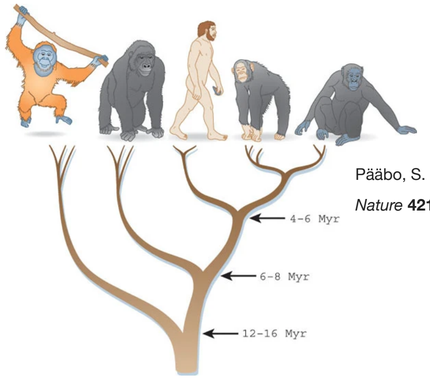
# BIOINFORMATICS IN POPULATION GENETICS

## *BIN784*

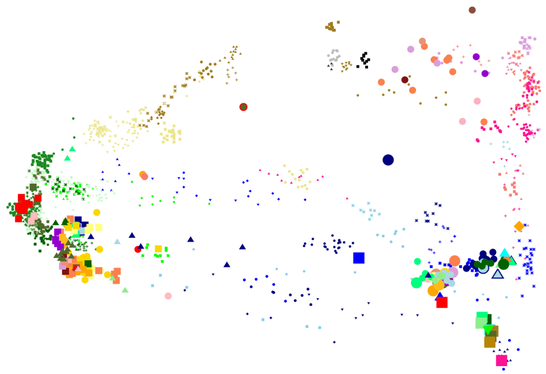

GRADUATE SCHOOL OF HEALTH SCIENCES
DEPARTMENT OF BIOINFORMATICS
INSTRUCTOR: GÜLŞAH MERVE KILINÇ

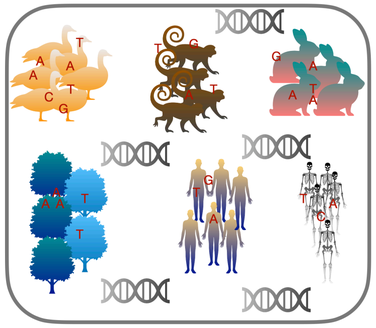
email: gulsahhdal@gmail.com

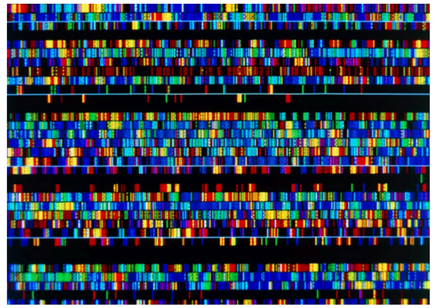

October 6th - January 5th
Tuesday, 13.30 am



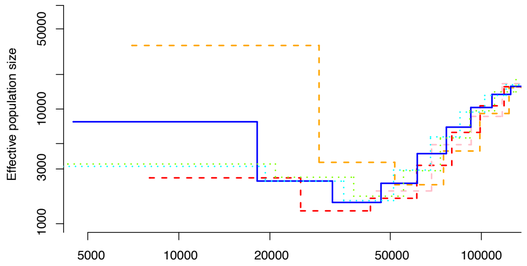

Pääbo, S. The mosaic that is our genome.
*Nature* **421**, 409–412 (2003). https://doi.org/10.1038/nature01400







James King-Holmes/Science Photo Library, Nature, News 2021



# COURSE OUTLINE

## SYLLABUS

| | | |
|---|---|---|
| Week 1 | 27.Feb.24 | Introduction to population genetics: Allele and genotype frequencies, Hardy-Weinberg Equilibrium, genetic drift and mutation Nielsen & Slatkin 2013 An Introduction to population genetics. Chapter 1 & 2 Mark Jobling, Edward Hollox, Matthew Hurles, Toomas Kivisild, Chris Tyler-Smith, Human Evolutionary Genetics, Chapter 5 |
| Week 2 | 5.Mar.24 | Population Genetics Analysis based on next generation sequencing data - DATA FORMATS [FASTQ, BAM, VCF, PLINK, EIGENSTRAT] Luikart, G., England, P., Tallmon, D. et al. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4, 981?994 (2003). https://doi.org/10.1038/nrg1226 |
| Week 3 | 12.Mar.24 | Considering the statistical uncertainty: Population genetics analysis based on genotype probabilities - ANGSD & ngsTools Korneliussen T, Albrechtsen A, Nielsen R BMC Bioinformatics. 2014 Nov 25;15(1):356 Fumagalli M, Vieira FG, Linderoth T, Nielsen R. Bioinformatics. 2014 May 15;30(10):1486-7 Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. . PLoS One. 2012;7(7):e37558. doi: 10.1371/journal.pone.0037558. Epub 2012 Jul 24. PMID: 22911679; PMCID: PMC3404070. |
| Week 4 | 19.Mar.24 | Preparing genomic datasets for population genetics analysis https://github.com/stschiff/sequenceTools https://zzz.bwh.harvard.edu/plink/data.shtml https://www.htslib.org |
| Week 5 | 26.Mar.24 | Principal component analysis |
| Week 6 | 2.Apr.24 | F3-statistics |
| Week 7 | 16.Apr.24 | F4-statistics |

DATA ANALYSIS - HANDS-ON

| | | |
|---|---|---|
| **Week 8** | **23.Apr.24** | **MIDTERM - Project-based** |
| Week 9 | 30.Apr.24 | Linkage Disequilibrium and Runs of homozygosity Slatkin, M. Linkage disequilibrium understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9, 477?485 (2008). https://doi.org/10.1038/nrg2361 Fox EA, Wright AE, Fumagalli M, and Vieira FG ngsLD: evaluating linkage disequilibrium using genotype likelihoods, Bioinformatics (2019) 35(19):3855 - 3856 https://github.com/fgvieira/ngsLD https://www.cog-genomics.org/plink/1.9/ld |
| Week 10 | 7.May.24 | qpAdm & qpGraph Analysis |
| Week 11 | 14.May.24 | Allele frequency trajectories |
| **FINAL EXAM** | **28.May.24** | **FINAL EXAM** |

*ATTENDANCE 5%*

*HANDS-ON 15%*

*MIDTERM - Week 8 30%*

*FINAL - Week 13 50%*

*MIDTERM 23th of April - Project*

*FINAL EXAM 28th of May*

*Week 1 - 27th of Feb*

*Weeks that we do not have courses:*
*9th of April - Holidays*
*23th of April - Midterm*
*21th of May*

**GitHUB** https://github.com/gulki/BIN784

**UNIX & UBUNTU**
ssh username@10.135.10.20

**WINDOWS**

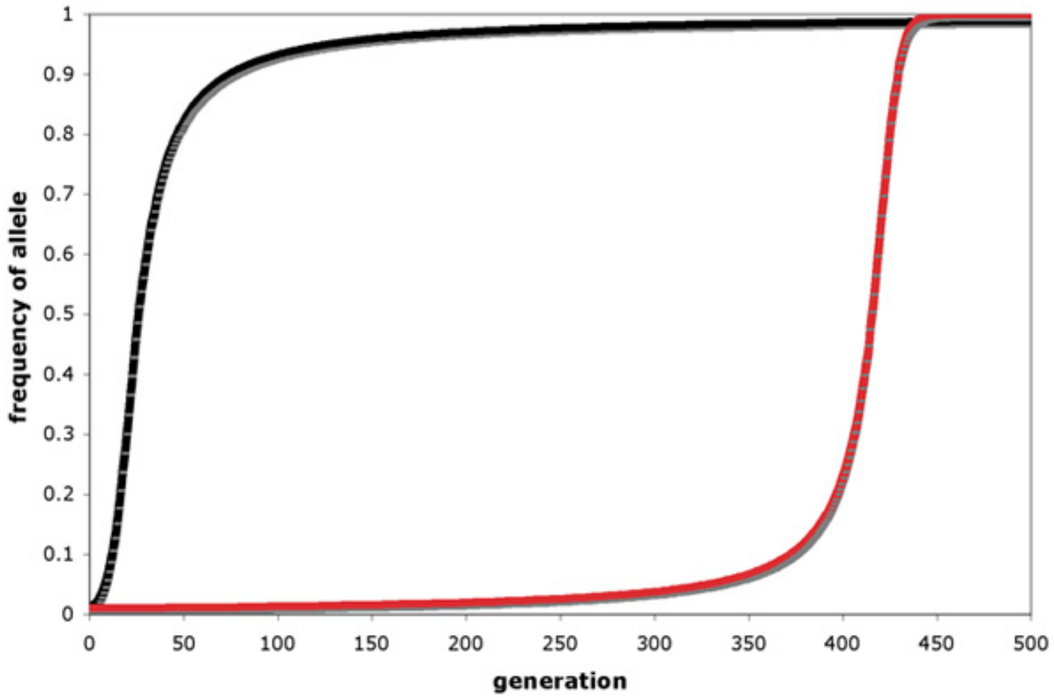



username@10.135.10.20

`/DATA/teaching/BIN784_2024`

# OUTLINE

- Terminology
- Allele frequency
- Genotype frequency
- Hardy-Weinberg Equilibrium
- Wright-Fisher  Model
- Mutation
- Mutation rate
- Genetic drift

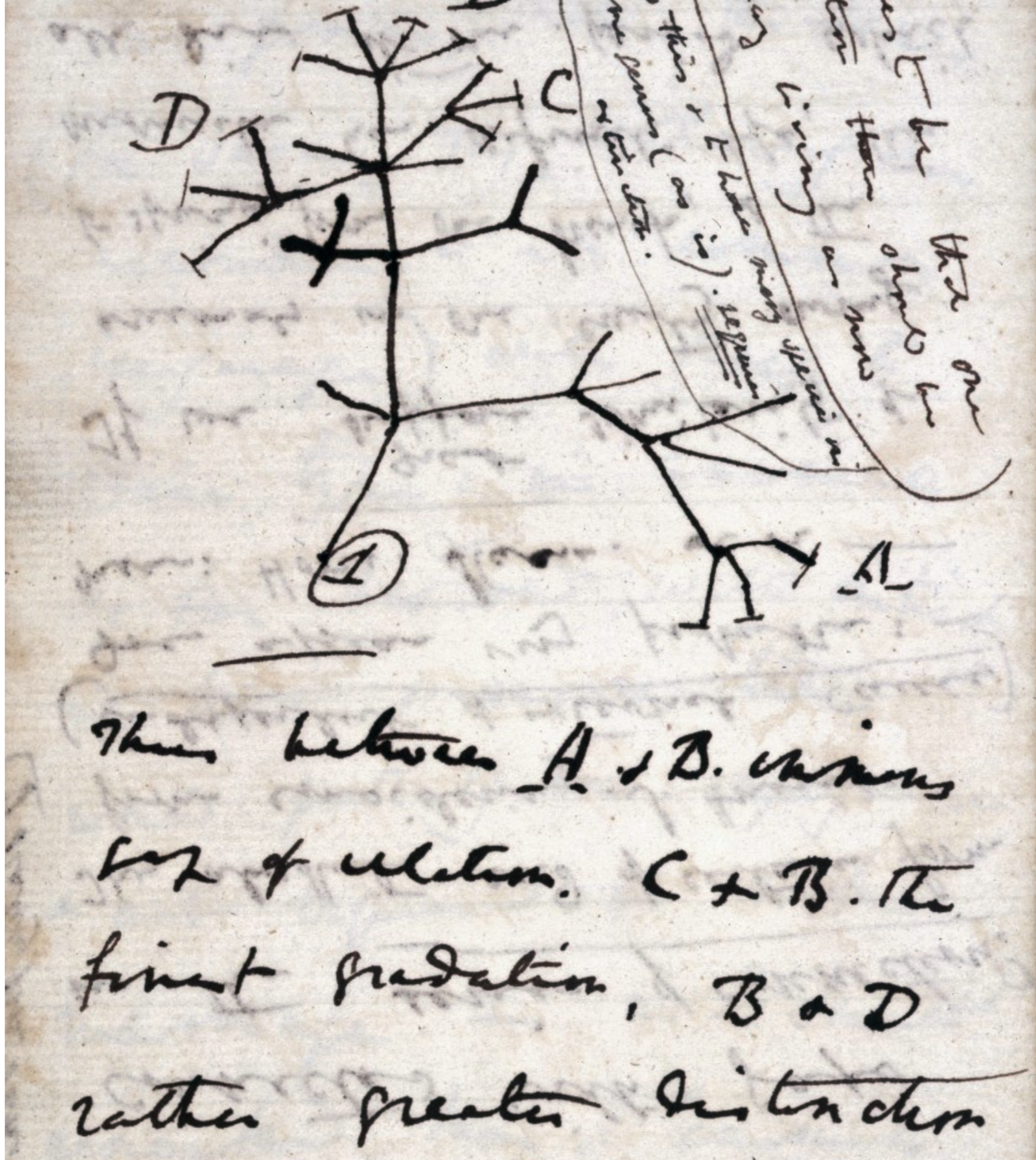



Andrews, C. A. (2010) Natural Selection, Genetic Drift, and Gene Flow
Do Not Act in Isolation in Natural Populations. Nature Education Knowledge 3(10):5

Week 1 is based on *Nielsen and Slatkin 2013 Chapter 1 and Chapter 2*

## Macro- and microevolution

- Macroevolution - evolution of species over time
- Microevolution - evolution of a population
- Microevolutionary processes -> Genetic diversity
- Population genetics -> studying allele frequencies

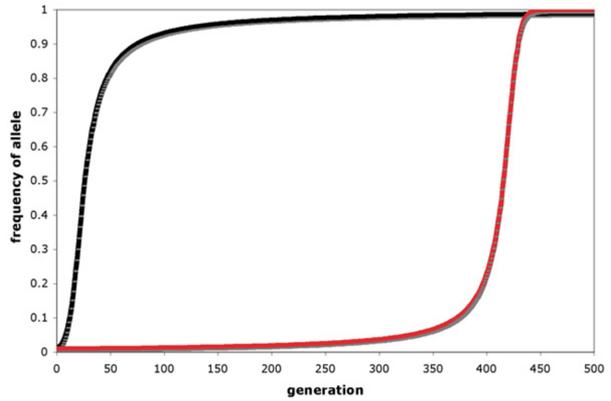


# How do we study evolutionary processes?

## *Spatiotemporal changes in allele frequencies in a population*

## *Population genetics*

- We need evolutionary models
- Infer past evolutionary processes from observed genetic diversity
- Distinct models -> test which one fits best to the observed data
- Explain patterns of **genetic diversity** inferred from data with these models: E.g. a prehistoric migration between populations? divergence during isolation? selection?




Andrews, C. A. (2010) Natural Selection, Genetic Drift, and Gene Flow Do Not Act in Isolation in Natural Populations. Nature Education Knowledge 3(10):5

# Terminology and basics

## *Population genetics*

- Some terms are defined differently in population genetics
- **Locus** (population genetics): Any position in the genome / any unit in the genome with one or more alleles

E.g. MC1R gene, microsatellite, SNP

- **Locus** (genetics): A coding gene (sometimes)
- **Genotype**: A combination of alleles carried by an individual in a particular locus

An individual is homozygous in position 2,345,678 - individual's genotype is TT

✳ Diploid species - 2 copies of each chromosome (such as humans)

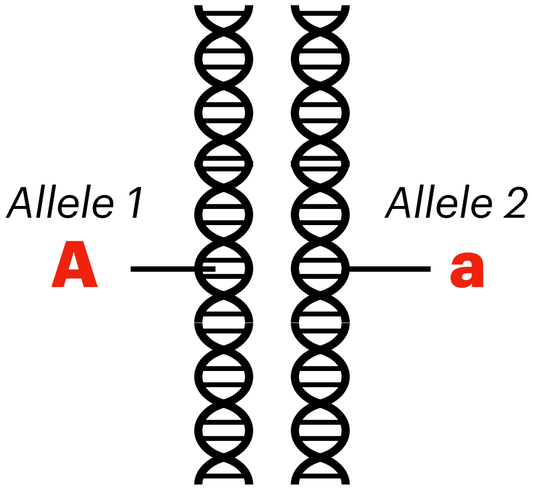
✳ For a collection of N diploid individuals there are 2N gene copies at each locus

***Population genetics' main objective is to study how allele frequency change over time***

# Allele Frequencies

Number of copies of the allele in the population divided by the total number of gene copies in the population

***Allele: Two or more versions of DNA sequence***




2N - 2 copies of chromosomes

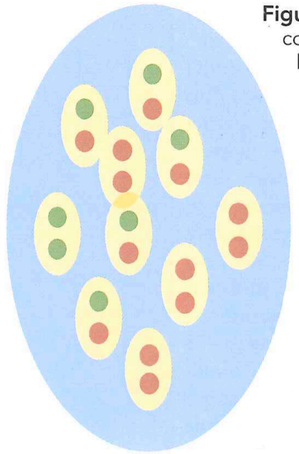


**Figure 1.1** A hypothetical population with $N = 10$ individuals, 20 gene copies, and a total of 7 copies of allele $A$ (green) and 13 copies of allele $a$ (red), i.e., $f_A = 7/20$ and $f_a = 13/20$. The genotype frequencies are $f_{AA} = 1/10$, $f_{Aa} = 5/10$, and $f_{aa} = 4/10$.

$$f_A = \frac{N_A}{2N} \text{ and } f_a = \frac{N_a}{2N} \qquad (1.1)$$

$$f_A + f_a = 1$$

*Nielsen and Slatkin 2013*

# Genotype Frequencies

Allele frequencies in the population can be calculated from the genotype frequencies.

Possible genotypes for a diallelic locus: AA, Aa, aa ; Number of copies for genotypes $N_{AA}$, $N_{Aa}$, $N_{aa}$

$$f_{AA} = \frac{N_{AA}}{N} \qquad f_{Aa} = \frac{N_{Aa}}{N} \qquad f_{aa} = \frac{N_{aa}}{N} \qquad (1.2) \qquad f_{aa} + f_{Aa} + f_{AA} = 1$$

N - Number of individuals

Frequency of allele A

$$f_A = \frac{2N_{AA} + N_{Aa}}{2N} = f_{AA} + f_{Aa}/2 \qquad (1.3)$$

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

**Heterozygosity**: Proportion of individuals that are heterozygous in the population $f_{Aa}$,

***Homozygosity***: *Proportion of individuals that are homozygous in the population* $1\text{-}f_{Aa} = f_{AA} + f_{aa}$

*Nielsen and Slatkin 2013*

# K-allelic Loci

A locus with k different alleles - k is any positive number -> ***k-allelic loci***

Microsatellite loci -> often with more than 2 alleles

Allele and genotype frequencies  for a k-allelic locus is similar to the di-allelic locus

For an allele $i \in \{1, 2, \ldots, k\}$ with $N_i$ copies in the population $f_i = N_i/2N$

For a genotype $ij\ (= ji)$, the genotype frequency is $f_{ij} = N_{ij}/N$.

*Nielsen and Slatkin 2013*

# Hardy-Weinberg Equilibrium - HWE

## *Relationship between allele and genotype frequencies in a population*

***Allele frequencies can be estimated based on genotype frequencies - is the opposite possible?***

*If the allele frequency for T in 478th locus of MC1R gene is 0.08 - what proportion of the pop can have TT genotype?*

- Infinite number of random mating
- Sexually reproducing diploid organisms
- No selection
- No mutation
- No overlap between generations
- No migration
- No substructure

***If genotype frequencies are inconsistent with HWE - one of these parameters might has been broken***

# Hardy-Weinberg Equilibrium - HWE

*If the allele frequency for T in 478th locus of MC1R gene is 0.08 - what proportion of the pop can have TT genotype?*

Assumptions:

1-Allele frequencies among males and females are same for a particular locus with two alleles: A and a

2-Mendel->Probability of allele A is transmitted to the next generation is frequency of allele A = $f_A$

3-Assumption of random mating -> we can multiply the probabilities from mother and father -> $f_A^2$

4- An individual can be heterozygous by getting an A from one parent and a from another parent $f_A f_a + f_a f_A = 2f_A f_a$

5- Homozygous for a -> $f_a^2$

**Expected heterozygosity in the population -> $2f_A f_a$**

**Expected homozygosity in the population -> $f_A^2 + f_a^2$**

$$f_A^2 + 2f_A f_a + f_a^2 = (f_A + f_a)^2 = 1 \qquad (1.5)$$

$$\begin{aligned} &\Pr(\textit{offspring genotype} = AA) \\ &= \Pr(\textit{paternal allele} = A) \times \Pr(\textit{maternal allele} = A) \qquad (1.6) \\ &= f_A f_A = f_A^2 \end{aligned}$$

*Nielsen and Slatkin 2013*

# HWE - Example

*Allele frequency for T in 478th locus of MC1R gene is 0.08 - what proportion of the pop can have TT genotype?*

*Based on HWE:*

*$TT = 0.08^2 = 0.06$*

*$CC = 0.92^2 = 0.85$*

*$CT = 2 \times 0.92 \times 0.08 = 0.15$*

*Nielsen and Slatkin 2013*

# Deviations from HWE

## 1 - Assortative mating

Individuals may be more likely to mate with other individuals with similar genotype

AA individuals prefer AA individual to mate, aa prefer to mate with aa, AA and aa rarely mate so lower number of expected heterozygotes than predicted by HWE

## 2 - Inbreeding

Mating between individuals that are related and have a common ancestor.

Similar to assortative meeting number of heterozygous individuals will be smaller than predicted by HWE.

## 3 - Population structure

Sampling ignoring the population structure

## 4 - Selection

Example: Genotype frequencies in HEXA gene: Homozygous individuals dies before adulthood, adult population is out of HWE with excess of heterozygotes

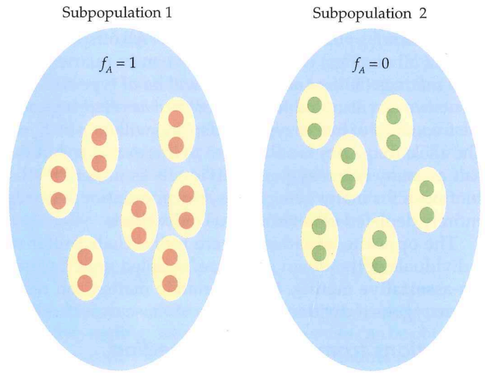



**Figure 1.2** Two subpopulations with allele frequencies $f_A = 0$ and 1, respectively. In the combined population, obtained by pooling individuals from subpopulation 1 and subpopulation 2, all individuals are homozygous and there is an apparent deficit of heterozygous individuals compared to the HWE expectation.

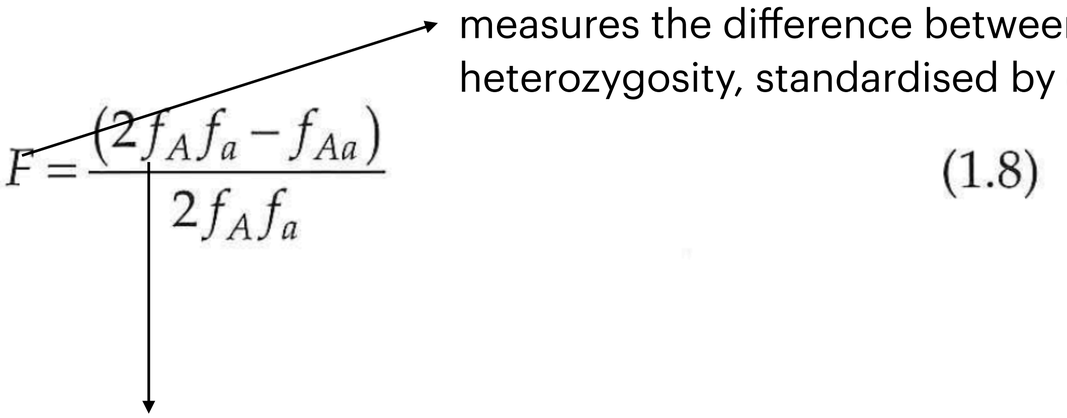
*Nielsen and Slatkin 2013*

# Why we can not detect selection based on deviation from HWE?

- There should be very strong natural selection to change allele frequencies enough to be detected
- One round of random mating restores the HWE - so HWE can detect selection in current generation
- Deviation from HWE is very rare in humans
- More sophisticated models are needed incorporating mutation rate, recombination rate, effective pop size

# The most common statistic to measure deviation from HWE:
## *The inbreeding coefficient, F*

- F is used to describe degree to which heterozygosity is reduced both in individuals and in population as a result of inbreeding.
- Decrease in heterozygosity in the population beyond that expected under HWE



$$F = \frac{(2f_A f_a - f_{Aa})}{2f_A f_a} \qquad (1.8)$$

measures the difference between expected and observed heterozygosity, standardised by expected heterozygosity

proportion of individuals expected to be heterozygous under HWE

F = 0 -> population is under HWE

F = 1 -> no heterozygotes in the population

# **Population genetic theory:** Focused on allele frequency changes over time

- How and why frequencies change over time?
- Two most important factors causing allele frequency changes: Natural selection and genetic drift.
- **Genetic drift**: Random changes in allele frequencies in populations of finite size

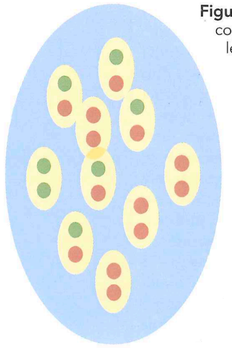


**Figure 1.1** A hypothetical population with $N = 10$ individuals, 20 gene copies, and a total of 7 copies of allele $A$ (green) and 13 copies of allele $a$ (red), i.e., $f_A = 7/20$ and $f_a = 13/20$. The genotype frequencies are $f_{AA} = 1/10$, $f_{Aa} = 5/10$, and $f_{aa} = 4/10$.

A scenario:

individuals are randomly mating in this population

some produce more some produce less offspring due to extrinsic factors

some might die before reproductive age

next generation not 7 A and 13 a alleles

if this change continuously over generations, it will cause large allele frequency changes

*Nielsen and Slatkin 2013*

# The Wright-Fisher Theory

- The most common model that is used to describe the genetic drift
- WF model - assumes a haploid population, w/o sexes appropriate for many bacterial populations but:
- Most of the dynamics of a diploid population with two sexes have identical properties with haploid population.
- WF model -> haploid population w/o sexes, discrete generations, constant population size

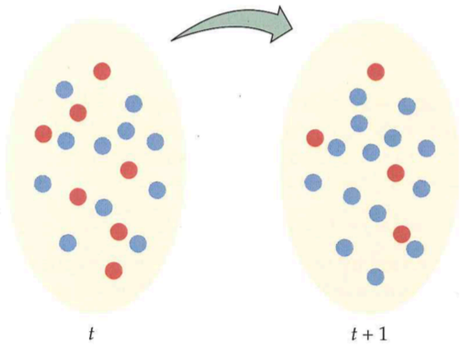



**Figure 2.1** An illustration of two generations of a Wright–Fisher population with $2N = 18$ gene copies. In generation $t$ the allele frequency of allele $A$ (red) is $7/18$, but due to genetic drift, the allele frequency is $4/18$ in generation $t+1$.

*Nielsen and Slatkin 2013*

# Genetic drift and expected allele frequencies

- Characterize AF changes using WF model:
- What is the probability that any particular gene copy in generation t+1 is of type A? $f_A(t)$
- Expected number of allele A in t+1 generation: 2N gene copies in generation t+1 $2Nf_A(t)$ *A alleles*
- If there are $2Nf_A(t)$ *A* alleles, frequency of allele A is $2Nf_A(t)/2N$

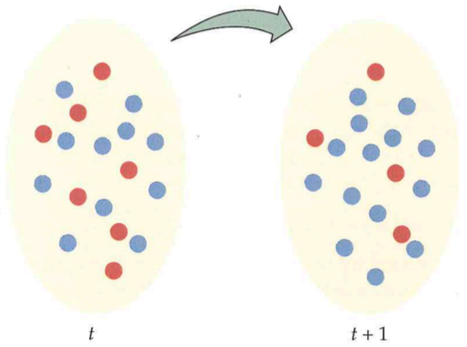



**Figure 2.1** An illustration of two generations of a Wright–Fisher population with $2N = 18$ gene copies. In generation $t$ the allele frequency of allele $A$ (red) is $7/18$, but due to genetic drift, the allele frequency is $4/18$ in generation $t+1$.

$$E[f_A(t+1)] = 2Nf_A(t)/2N = f_A(t)$$

*Nielsen and Slatkin 2013*

# Continue WF sampling over generations

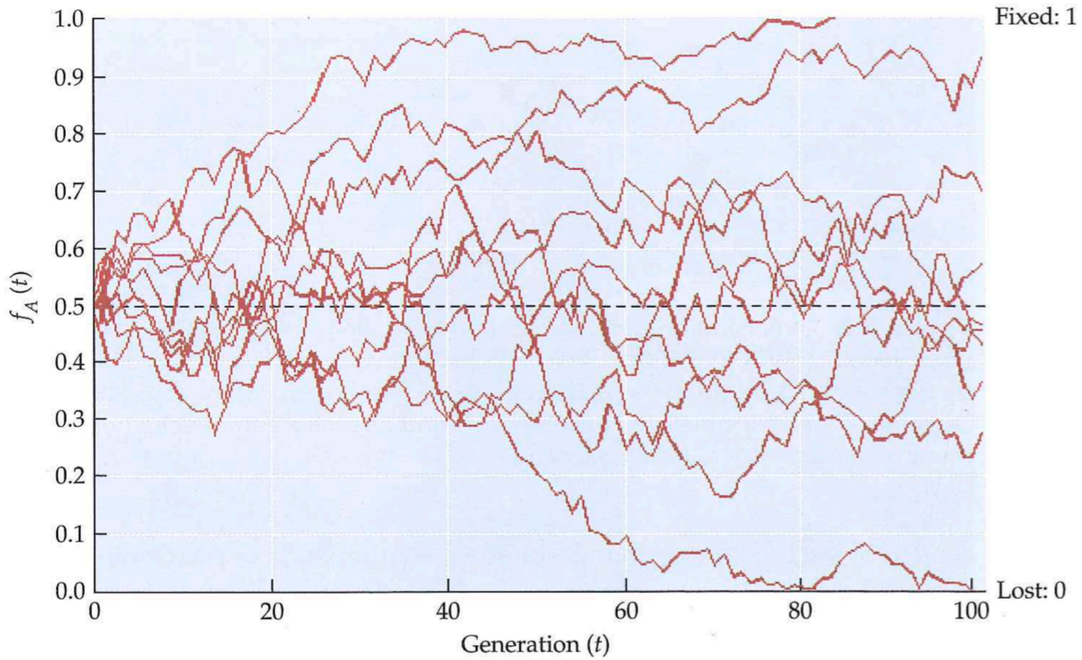



**Figure 2.2** The Wright–Fisher model simulated for 10 populations, with $2N = 100$, over 100 generations (solid lines) for an initial allele frequency of 50%. Allele frequencies change randomly due to genetic drift. The expected (mean) allele frequency is shown by the dashed line.

- In each generation allele frequency might change slightly
- Small changes add up -> results in a significant change
- Many small changes result in large changes
- Increase/decrease -> expected because none of the alleles is favoured

*Nielsen and Slatkin 2013*

# How fast genetic drift can change allele frequencies?

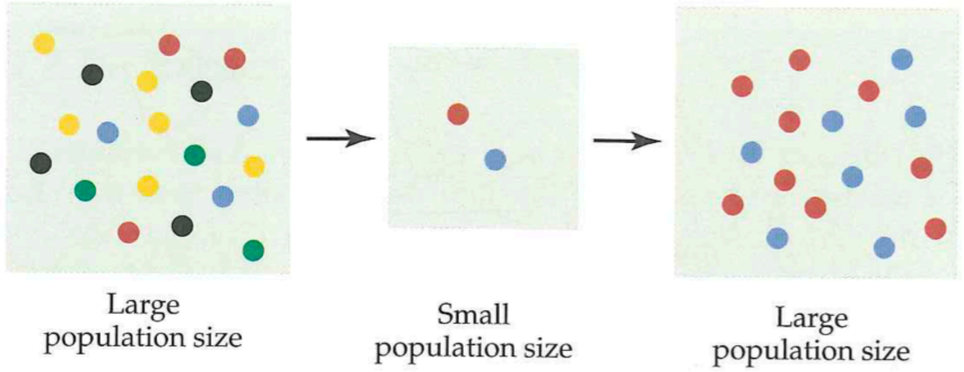


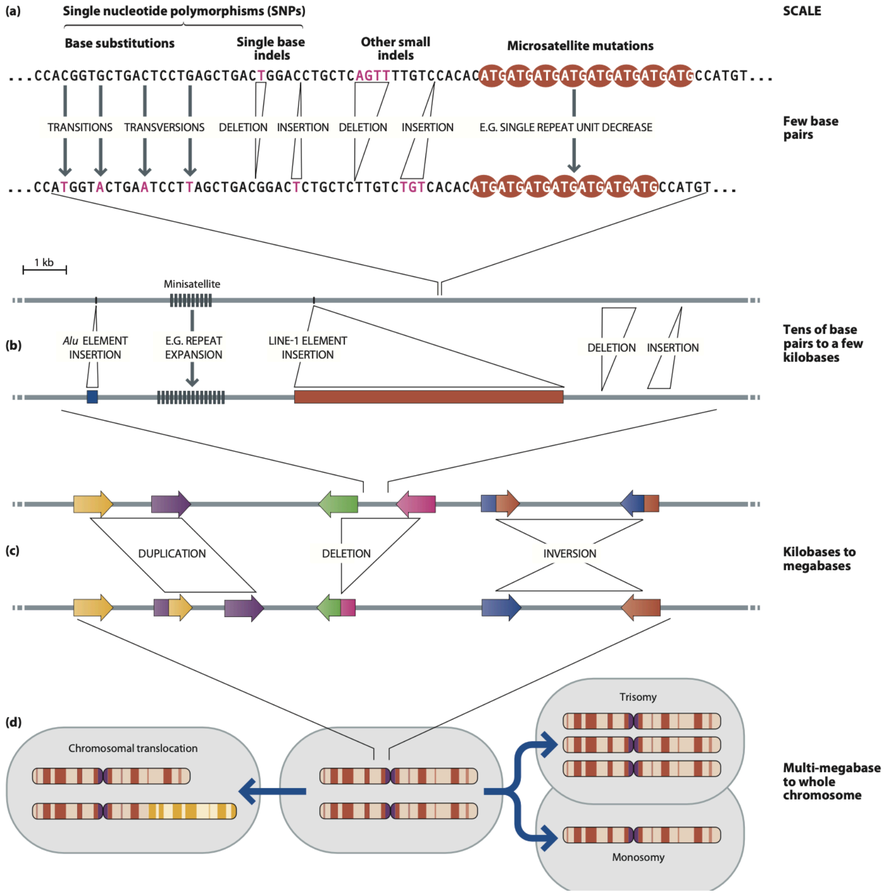

**Figure 2.3** An illustration of the effect of a bottleneck in population size on genetic variability. The initial population has a high degree of variability, illustrated by the variety of colors of the balls in the box. The population goes through a bottleneck—a temporary decrease in size—and after the bottleneck, there are many fewer different alleles present.

- Population size (N)
- Small populations -> large changes

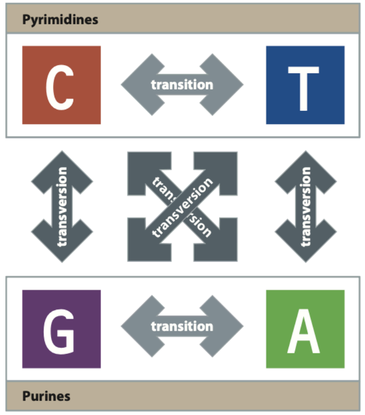
*Nielsen and Slatkin 2013*

# MUTATION
## Human Genome Variation

- Substrate of human evolution
- Large scale - from SNPs to large structural variants
- Different types - transitions (pyrimidine to pyrimidine - purine to purine), transversions (pyrimidine to purine vice versa)




Mark Jobling et al, Human Evolutionary Genetics, Chapter 3



# Mutation changes allele frequencies

- Mutation changes allele frequency
- Mutation pressure: frequency of an allele decreases over time
- Mutation pressure is a weak force- long time is needed to see its impact
- There are multiple mutation models - from simple to complex
- These models are to explain probability of a change at a given locus e.g. Jukes-Cantor model, General reversible model

# Mutation and expected allele frequencies

$$E[f_A(t+1)] = 2Nf_A(t)/2N = f_A(t) \quad (2.1)$$

$$E[f_A(t+1)] = f_A(t) + \mu f_a(t) \quad (2.2)$$

- All individuals will be A in long time

$$E[f_A(t+1)] = (1-\mu_{A\to a})f_A(t) + \mu_{a\to A} f_a(t) \quad (2.3)$$

- Mutation occurs in both directions/ expected frequency in the next generation

# Probability of fixation

- If there is no mutation - each allele can be lost or fixed
- In the absence of selection or drift probability of fixation is the frequency of the allele

$$\Pr(\textit{fixation of allele } A) = N_A \times 1/(2N) = f_A(t) \qquad (2.6)$$

# Hands-on

- Please see the exercises section in GitHub

  https://github.com/gulki/BIN784