Key concepts in Population & Quantitative Genetics Theory

Loic Yengo, PhD

Institute for Molecular Bioscience

The University of Queensland

l.yengo@imb.uq.edu.au

Two parts…

- Part I: Hardy-Weinberg Equilibrium and Linkage Disequilibrium

- Part II: Heritability

Part I: Population Genetics

Population Genetics Theory is concerned with characterizing and quantifying **genetic variation** within and between **populations**.

What is a Population?

"All inhabitants of particular place" – Oxford dictionary

[BIOLOGY] – "A community of animals, plants, or humans among whose members interbreeding occurs" – Oxford dictionary

"Any complete group with at least one characteristic in common" – Australian Bureau of Statistics

"All the inhabitants of a country, territory, or geographic area, total or for a given sex and/or age group, at a specific point of time. In demographic terms it is the total number of inhabitants of a given sex and/or age group that actually live within the border limits of the country, territory, or geographic area at a specific point of time, usually mid-year. The mid-year population refers to the actual population at July 1st." – World Health Organization

"All inhabitants of **particular place**" – Oxford dictionary

[BIOLOGY] – "A community of animals, plants, or humans among whose members ***interbreeding*** occurs" – Oxford dictionary

"Any complete group with at least one characteristic in common" – Australian Bureau of Statistics

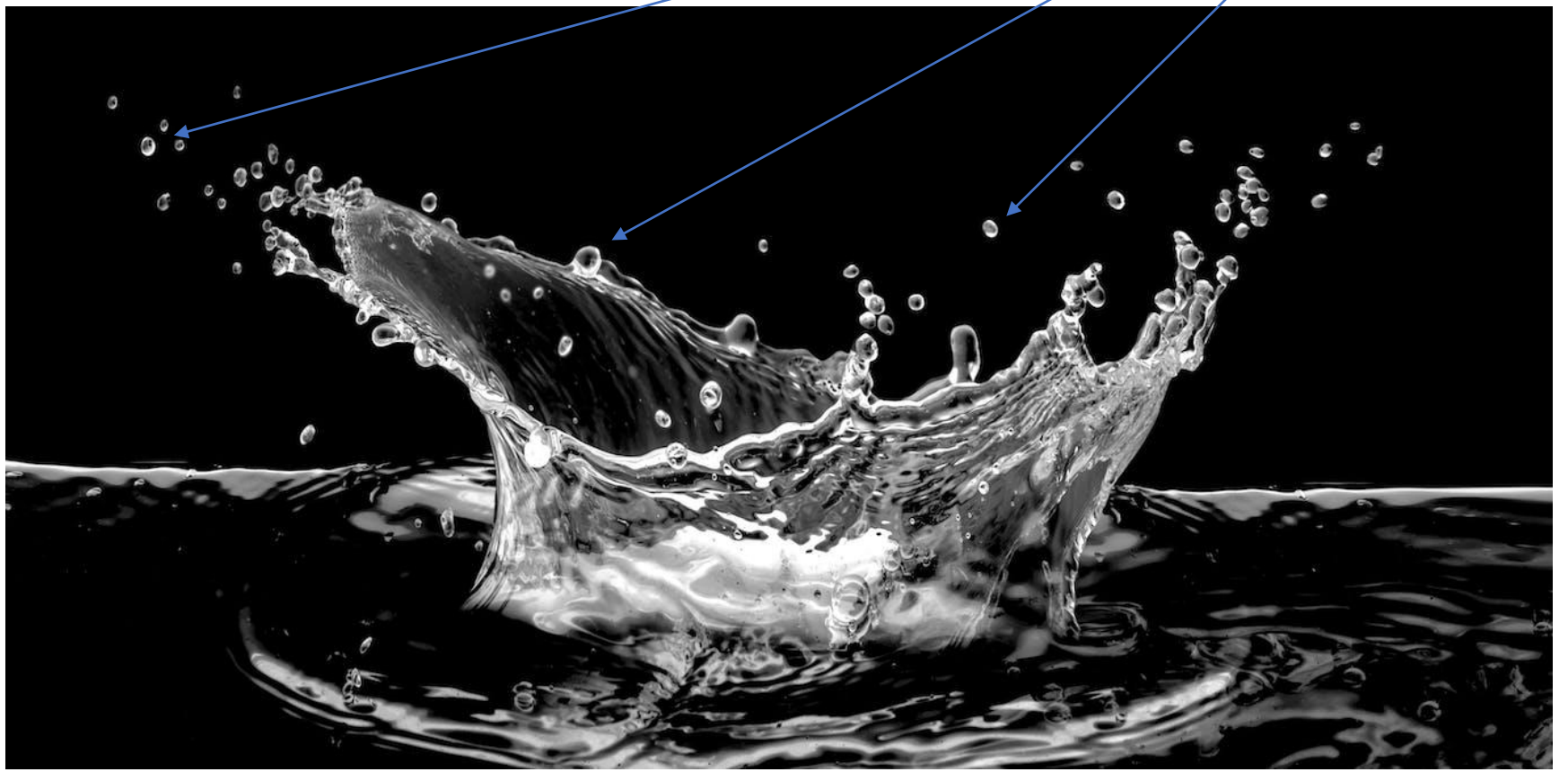
"All the inhabitants of a **country, territory, or geographic area**, total or for a given sex and/or age group, at a specific point of time. In demographic terms it is the total number of inhabitants of a given sex and/or age group that actually live within the border limits of the country, territory, or geographic area **at a specific point of time**, usually mid-year. The mid-year population refers to the actual population at July 1st." – *World Health Organization*

Source: https://www.shutterstock.com/image-photo/water-splashwater-splash-isolated-on-backgroundwater-573666874?src=sv-vkrkfiyUnpA-vYq4Oog-1-16

A population is a concrete abstraction…

1. Population = Group of individuals
2. Who is *inside* or *outside* is arbitrary, i.e., depends on your research question

Are these drops from the same population?



Time

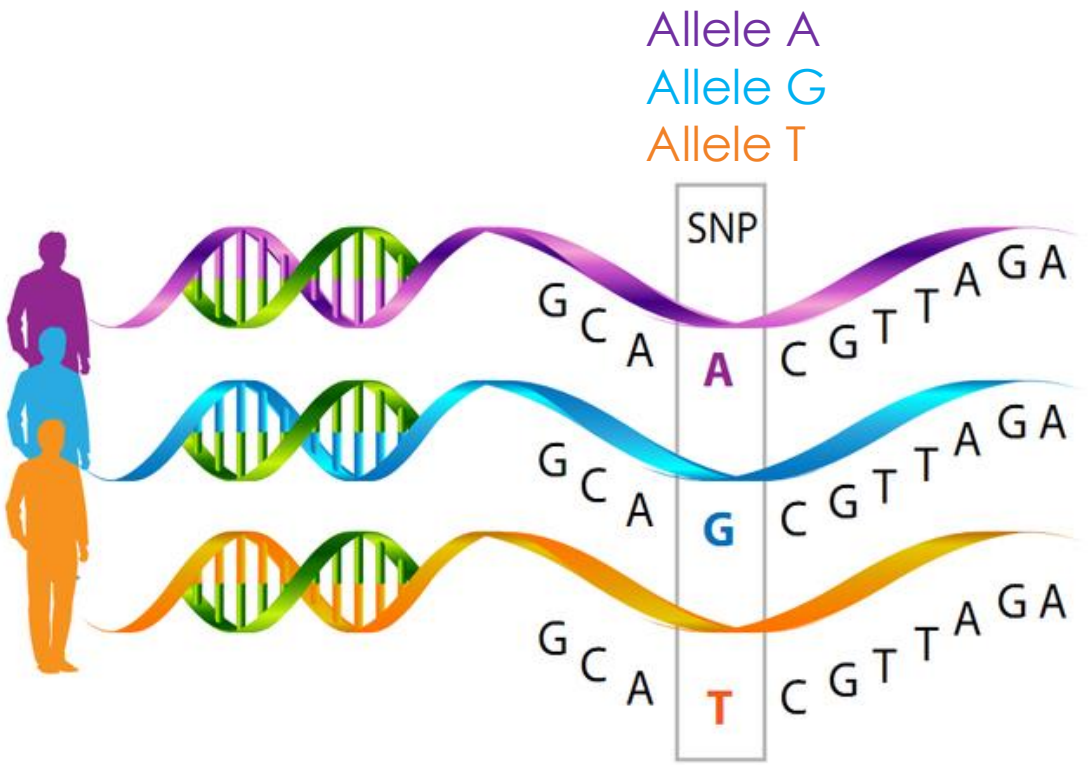
Space

What is Genetic Variation?

Genetic variation…

There is a near infinite number of ways to measure a distance between DNA sequences of two individuals

(1) Number of repeats
(2) Inversions
(3) Deletions
(4) Single letter (nucleotide) changes
(5) …

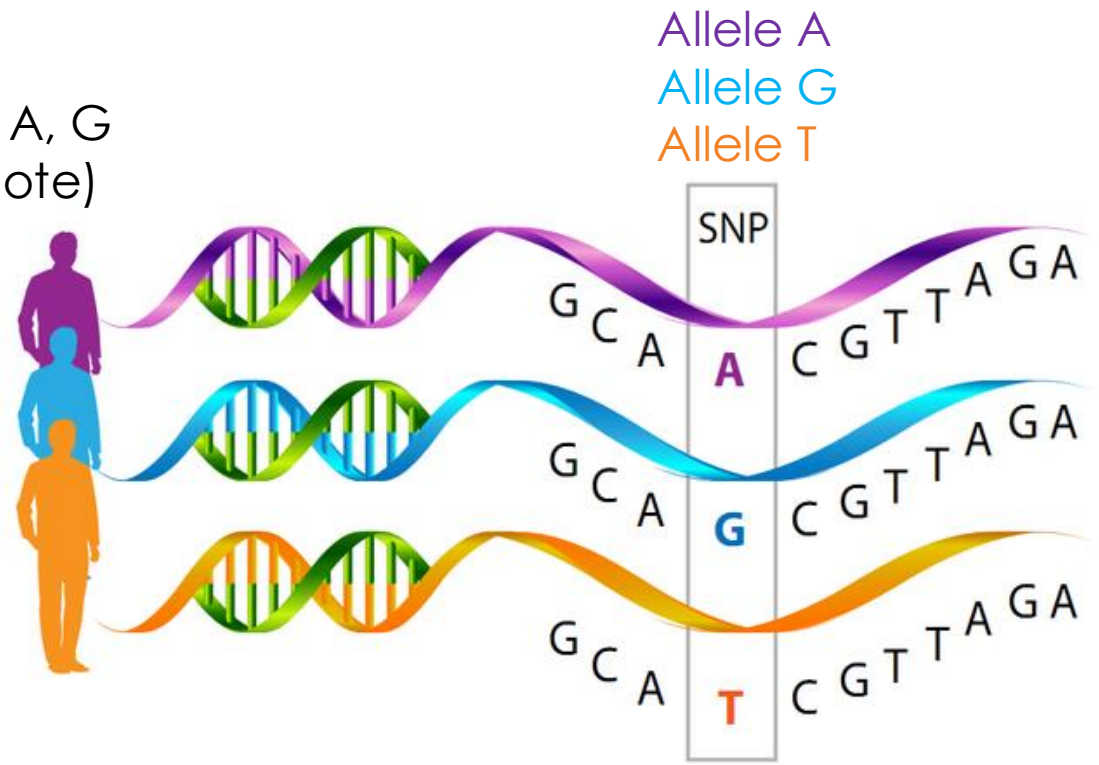


Single Nucleotide Polymorphisms

Definitions

Allele = possible state of the DNA sequence at a given locus.
Genotype = defined by the states of all alleles at the locus.

Examples: In a diploid individual. If there are three alleles A, G and T, then possible genotypes are AA, GG, TT (homozygote) and AG, AT, GT (heterozygote)



Single Nucleotide Polymorphisms

Population Genetics Theory is the theory of **alleles and genotypes frequencies** within and between **groups of individuals**.

Outline

- Hardy-Weinberg Equilibrium
- What drives changes in allele/genotype frequencies?
- How can you measure genetic distance between populations?
- Linkage Disequilibrium

Learn more: **https://www.colorado.edu/ibg/isg-workshop/isgw-online-resources/2-introduction-population-genetics**

Hardy-Weinberg Equilibrium

G. H. Hardy (1877 – 1947)
W. Weinberg (1862 - 1937)

Relationship between alleles and genotypes frequencies

Let's consider a population of N diploid individuals at a particular locus with two alleles 0 and 1.

We denote n_{00}, n_{01} and n_{11} the genotypes counts and n_0 and n_1 the allele counts in the population. So, $n_{00} + n_{01} + n_{11} = N$.

We have the following relationships:

=> $n_0 = 2n_{00} + n_{01}$ and $n_1 = 2n_{11} + n_{01}$.

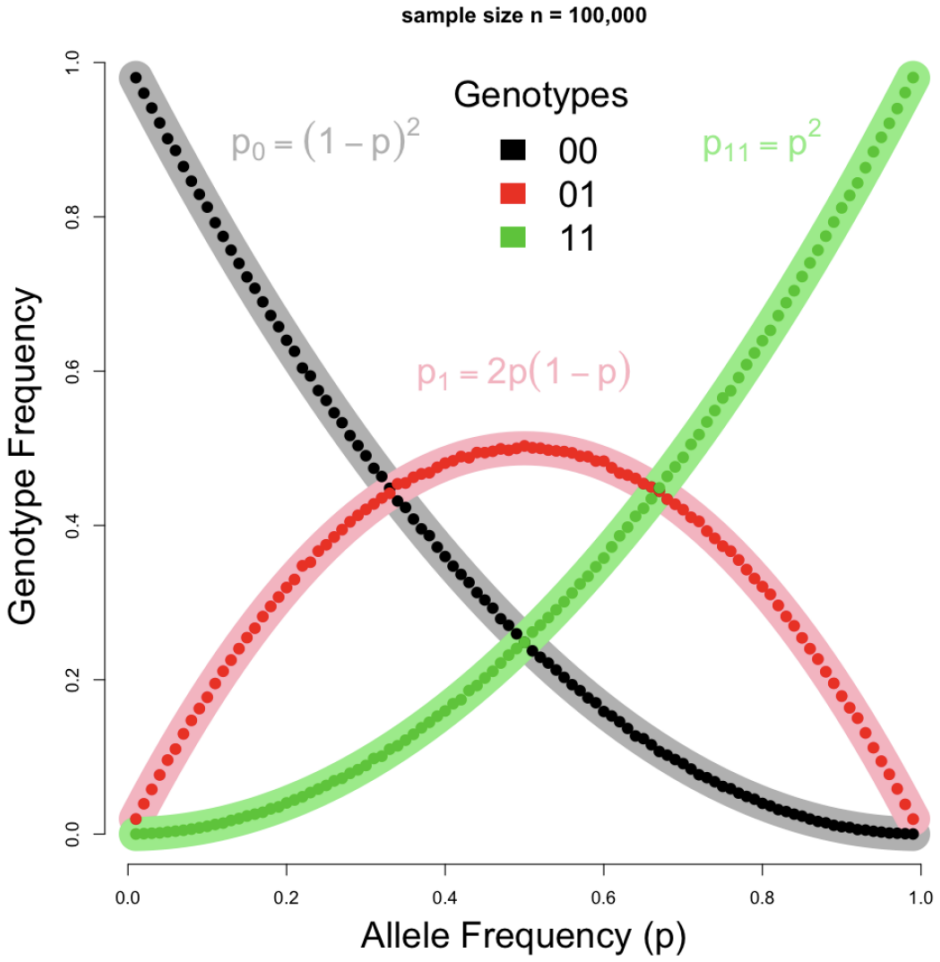
=> $p_0 = n_0 /(2N) = (n_{00} /N) + 0.5(n_{01} /N)$ and $p_1 = 1- p_0$.

$\Rightarrow$ $p_0 = p_{00} + 0.5p_{01}$ and $p_1 = 1- p_0$.

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

If genotypes and alleles frequencies are **constant from one generation to the next**, then the population is said to be under Hardy-Weinberg Equilibrium.

If genotypes and alleles frequencies are **constant from one generation to the next**, then the population is said to be under Hardy-Weinberg Equilibrium*.



*This is the diploid / autosomal version of the HWE.

Under which assumption(s) does HWE holds?

Constant allele frequency:

- no migration,
- no mutation,
- no natural selection

Random mating (diploid individuals, sexual reproduction, allele frequencies are the same between sexes)

Large population size

Testing HWE (1/3)

Deviation from HWE can be detected using a χ^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

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

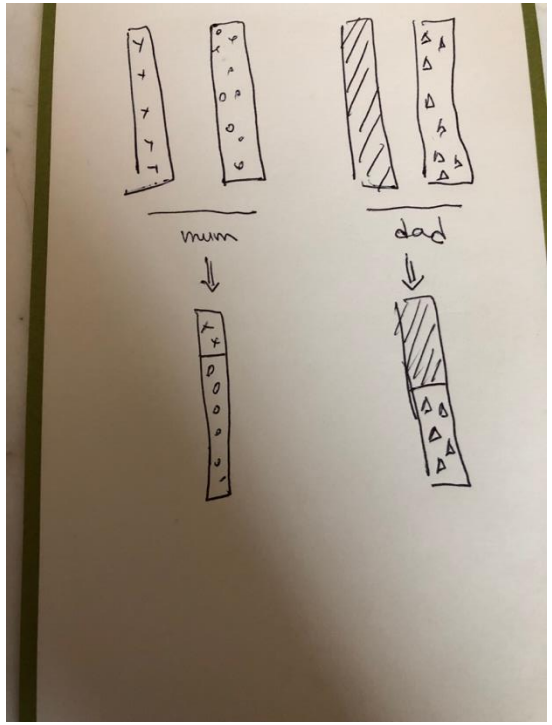
Summary

- Population can be characterized by the frequency distribution of alleles and genotypes
- Under certain assumptions, genotype frequencies can be predicted from allele frequencies (Hardy-Weinberg Equilibrium)
- HWE can be extended to sex-linked loci and polyploid individuals
- Deviation from HWE can be used to inform non-random mating (e.g., inbreeding), population history or quality of genetic data (see GWAS lectures)

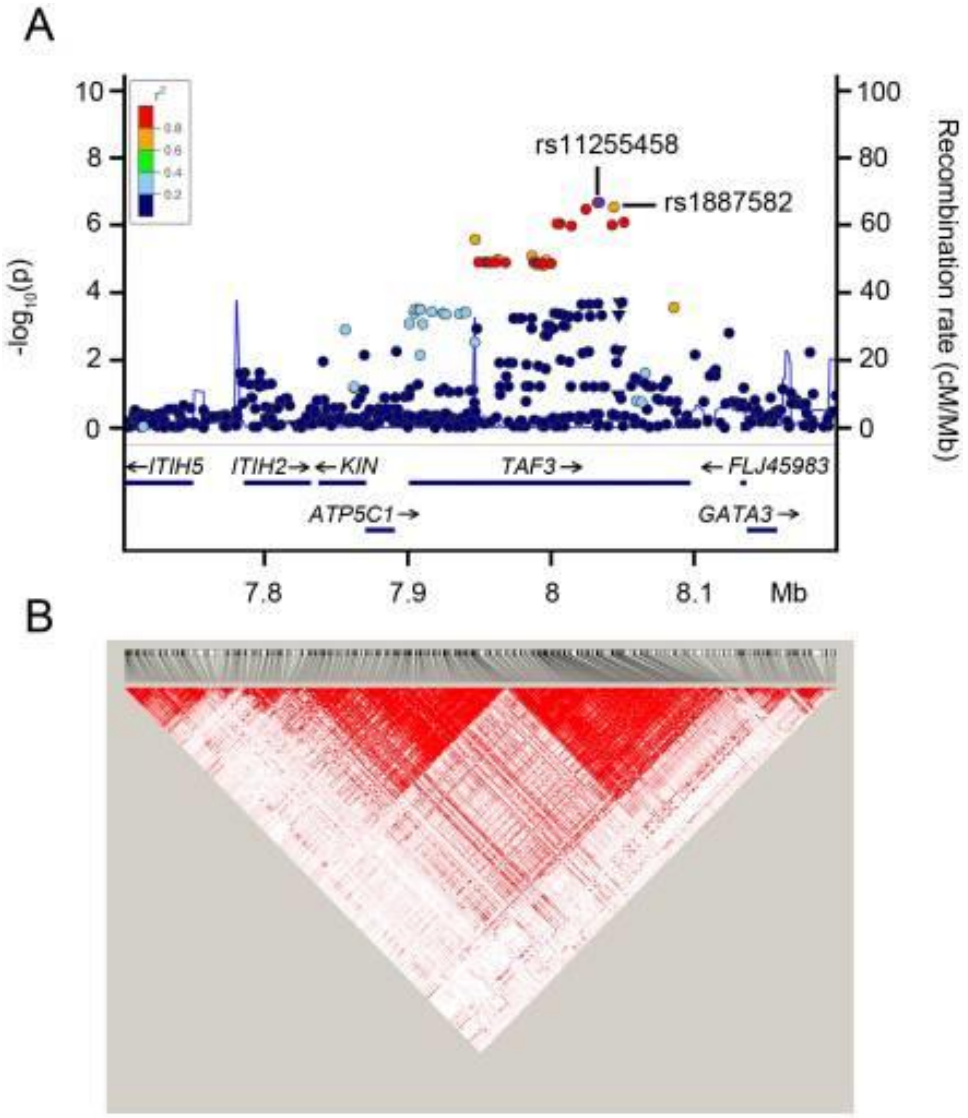
Linkage Disequilibrium

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 homologous chromosomes: recombination.
- Recombinant chromosomes are then transmitted to the offspring.
- => Close DNA sequences on a chromosome will tend to be transmitted together: linkage disequilibrium (LD).

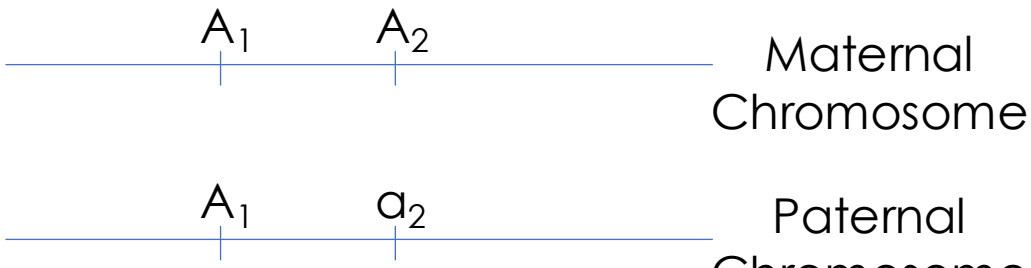


Haplotype



Pistis et al. (2013) Plos One

Haplotype = Group of alleles within a chromosome (e.g., within a LD block)



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 calculation

Allele counts	a_j	A_j	Total
a_k	1000	1500	2500
A_k	1500	1000	2500
	2500	2500	5000

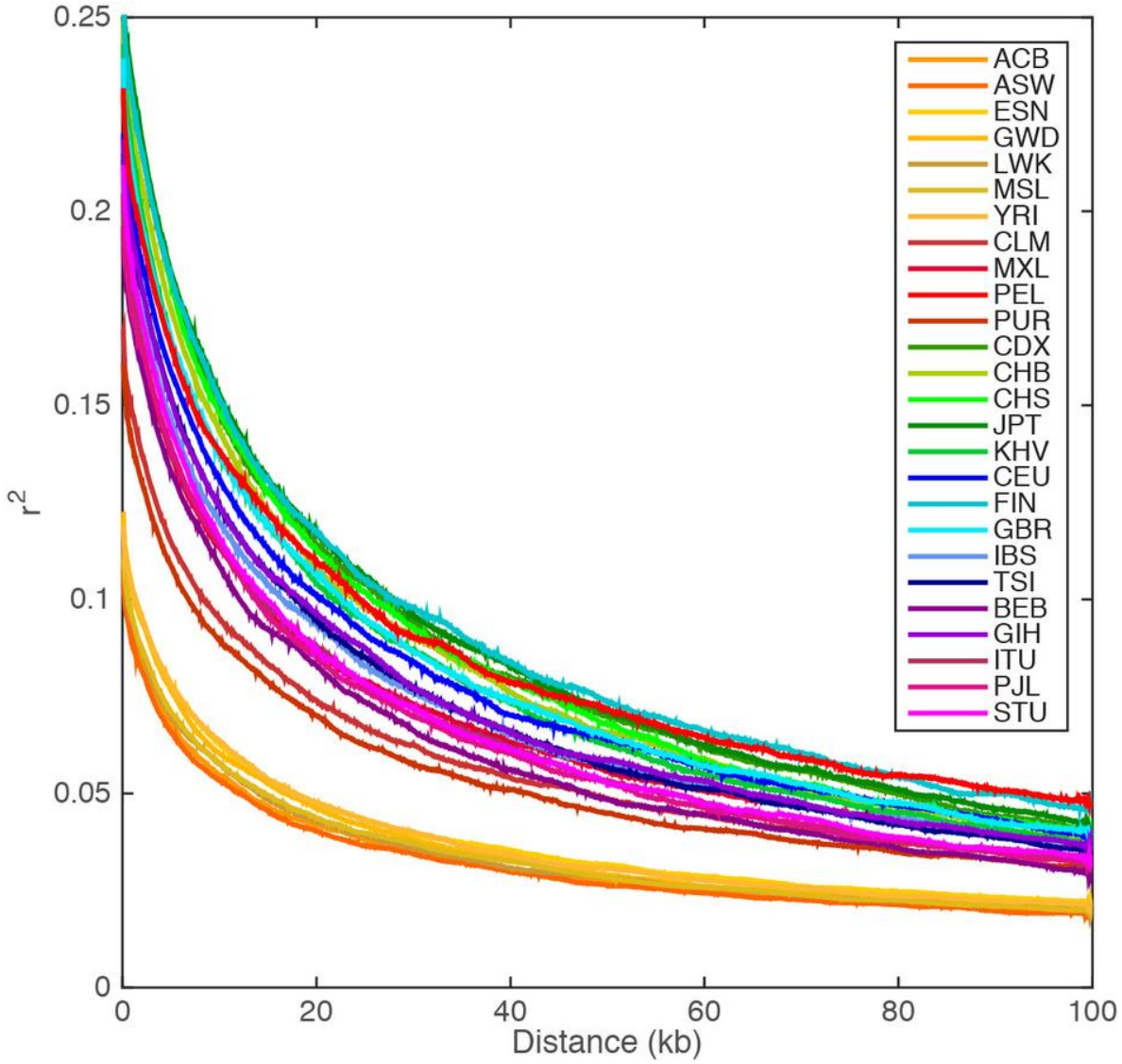
$p(a_j) = 0.5$, $p(a_k) = 0.5$, $p(a_j a_k) = 1000/5000 = 0.2$.

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

$D(a_j, A_k) = D(a_k, A_j) = 0.3 - 0.25 = +0.05$ 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 decay with physical distance



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

Part II: Quantitative Genetics

Quantitative Genetics Theory is concerned with characterizing and quantifying the **genetic causes** of continuously varying traits (e.g., height and weight) within and between **populations**.

Definitions

Heritability (h^2) quantifies the degree to which inter-individual differences and resemblance in the population are due to genetic factors.

Chial, H. (2008) Polygenic inheritance and gene mapping. Nature Education 1(1):17

Definitions

If the value, Y, of trait (=phenotype) can be modelled as

$$Y = G + E$$

Genetic factors

Non-genetic factors

Chial, H. (2008) Polygenic inheritance and gene mapping. Nature Education 1(1):17

then h^2 = var(G) / var(Y), i.e. proportion of trait variance explained by genetic factors.

Nice definition but not very useful unless we can observe G!

Definitions

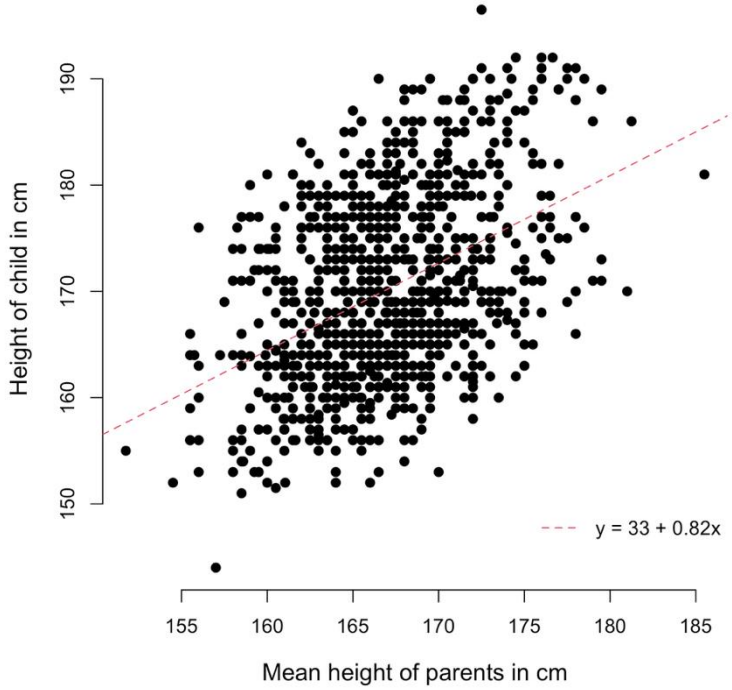
Another view of h^2 is the given by the phenotypic correlation between relatives.

E.g.,

- How much is the height of siblings correlated?
- How much having a family history of schizophrenia predisposes you to also develop it?

Quantitative Genetics theory answers those questions using the following Equation

$$corr(Y_i, Y_j) = h^2 R_{ij} + Residual$$



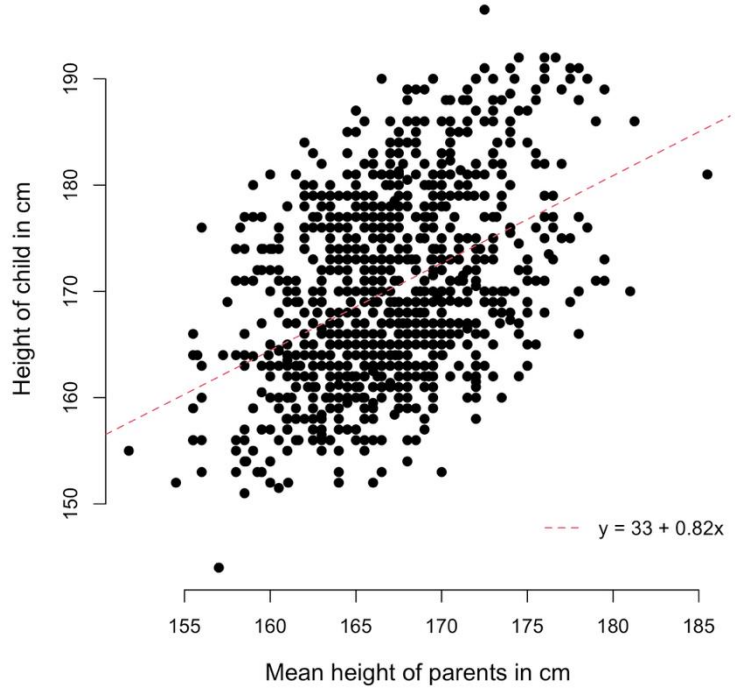
Data from UK Biobank participants (Application number 12505)

Definitions

Quantitative Genetics theory answers those questions using the following Equation

$$corr(Y_i, Y_j) = h^2 R_{ij} + Residual$$

where R_{ij} is the **coefficient of genetic relationship** between individual i and individual j (e.g., R_{ij}=0.5 for full siblings).



Data from UK Biobank participants (Application number 12505)

this definition is a bit more useful As we can observe both $corr(Y_i, Y_j)$ ***and*** R_{ij} **(Part 2).**

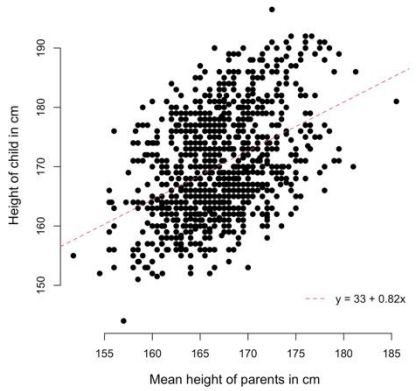
Definitions

Heritability (h^2) quantifies the degree to which inter-individual differences and resemblance in the population are due to genetic factors.

Heritability can be approached in terms of

- Differences between people in the population: h^2 = var(G) / var(Y),

- Resemblance between relatives (in families): corr(Y_i,Y_j) = h^2R_{ij} + Residual



Is heritability a universal constant?

No!

Heritability is a property of a trait, in a population, at a given time.

Heritability can change over time (e.g., Rimfeld et al. NHB 2018, heritability of educational attainment in Estonia before/after the cold war).

Larger is not necessarily better.

Genetic correlation

The genetic correlation (r_g) between two traits Y_1 and Y_2 can be defined similarly as the heritability

Population-based definition

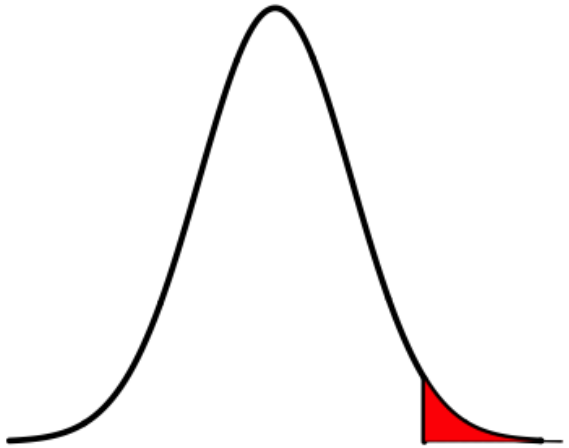
$$Y_1 = G_1 + E_1$$
$$Y_2 = G_2 + E_2$$
$$\Rightarrow r_g = \text{corr}(G_1, G_2)$$

(Co)variance-based definition

$$\text{corr}(Y_{i1}, Y_{j2}) = r_g \sqrt{h_1^2 h_2^2} R_{ij} + \text{Residual}$$

Heritability of binary traits

A binary trait (e.g., a disease) can be modelled as an extreme form of a certain liability (liability threshold model).



Obesity = BMI > 30 kg/m^2

Type 2 diabetes = Fasting glucose > 7mmol/L.

Therefore, heritability is well defined as the **heritability of the continuous liability**.

Falconer (1965): The inheritance of liability to certain diseases, estimated form incidence among relatives.

A few applications

1) The heritability of a trait gives an upper bound for the accuracy of genetic predictors of that trait.

2) The heritability predicts the response to (natural) selection.

3) The heritability predicts an individual's risk to develop a certain disease knowing they have affected relatives.

4) The heritability influences the statistical power of genome-wide association studies (GWAS)

Examples of ongoing research around heritability

- Computational efficiency (biobank scale methods)
- Incorporate biological information (functional annotation)
- Estimation of non-additive genetic variance (e.g., dominance)
- Estimation of SNP-based heritability using whole-genome sequence data (integrating rare and ultra-rare variants)
- The impact of assortative mating on heritability estimates