Introduction to Bioinformatics for Computer Scientists

Lecture 11

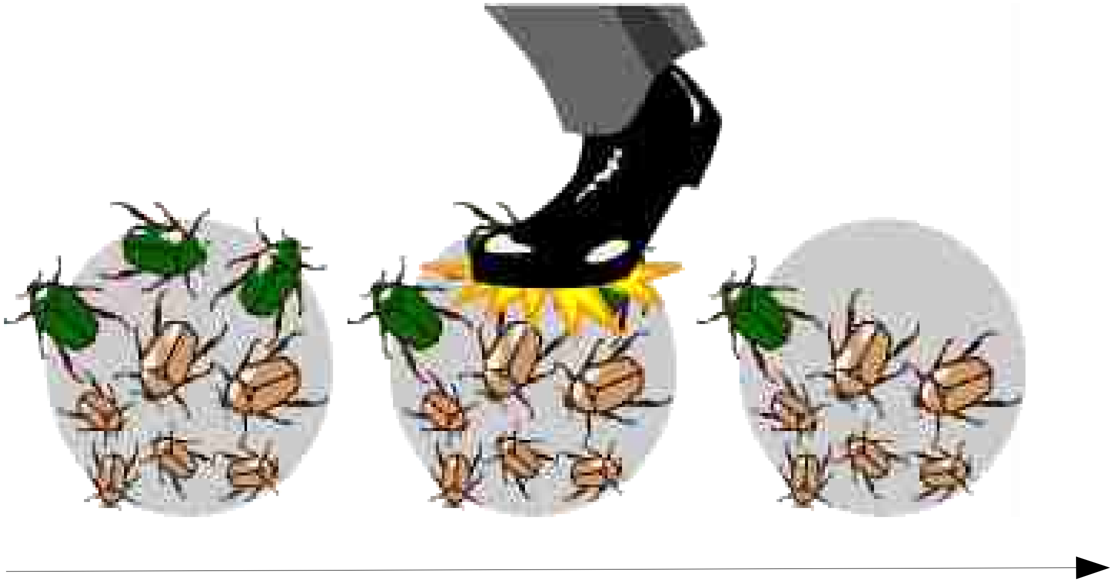
Outline

- Last Time
    - Introduction to Population genetics
    - Hardy’s model (null model – nothing happens – no forces act)
        - for **infinite** population sizes
- Today
    - Simple models for **finite** populations sizes
    - Course Revision and Exam Preparation

Changes in Population Features

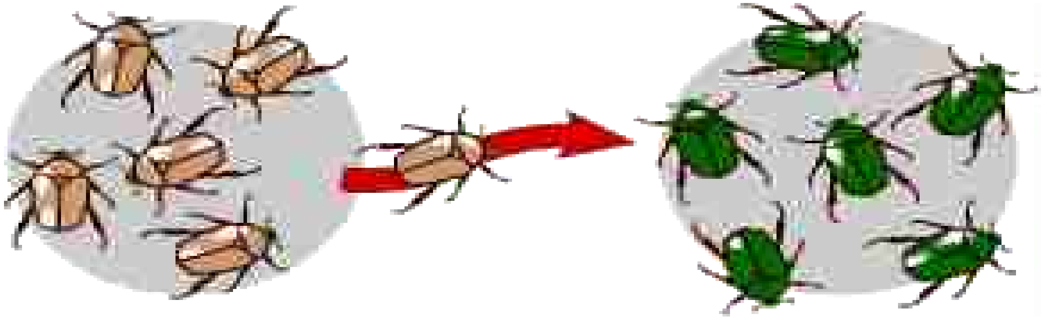
- Feature frequency can change due to
  1. *Genetic Drift*: Chance (other than a random mutation)
  2. *Migration*
  3. *Mutation*
  4. *Natural Selection*: Response to some pressure (e.g., antibiotics, climate change)
- Features can be
  - *Genotype*
  - *Phenotype*

Genetic Drift

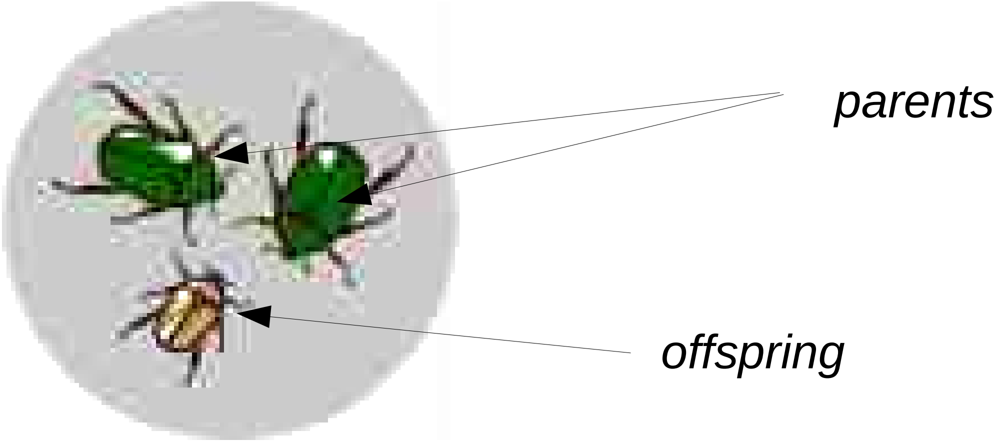


Composition of population changes by some random event

Migration

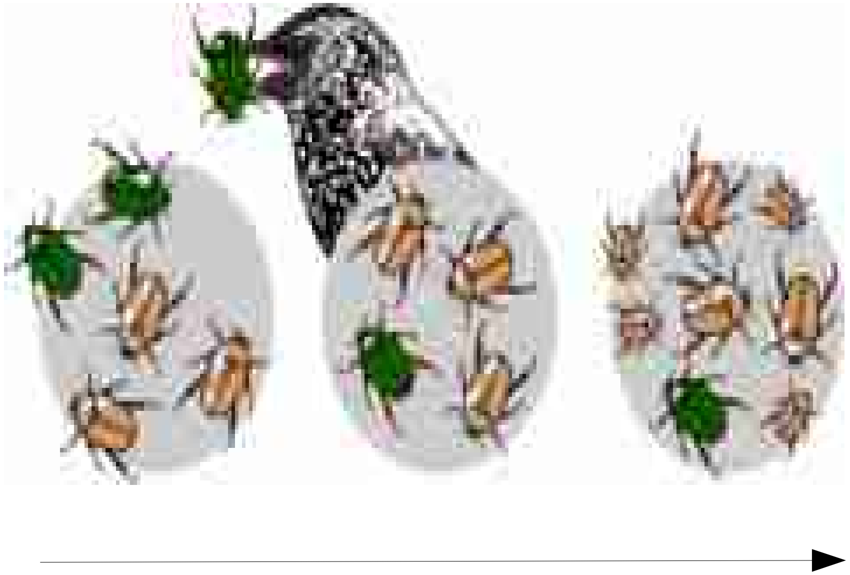


Mutation



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

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*
- Question:
  - 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

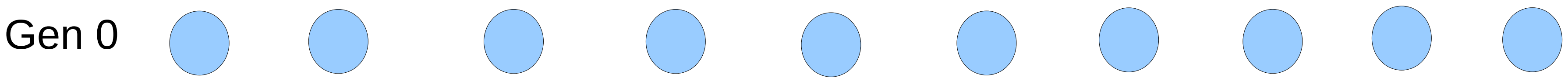
Wright-Fisher Rules/Simulation Example

- 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

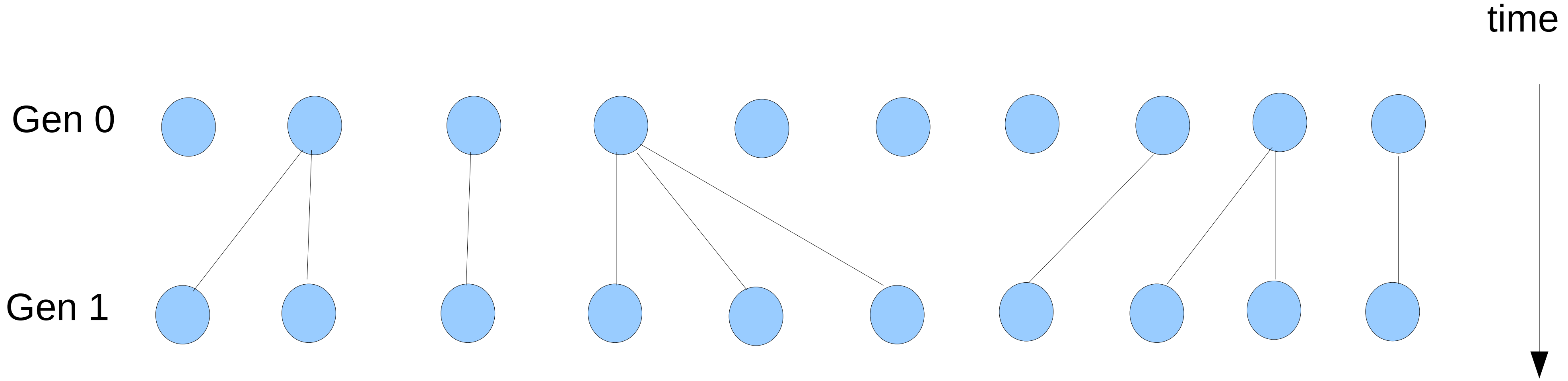
  → all parents have equal probability to be picked

  → 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

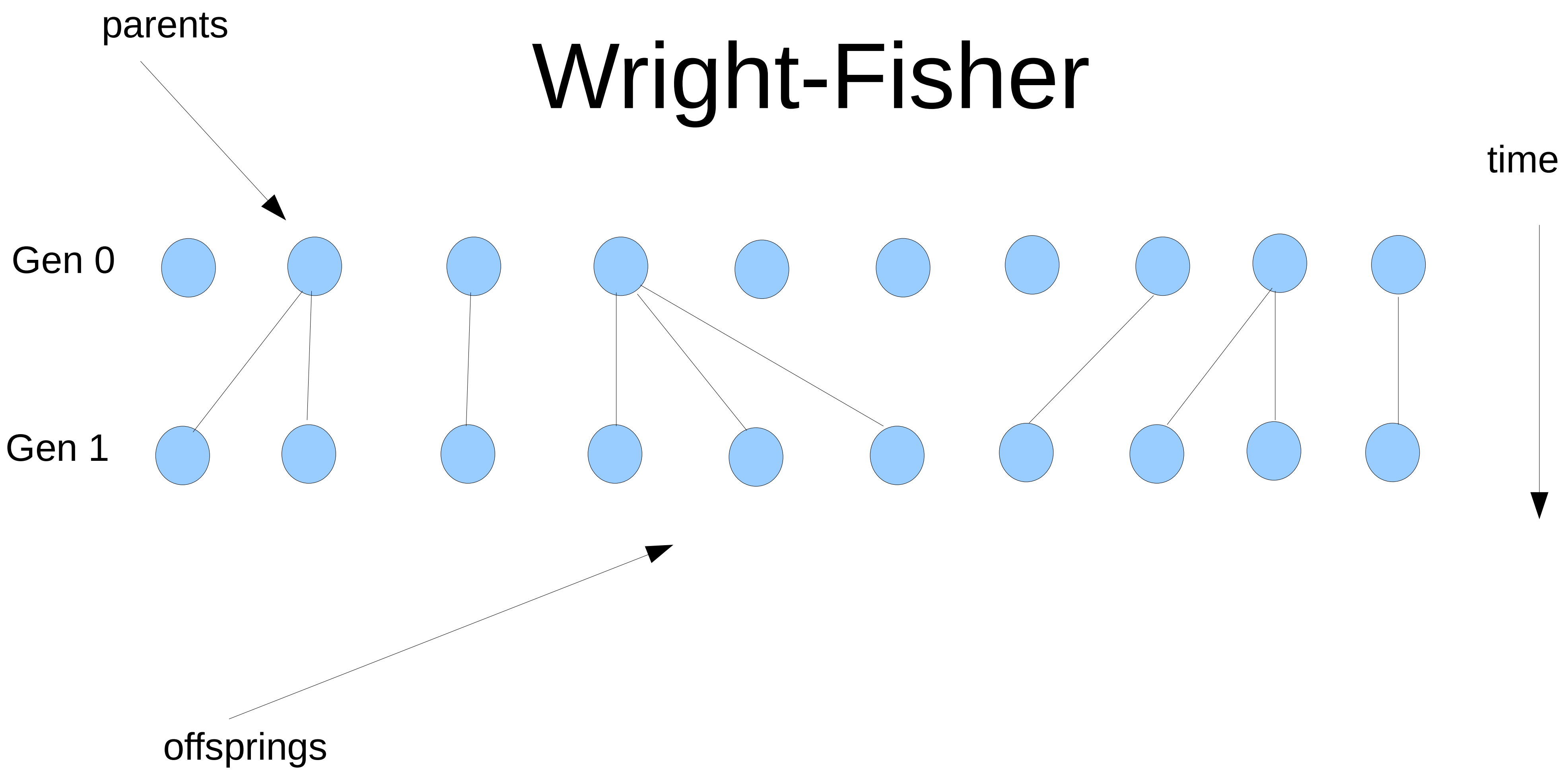
Wright-Fisher



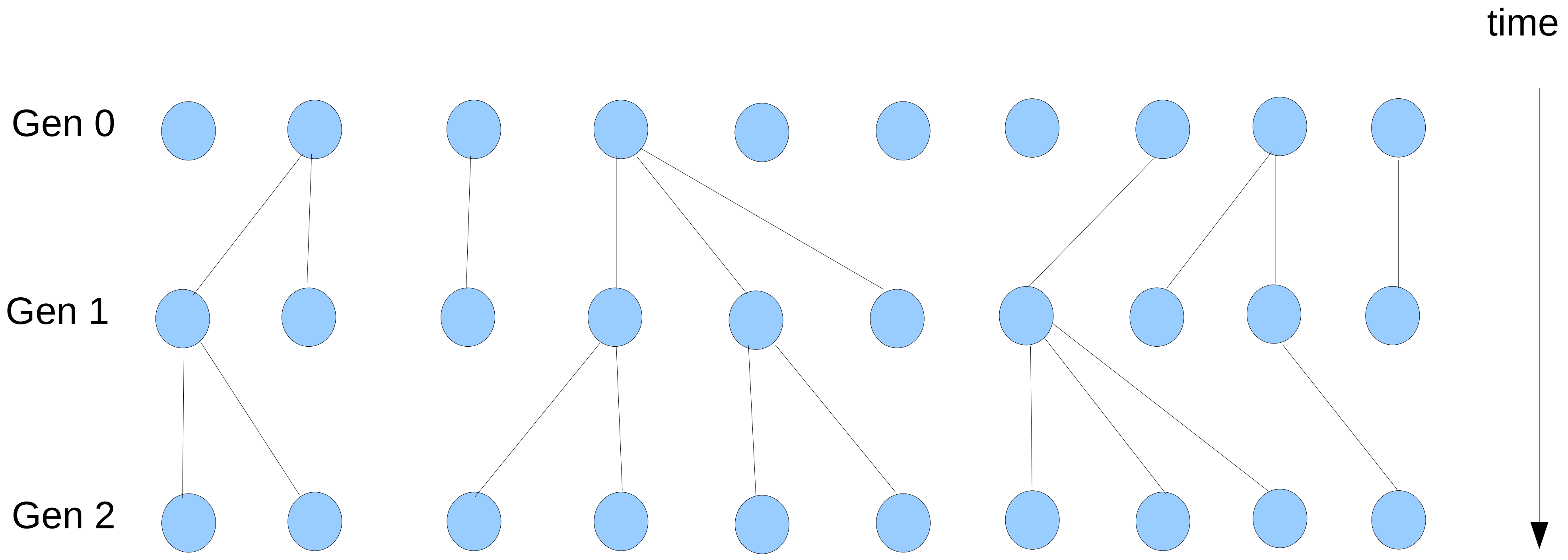
Wright-Fisher



Wright-Fisher



Wright-Fisher



Wright-Fisher
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*

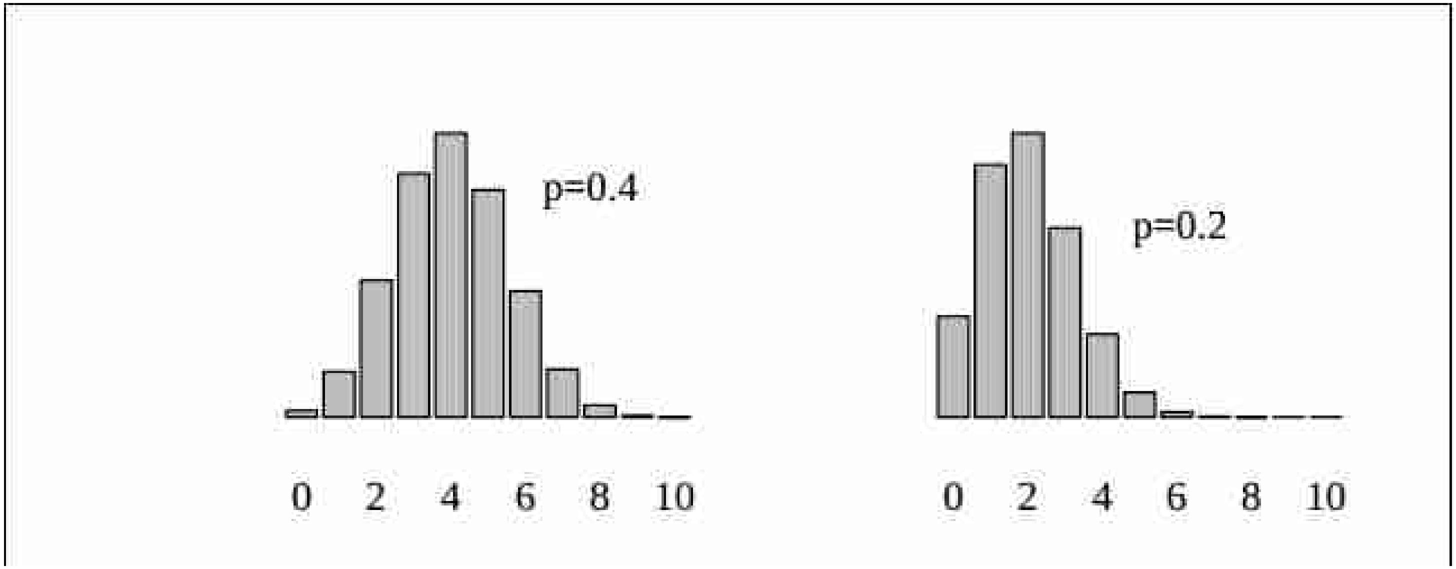
Binomial Random Sampling

- 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 of 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

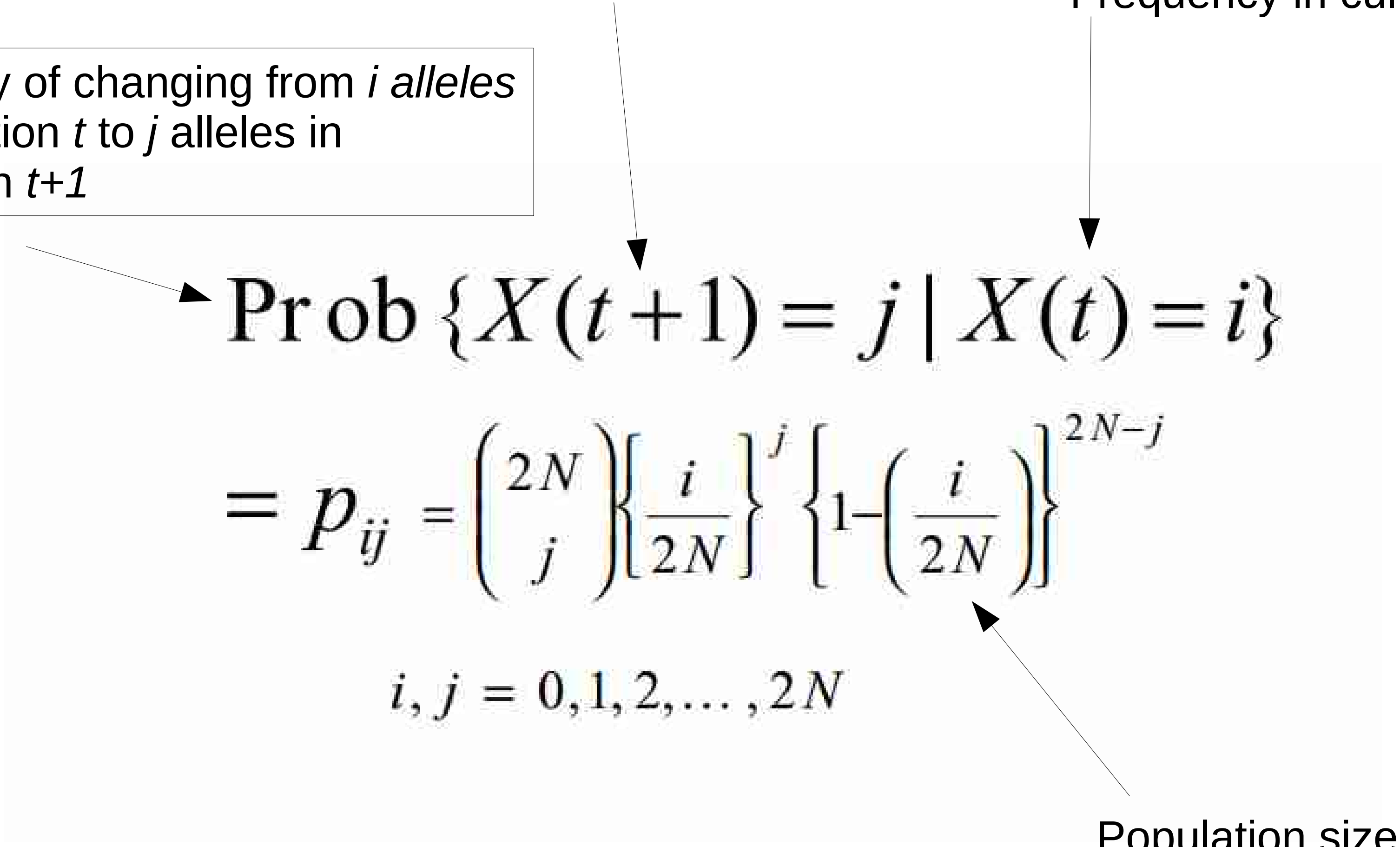
- From the properties of the binomial distribution we obtain
  - $E(\#A) = 2N * p$
  - $Var(\#A) = 2N * p * (1 – p)$
  - **Remember** $p := \#A/1N$

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 Wright-Fisher



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*

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

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

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

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

- **Remember** 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

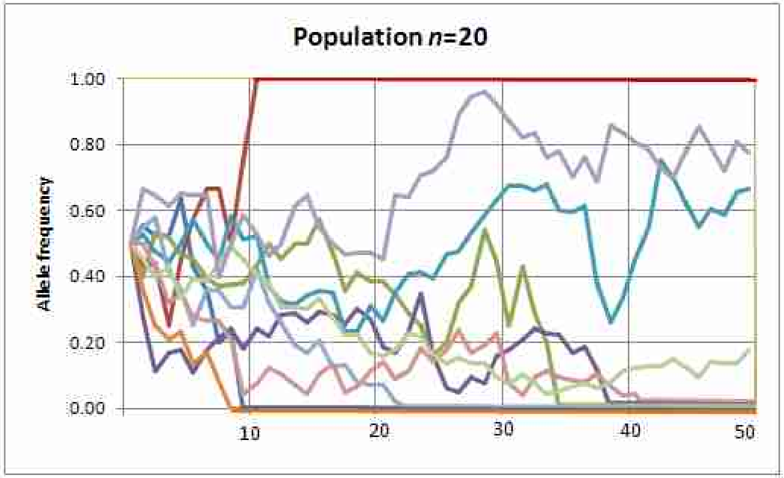
- Reduction of polymorphism is quantified by the degree of homozygosity → The probability that two alleles are identical

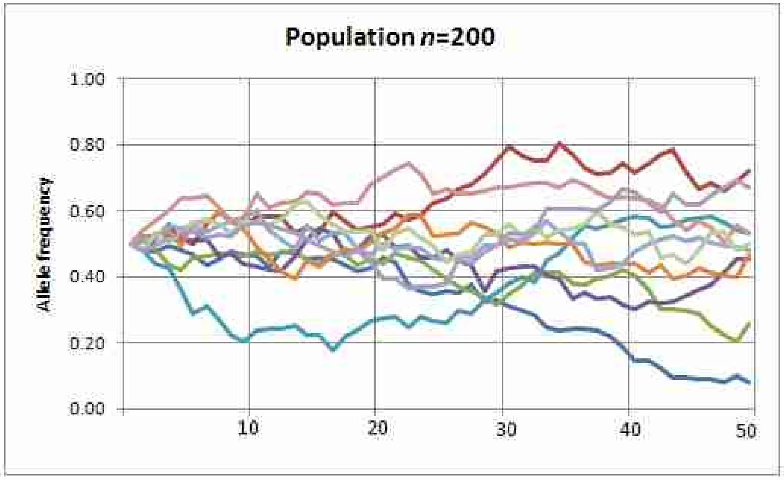
  → *heterozygosity = (1 – homozigosity)* at generation *t* is defined as: Het_t

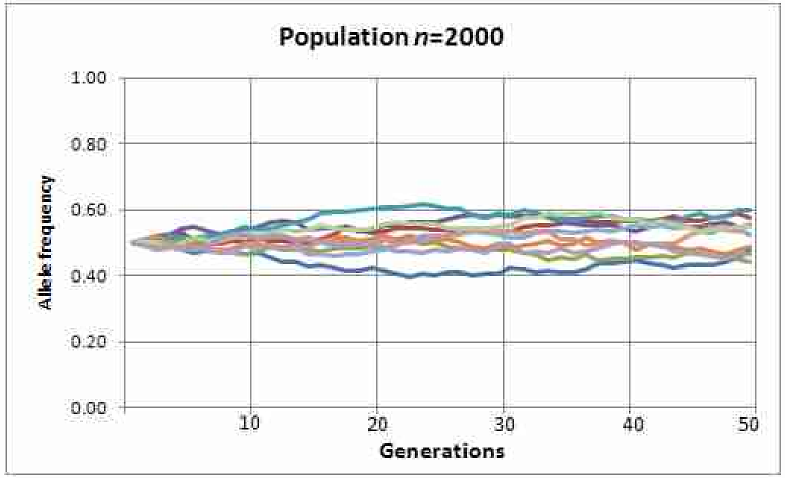
- Assume a population of size $2N$
- We can define the heterozygosity recursively as $Het_t = Het_{(t-1)}\ (1 - 1/2N)$
- Thereby we obtain: $Het_t = Het_0\ (1 - 1/2N)^t$

Probability that two randomly chosen Alleles are different

Initial allele frequency
f(A) = 0.5


Population *n*=20


Population *n*=200


Population *n*=2000

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

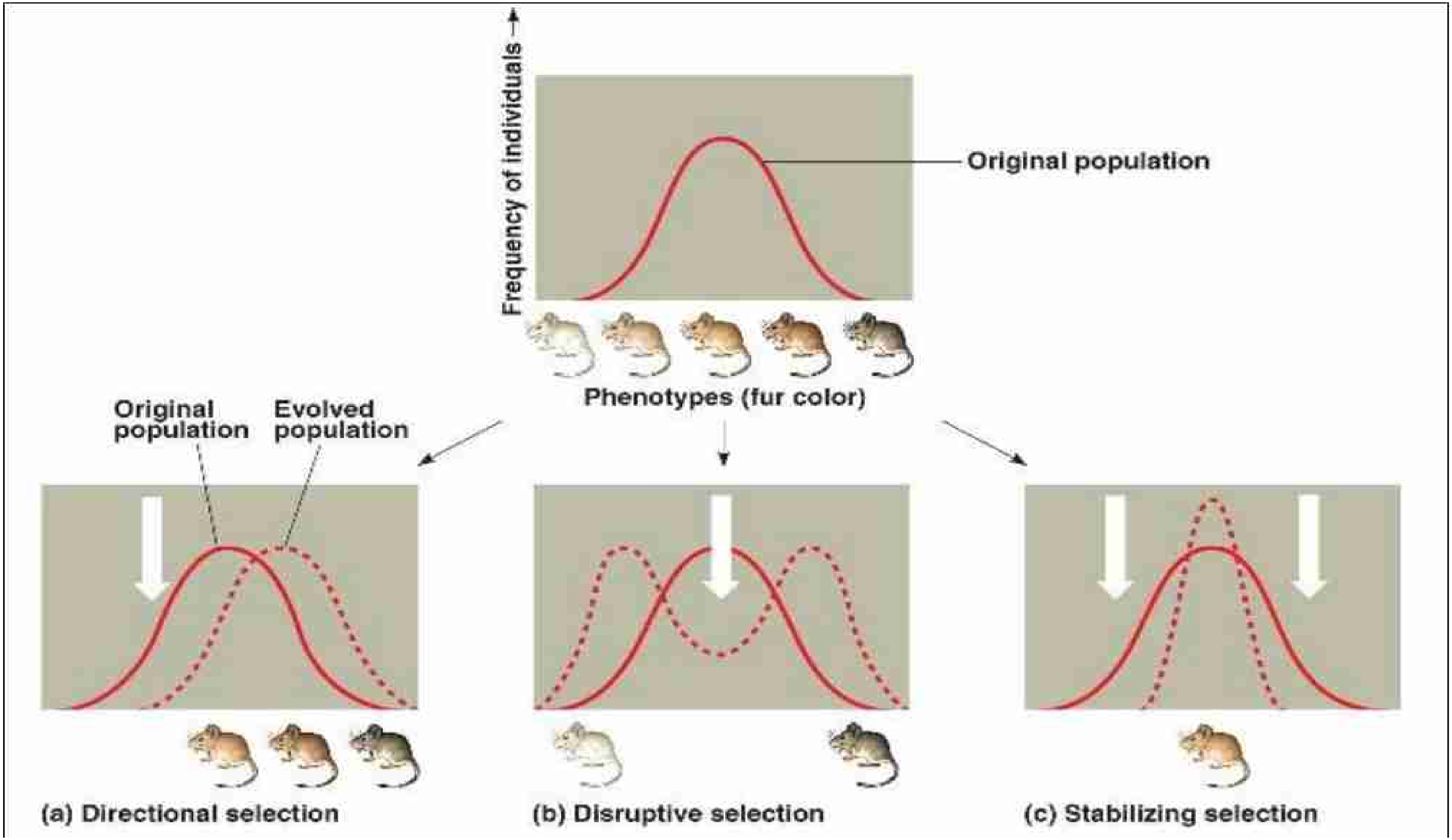
Natural Selection

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

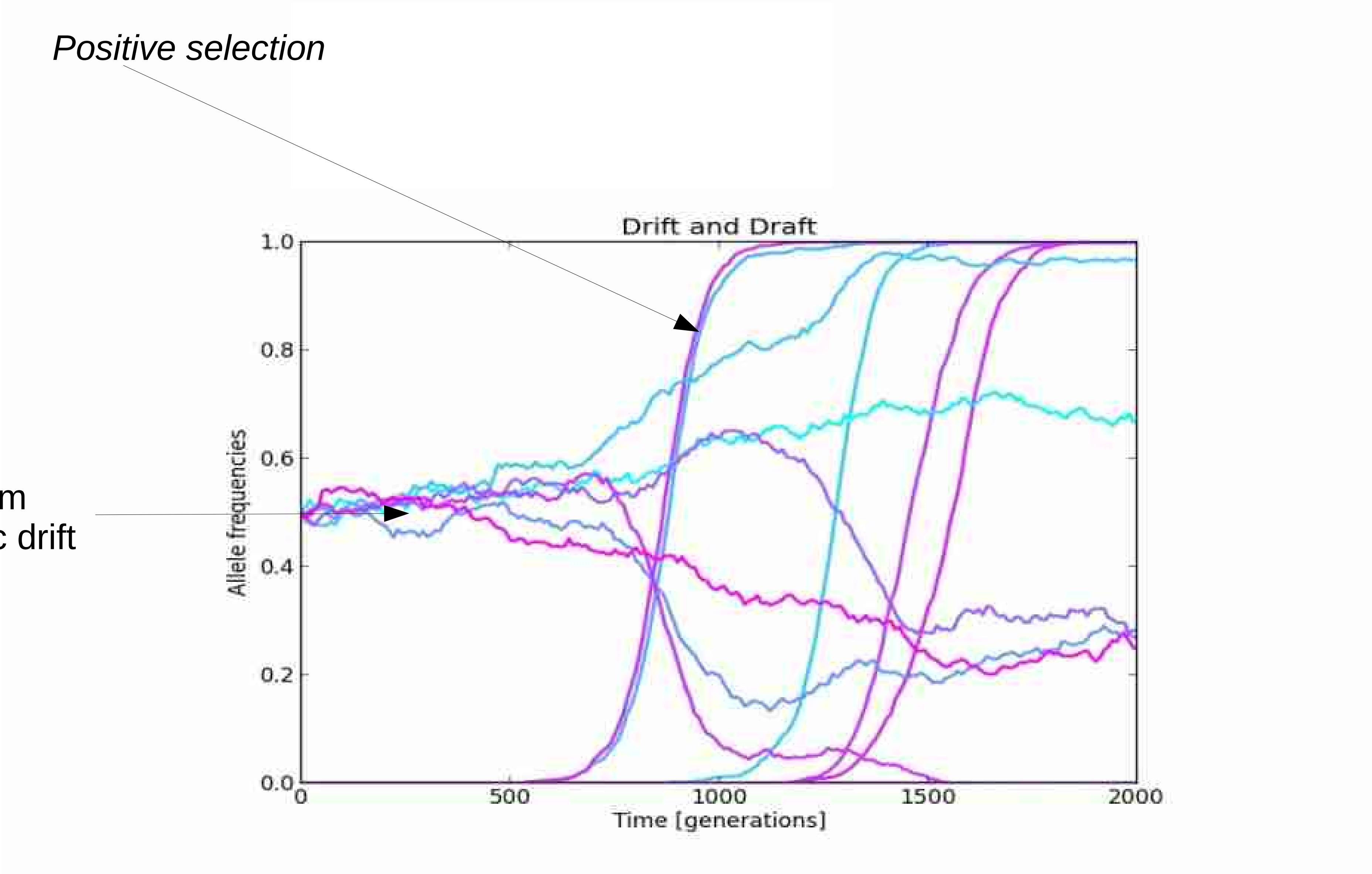
  → it will fix in the population after several generations

  → 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

Outline

- Last Time
    - Introduction to Population genetics
    - Hardy’s model (null model – nothing happens – no forces act)
        - for **infinite** population sizes
- Today
    - Simple models for **finite** populations sizes
    - **Course Revision and Exam Preparation**

Exam

- Don't underestimate the exam!
- In general, if you get equations wrong, that's not a catastrophe
- You should always know and be able to explain how things work in principle though!
- **Register for the exam via the KIT campus system!!!!!!**
- You can chose to do the exam in English or German or Greek

Exam II

- 20 minutes oral exam
- **Be ready to show your ID and student card!**
- **KIT: Will take place in my office at KIT**

  Office 234 in the CS building (2nd floor),

  Am Fasanengarten 5, 76131 Karlsruhe
- **UoC: Will take place in my office Γ106 in the main building of FORTH**

Course Overview

- Biological background knowledge
- Pair-wise alignment
- Sequence Assembly
- Multiple Alignment
- Markov models
- Phylogenetics
- MCMC
- Population genetics

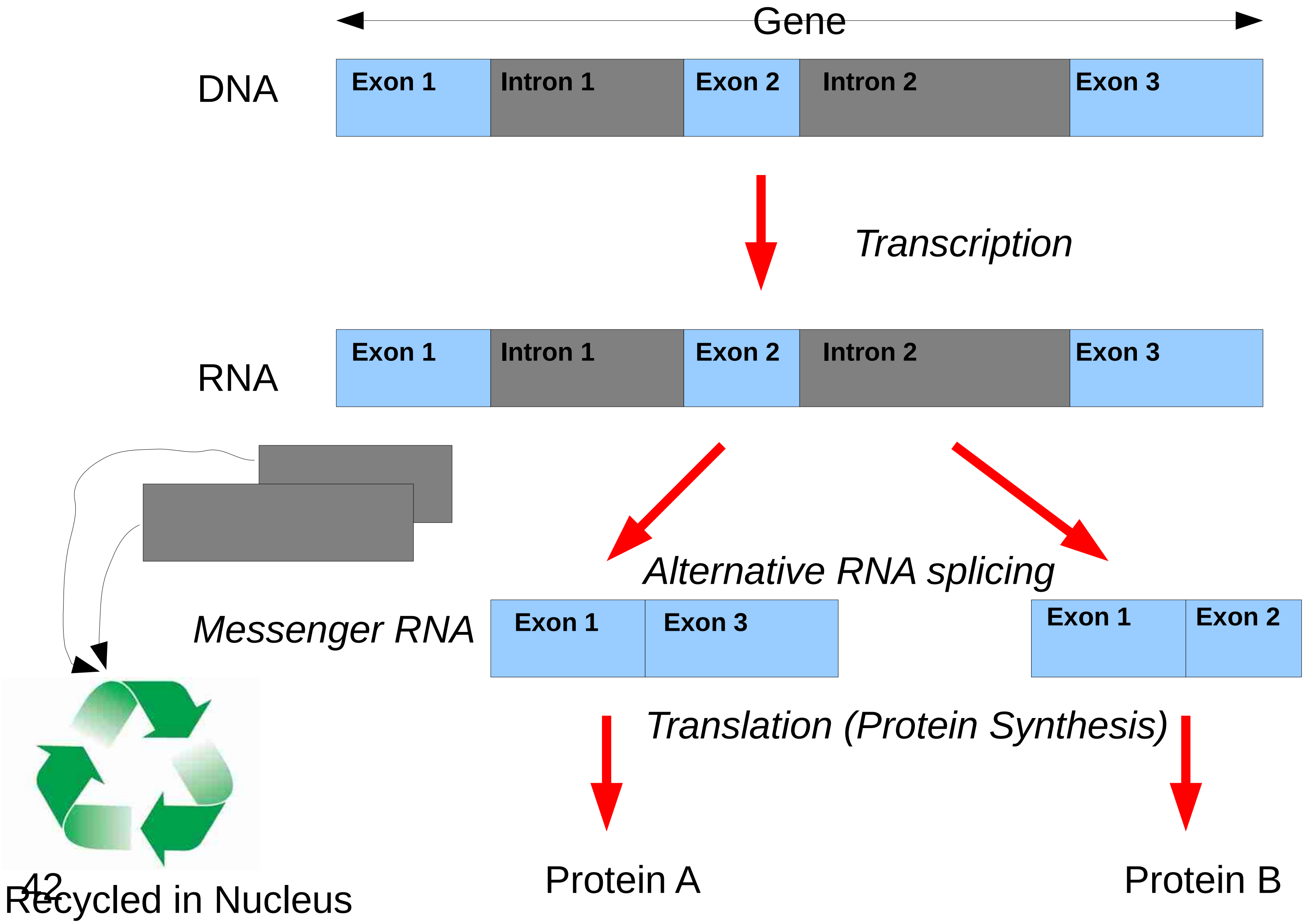
Biological Knowledge

- DNA and AA alphabets
- What are paired-end reads?
- What's a genome?
- Name some model organisms
- Why do we use model organisms?
- Coding versus non-coding sequence data
- What's a transcriptome?
- Is the transcriptome constant or does it change?
- What's a gene?

Biological Knowledge

- What is RNA data?
- What's a Codon?
- 1st & 2nd versus 3rd Codon position
- Synonymous versus non-synonymous substitutions
- Where do genes start and end?
- DNA: what's the *3′* and *5′* end?
- What are the three domains of life?
- What's the difference between Prokaryota and Eukaryota?

Alternative Splicing



Gene

DNA

Exon 1 | Intron 1 | Exon 2 | Intron 2 | Exon 3

Transcription

RNA

Exon 1 | Intron 1 | Exon 2 | Intron 2 | Exon 3

Alternative RNA splicing

Messenger RNA

Exon 1 | Exon 3

Exon 1 | Exon 2

Translation (Protein Synthesis)

Protein A

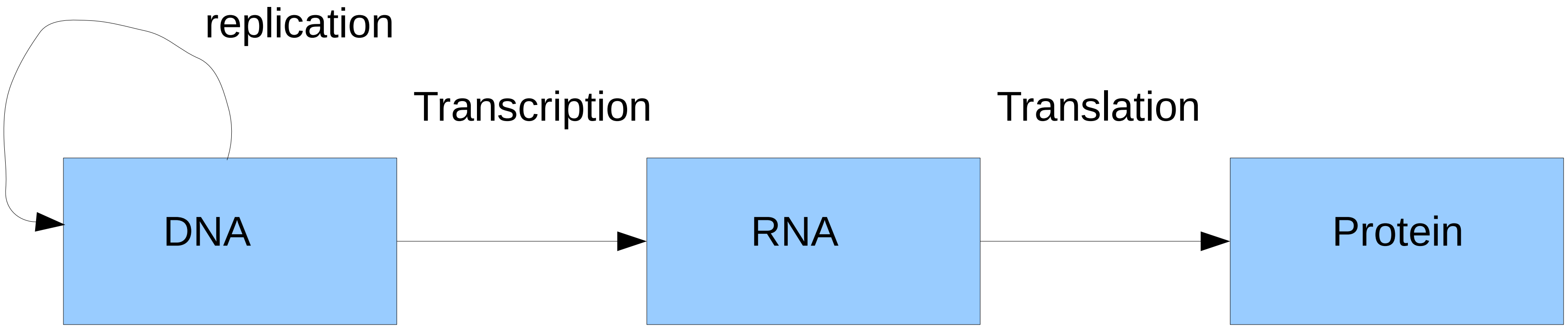
Protein B

Recycled in Nucleus

RNA

- Which types of RNA do you know, what do they do?
- Why is RNA interesting for building phylogenies?
- Name some interesting RNA genes!
- Why is the RNA secondary structure interesting for RNA evolution?

Central Dogma of Molecular Biology



replication

Transcription

Translation

DNA

RNA

Protein

Biological Knowledge

- What is a meta-genome?
- What's a chromosome?
- What's a taxonomy?
- What's a phylogeny?
- What is an outgroup?

Pairwise Sequence Alignment

- Name some distances for comparing strings
- What is the difference between local and global pair-wise sequence alignment?
- How is the edit distance defined?
- What's the definition of the Hamming distance?
- How do we define an optimal pair-wise sequence alignment?
- Outline how a pair-wise sequence alignment algorithm that uses dynamic programming works!

Pair-wise Sequence Alignment

- What's the difference between the Needleman-Wunsch and Smith-Waterman algorithms?
- What is their time and space complexity?
- What is a substitution matrix?
- How does the backtracking work?
- Can there be multiple, distinct, equally optimal pair-wise sequence alignments?

Blast & Genome assembly

- By reference versus de novo assembly (mapping)
- What is BLAST?
- What is BLAST good for?
- Why not use Smith-Waterman instead?
- How does BLAST work → *seed, extend, evaluate*!
- What is Genome assembly?

De novo Genome Assembly

- How does *de novo* assembly work?
- What is an *overlap graph*?
- How do we traverse this graph to assemble a genome?
- What is a *de Bruijn* graph?
- What is a *k-mer*?
- How do we traverse a *de Bruijn* graph?

By reference assembly - Mapping

- Which problem are we trying to solve?
- Why not use Blast?
- Why not use pair-wise sequence alignment?
- What techniques can we apply to accelerate mapping?
- How does mapping with hashing work?
- How do we select a *k-mer* representing a read?
- How do we extend the read?
- What's the drawback of the hashing procedure?
- How can this be improved?

Multiple Sequence Alignment

- What is homology?
- How can we assess the quality of an MSA?
- How do we compute the *SP* score?
- What are MSAs good for?
- What's a gene duplication?
- Does sequence similarity induce homology?

MSA

- Can we build an MSA with an optimal SP score?
- What's the time complexity?
- How does the star alignment approximation work?
- How is the tree alignment problem defined?

  → can you compute a tree alignment score on a given tree?

  → students always get this wrong … this is not a guide tree method, but an explicit criterion!!!!!
- How do practical approaches for MSA work?
- Describe how progressive MSA methods work in principle
- How does **pair-wise profile** alignment work?
- What are the shortcomings of progressive alignment methods?
- Solutions to overcome these?
- How do we benchmark MSA programs?

Phylogenetics

- Why is an appropriate *outgroup* choice important?
- What is an *ultrametric* tree?
- How do we put real times on a phylogeny?
- What input data can be used to build phylogenies?
- What can we do with phylogenies?
- How many unrooted binary trees exist for *n* taxa?
- How can we come up with this formula? → draw an image/graph

Phylogeny Reconstruction Methods

- Name the two basic classes of reconstruction methods!
  - Distance- versus character-based methods students often seem to be confused by this question
- Name some methods that are NP-hard
- How do Neighbor Joining/UPGMA work **in principle**?

  → run time & space complexities!!!!
- How does the least-squares algorithm work?

  → suggest a tree search heuristic for the least-squares criterion
- How is the minimum evolution criterion defined?
- How does parsimony work?
- What's the time and space complexity for computing the parsimony score on a tree?
- What's the underlying principle?
- Given a small tree and alignment, calculate the parsimony score!

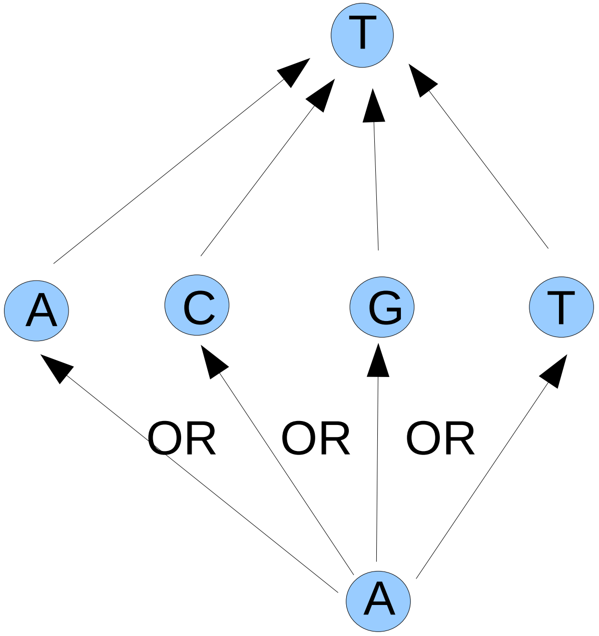
Search Strategies

- How can we build starting trees?
  - Random
  - Stepwise addition
  - NJ, UPGMA, parsimony
- How can we change a given comprehensive topology to find a better tree?
  - NNI moves, SPR moves, TBR moves, etc.

Markov chains

- What is a discrete Markov chain?
- What is its key property?
- Draw an example of a Markov chain (e.g., flea hopping)
- Draw the transition matrix for your example
- Why do the rows of the transition matrix need to sum to 1?

What is the transition probability for getting from A(0) to T(2)



2 $X_2 = j$

1 $X_1 = k$ Sum over k

0 $X_0 = i$

Markov chains

- What does the π vector denote?
- What is the equilibrium distribution?
- How do we compute a transition matrix for a continuous time Markov Chain $P(t)$?

Maximum Likelihood

- What's the long branch attraction problem?
- Why shall we model distances as stochastic processes?
- What does a substitution matrix look like?
- What is time-reversibility?
- How does ML work in principle? Remember: **AND** and **OR**
- How does the Felsenstein pruning algorithm work?
- Which parameters do we have (to optimize)?
- What's the time & space complexity for evaluating one tree?
- How can we optimize branch lengths?

ML continued

- How can we model rate heterogeneity among sites?
- How does the Γ model work?
- How are protein substitution models obtained?
- How can we obtain the base frequencies?

MCMC Methods

- How do they differ from ML methods?
- What are we trying to approximate?
- What are the computational difficulties?
- How do they work in principle?

  → robot metaphor (you can use this to explain everything)
- How do we compute if we want to accept or reject a proposal?
  - Why does this ratio solve a lot of problems?
- Where do we get the priors from?
- How can we summarize samples?
- How does MCMC work in practice for phylogenetics?

MCMC Methods

- What's the difference between the proposal and the target distribution?
- What does the term “good mixing” mean?
- What is the Hastings ratio and why do we need it? → drunk robot
- What is Metropolis-Coupled Markov Chain Monte Carlo?

  → multiple robots on our planet
- What is thinning?
- For DNA under GTR+Gamma what types of proposals do we need?

  → which proposal type would you apply most frequently?
- Can we use MCMC to integrate over different models?

Population genetics

- How can a population evolve?

  → four main evolutionary forces
- Difference: Genotype versus Phenotype?
- Dominant versus Recessive?
- How are alleles inherited?
- How is polymorphism defined?
- What are SNPs?
- Can you describe Hardy's model?

  →assumptions?

  → what's amazing about this model?
- What is the Wright-Fisher Model?

  → how can we simulate a population under it?

  → which assumptions does it make
- What is random genetic drift?

KIT Course in Summer

Seminar *Hot Topics in Bioinformatics*

- 2 hours per week seminar (live at KIT – two meetings)
- We select interesting Bioinformatics papers and present them

  → select any subject/paper mentioned in the course you find interesting

  → ask us if you are interested in a different topic

  → one of my lab members will help you to understand the paper, prepare the presentation and the report

  → report & presentation language: either English or German, English much preferred though!
- 35 Minute presentation of paper
- Submit a report of **8** pages at the end of the semester
- 3 ECTS

UoC Course in Spring

Seminar *Reproducibility in Bioinformatics*

- 2 hours per week seminar (live at UoC – two meetings)
- We select interesting Bioinformatics papers **that have a chance of being reproducible** and present them

  → select any subject/paper mentioned in the course you find interesting

  → ask us if you are interested in a different topic

  → one of my lab members will help you to understand the paper, prepare the presentation and the report

  → report & presentation language: either English or Greek, English much preferred though!
- 35 Minute presentation of paper
- Submit a report of **8** pages describing your attempts to reproduce the results
- 3 ECTS

Prerequisites for Seminar

- Attended & passed Introduction to Bioinformatics
- To register,
  - UoC & KIT: write me an email
  - KIT: registration via the campus system will only be possible AFTER you have passed the exam
- Maximum of 10 places available
- Who is interested in the Seminars?