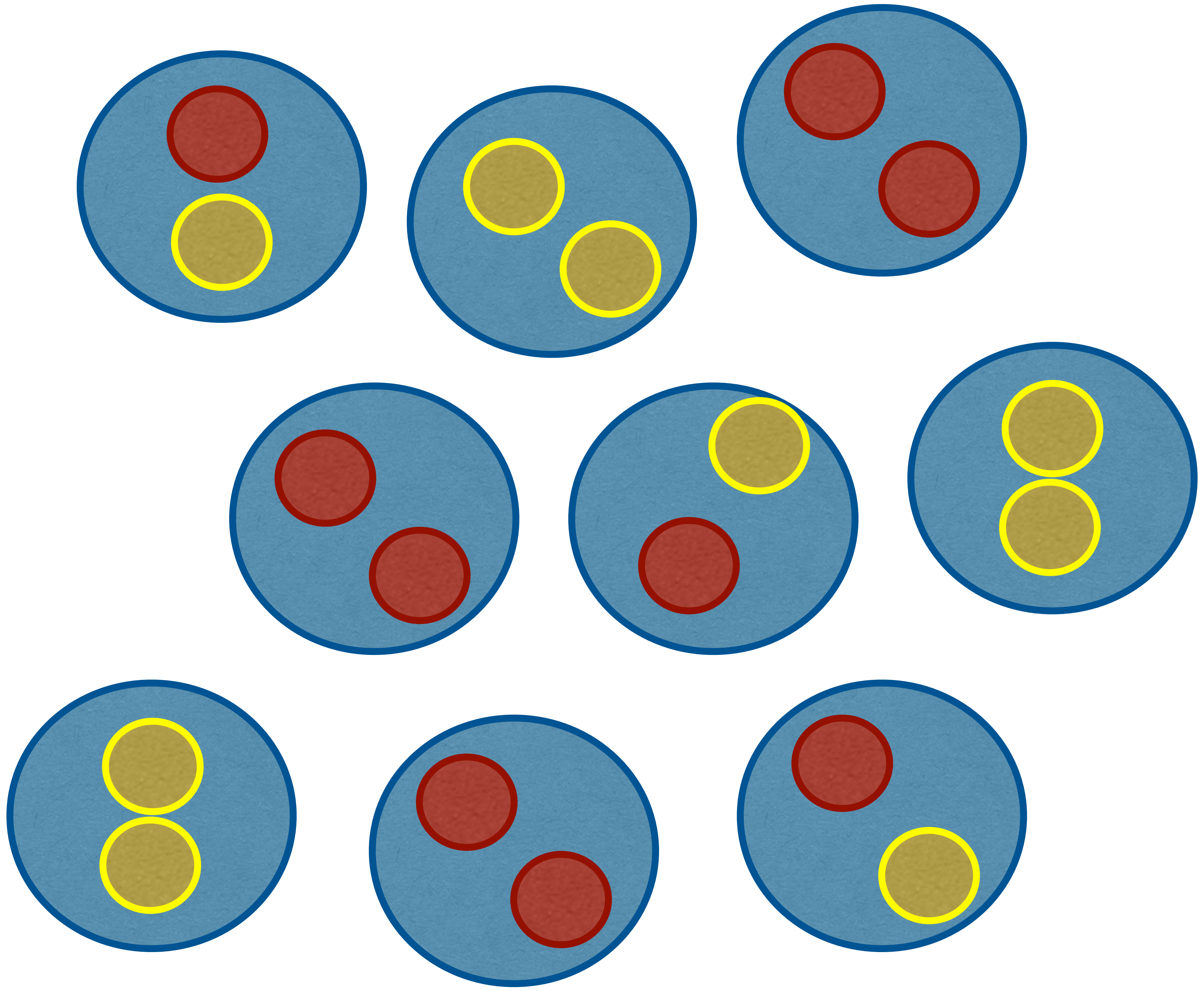
Week-2
Allele and Genotype Frequencies



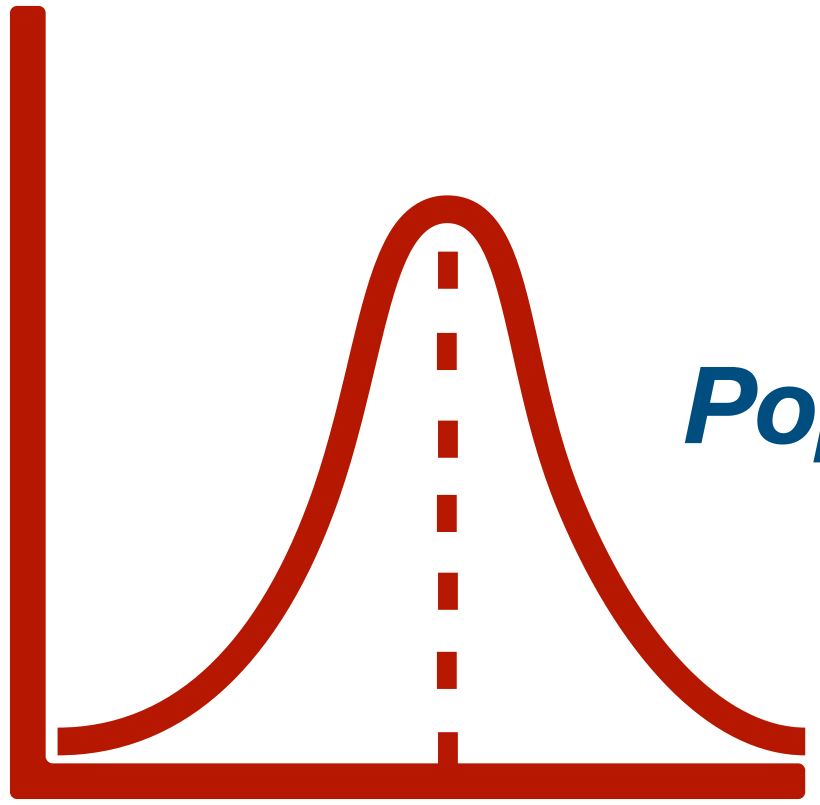
Aim: *Learning about the sequencing data, data formats, programs, softwares, tools to prepare genomic datasets for population genetics analyses*

Hands-on: *Preparing a toy dataset and computing allele frequencies and measuring the HWE*

Reading suggestions: *Nielsen and Slatkin 2013 An Introduction to Population Genetics Chapter 1*
https://evolutionarygenetics.github.io/Chapter3.html *https://cooplab.github.io/popgen-notes/*

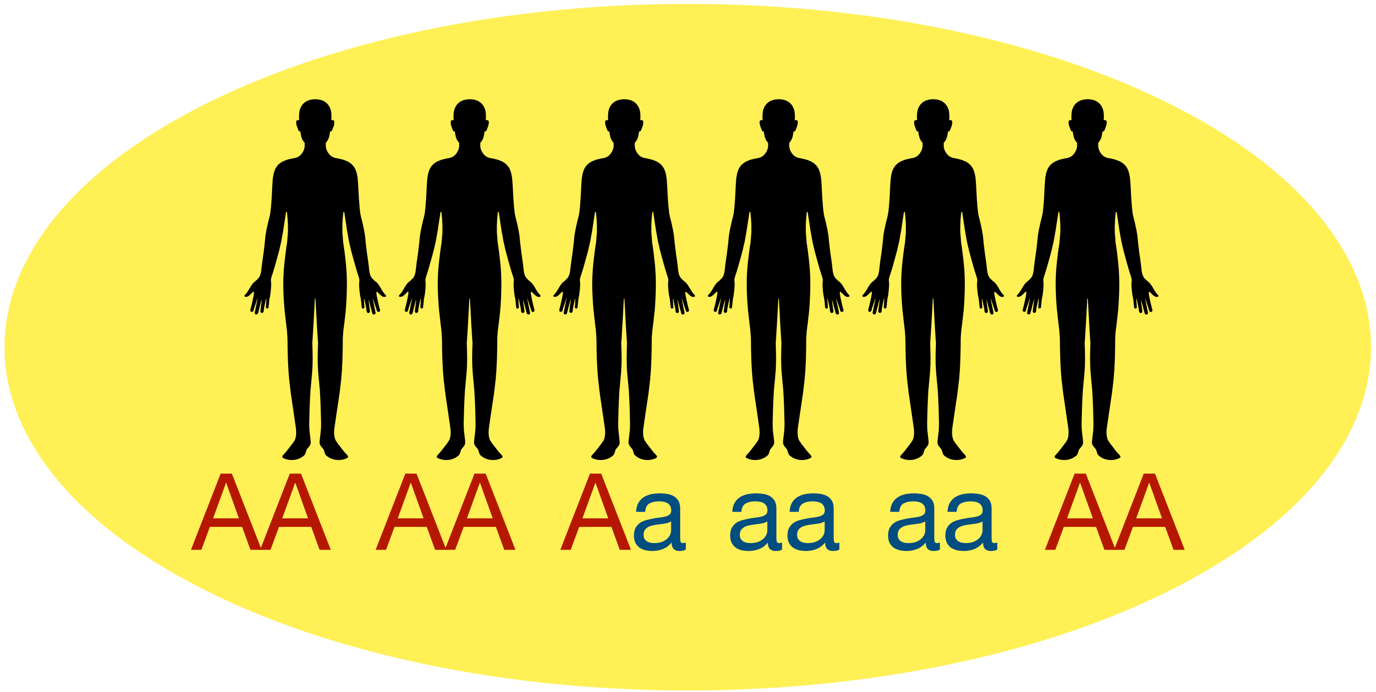
Basics

- **Locus (p. Loci):** Any segregating position(s) in the genome / should not be necessarily coding. -> MC4R, rs373838, Chr1 3874902, C, CTCTCT….
- **Genotype:** Combination of alleles for the given position: Chr17-9,158,696, ->TT, TC, CC
- ***Diploid species ->*** Two copies of all chromosomes / N diploid individuals > 2N copies of each locus



Population genetics: How allele frequencies change over time?

Allele and genotype frequencies



Di-allelic model -> A, a

N= 6 individuals
Gene copies: 12
Allele A: 7
Allele a: 5

Allele frequency: Number of allele copies in the population / Number of gene copies in the population

$f_A = N_A/2N$

$f_a = N_a/2N$

$f_A + f_a = 1$

Genotype frequency: Number of individuals carrying the genotype / Total number of individuals

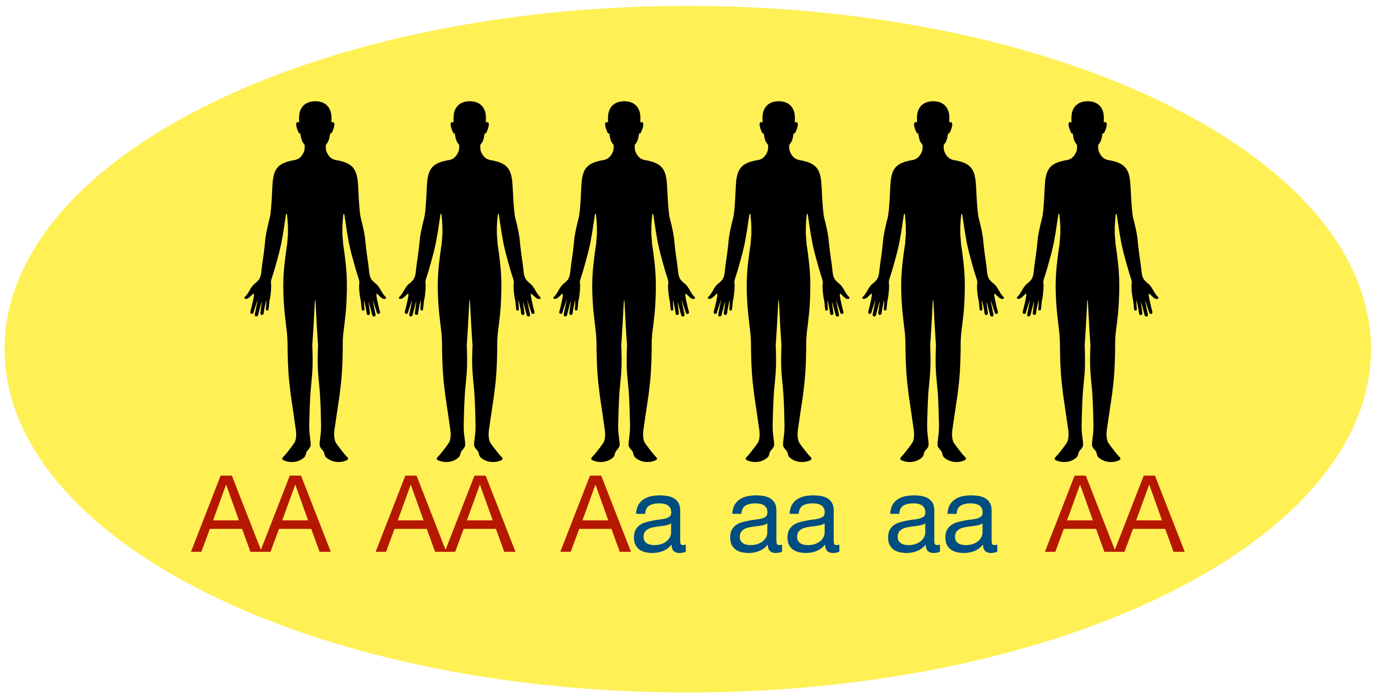
$f_{AA} = N_{AA}/N$

$f_{aa} = N_{aa}/N$

$f_{Aa} = N_{Aa}/N$

$f_{AA} + f_{aa} + f_{Aa} = 1$

Genotype frequencies and heterozygosity



Di-allelic model -> A, a

N= 6 individuals
Gene copies: 12
Allele A: 7
Allele a: 5

We can compute allele frequency based on genotype frequency

$$f_A = N_A/2N$$

$$f_A= 2N_{AA} + N_{Aa} /2N \quad = \quad f_{AA} + f_{Aa} /2$$

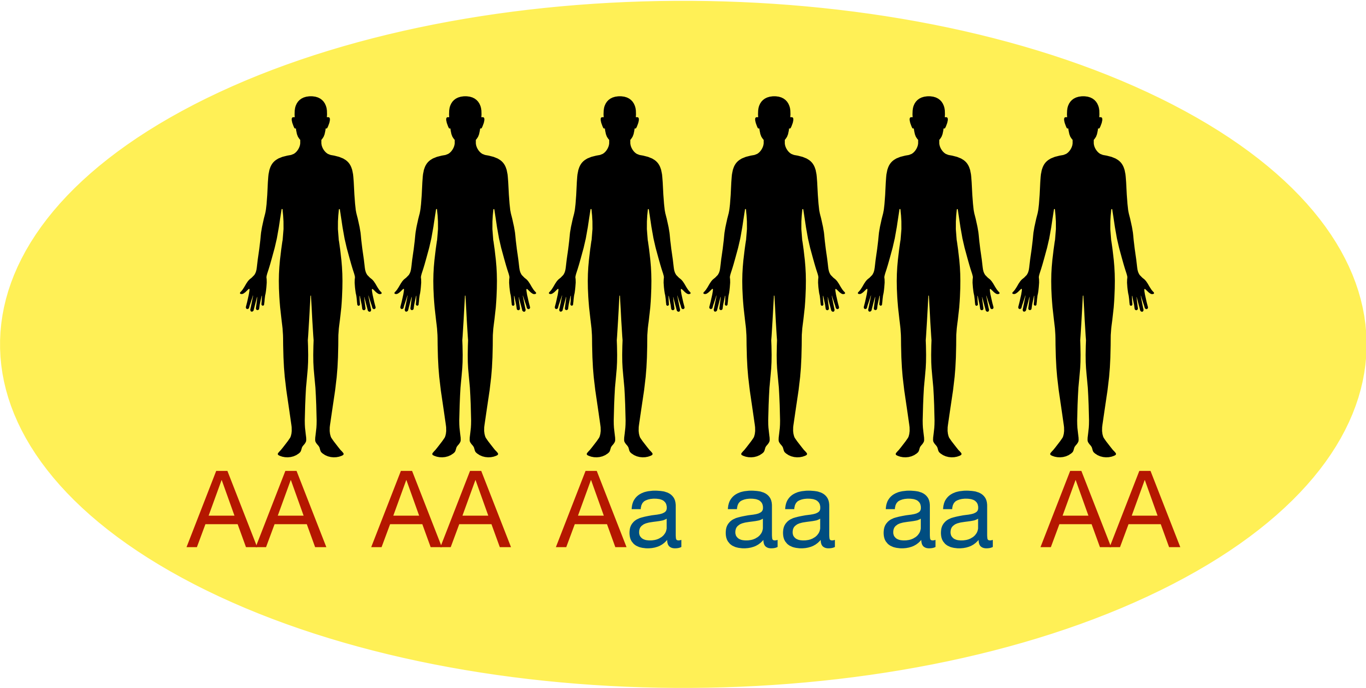
$$f_a \; = \; f_{aa} + f_{Aa} /2$$

f_{Aa} -> Proportion of heterozygous individuals in the population => HETEROZYGOSITY

$1\text{-}\ f_{Aa} = f_{aa} + f_{AA}$ HOMOZYGOSITY

If heterozygosity is high in the population -> more diverse

K-allelic loci -> k different alleles



Di-allelic model -> A, a

N= 6 individuals
Gene copies: 12
Allele A: 7
Allele a: 5

$$f_A = N_A/2N$$

$$f_A = 2N_{AA} + N_{Aa} /2N \quad = \; f_{AA} + f_{Aa} /2$$

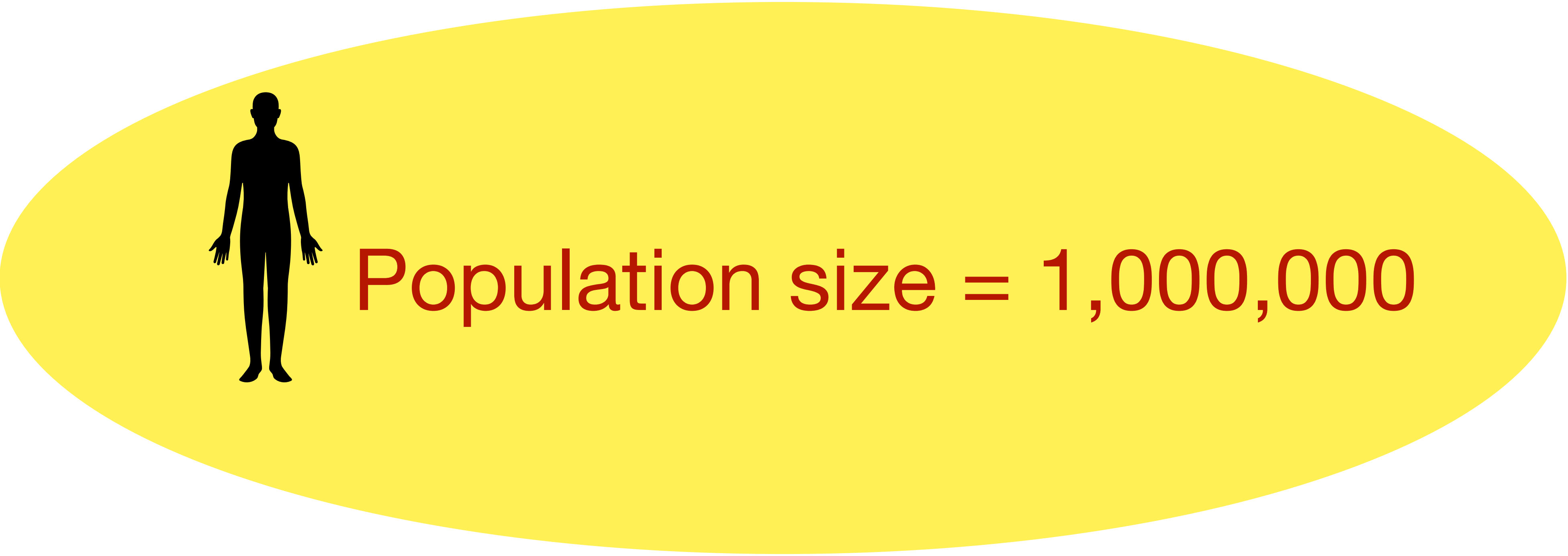
$$f_A = f_{AA} + f_{Aa} /2$$

$$f_i = f_{ii} + \sum_{j:j\neq i} f_{ij} /2$$ ***k-allelic***

Homozygosity: $\sum_{i} f_{ii}$

Heterozygosity: $\sum_{(i,j):i<j} fj$

Estimate allele frequency



Population size = 1,000,000

Random sample: 30 individuals
Chr 2, 122839, C/T

$f_{CC} = 25/30 = 0.833$

$f_{CT} = 5/30 = 0.167$

$f_{TT} = 0$

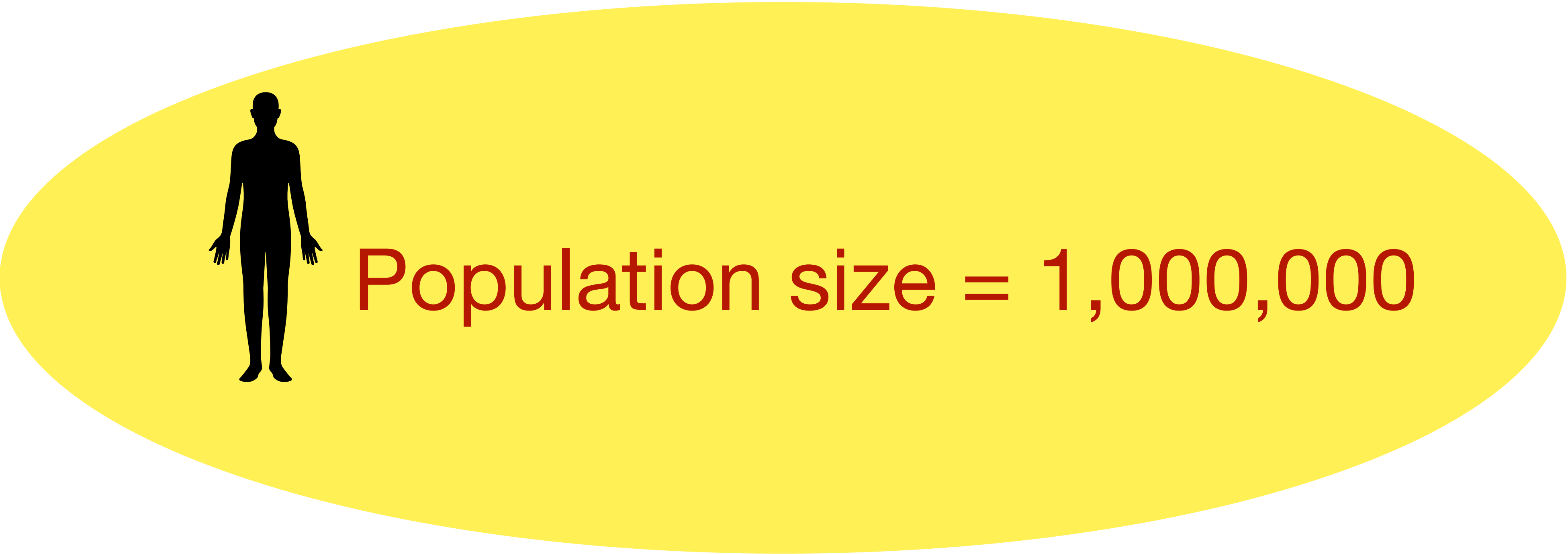
$f_{C} = 0.833+0.167 /2 = 0.917$

$f_{T} = 1- 0.917 = 0.083$

Allele frequencies can be computed from genotype frequencies

Can we estimate genotype frequencies from allele frequencies?

Hardy-Weinberg Model - how can these frequencies change?



Random sample: 30 individuals
Chr 2, 122839, C/T

If f_T = 0.08 , what proportion of the population is expected to have TT genotype?

A way of explaining relationships between genotype and allele frequencies

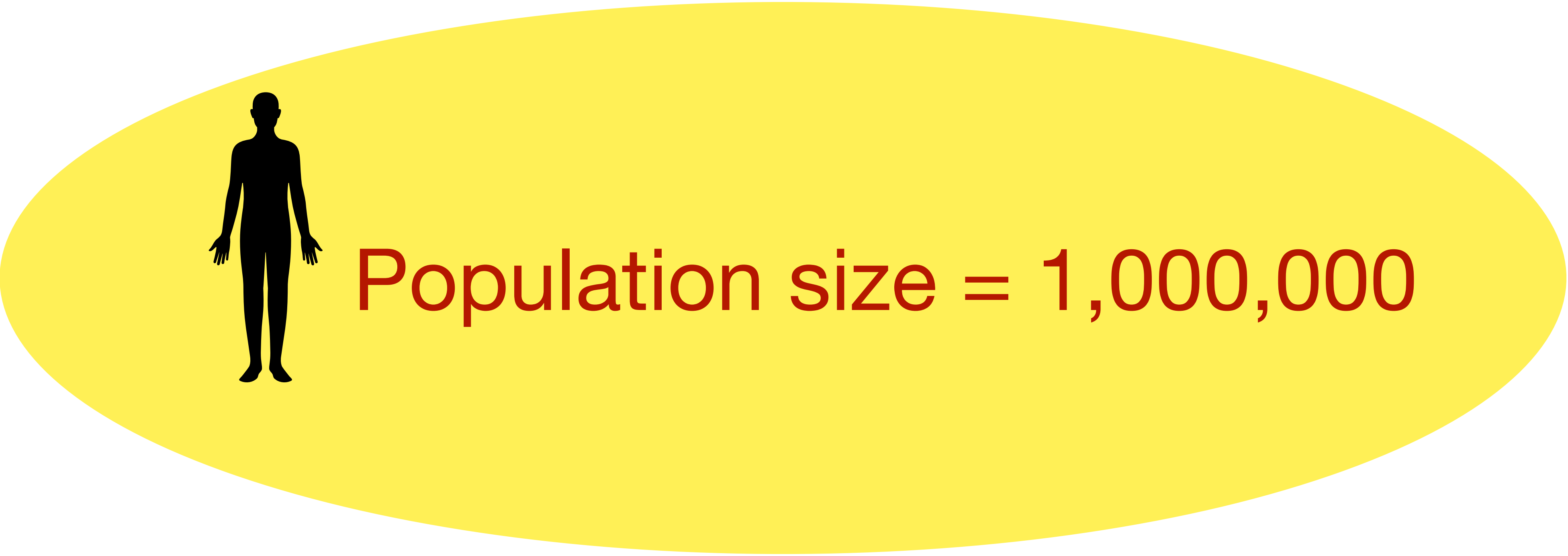
Assumptions:
Random mating - without regard genotypes
Infinitely large population
No selection/mutation/gene flow

-> Genotype frequency <- probability

Deviations:

Assortative mating
Inbreeding
Selection
Population structure

Hardy-Weinberg Model - how can these frequencies change?



Population size = 1,000,000

Random sample: 30 individuals
Chr 2, 122839, C/T

If f_T = 0.08 , what proportion of the population is expected to have TT genotype?

Random mating

-> Genotype frequency <- probability

Probability of A allele transmitted to next generation f_A

Probabilities of genotype frequencies under HWE

Genotype AA -> Probability from mother and father ->

$$f_A f_A = f_A^2$$

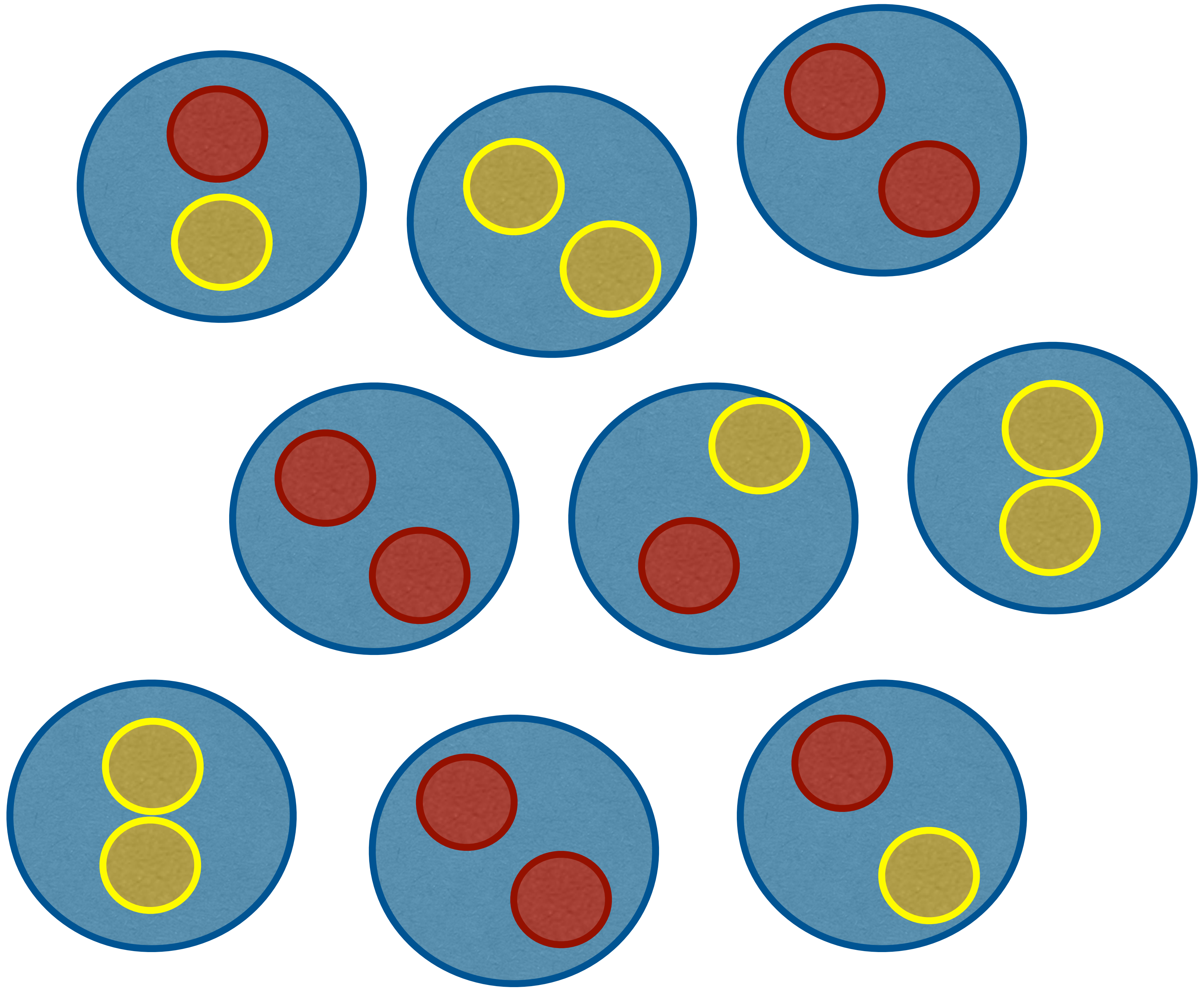
Genotype aa f_a^2

Genotype Aa $2f_af_A$

Expected heterozygosity: $2f_af_A$
Expected homozygosity: $f_A^2 + f_a^2$

Week-2
Hands-on

Aim: *Preparing a toy dataset using pileupCaller, computing allele frequencies, measuring the HWE, visualising results with R*



https://evolutionarygenetics.github.io/Chapter3.html *https://cooplab.github.io/popgen-notes/*

PLAN

1- Prepare a dataset from low coverage bams

use samtools and mpileupcaller + scripts in our Github page

2- Merge new dataset with the existing one

use AdmixTools mergeit + scripts in our Github page

3- Convert eigenstrat files to ped files

use AdmixTools convertf, + scripts in our Github page

4- Use Plink to compute allele frequencies

use Plink —freq, + scripts in our Github page

5- Use R to understand allele and genotype frequencies

Our Github page

Preparing a dataset - starting from BAM files

SAMTools	GATK	SAMTools + PileupCaller
Low to high coverage	Medium to high coverage	Low to high coverage
Fast	Comparably slow	Fast
output: bcf, mpileup	output: vcf	output: eigenstrat

Install softwares and tools

Samtools:
http://www.htslib.org/download/

Pileupcaller
https://github.com/stschiff/sequenceTools

Plink:
https://www.cog-genomics.org/plink/

Prepare a dataset

Reference genome: Chromosome 22 (*v. hs37d5*): https://www.dropbox.com/sh/ys3ud3jvu2hk0jo/AAA1YDv8L9z4uEYyusLCxWkna?dl=0

Bam files: https://github.com/gulki/BIN784

Eigenstrat files: https://github.com/gulki/BIN784

Bed file: https://github.com/gulki/BIN784

Make a dataset using pileupcaller

Example code: https://github.com/gulki/BIN784/tree/main/scripts/pileupcaller.sh

We need:

1- bamlist: bamlist.txt - one bam file per line
2- reference genome
3- bed file
4- eigensnpfile

Compute allele frequencies

Example code: https://github.com/gulki/BIN784/blob/main/scripts/plink_allele_freq.sh