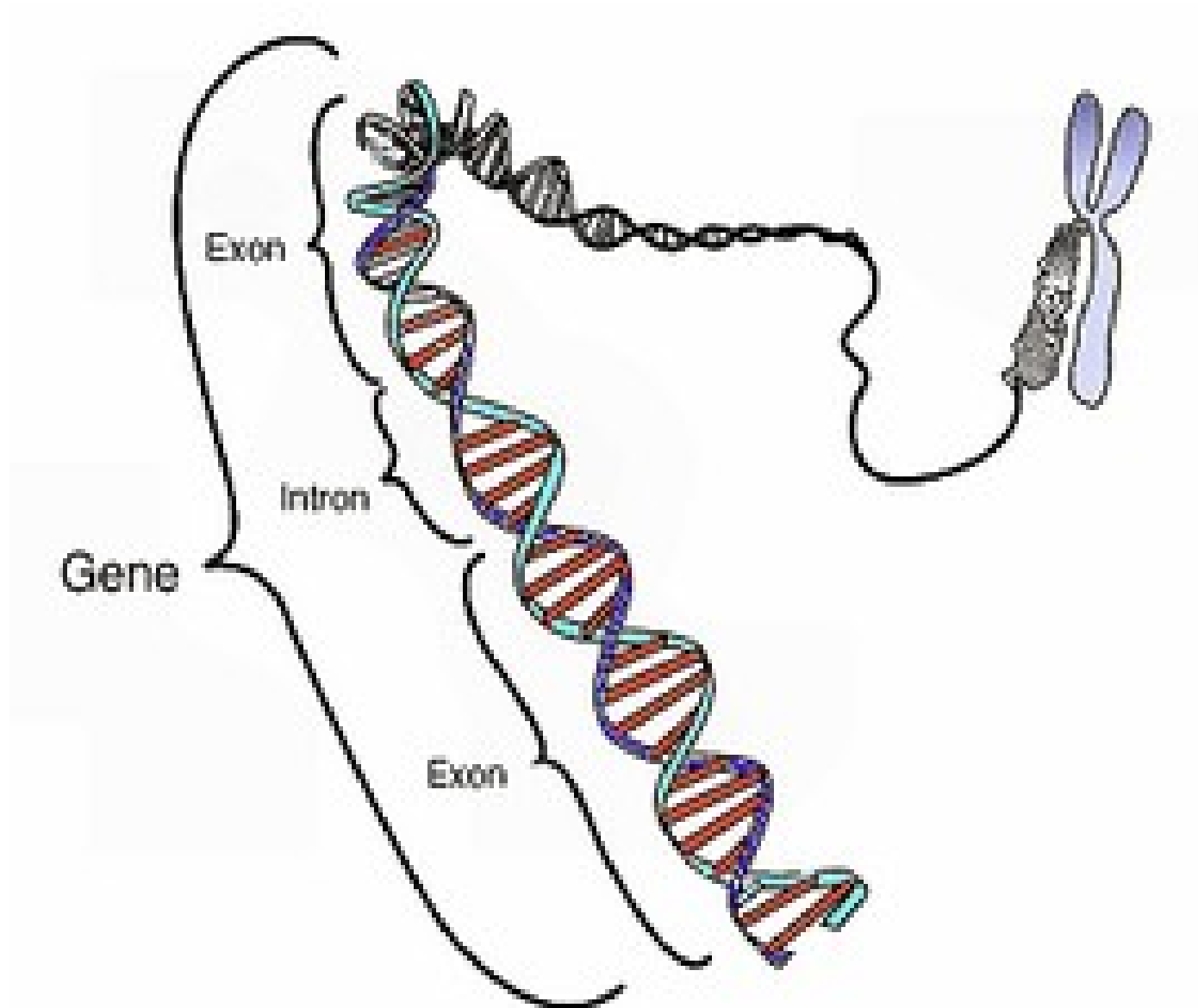


# Introduction to genetics

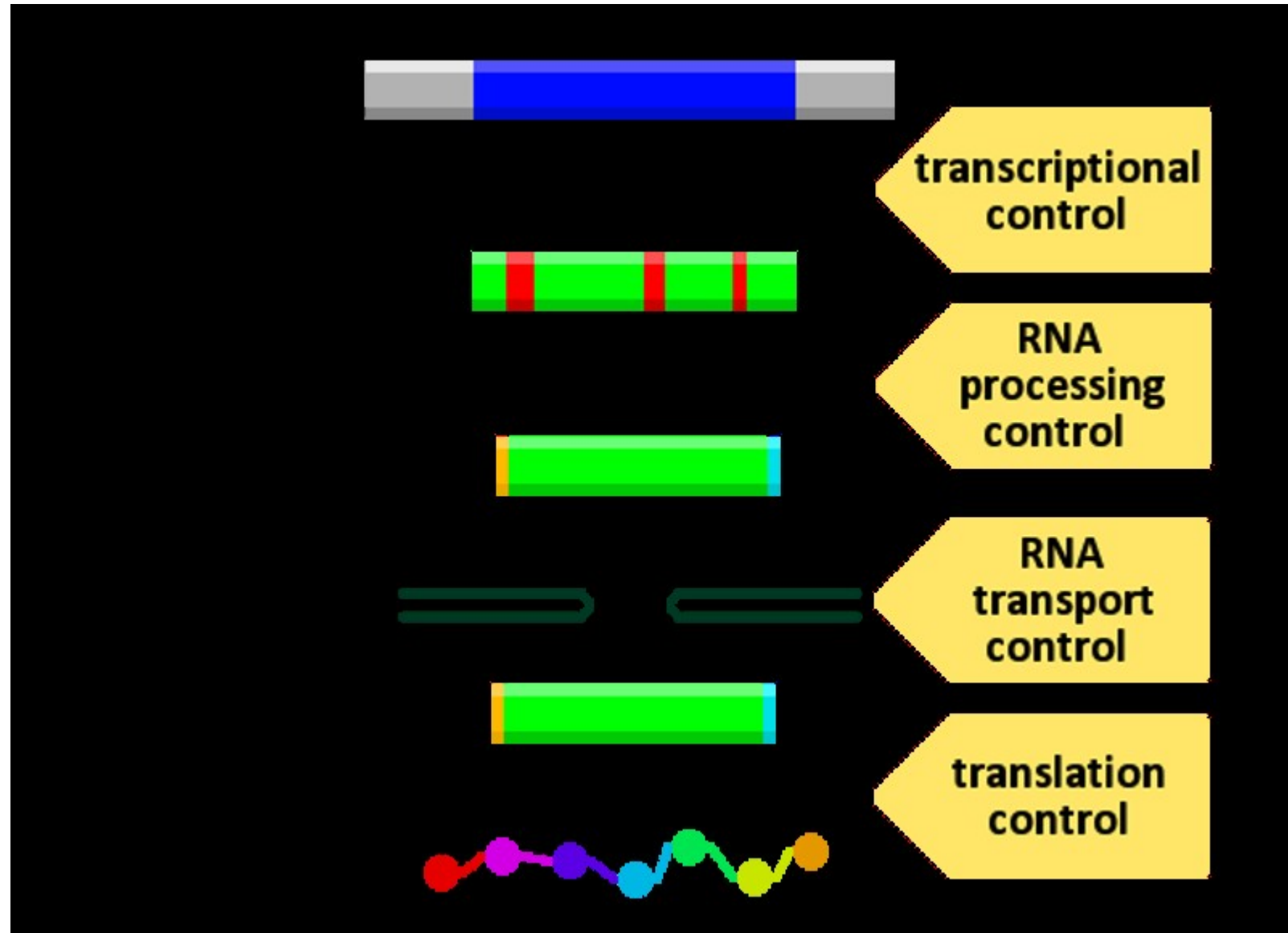
# The genome



# Chromosomes and Genes



# Genes are known for ...



# Genes as units of information

- Genes encode some information
- For example, a gene **A** encodes the color of the eye  
(about 15 genes determine the eye color, but let's assume that there is only one)
- If **A** has some specific form  $A \rightarrow \text{color} = \text{blue}$
- If **A** has some other form  $a \rightarrow \text{color} = \text{brown}$
- What does different form mean?

# Genes as units of inheritance

- Genes are pieces of DNA
- They are inherited from generation to generation, from parents to kids
- Humans inherit one allele from the father and one allele from the mother
- Such organisms are called diploid
- Do you know other diploid or haploid organisms?

# Mendel and inheritance

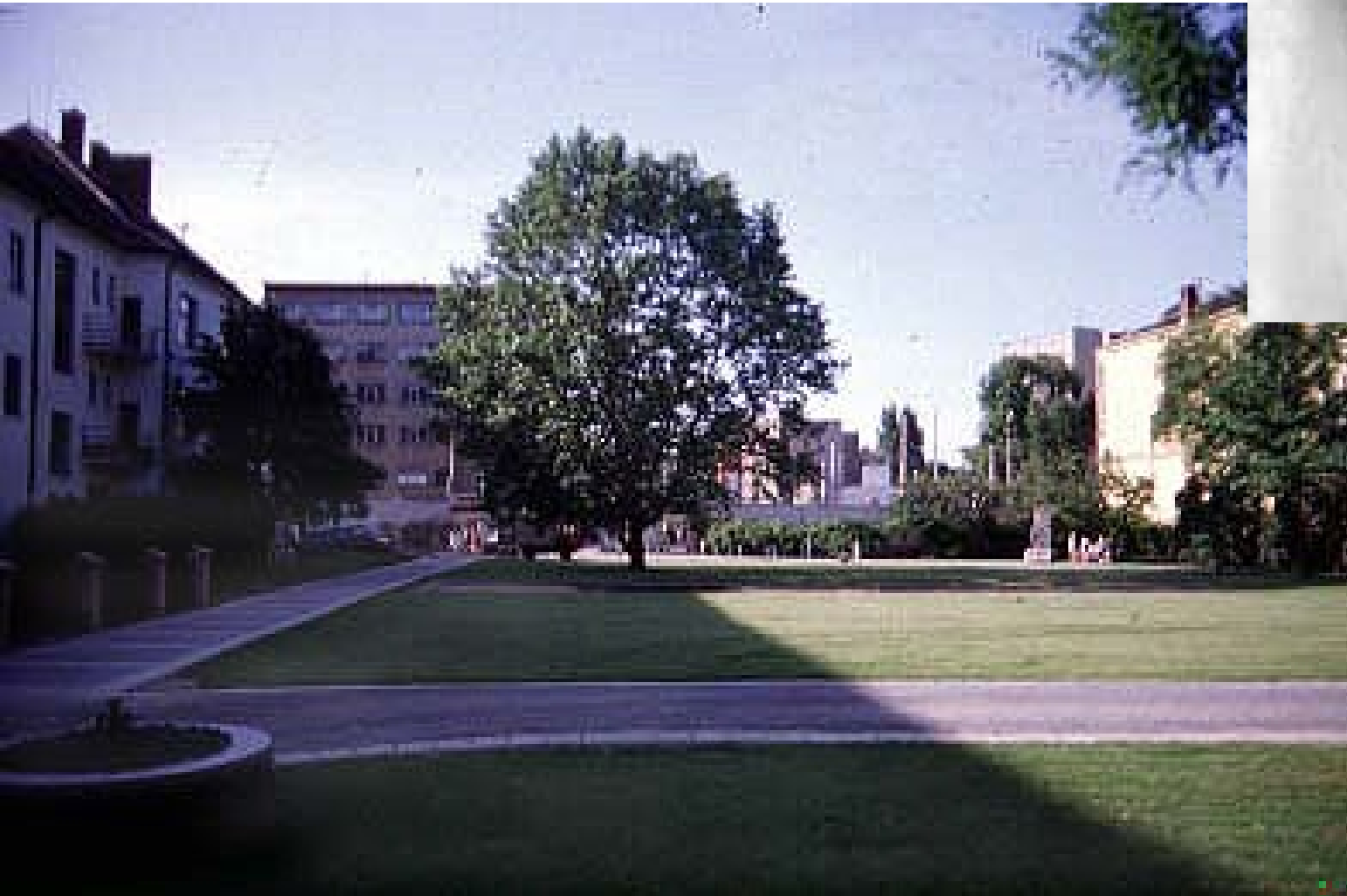
( slides from

<http://science.kennesaw.edu/~echen1/> )

# Gregor Mendel



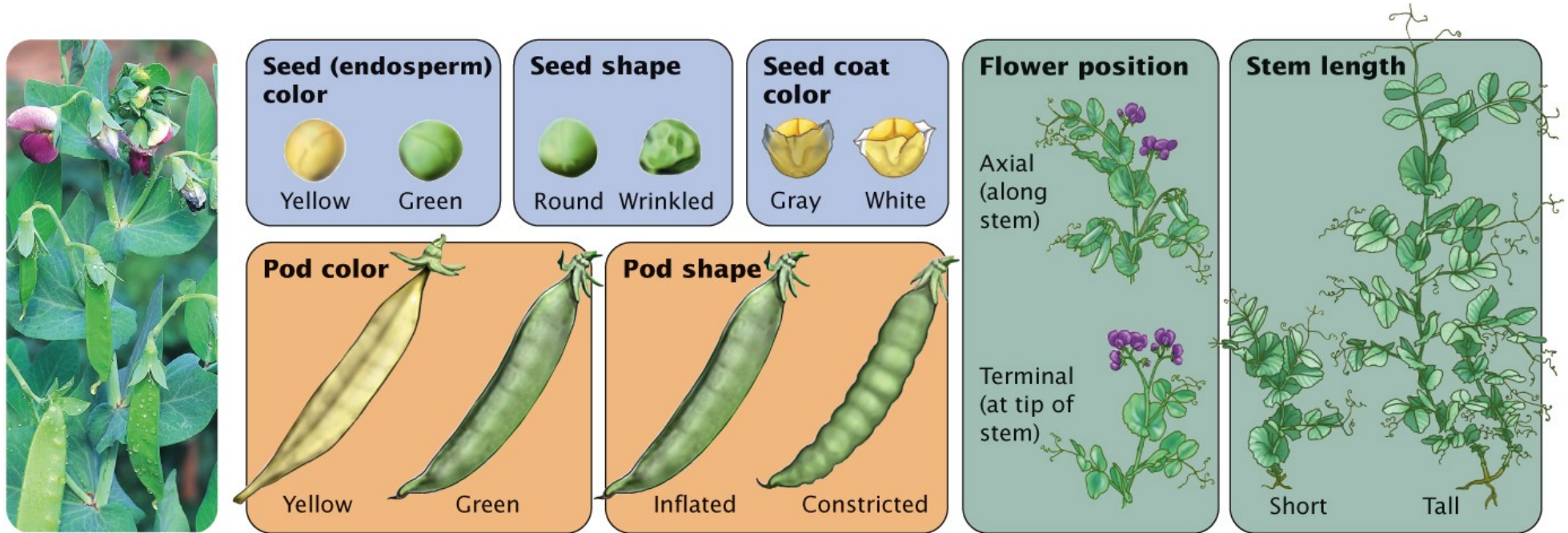
Culver Pictures



The garden of the Augustinian Convent in Brno. This view is looking towards the entrance to the garden, with *Mendlovonamesti*, Mendel Square, beyond. In the shadows in front of the tree can just be seen part of the foundations of the greenhouse that Mendel used. His peas were planted in the beds against the building on the left. *A MENDEL PHOTO-ESSAY –simonmawer.com*



# Traits of Pea Plant Studied by Mendel



Fig\_03-01 *Genetics, Second Edition* © 2005 W.H. Freeman and Company

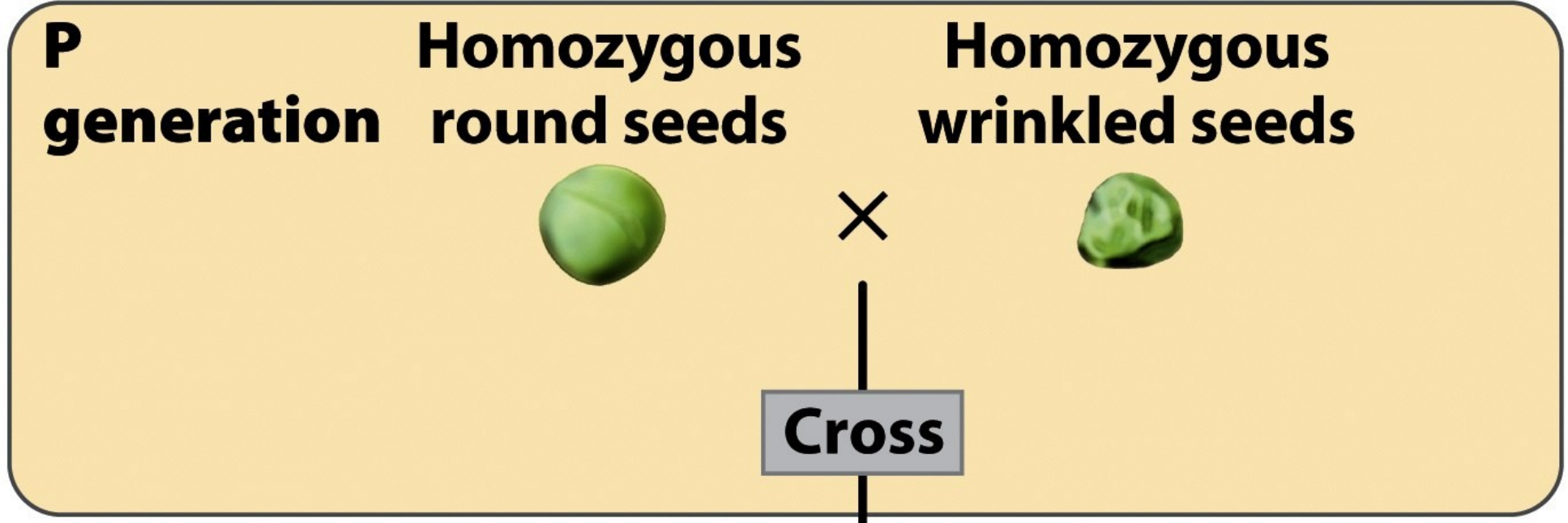
# Some terms

- Genotype (for a certain gene):
  - The set of the two corresponding alleles that the organism carries.
- Phenotype (for a certain gene):
  - What we observe for the trait that the gene controls: e.g. brown eyes, yellow color of the pea etc

# Dominance

- In Mendel's experiment, when an individual has the RoundWrinkled Genotype → Round
- Round allele is **dominant** and Wrinkle allele is **recessive**
- What is the phenotype of
  - RR
  - RW
  - WR
  - WW
- If there is no dominance then the phenotype is intermediate

Mendel observes that the recessive trait “disappears” in the 2<sup>nd</sup> generation and reappears in the next



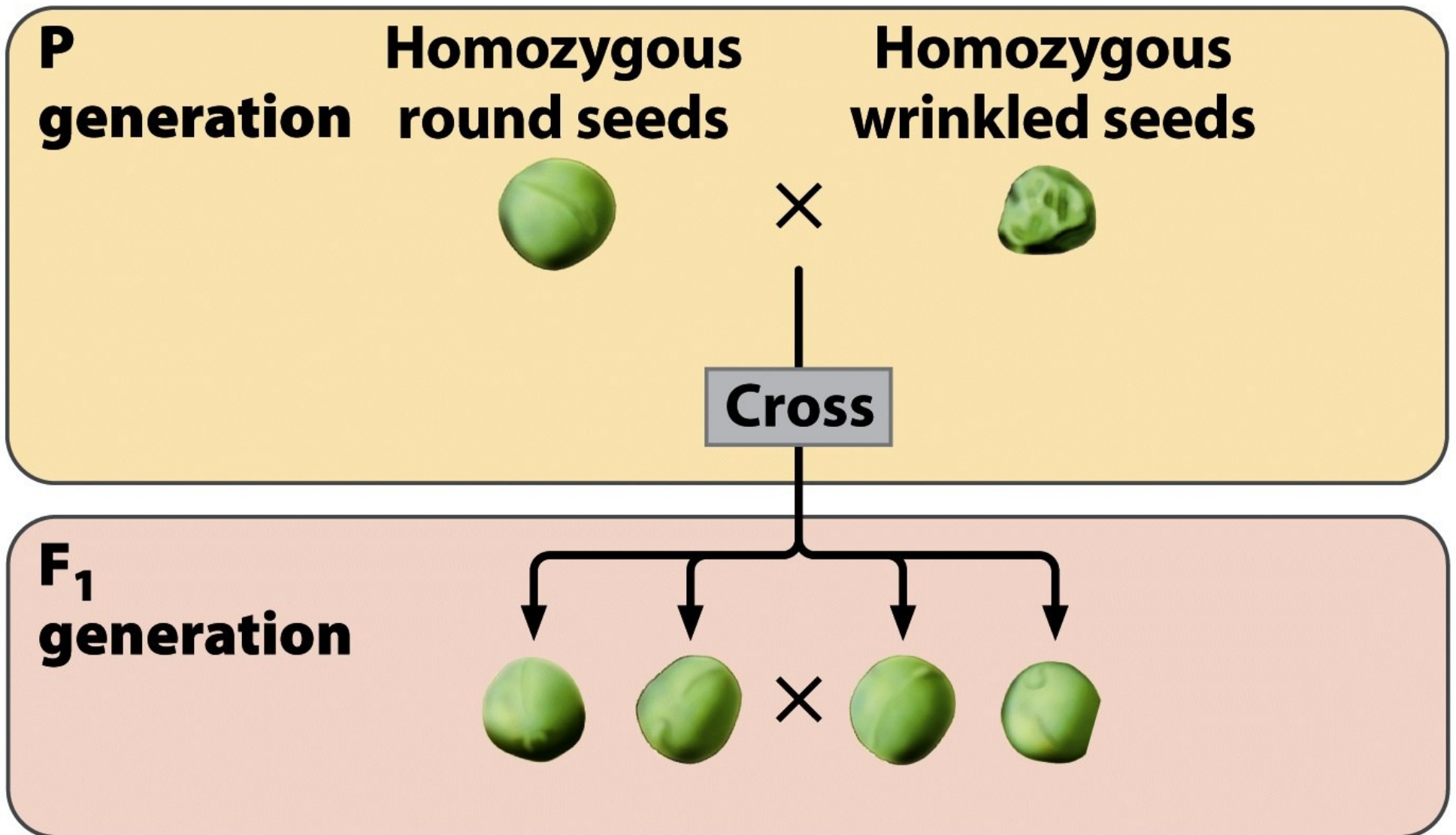


Figure 3-3 part 2  
*Genetics: A Conceptual Approach, Third Edition*  
© 2009 W. H. Freeman and Company

**F<sub>1</sub>**  
**generation**



**Self-  
fertilize**

**Results**

**F<sub>2</sub> generation**

**5474 round**



**1850 wrinkled**



**Conclusion:** The traits of the parent plants do not blend. Although F<sub>1</sub> plants display the phenotype of one parent, both traits are passed to F<sub>2</sub> progeny in a 3 : 1 ratio.

Figure 3-3 part 3

*Genetics: A Conceptual Approach, Third Edition*

© 2009 W.H. Freeman and Company

# Mendel's 1<sup>st</sup> Law—Principle of Segregation

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

Observation: the F<sub>2</sub> generation shows all traits from P<sub>0</sub>, even though the F<sub>1</sub> parents do not show all traits.

Conclusion: F<sub>1</sub> must receive *something* that causes the “hidden” trait revealed in F<sub>2</sub>, in addition to the trait that F<sub>1</sub> show.

# Factor inheritance vs blending inheritance

- Mendel showed that “factors” (now we call them genes) are responsible for the inheritance.
- What would happen in case that the children would just be a blend of the parents?



# Inheritance of alleles

- A diploid organism has 2 alleles for every gene
- One allele is inherited by his mother and one by his father
- Each parent will give only *one* of his alleles to his kids
- For a certain allele, there is a 50 % probability to have obtained it from the mother and 50 % to have obtained it from the father

# Inheritance of alleles

- It's typical to denote a gene with the capital letter **A** and alleles as 'A', 'a' if two alleles exist or as A1, A2, A3, ... if more alleles exist  
(**'A'** denotes both an allele and the gene)

# Why are we interested in alleles?

- In population genetics we study the evolution of populations, i.e., how the frequency of alleles changes over time AND the reasons for these changes
- This means that we are very interested in the evolution of POLYMORPHISMS

# Polymorphisms

- Polymorphic locus (gene)
  - A gene is polymorphic in the population when there are multiple alleles present e.g. A, a
- Polymorphic site
  - We can now sequence the whole DNA sequence of organisms.
  - After alignment we can observe sites with more than one state. Such sites are called polymorphic.



# Polymorphic sites

- After alignment, we observe that at some sites, there are more than one nucleotide state.
- Polymorphic sites: SNPs = Single Nucleotide Polymorphism
- Modern population genetic analyses consider SNPs

# Population genetics

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

# A very simple first Hypothesis

- Dominance affects the frequency of alleles
  - This hypothesis was first tested by the famous mathematician G. Hardy in the beginning of the 20<sup>th</sup> century
  - Assume infinite population size, **random mating**, diploid population, ....
  - A **gene** with 2 alleles: A and a
  - Current frequencies of AA, Aa, aa is x, 2y, z, respectively
  - Does the frequency of the 'A' allele change over generations?



# The effect of finite population size

## RANDOM GENETIC DRIFT

- Populations are not of infinite size.
- Does the finite population size affect the evolution of the allele frequencies over generations?
- Assume that there are  $N$  individuals in a population  
→  $2N$  chromosomes (individuals are diploid)
- Frequency of 'A' alleles  $p$ .
- What will be the frequency of 'A' alleles in the next generation?

# The Wright-Fisher model

- Assume a diploid population:
  - Size:  $N$  (number of chromosomes  $2N$ )
  - Random mating
  - Non-overlapping generations
  - No selection (we will discuss that later)

This model is called the Wright-Fisher model, and is the simplest model for the evolution of a population

# simulation

- <http://darwin.eeb.uconn.edu/simulations/drift.html>
- What is the effect of the population size?
- What is the meaning of the values 0 and 1 on the y axis?
- What happens if the frequency of the alleles becomes 0 or 1?
- How can the frequency of 'A' assume values between 0 and 1 again after it became either 0 or 1 ?

# Binomial random sampling

- The probability to pick an 'A' allele is  $p = (\#A)/2N$
- If the population remains constant then you repeat sampling  $2N$  times.
- For every sample, the probability to pick an 'A' is still  $p$ .
- Thus, the probability to pick  $k$  'A' alleles is:  
$$P(k;N, p) = p^k (1 - p)^{2N-k} \binom{2N}{k}$$

# Mean and Variance of allelic frequency due to drift

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

# The evolution of the frequency of 'A' as a Markov Chain

- The evolution of the frequency of 'A' is a stochastic process. Even if we know everything about the population we cannot predict *for sure* the state at the next generation
- One important property of the process: the next state depends only on the current state
- The process can be modelled as a Markov Chain.

# Transition probability in the “simple” Wright-Fisher model

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

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

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

# 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?



# The probability to enter an absorbing state

- What is the probability that the population will end up in the absorbing state where the frequency of 'A' is 1 (= fixation)?
- Given that the frequency of 'A' is  $i/2N$ , the probability that 'A' will be **fixed** is  $i/2N$ .
- In general it is useful to study the evolution in a Wright-Fisher model as a Markov Chain because you can get several insights by using the standard theory of Markov Chains.

A retrospective approach to get the probability that the 'A' will be fixed

Eventually all genes in the population will have descended from one gene in the parental generation. The probability that this is an *A* gene is, by symmetry, simply the initial proportion  $i / 2N$  of *a* genes in the population.

# Random Genetic Drift

- The change of the allelic frequency due to stochasticity (resampling) 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 homozygosity = The probability that two random alleles are identical (both A or both a)
  - Heterozygosity ( 1 - homozygosity) at generation  $t$ :  $H_t$
- Assume a population of size  $2N$ , and let it be in generation 1
  - Pick two alleles. The probability of picking the same allele twice is  $1/2N$ . In this case  $H_1 = 0$ .
  - With probability  $(1 - 1/2N)$  they will be different. In this case, the probability that they have different state is  $H_0$ . Thus,  $(1 - 1/2N)$ .
  - In total,  $H_1 = H_0 (1 - 1/2N)$
  - Continuing the argument over generations:  $H_t = H_0 (1 - 1/2N)^t$

# Mutation-drift balance

- Drift removes polymorphisms from the population
- Mutations introduce polymorphism in the population
- Is there some balance?

# Heterozygosity at mutation – drift balance

- $H$ : Heterozygosity
- $-1/2N * H$  : Loss of Heterozygosity due to drift
- $\mu$  is the mutation rate per gene and per generation
- $2 \mu * (1 - H)$  : gain due to mutation
  - There are two alleles. The probability that they are the same is  $(1-H)$ . Given that they are the same, the probability that either one or the other will mutate is  $2 \mu$ . Thus  $2 \mu (1 - H)$  is the gain of heterozygosity due to mutation
- In total:  $H' = H - 1/2N * H + 2 \mu (1-H)$   
 $\Delta H = -1/2N H + 2\mu(1-H)$
- $\Delta H = 0 \rightarrow H = (4 \mu N) / (1 + 4 \mu N)$

# Rate of Evolution by mutation and drift

- Rate of Evolution = The probability of a new mutation to arise in the population and to eventually be fixed
- Assume  $\mu$  the probability of mutation per generation and per individual.
  - $2N$  individuals  $\rightarrow 2N \mu$  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

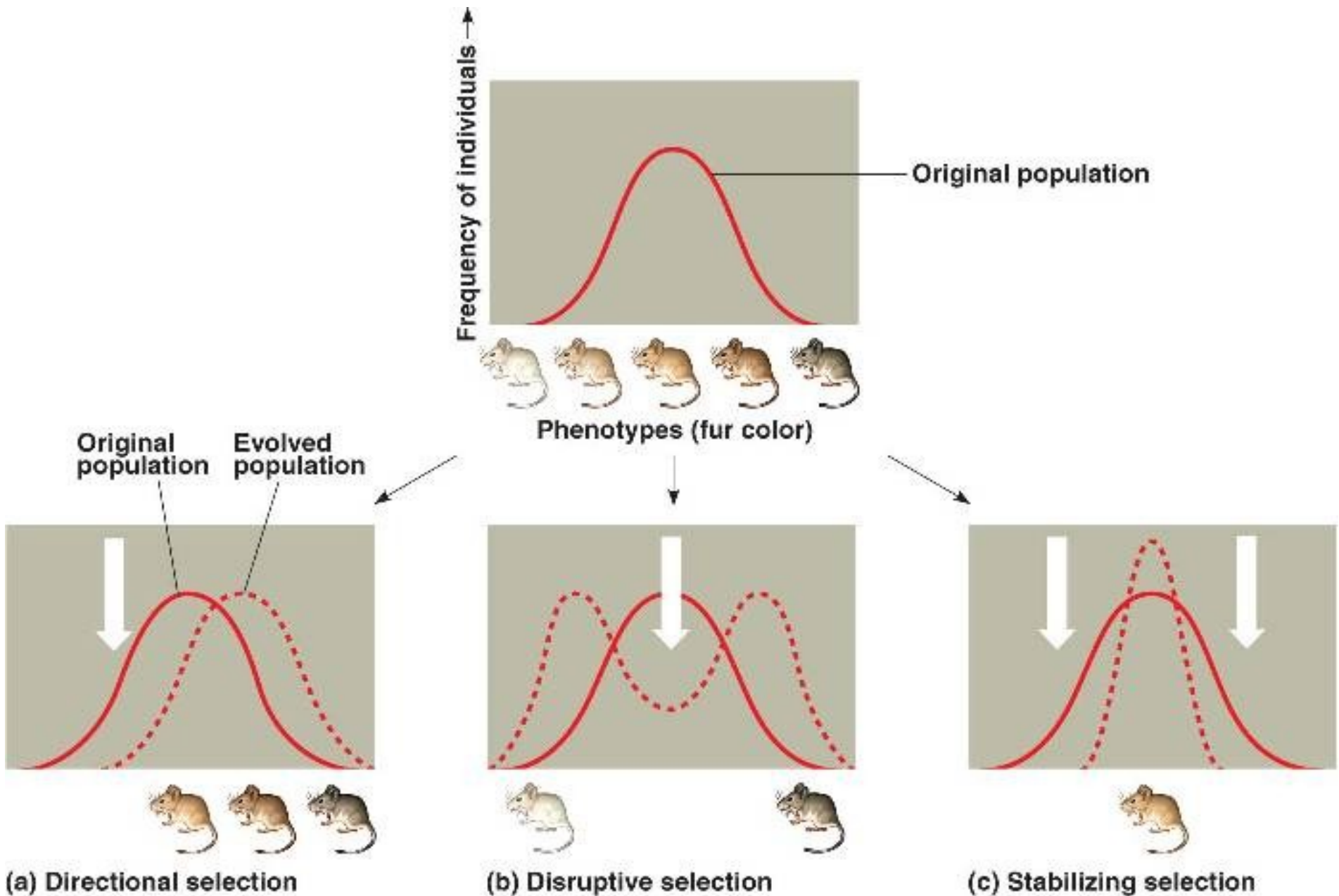
- We have assumed that the probabilities of viability and reproduction is the same for each individual, independently of its genotype
- Equivalently, a random individual at generation  $(t+1)$ , has been descended by any individual from generation  $t$ , with the same probability
- We quantify the ability of an individual “to survive and reproduce” as fitness
- We assume that fitness depends on the genotype.
- If there are 2 alleles for one gene, how many fitness values exist for this gene?



# Natural selection

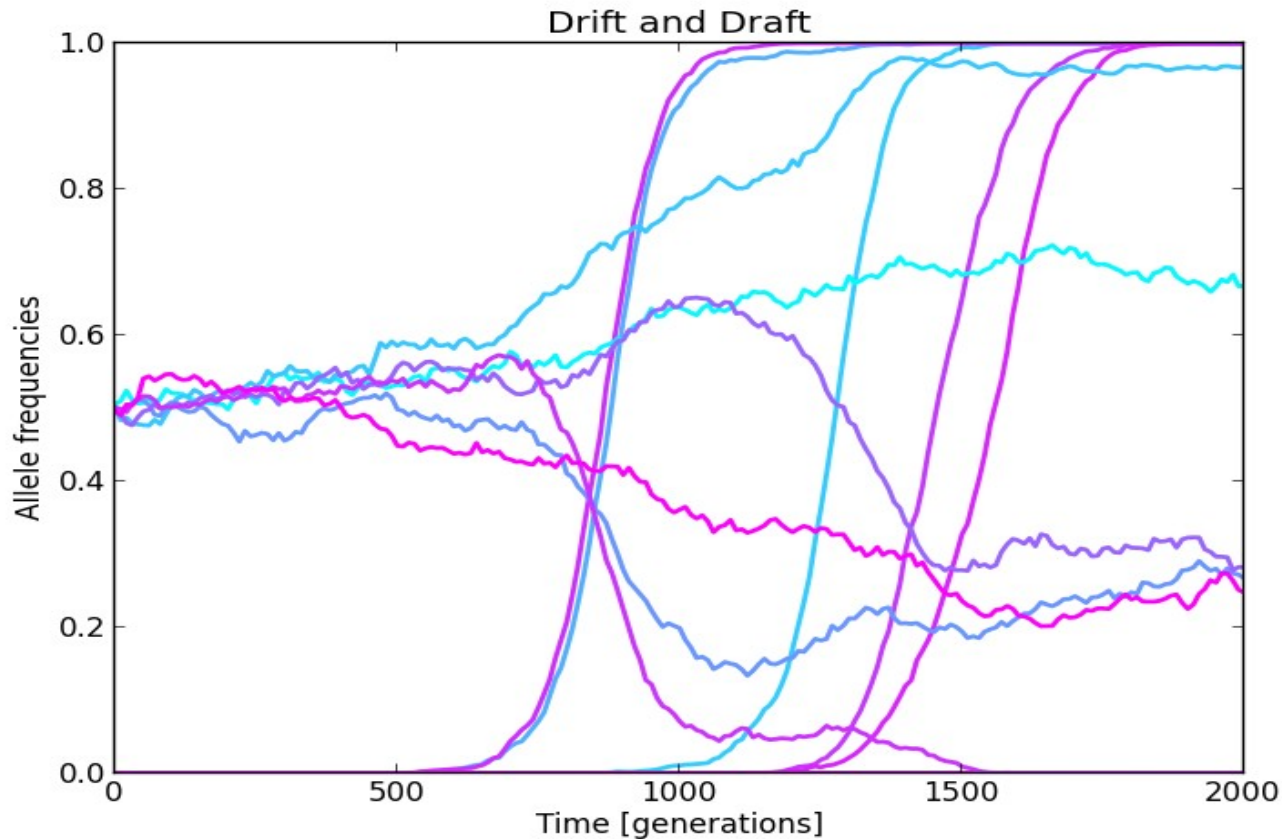
- The term 'selection' means that a genotype is 'chosen' by nature to reproduce more often than the rest.
- If a certain genotype e.g. 'AA' has higher fitness, then after several generations it will fix in the population (consequently, the allele 'A' will fix as well).
- Natural selection has favored allele 'A'.
- In this case, selection on 'A' is termed “Positive Selection”

# Different modes of selection



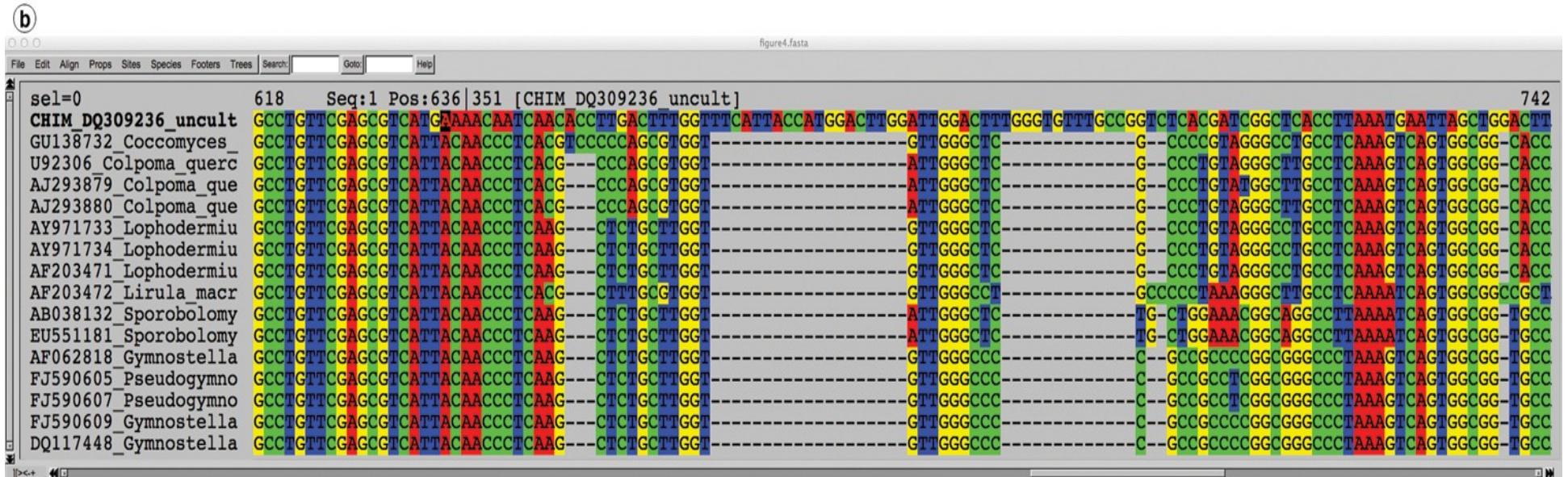
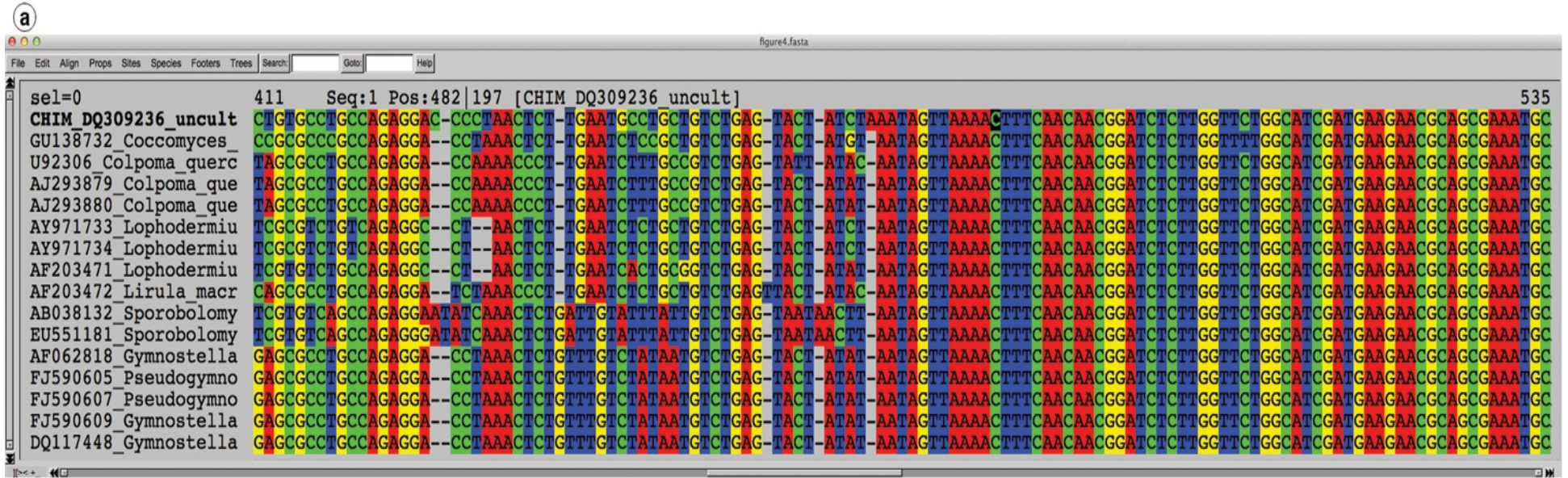
# The frequency evolution of 'A', when it is affected by positive selection

$$X_1(T) = \frac{\varepsilon}{\varepsilon + \exp(-sT)}$$



# Summary Statistics

- Summary statistics provide a summarized description of the dataset
- For example: the number of polymorphic sites
- Summary statistics are important because:
  - They help us to estimate parameters of the population  
For example, the  $4N\mu$ .
  - They help us to estimate whether positive selection has occurred or not in the recent past.



# The site frequency spectrum

The site frequency spectrum is a histogram, of the frequency of mutations

- It is important because in the Wright – Fisher model without selection it has a specific expectation
- $SFS(i) = (4 N u) / i$

# Online simulations for population genetics

- [exelixis-lab.org/pavlos](http://exelixis-lab.org/pavlos)

Thank you for your attention !!!