Population genetics with PopGen

Jakob Nybo Nissen, 2021-10-21

Today and tomorrow

- Presentation of concepts (15-30 mins?), then exercises.
    - Concepts explained superficially, only enough to use and learn the software!
- Today: Population genetics with PopGen.jl
- Tomorrow morning: Biological sequencing and BioJulia parsers
    - BioJulia as an organisation
    - BioSequences.jl
    - FASTX.jl
- Tomorrow afternoon: Sequence alignment
    - BioAlignments.jl
    - XAM.jl

Population genetics

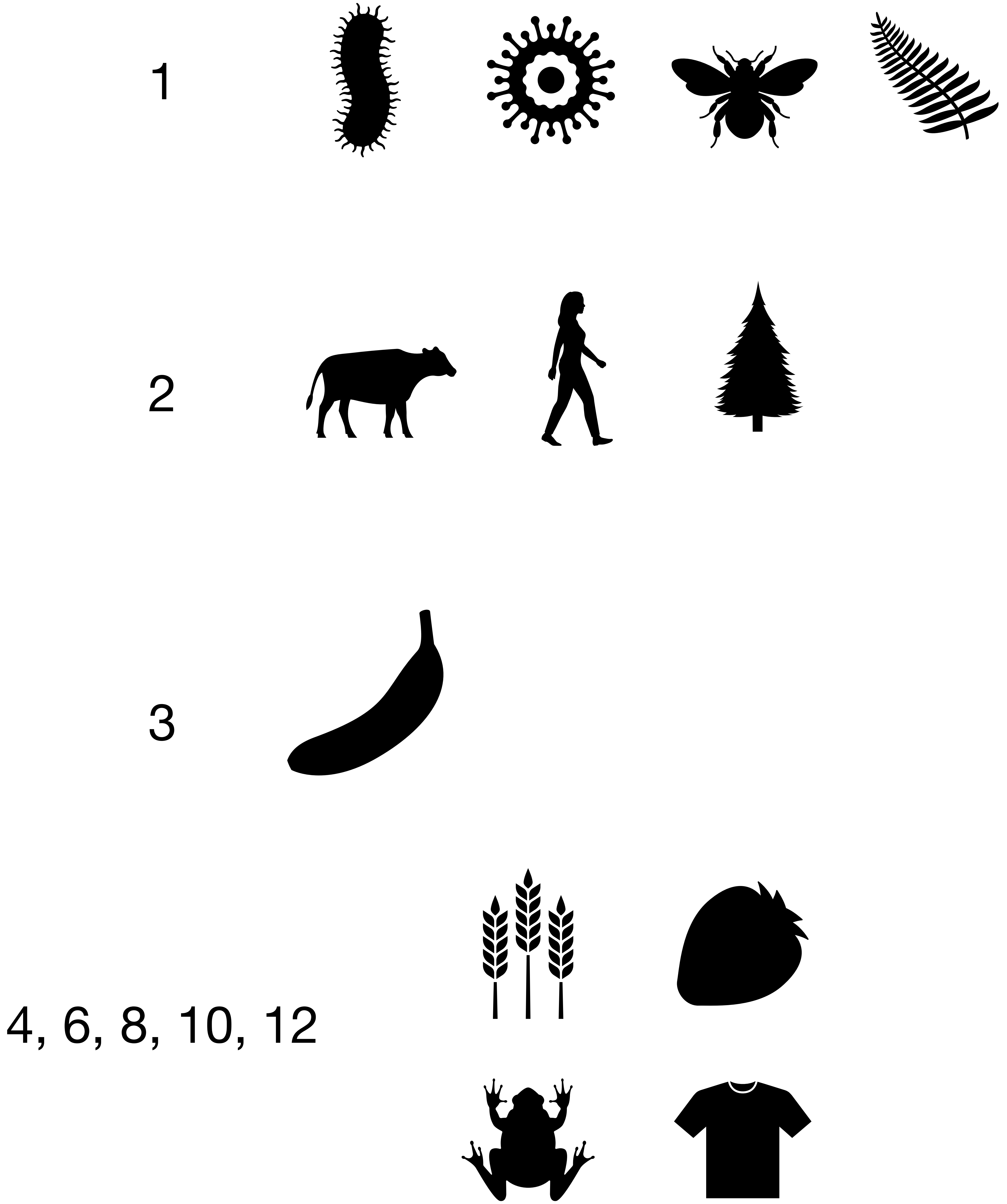
- Population genetics: Genetic differences viewed as *statistics:* Population structure, adaptation, speciation.

- 1880's: Evolution widely accepted, but fundamental mechanism uncertain.
- 1900's: Genetics and inheritance patterns re-discovered
- 1920's: Quantitative genetics + population genetics
- 1930's: Modern synthesis: Evolution = Darwinism + population genetics + inheritance

I am not an expert in population genetics!

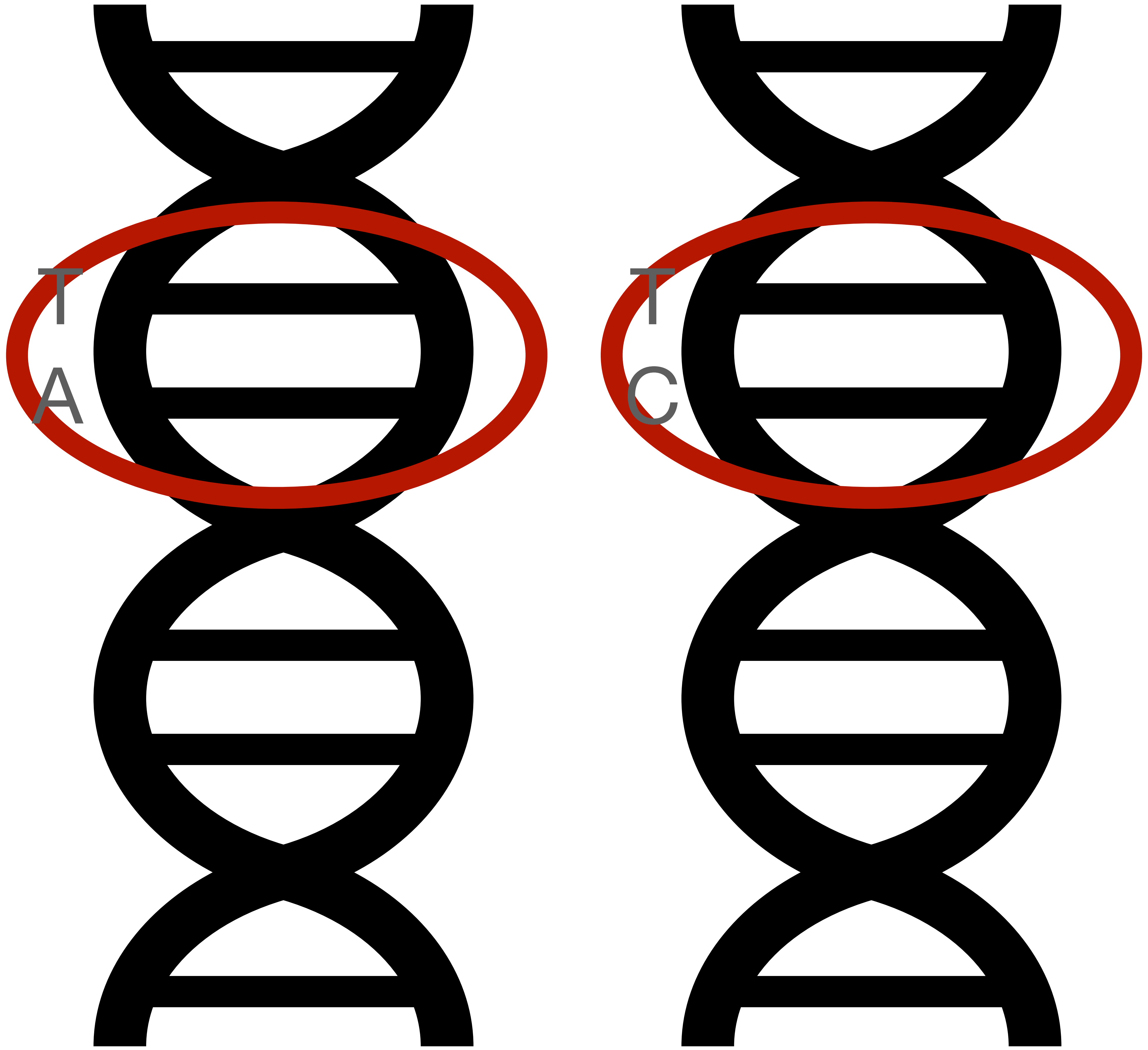
Genetics 101

- All inherited information is stored in the *genome*
- Every organism has N copies of the genome. N is the organism's *ploidy*
- Most organisms are haploid (1) or diploid (2).
- Odd-numbered N>1 ploidy usually cause sterility. Non-sexual organisms only
- Some organisms, especially cultivated plants, are polyploid



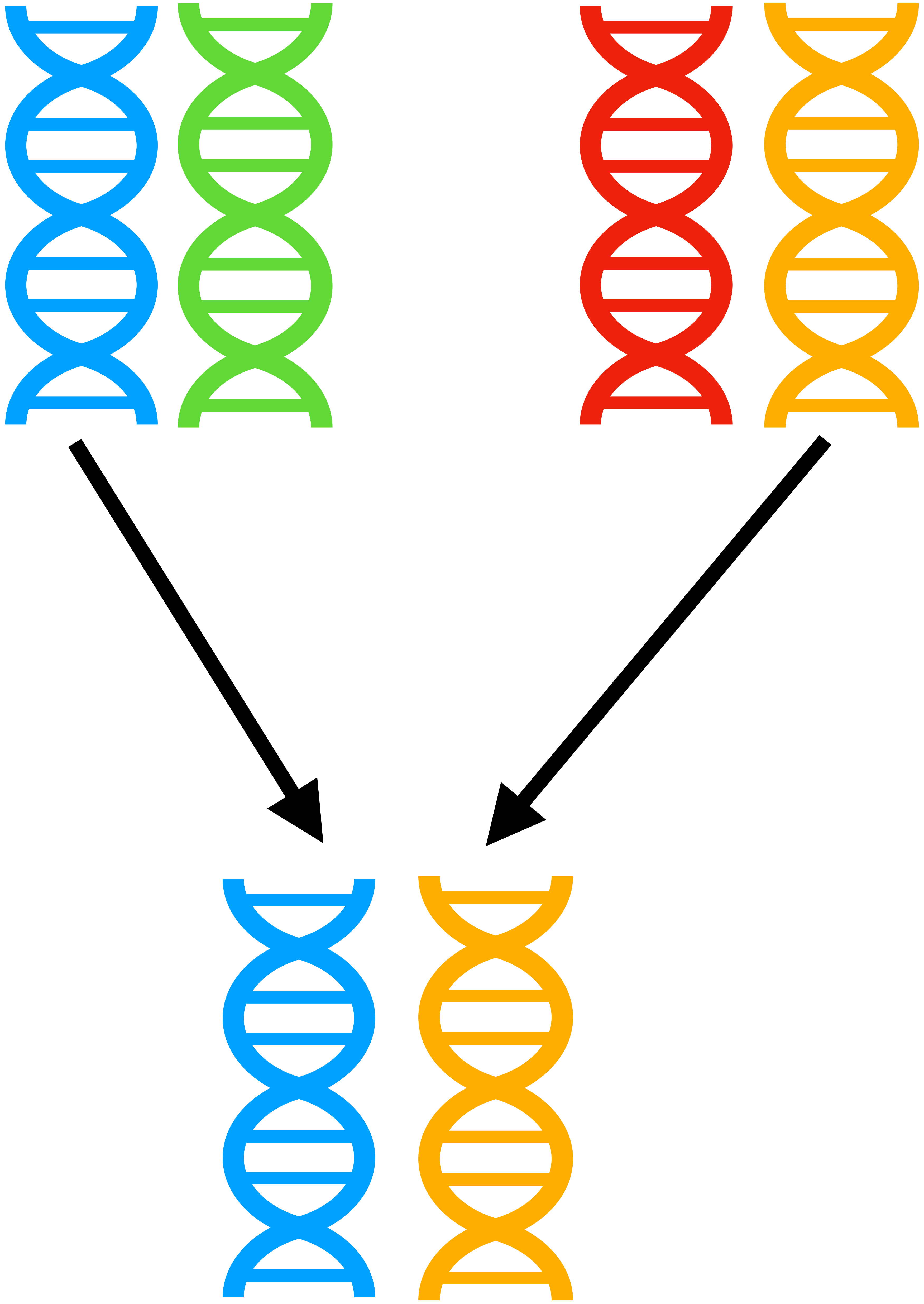
Genetics 101

- A location on the genome is a *locus*. This may or may not be the location of a gene.
- A specific sequence at a locus is called an *allele*. Genetic variation means different alleles of the same locus exists.
- Humans are diploid, so we have 2 of each locus.
- At any locus, we can have two of the same alleles (homozygous), or two different alleles (heterozygous)



Genetics 101

- With sexual reproduction, the offspring inherits one locus at random from each parent
- Inheritance is heavily influenced by the physical organisation of loci in the genome
    - Meaning is not really random
    - But we don't go into that here
- Half genome inherited from each parent
    - Relatedness: Siblings, parents have 50% of out genome = 0.5
    - Half-siblings, uncles/nephews = 0.25
    - First cousins = 0.125 etc.



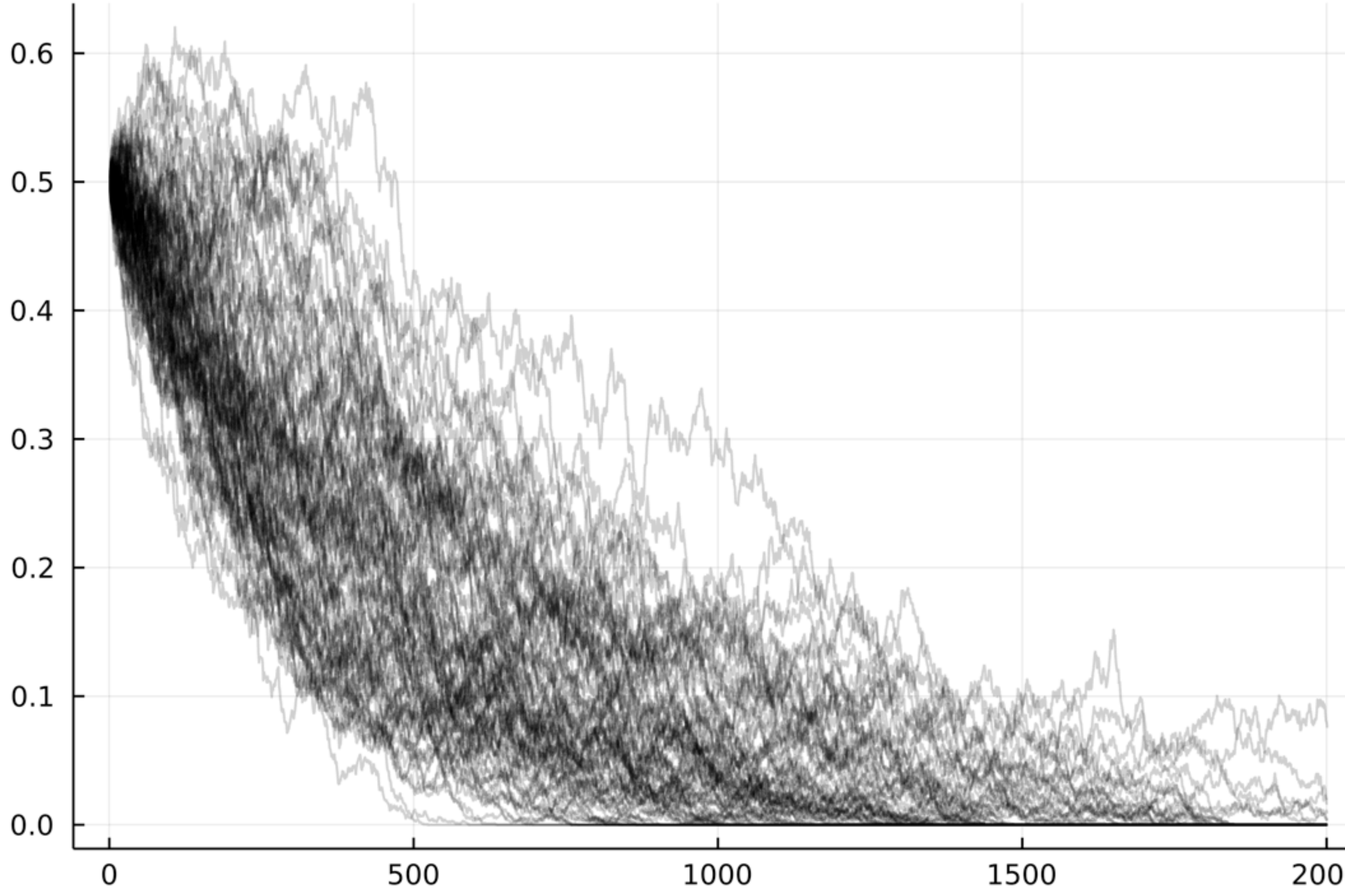
Population genetics

- Individuals are ephemeral, alleles last
    - On evolutionary scale, each generation shuffles genes together fast, and individuals die in the blink of an eye.
- One can think of each individual genome being sampled randomly from a *gene pool* that belongs to the population
- Sex maintains the integrity of the gene pool
- If two populations are not linked through sex, their pools drift apart due to mutation. This causes speciation
- Modeling shows only 2 migrations / generation fuses two gene pools
- Practically speaking, "populations" and "species" blend together, not black/ white.

Population genetics

- Allele frequency: How common is a certain allele among all alleles (ploidy x population)
  - 100 heterozygotes in a pop of 250 diploids mean (100 / 2*250 = 0.2)
- It takes almost no selection pressure to cause quick evolution
- Gene pools, populations, selection... all depends on *random mating*
- Can we quantify random mating?

Simulation: Allele causes 0.1% lower mortality



Hardy-Weinberg equilibrium

- If mating is random and two alleles has frequency *p* and *q*, homo/ heterozygosity should have frequencies

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

- Allele pairs with this frequency are in *Hardy-Weinberg equilibrium*.
- We can test deviations of this using a chi-square test
- Higher homozygosity = inbreeding relative to random
- Lower homozygosity = outbreeding relative to random

Wright's fixation index Fst

- How distinct are two populations?
- Fst was originally defined as

$$\frac{var_{between} - var_{within}}{var_{between}}$$

- Without any definition of "variation".
- We can define it using allele frequencies!
    - Example: "Heterozygocity": $1 - \sum p_i^2$
- ... but we don't have access to allele frequencies, only estimates from finite samples from populations.
- Various statistical estimators have been developed including Weir & Cockerham

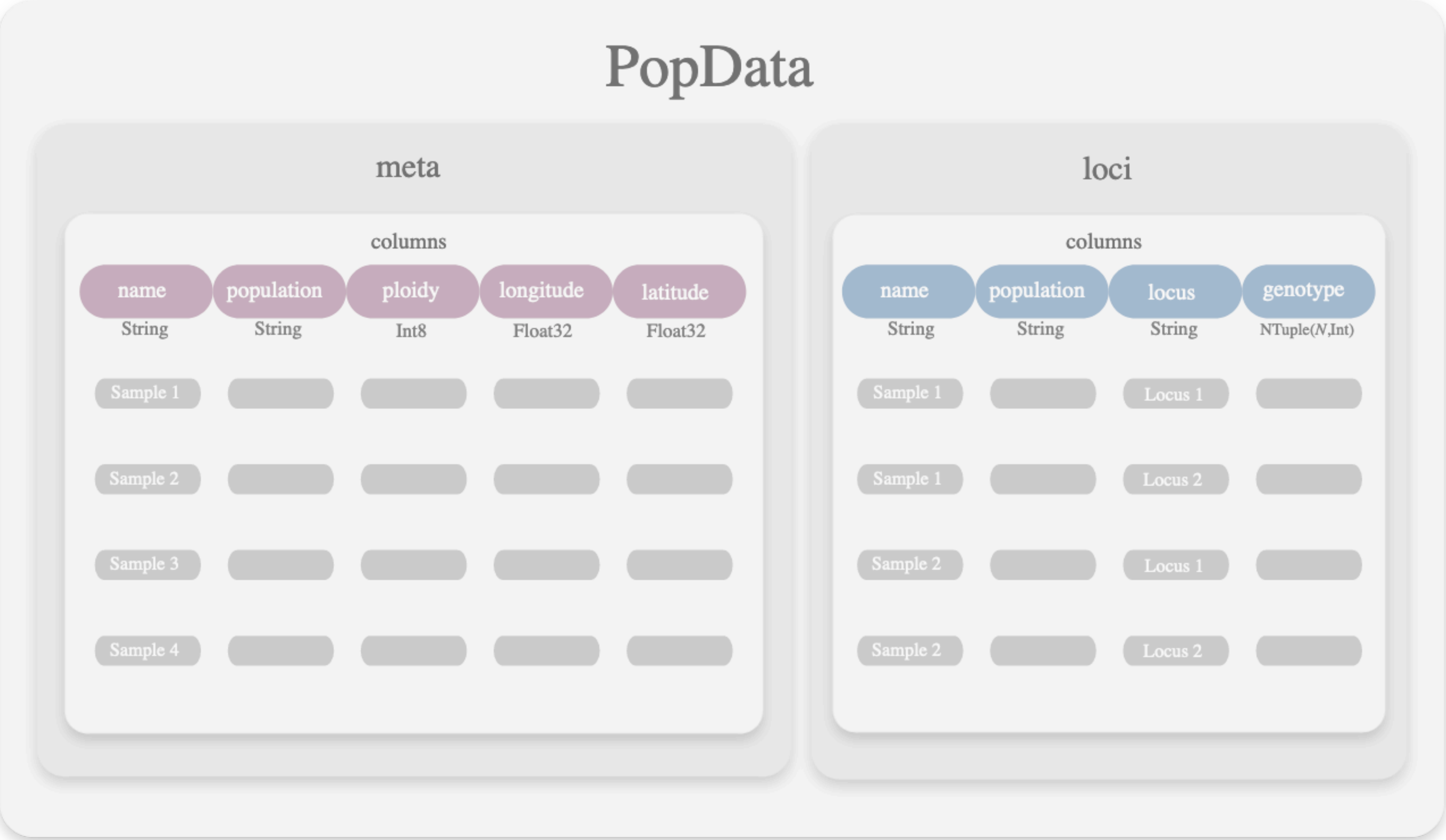
PopGen.jl

- Relatively new package, 2020-06-21
- Takes a "data science" approach, less so a software engineering approach.
    - Suitable, because population genetics deals with masses of data by definition!
- A single core data type, based on the DataFrame.
    - Not generic, abstract interfaces!
- Efforts seem to be: Standardised *functions* for common pop gen operations on the single data type.

Pavel Dimens

PhD student, population genomics of fish species

PopGen.jl



PopData

meta

columns

name	population	ploidy	longitude	latitude
String	String	Int8	Float32	Float32
Sample 1				
Sample 2				
Sample 3				
Sample 4				

loci

columns

name	population	locus	genotype
String	String	String	NTuple(*N*,Int)
Sample 1		Locus 1	
Sample 1		Locus 2	
Sample 2		Locus 1	
Sample 2		Locus 2	

Today's exercise

- Load population genetics data into DataFrame like object
- Calculate allele frequencies and do a population comparison
- Measure relatedness between individuals in the dataset
- Test for HWE and measure Fst in the population

- First: `git clone` https://github.com/jakobnissen/julia_bio_phdcourse
- Download the data: Link in the Slack and on mail
- Use environment: Go to exercise/popgen directory
    - `pkg> activate .`
    - `pkg> instantiate`

Questions?

Exercise 1