# Genetic constitution of a population

Dr Loïc Yengo

Institute for Molecular Bioscience
The University of Queensland

August 6$^{th}$, 2019

# Contents

# Definitions

## Allele(s) and Genotypes(s)

Each "gene" in the genome has multiple possible states: the alleles. In diploid species, individuals at a particular locus can be homozygote if the two copies of the genes are the same or heterozygote if those copies differ.

The genotype of an individual at a particular locus is defined by the states (alleles) of all copies that individual has of the gene at that locus.

Example: In a diploid individual if there are two alleles A or B, then the possible genotypes are AA, BB (homozygote) and AB (heterozygote).

In a triploid individual (e.g. watermelon), the possible genotypes are AAA, BBB, AAB=ABA=BAA (no order) and ABB.

# Definitions

## Allele(s) and Genotypes(s)

Each "gene" in the genome has multiple possible states: the alleles. In diploid species, individuals at a particular locus can be homozygote if the two copies of the genes are the same or heterozygote if those copies differ.

The genotype of an individual at a particular locus is defined by the states (alleles) of all copies that individual has of the gene at that locus.

Example: In a diploid individual if there are two alleles A or B, then the possible genotypes are AA, BB (homozygote) and AB (heterozygote).

In a triploid individual (e.g. watermelon), the possible genotypes are AAA, BBB, AAB=ABA=BAA (no order) and ABB.

# Definitions

## Allele(s) and Genotypes(s)

Each "gene" in the genome has multiple possible states: the alleles. In diploid species, individuals at a particular locus can be homozygote if the two copies of the genes are the same or heterozygote if those copies differ.

The genotype of an individual at a particular locus is defined by the states (alleles) of all copies that individual has of the gene at that locus.

Example: In a diploid individual if there are two alleles A or B, then the possible genotypes are AA, BB (homozygote) and AB (heterozygote).

In a triploid individual (e.g. watermelon), the possible genotypes are AAA, BBB, AAB=ABA=BAA (no order) and ABB.

# Characterization of a population

"To describe the genetic constitution of a [population] we should have to specify their genotypes and say how many of each genotype there were".
*D.S. Falconer*[1]

Population Genetics is a theory of alleles and genotypes frequencies in populations.

Questions

- What is the relationship between alleles and genotypes frequencies?
- How do alleles and genotypes frequencies change?

[1] *Introduction to Quantitative Genetics*, 4th Edition (1996)

# Characterization of a population

"To describe the genetic constitution of a [population] we should have to specify their genotypes and say how many of each genotype there were".
*D.S. Falconer*[1]

Population Genetics is a theory of alleles and genotypes frequencies in populations.

## Questions

- What is the relationship between alleles and genotypes frequencies?
- How do alleles and genotypes frequencies change?

[1] *Introduction to Quantitative Genetics*, 4th Edition (1996)

# Relationship between alleles and genotypes frequencies

Let's consider a population of $n$ diploid individuals at a particular locus with two alleles A and B.

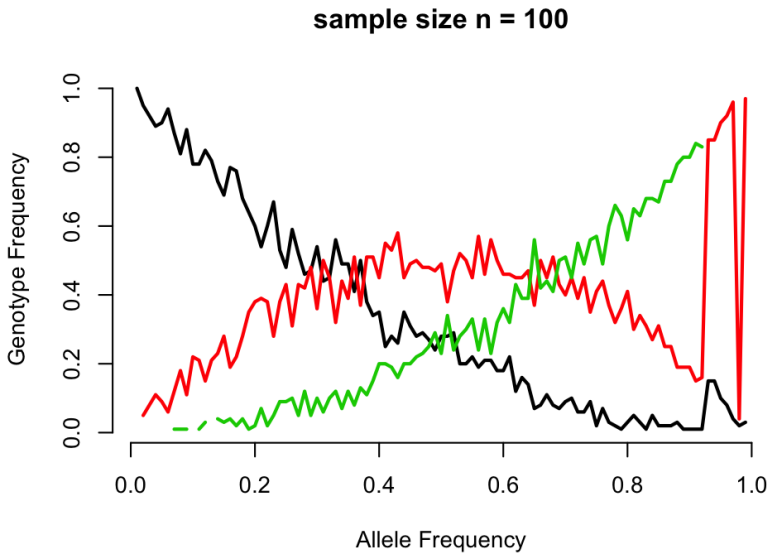
We denote $n_{AA}$, $n_{AB}$ and $n_{BB}$ the genotypes counts and $n_A$ and $n_B$ the allele counts in the population.

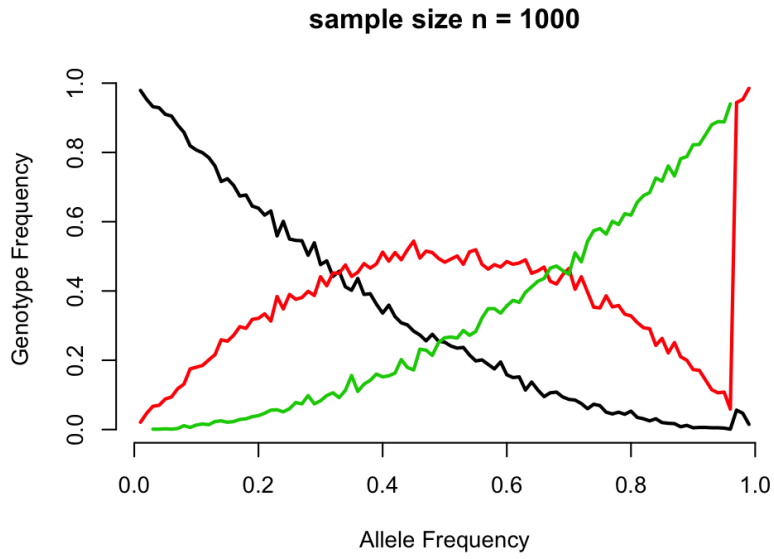
We have the following relationships:

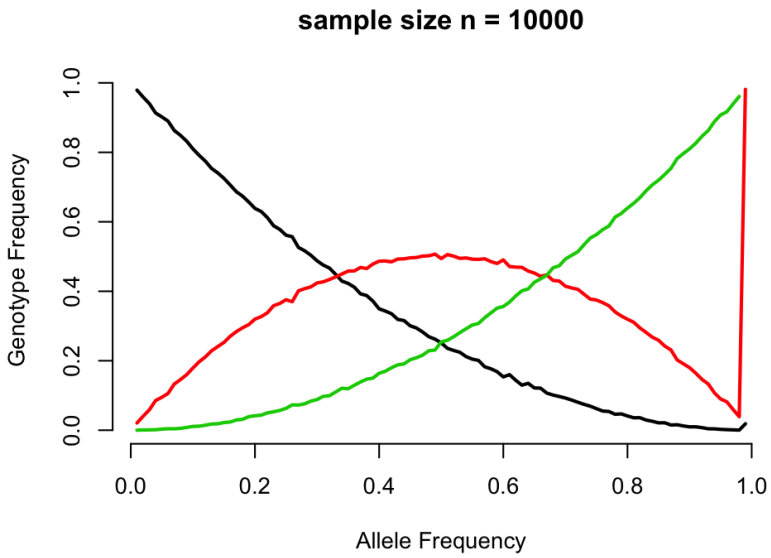
$$n_A = 2n_{AA} + n_{AB} \text{ and } n_B = 2n_{BB} + n_{AB}.$$

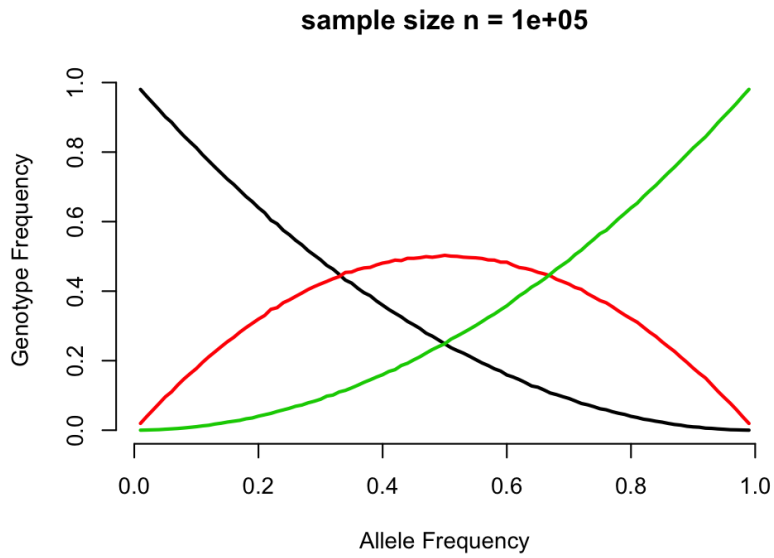
In general, we can not predict the genotype frequencies from the allele frequencies (under-determined).

# Simulation

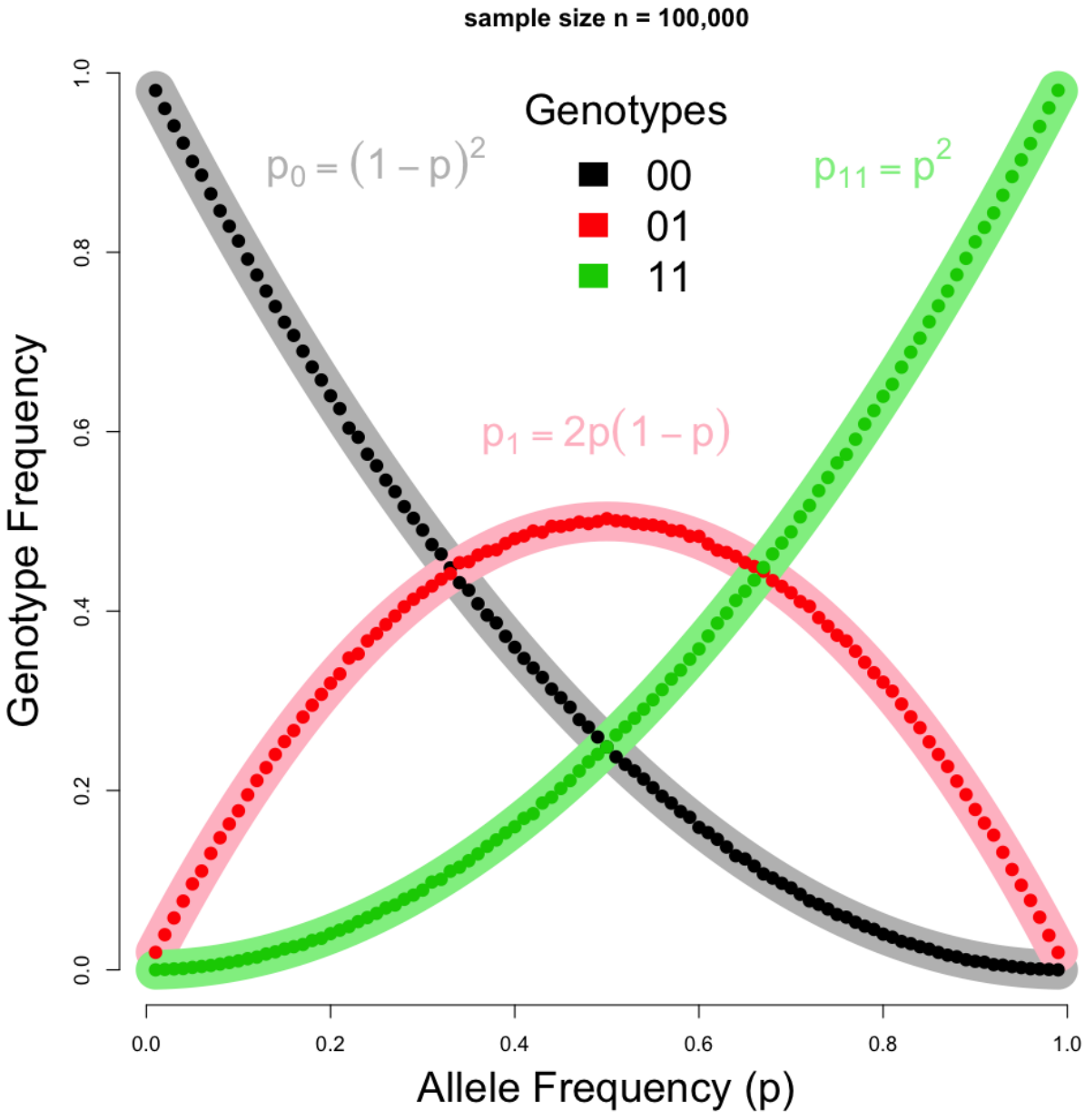

sample size n = 100
Genotype Frequency
0.0
0.2
0.4
0.6
0.8
1.0
Allele Frequency


sample size n = 1000
Genotype Frequency
0.0
0.2
0.4
0.6
0.8
1.0
Allele Frequency


sample size n = 10000
Genotype Frequency
0.0
0.2
0.4
0.6
0.8
1.0
Allele Frequency


sample size n = 1e+05
Genotype Frequency
0.0
0.2
0.4
0.6
0.8
1.0
Allele Frequency

# Prediction from the simulation study



# What assumptions underlie the simulation study?

## Assumptions

- Large sample size
- Independence between mates (random mating)
- Diploid individual
- Sexual reproduction
- allele frequencies are equal between sexes
- there is no migration, no mutation or selection

## Hardy-Weinberg Equilibrium (HWE)

Under aforementioned assumptions, we have

$$p_{AA} = p_A^2,\ p_{AB} = 2p_A(1 - p_A) = 2p_Ap_B \text{ and } p_{BB} = p_B^2.$$

# Testing HWE (1/3)

Deviation from HWE can be detected using a $\chi^2$ test with 1 degree of freedom.

Example: Diploid population with the following genotypes counts

| Observed | | | |
|---|---|---|---|
| AA | AB | BB | $N_{total}$ |
| 125 | 225 | 150 | 500 |

$p_A = (2 \times 125 + 225)/(2 \times 500) = 0.475 \Longrightarrow p_B = 0.525.$

$p_{AA} = 125/500 = 0.25$, $p_{AB} = 225/500 = 0.45$ and $p_{BB} = 150/500 = 0.3$.

Expectation under HWE: $E[n_{AA}] = p_A^2 \times N_{total} = 112.8125$.

# Testing HWE (2/3)

| Observed | | | | Expected | | |
|---|---|---|---|---|---|---|
| AA | AB | BB | $N_{total}$ | E[AA] | E[AB] | E[BB] |
| 125 | 225 | 150 | 500 | 112.8 | 249.4 | 137.8 |

Test Statistic

$$\chi^2 = \frac{(125-112.8)^2}{112.8} + \frac{(225-249.4)^2}{249.4} + \frac{(150-137.8)^2}{137.8}$$
$$= 4.78 > 3.84.$$

This example illustrates a **significant** deviation from HWE.

# Testing HWE (3/3)

## General form of the test statistic

$$\chi^2 = \frac{(n_{AA} - E[n_{AA}])^2}{E[n_{AA}]} + \frac{(n_{AB} - E[n_{AB}])^2}{E[n_{AB}]} + \frac{(n_{BB} - E[n_{BB}])^2}{E[n_{BB}]} \tag{1}$$

$$E[n_{AA}] = N_{total} \times p_A^2$$
$$E[n_{AB}] = N_{total} \times 2p_A p_B$$
$$E[n_{BB}] = N_{total} \times p_B^2.$$

with

$$p_A = (2n_{AA} + n_{AB}) / (2N_{total}) \text{ and } p_B = 1 - p_A.$$

# Summary

- Populations can be characterized by the distribution of alleles and genotypes.
- Under certain assumptions, genotype frequencies can be predicted from alleles frequencies (Hardy-Weinberg Equilibrium)
- HWE can be extended to sex-linked loci and polyploid individuals.
- Deviation from HWE can be used to inform mating types (e.g. inbreeding), population history or the quality of genetic data (see further lecture on Genome-Wide Association Studies - GWAS).

We briefly introduce in this section some of the forces that drive change in alleles frequencies in the population.

# Random fluctuation: allelic drift (S. Wright)

Let us consider a population of $2N$ individuals ($N$ males and $N$ females). At each generation we randomly form $N$ pairs of mates. Each pair to generate 2 progenies. How do allele frequencies change over time in that population?

Let's run some simulations...

# Random fluctuation: allelic drift (S. Wright)

Let us consider a population of $2N$ individuals ($N$ males and $N$ females). At each generation we randomly form $N$ pairs of mates. Each pair to generate 2 progenies. How do allele frequencies change over time in that population?

Let's run some simulations...

# Random fluctuation: allelic drift (S. Wright)

- In the absence of mutation, alleles frequencies tend to fixation ($p = 0$ or $p = 1$) as a consequence of random sampling of mates.
- The time to fixation depends on the allele frequencies in the founding population and of the size of that population.

# Other forces driving changes in alleles frequencies

- (Random) mutation
- Migration
- Selection (differential fertility and/or mortality)

# Meoisis and genetic linkage

- In sexual reproduction gametes are produced during specialized cell division called meiosis.
- Meiosis involves multiple phases (prophase, meiosis I and II) in which genetic information is exchanged between homolgous chromosomes: recombination.
- Recombinant chromosomes are then transmitted to the offspring.
- $\Longrightarrow$ Close DNA sequences on a chromosome will tend to be transmitted together: linkage disequilibrium (LD).

# Statistical measures of linkage disequilibrium (1/3)

Let us consider two loci $j$ and $k$ with alleles $a_j / A_j$ and $a_k / A_k$.

Linkage disequilibrium between alleles $a_j$ and $A_k$ (for example) is often measured as using the coefficient $D(a_j, A_k)$

$$D(a_j, A_k) = p(a_j A_k) - p(a_j)p(A_k)$$

where $p(a_j A_k)$ is the proportion of individuals in the population with both alleles $a_j$ and $A_k$; and $p(a_j)$ and $p(A_k)$ the proportion of individuals with allele $a_j$ and $A_k$ respectively.

# Statistical measures of linkage disequilibrium (2/3)

| | $a_j$ | $A_j$ | |
|---|---|---|---|
| $a_k$ | $p_j p_k + D_{jk}$ | $(1-p_j)p_k - D_{jk}$ | $p_k$ |
| $A_k$ | $(1-p_k)p_j - D_{jk}$ | $(1-p_j)(1-p_k) + D_{jk}$ | $(1-p_k)$ |
| | $p_j$ | $(1-p_j)$ | 1 |

Table: Joint distribution of allele frequencies, as a function of the linkage disequilibrium parameter $D_{jk}$. $a_j$ and $a_k$ are the minor alleles at locus $j$ and $k$, and $A_j$ and $A_k$ the corresponding major alleles respectively.

$D_{jk} > 0$ (positive LD) $\Longrightarrow$ alleles $a_j$ and $a_k$ or $A_j$ and $A_k$ are often "transmitted" (observed) together.

# Statistical measures of linkage disequilibrium (3/3)

Common measures of linkage disequilibrium (LD) are $D'_{jk}$

$$D'_{jk} = D_{jk} / D_{max} \tag{2}$$

and the squared correlation $r_{jk}$ between allele counts

$$r^2_{jk} = \frac{D^2_{jk}}{p_j(1-p_j)p_k(1-p_k)} \tag{3}$$

where

$$D_{max} = \begin{cases} min\left(p_j p_k, (1-p_j)(1-p_k)\right) & \text{when } D_{jk} < 0 \\ min\left(p_j(1-p_k), (1-p_j)p_k\right) & \text{when } D_{jk} > 0 \end{cases}$$

$D'$ takes values between -1 and 1 and $r^2$ between 0 and 1.

**LD depends on alleles frequencies and population size.**

# LD calculations (1/2)

| Allele counts | $a_j$ | $A_j$ | Total |
|---|---|---|---|
| $a_k$ | 1000 | 1500 | 2500 |
| $A_k$ | 1500 | 1000 | 2500 |
| | 2500 | 2500 | 5000 |

$p(a_j) = 0.5$, $p(a_k) = 0.5$, $p(a_j a_k) = 1000/5000 = 0.2$.

$D(a_j, a_k) = p(a_j a_k) - p(a_j)p(a_k) = 0.2 - 0.25 = -0.05$.

$D(a_j, A_k) = D(a_k, A_j) = 0.3 - 0.25 = +0.05$ and
$D(A_j, A_k) = -0.05$.

$$\Longrightarrow r^2(a_j, a_k) = r^2(A_j, A_k) = r^2(a_j, A_k) = r^2(A_j, a_k) = \frac{0.05^2}{0.5^4} = 0.04.$$

# LD calculations (1/2)

| Allele counts | $a_j$ | $A_j$ | Total |
|---|---|---|---|
| $a_k$ | 1000 | 1500 | 2500 |
| $A_k$ | 1500 | 1000 | 2500 |
| | 2500 | 2500 | 5000 |

$p(a_j) = 0.5$, $p(a_k) = 0.5$, $p(a_j a_k) = 1000/5000 = 0.2$.

$D(a_j, a_k) = p(a_j a_k) - p(a_j)p(a_k) = 0.2 - 0.25 = -0.05$.

$D(a_j, A_k) = D(a_k, A_j) = 0.3 - 0.25 = +0.05$ and
$D(A_j, A_k) = -0.05$.

$$\Longrightarrow r^2(a_j, a_k) = r^2(A_j, A_k) = r^2(a_j, A_k) = r^2(A_j, a_k) = \frac{0.05^2}{0.5^4} = 0.04.$$

# LD calculations (2/2)

| Allele frequency | $a_j$ | $A_j$ | |
|---|---|---|---|
| $a_k$ | $p(a_j a_k)$ | $p(A_j a_k)$ | $p(a_k)$ |
| $A_k$ | $p(a_j A_k)$ | $p(A_j A_k)$ | $p(A_k)$ |
| | $p(a_j)$ | $p(A_j)$ | 1 |

$$D(a_j, a_k) = p(a_j a_k)p(A_j A_k) - p(A_j a_k)p(a_j A_k) \quad (4)$$

$D(a_j, a_k)$ is the determinant of the matrix of allele frequencies.

# Change of LD over time

Let us consider two loci with alleles $A_1$ and $A_2$. We denote $r_n$ the frequency of $A_1A_2$ in the current generation and $c$ the probability of recombination between locus 1 and locus 2.
Under random mating the frequency of $A_1A_2$ in the next generation can be written

$$r_{t+1} = r_t(1-c) + cp(A_1)p(A_2)$$

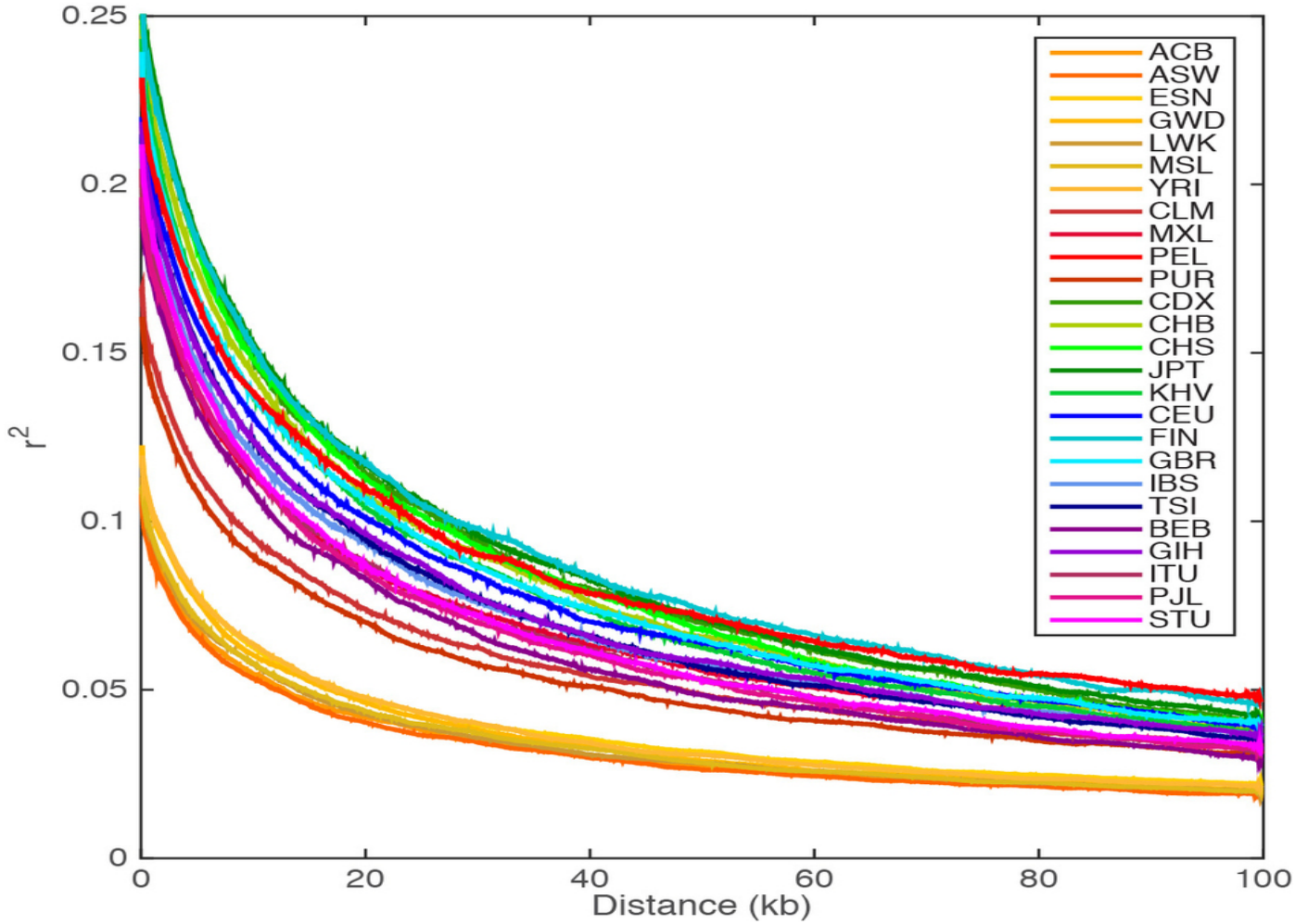
which in terms of of disequilibrium can be expressed as

$$\begin{aligned} D_{t+1}(A_1, A_2) &= r_{t+1} - p(A_1)p(A_2) \\ &= (r_t - p(A_1)p(A_2))(1-c) \\ &= D_t(A_1, A_2)(1-c) \end{aligned}$$

This can be generalized as

$$D_t(A_1, A_2) = D_0(A_1, A_2)(1-c)^t \tag{5}$$

# Example of LD decay with physical distance




From A. Auton et al. (1000 Genomes Consortium) *A Global Reference for Human Genetic Variation*. Nature (2015).

# Summary

- Linkage disequilibrium (often) derives from the block property of genetic recombination during meiosis.
- LD is a property of a population that depends on its size, on alleles frequencies and may change over time.
- LD can be measured using coefficients such as $r^2$ or $D'$.
- LD between two loci declines with physical distance and with recombination rate between them.

We now illustrate through small examples how the concepts introduced so far can be used for

- Comparing populations
- Inferring population history
- Compress genetic information

# Comparing populations with alleles frequencies

Differences in allele frequencies are often used to compare genetic constitutions of populations.

In particular, the fixation index $F_{ST}$ [Wright (1921)] is used to measure genetic distance between two populations.

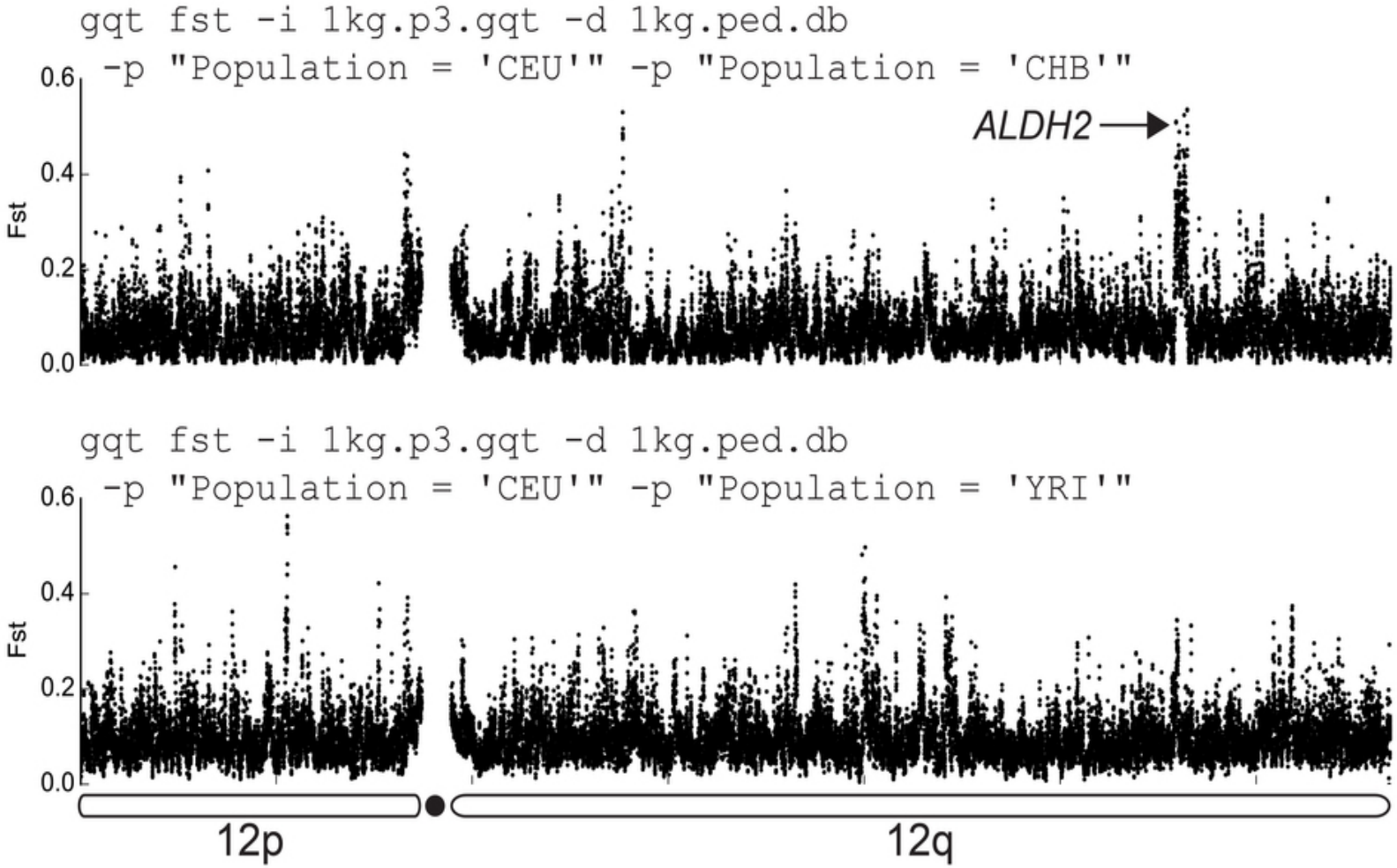
## Definition

Let us denote $p_1$, $p_2$ and $p_{12}$ the frequency of a given allele in Population 1, Population 2 and in the combined Population 1+2 respectively. We also denote $w_1$ and $w_2$ the relative (to the combined population) sizes of Populations 1 and 2.
We define the fixation index $F_{ST}$ at that locus as

$$F_{ST} = 1 - \frac{w_1 p_1 (1 - p_1) + w_2 p_2 (1 - p_2)}{p_{12}(1 - p_{12})} \tag{6}$$

Large $F_{ST} \iff$ Strong differentiation between populations.

# Example of $F_{ST}$ on Human chromosome 12




$F_{ST}$ for 1,000 Genomes phase 3 of Europeans versus East Asians and Europeans versus Africans on chromosome 12.

Layer R.M. et al. *Efficient genotype compression and analysis of large genetic-variation data sets*. *Nature Methods* (2016).

# Application of LD

## Population sizes

LD decays with the number of meioses across time. We can infer population sizes by comparing features of LD in different populations.

A known relationship often used is

$$E[r^2] \approx 1/(\alpha + 4N_e c) + 1/n \text{ [Tenesa et al. (2007)]} \quad (7)$$

## GWAS

LD also measures redundancy in the genome. This property has allowed the avent of the GWAS era, years before sequencing become affordable. Few SNPs can be used to represent the $\sim 3 \times 10^9$ base pairs in the genome

# Summary

Key parameters to describe the genetic constitution of a population are

- Alleles and genotypes frequencies
- Correlations between alleles: linkage disequilibrium (LD)

We discussed the key concept of **Hardy-Weinberg Equilibrium** and its use, and illustrate some applications of **linkage disequilibrium**.

# References

## Textbooks

Falconer, S.D. and Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*, 4$^{th}$ Edition.

Lynch, M. and Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer. Bulmer, M. (1980) *The Mathematical Theory of Quantitative Genetics*. Clarendon Press.

## Article(s)

Auton, A. et al. (2015) *A Global Reference of Human Genetic Variation*, *Nature*.

Tenesa, A. et al. (2007) *Recent human effective population size estimated from linkage disequilibrium*, *Genome Research*.