Central concepts in population genetics

PaNomics Summer School 2022

Marta Riise Moksnes

K.G. Jebsen Center for Genetic Epidemiology

Department of Public Health and Nursing,

NTNU, Norwegian University of Science and Technology

NTNU
K.G. Jebsen Center for Genetic Epidemiology

STIFTELSEN KRISTIAN GERHARD JEBSEN

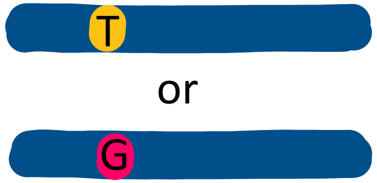
Slides revised from prof. Frank Skorpen at NTNU

Overview

- (Short repetition)
- The Hardy-Weinberg equilibrium (HWE)
- Linkage disequilibrium (LD)
- Heritability, H^2
- Identity by state & Identity by descent

Alleles, genotypes and phenotypes

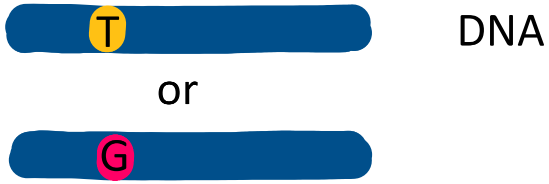
Alleles: The genetic variants at a locus that exist in a population



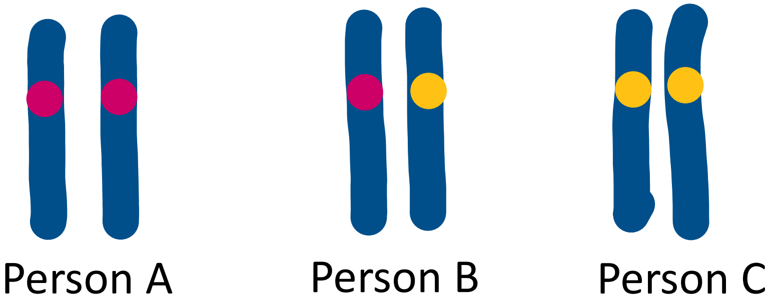
DNA

Alleles, genotypes and phenotypes

Alleles: The genetic variants at a locus that exist in a population

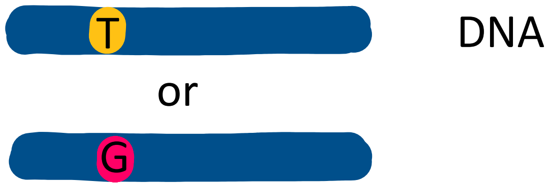


Genotype: The set of two alleles that an individual carries at a locus

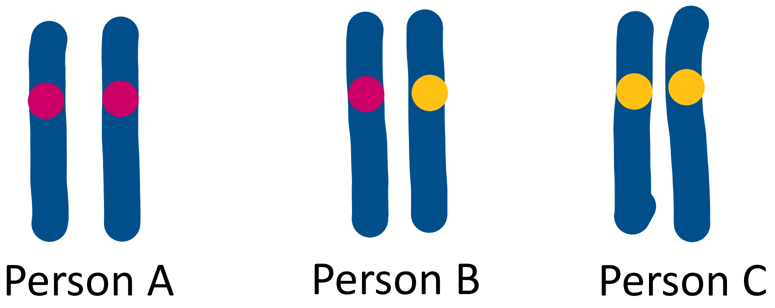


Alleles, genotypes and phenotypes

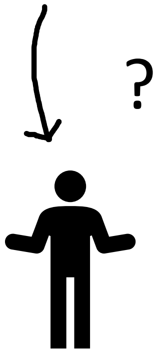
Alleles: The genetic variants at a locus that exist in a population



Genotype: The set of two alleles that an individual carries at a locus

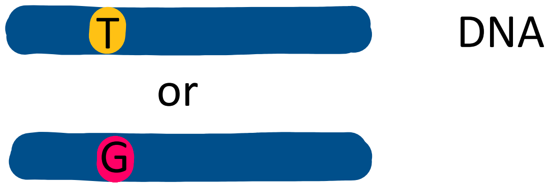


Phenotype: The observable characteristics (e.g. disease) of the individual.

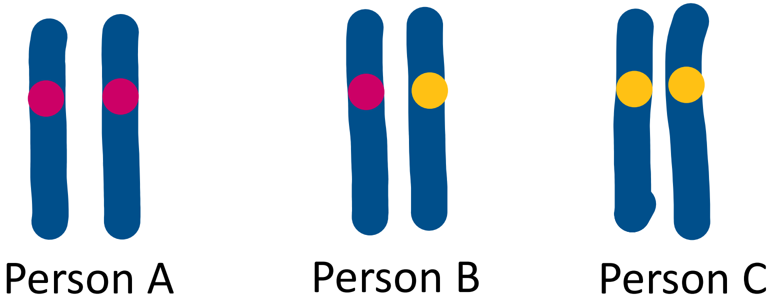


Alleles, genotypes and phenotypes

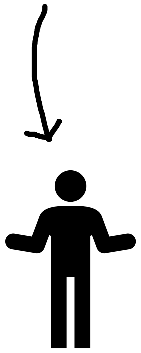
Alleles: The genetic variants at a locus that exist in a population



Genotype: The set of two alleles that an individual carries at a locus



Phenotype: The observable characteristics (e.g. disease) of the individual.



Type of trait: single gene or complex?

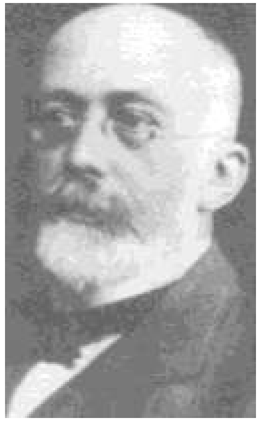
Inheritance pattern:
autosomal/X-linked/mitochondrial?
dominant/recessive (single gene traits)?

Overview

- The Hardy-Weinberg equilibrium (HWE)
- Linkage disequilibrium (LD)
- Heritability, H^2
- Identity by state & Identity by descent

The Hardy-Weinberg equilibrium (HWE)

Godfrey Harold Hardy, England, 1877- 1947, Mathematician



Wilhelm Weinberg Germany, 1862-1937 Physician

- In 1908, Godfrey Hardy and Wilhelm Weinberg independently pointed out that, under certain conditions, one could easily predict genotype frequencies from allele frequencies (or vice versa) in human populations

The Hardy-Weinberg equilibrium (HWE)

Autosomes: All chromosomes that are not sex chromosomes (not X or Y)

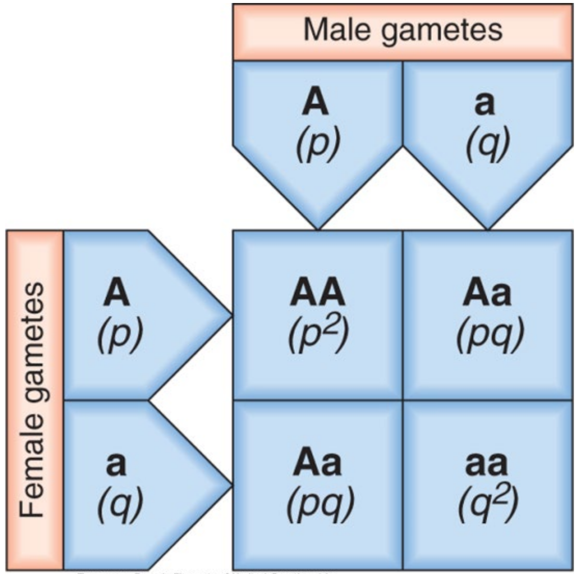
Assume an «ideal» population:

- The population is (infinitely) large
- Random mating with regard to the locus in question
- Allele frequencies remain constant over time (i.e. no new mutations, no selection against any genotypes, no migration)

Model: Autosomal locus with 2 alleles, A and a

- Assume that alleles combine into genotypes randomly
- Allele frequencies: *p* (A) and *q* (a)

Allele frequencies: $p + q = 1$
Genotype frequencies: $p^2 + 2pq + q^2 = 1$



	Male gametes	
Female gametes	**A** (p)	**a** (q)
A (p)	**AA** (p^2)	**Aa** (pq)
a (q)	**Aa** (pq)	**aa** (q^2)

Turnpenny: Emery's Elements of Medical Genetics, 14e
Copyright © 2011 by Churchill Livingstone, an imprint of Elsevier Ltd. All rights reserved.

The Hardy-Weinberg equilibrium (HWE)

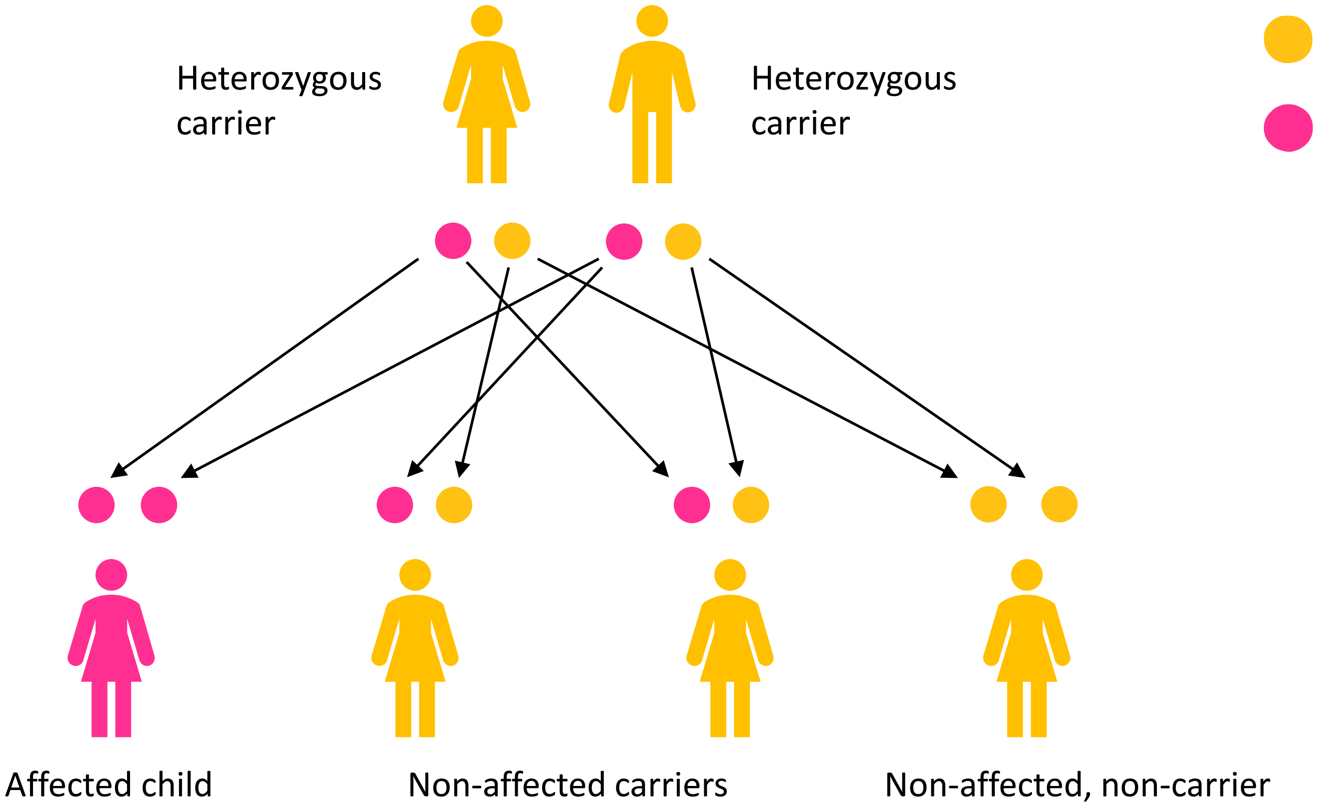
- The HWE does not specify any particular values for *p* and *q*.
- As long as allele frequencies do not change, and there is random mating, the genotype frequencies will remain in the same relative proportions, $\boldsymbol{p^2 : 2pq : q^2}$, from generation to generation.

Real populations may deviate more or less from such an “ideal” population

Some applications of the HWE

- Estimating allele frequencies from disease prevalence data
- Calculating carrier frequency from autosomal and X-linked recessive disease frequencies
- Checking whether observations in a population answer HWE expectations (and if they don’t – why not?)
- Checking whether genotype distributions in GWA studies answer HWE expectations
- Genetic counceling for autosomal recessive disorders

HWE in Autosomal Recessive Disorders



Heterozygous carrier

Heterozygous carrier

Affected child

Non-affected carriers

Non-affected, non-carrier

● Dominant non-disease allele (working gene)

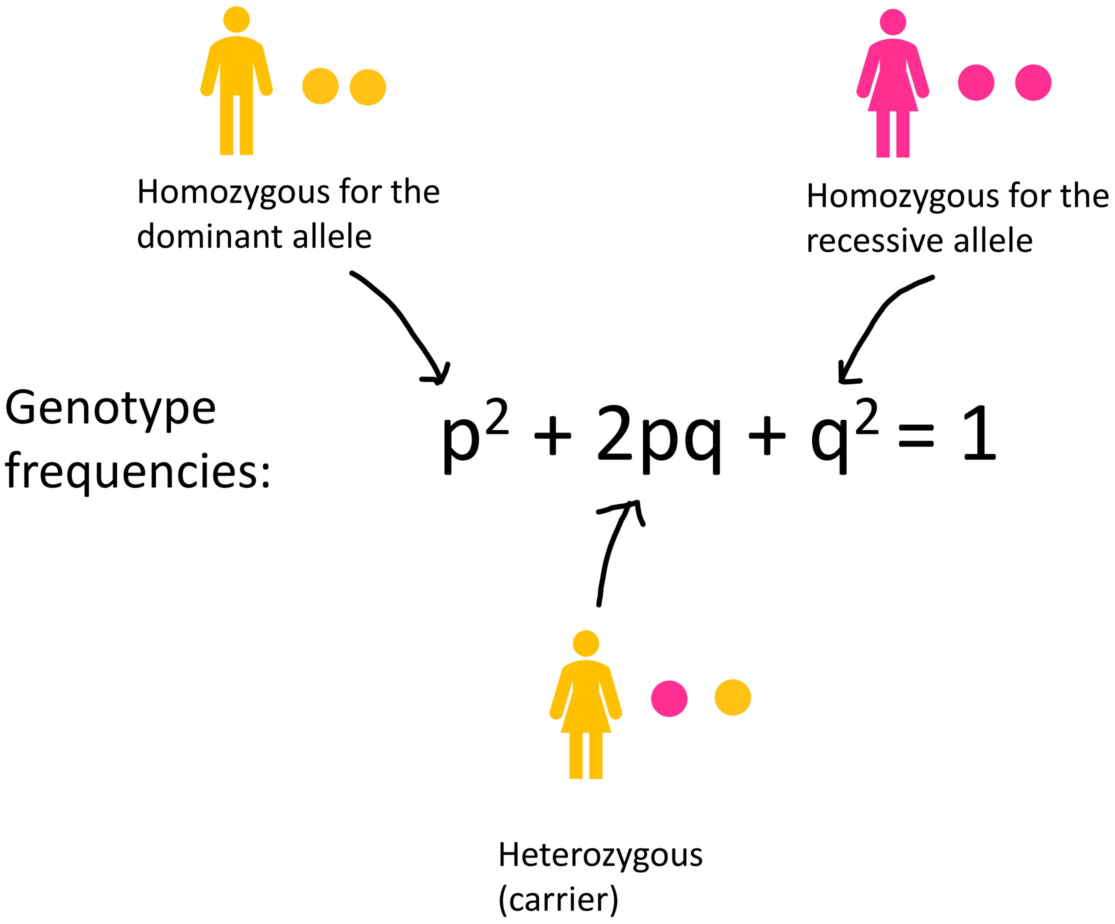
● Recessive disease allele (non-working gene)

Autosomal recessive disorders:

Affected individuals have inherited two copies of the disease allele.

Heterozygous individuals are usually unaffected

Calculating carrier frequency from autosomal recessive disease frequencies



Homozygous for the dominant allele

Homozygous for the recessive allele

Genotype frequencies:

$$p^2 + 2pq + q^2 = 1$$

Heterozygous (carrier)

Among 1000, two individuals are observed with a recessive disease.

How many are carriers?

$q^2 = 2/1000 = 0.002$

$q = \sqrt{0.002} = 0.045$

$p = 1 - q = 1 - 0.045 = 0.955$

Carriers = $2pq$ = 2 x 0.955 x 0.045 = 0.086

Number of carriers: 1000 x 0.086 = **<u>86</u>**

HWE as quality control in GWA studies

- In GWAS, it is generally assumed that deviations from HWE are the result of genotyping errors.
- Note: The violation of the HWE law in cases can be indicative of a true genetic association with disease risk

SNP

Genotype	AA	AG	GG	Total
Number obs.	360	470	230	1060

Are the genotype frequencies in our sample in HWE?

Calculating *expected* genotype distribution

Genotype	AA	AG	GG	Total
Number obs.	360	470	230	1060
Frequency exp.	p^2	$2pq$	q^2	1
Number exp.	334	522	204	1060

Allele frequencies:

A: p = [(360 x 2) + 470] / 2120 = 0.56
G: q = [(230 x 2) + 470] / 2120 = 0.44

Number of individuals: N = 1060

Expected genotype distribution:

AA : p^2 x N = 0.56^2 x 1060 = 334
AG : $2pq$ x N = 2 (0.56 x 0.44) x 1060 = 522
GG : q^2 x N = 0.44^2 x 1060 = 204

Test for deviation from HWE (usually a χ^2 test)

Genotype	AA	AG	GG	Total
Number obs.	360	470	230	1060
Frequency exp.	p^2	$2pq$	q^2	1
Number exp.	334	522	204	1060
Deviation	26	-52	26	
$(x_i - m_i)^2/m_i$	2.02	5.18	3.31	χ^2 = **10.51**

x: observed number

m: expected number

H0: There is no difference in distribution
H1: There is a difference between the distributions

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} = 10.51$$

Test for deviation from HWE (usually a χ^2 test)

Genotype	AA	AG	GG	Total
Number obs.	360	470	230	1060
Frequency exp.	p^2	$2pq$	q^2	1
Number exp.	334	522	204	1060
Deviation	26	-52	26	
$(x_i - m_i)^2/m_i$	2.02	5.18	3.31	χ^2 = **10.51**

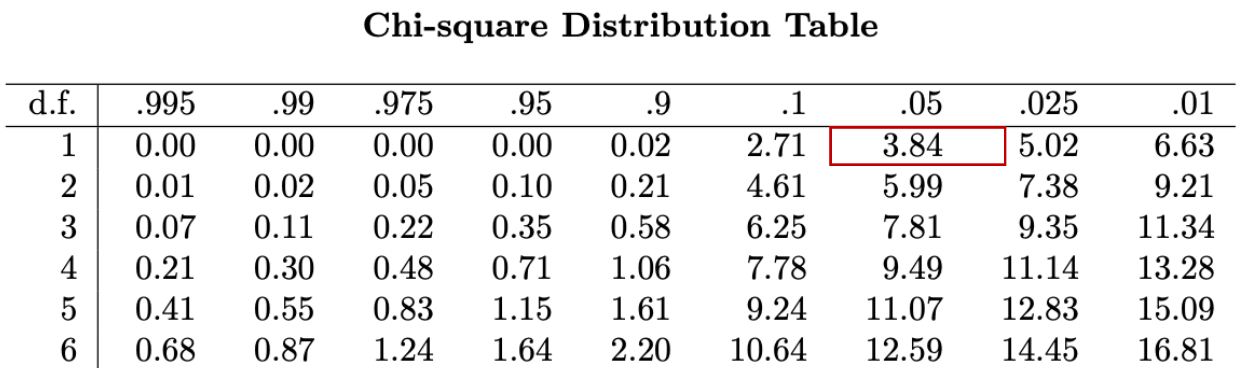
x: observed number

m: expected number

~~H0: There is no difference in distribution~~
H1: There is a difference between the distributions

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} = 10.51$$

$\chi^2 > 3.84$, 1 d.f. (**α** = 0.05)



Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81

Overview

- The Hardy-Weinberg equilibrium (HWE)
- Linkage disequilibrium (LD)
- Heritability, H^2
- Identity by state & Identity by descent

Linkage disequilibrium (LD)

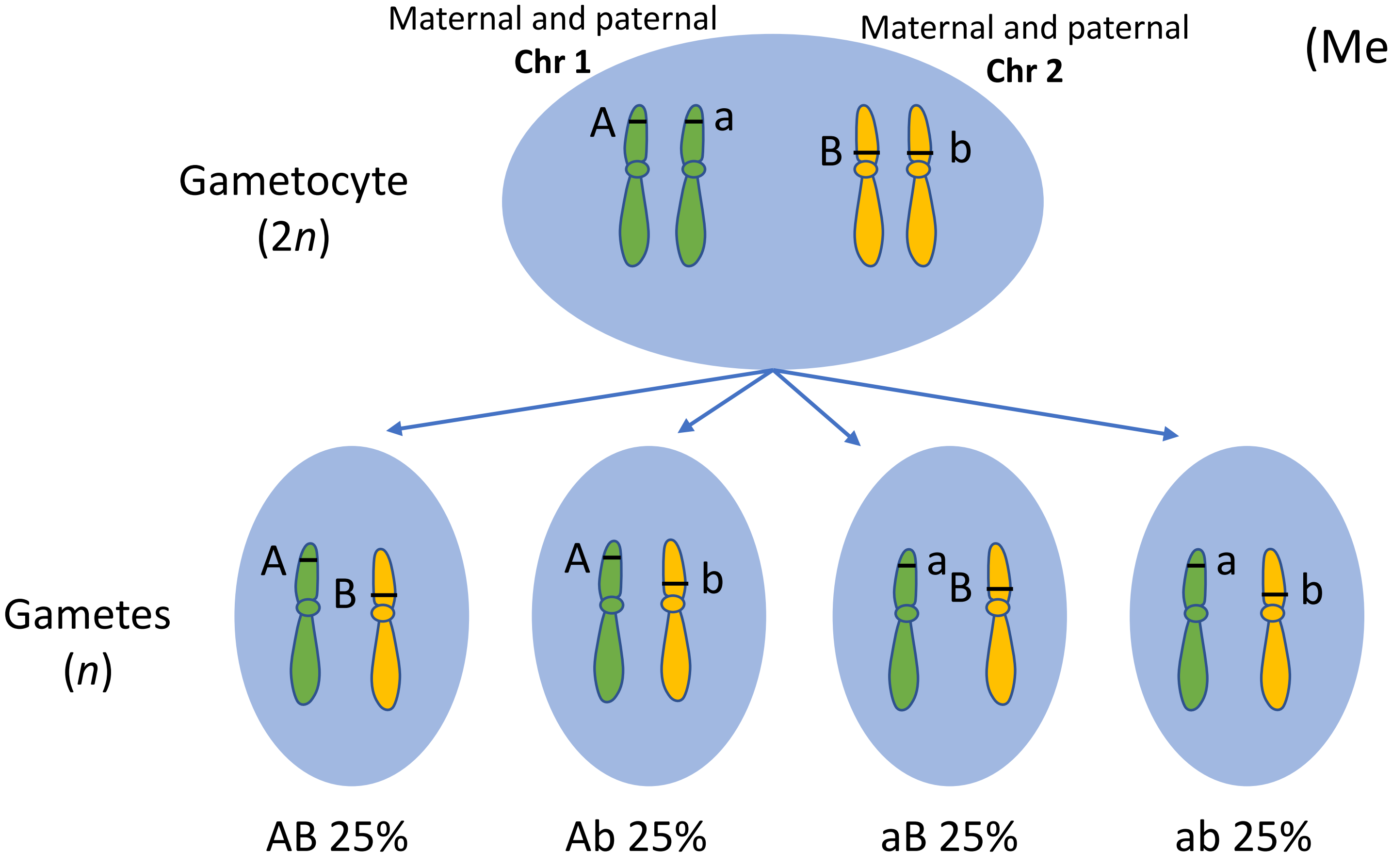
- LD is is a measure of non-random association between alleles at different loci at the same chromosome in a given population.
- Seen as a tendency of alleles to be transmitted together more often than expected by chance alone.
- Usually caused by close proximity of genes on the same chromosome.
- LD is influenced by many factors (e.g. selection, the rate of genetic recombination, mutation rate, genetic drift, the system of mating, population structure and genetic linkage).

Linkage equilibrium (LE)

Meiosis: The formation of reproductive cells (gametes)

- Chromosomes assort independently of one another during meiosis

(Mendel's law of independent assortment)

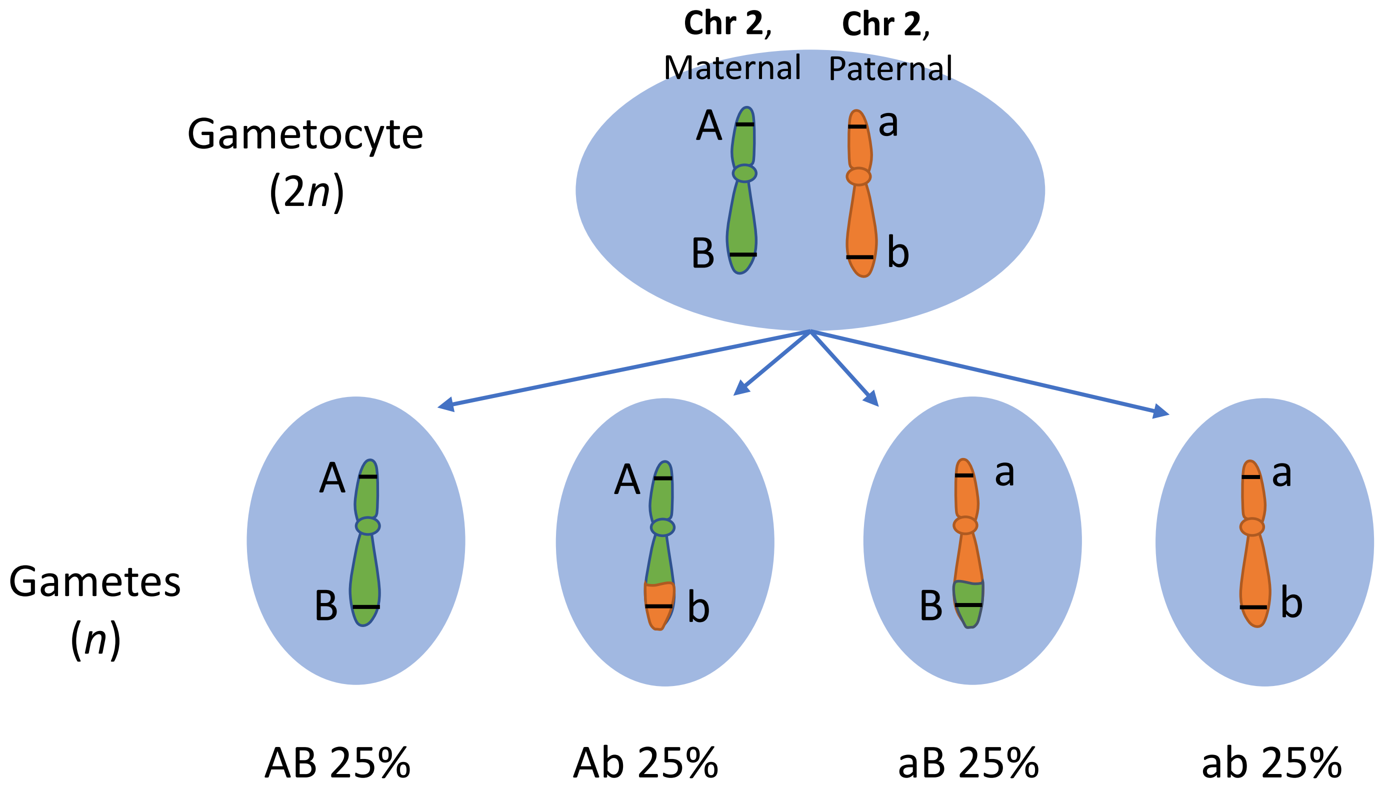


- Alleles that are located on separate chromosomes get sorted into gametes independently of one another.

Homologues: The maternal and paternal chromosomes from the same pair, e.g. chr 2.

Linkage equilibrium (LE)

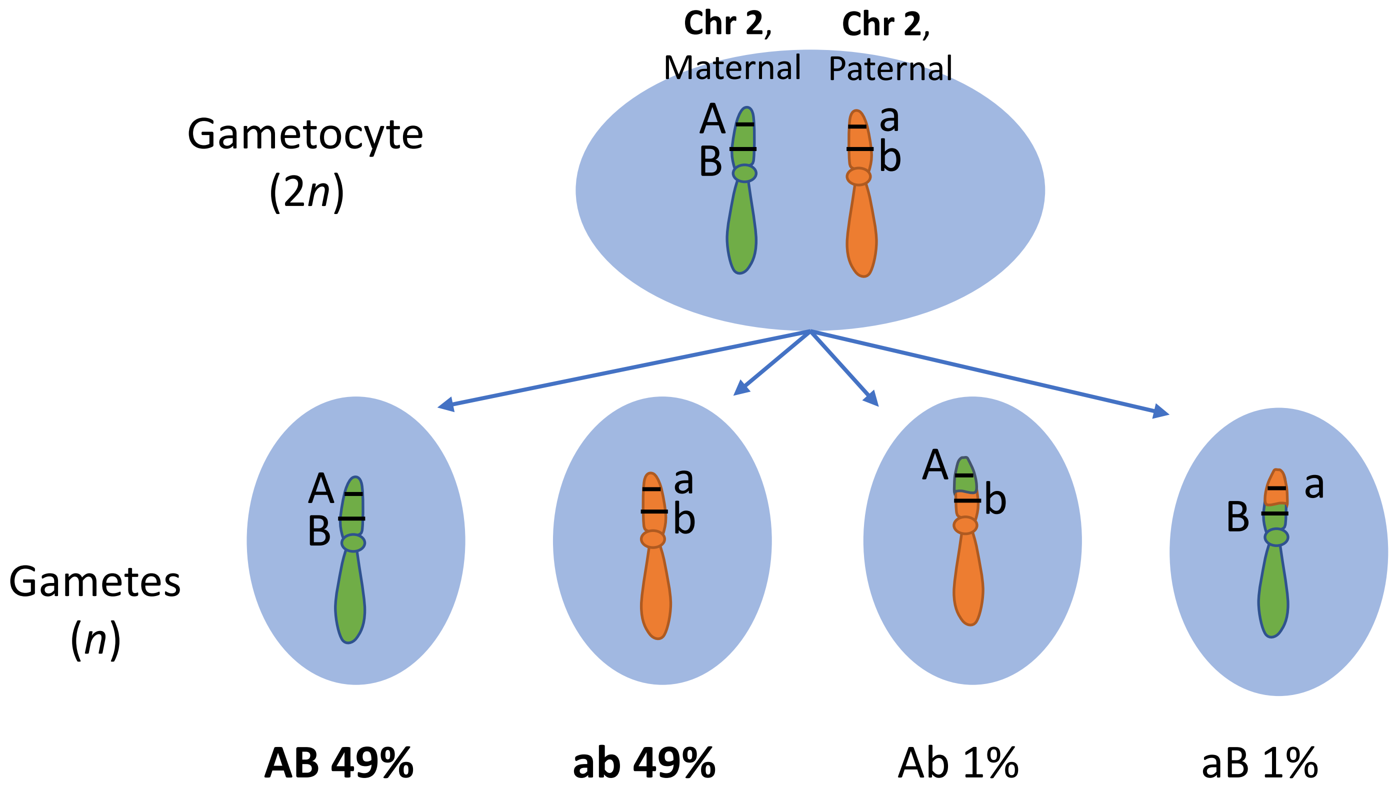
- Alleles at loci far apart on the same chromsome may behave as they were on separate chromosomes



- During meiosis I, homologous chromosomes line up in pairs along the meiotic spindle.
- The paternal and maternal homologues exchange homologous segments by crossing over (i.e. **recombination**), creating new chromosomes that are a mosaic of the original homologous chromosomes.

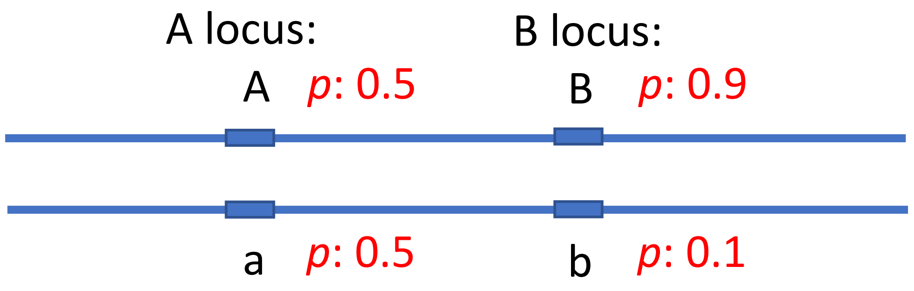
Linkage *dis*equilibrium (LD)

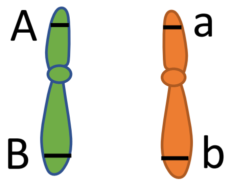
- Loci that are close together on the same chromosome tend to be transmitted together, as an intact unit, through meiosis



- The closer loci are to each other on a chromsome, the less likely they are to get separated by recombination
- Hence, close loci tend to be transmitted together as an intact unit.
- The alleles on one chromosome will occur together in gametes **more often than expected from independent assortment**.

How LD is measured



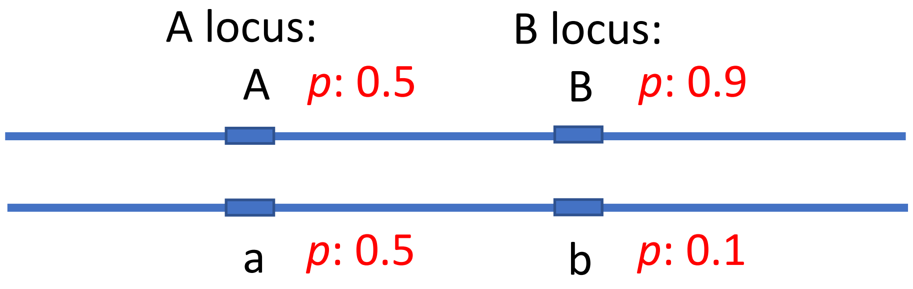


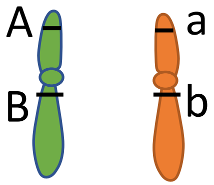
- *p* is the frequency of the alleles in the population.
- The *combination* of alleles you carry on a chromsome is called a **haplotype**.
- Here there are 4 possible haplotypes: AB, Ab, aB, ab.

At linkage equilibrium (LE)

		Allele frequency at locus B	
		p B = 0.9	*p* b = 0.1
Allele freq. at locus A	*p* A = 0.5	Haplotype AB freq. = **0.45**	Haplotype Ab freq. = **0.05**
	p a = 0.5	Haplotype aB freq. = **0.45**	Haplotype ab freq. = **0.05**

How LD is measured





- p is the frequency of the alleles in the population.
- The *combination* of alleles you carry on a chromsome is called a **haplotype**.
- Here there are 4 possible haplotypes: AB, Ab, aB, ab.

At linkage **dis**-equilibrium (LD)

		Allele frequency at locus B	
		p B = 0.9	p b = 0.1
Allele freq. at locus A	p A = 0.5	Haplotype AB freq. = **0.5**	Haplotype Ab freq. = **0**
	p a = 0.5	Haplotype aB freq. = **0.4**	Haplotype ab freq. = **0.1**

Quantification of LD

- The level of linkage disequilibrium between alleles A and B can be quantified by the coefficient of linkage disequilibrium D_{AB}, which is defined as:

$$D_{AB} = p_{\mathrm{AB}} - (p_{\mathrm{A}} \text{ x } p_{\mathrm{B}}) \qquad D_{AB} = 0.5 - (0.9 \text{ x } 0.5) = 0.05$$

Product of the frequencies of individual alleles

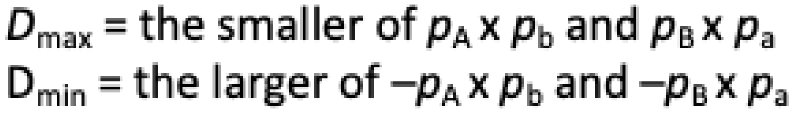
Frequency of the haplotype

- $D \neq 0$ is equivalent to saying the alleles are in LD

Standardization and other measures used for LD

- The coefficient of linkage disequilibrium *D* is not always a convenient measure of LD because its range of possible values depends on the frequencies of the alleles it refers to.
- This makes it difficult to compare the level of LD between different pairs of alleles. D' is standardized to become between 0 and 1:

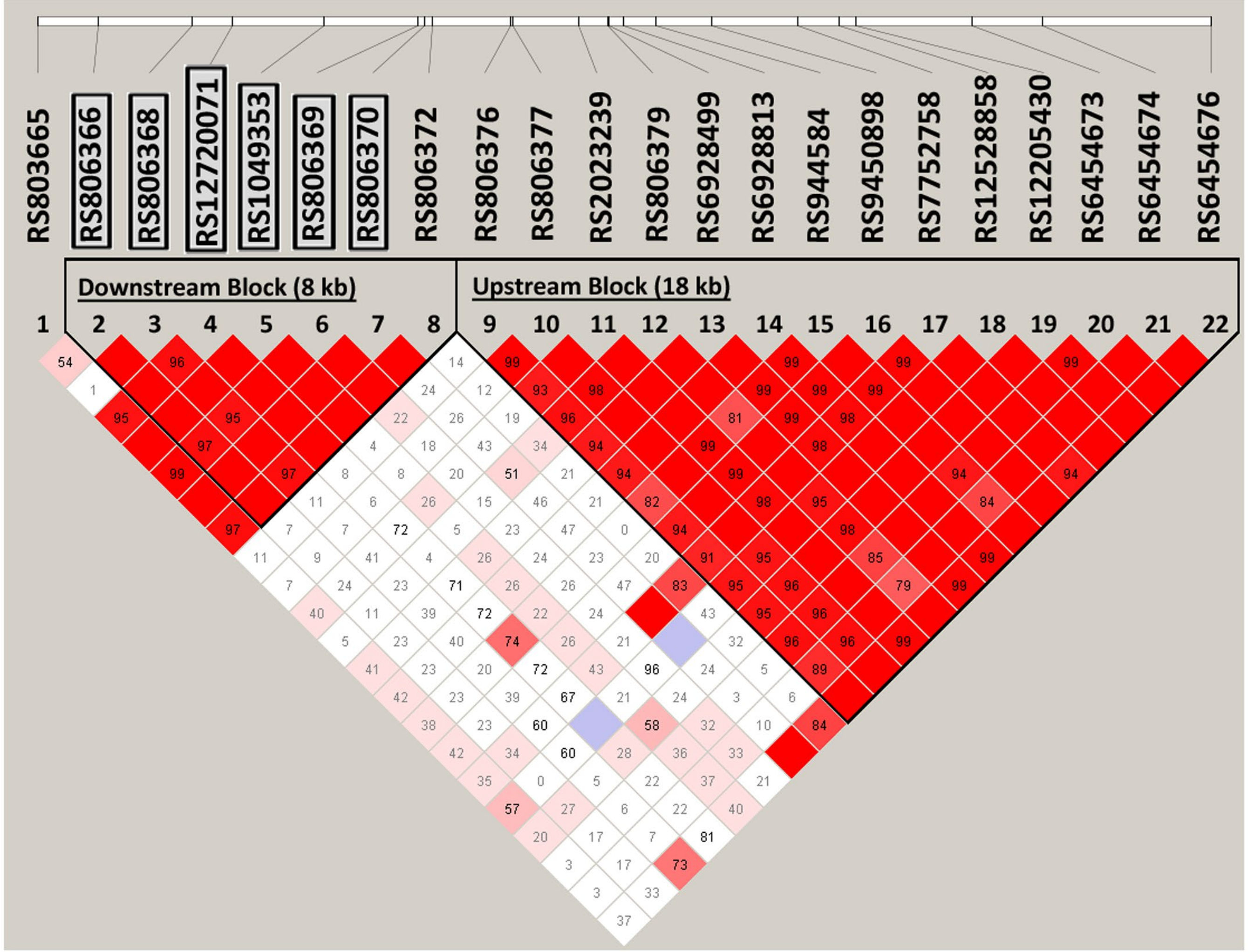
$D' = D/D_{\text{max}}$ (or $D' = D/D_{\text{min}}$ if D is negative)



D_{max} = the smaller of $p_A \times p_b$ and $p_B \times p_a$
D_{min} = the larger of $-p_A \times p_b$ and $-p_B \times p_a$

r^2 : Pearson coefficient of correlation, squared

LD visualization



- Color intensity increases with increasing LD
- Number within each diamond indicate r^2 (in %)
- r^2 = 1 (100%) for diamonds lacking a specified number
- The low LD between SNPs 8 and 9 indicates a site of recombination, flanked by two LD blocks

Practical application of LD in GWAS

- Genome-wide association studies (GWAS) aim to test for association of *all* known variants accross the genome, with a particular phenotype (e.g. a disease).
- **If a series of SNPs are in strong LD and one of them is associated with e.g. disease, a positive association should be detectable between the disease and *all* of the alleles within the LD block.**
- Thus, you only have to test a subset of SNPs («tag SNPs)» within the LD block to specify a particular haplotype. In this way the number of SNPs that need to be tested to cover the human genome can be significantly reduced.
- When interpreting GWAS findings, it can be difficult to know exactly which variant in an LD block that is causal for a disease.

Overview

- The Hardy-Weinberg equilibrium (HWE)
- Linkage disequilibrium (LD)
- Heritability, H^2
- Identity by state & Identity by descent

Understanding Heritability, H^2

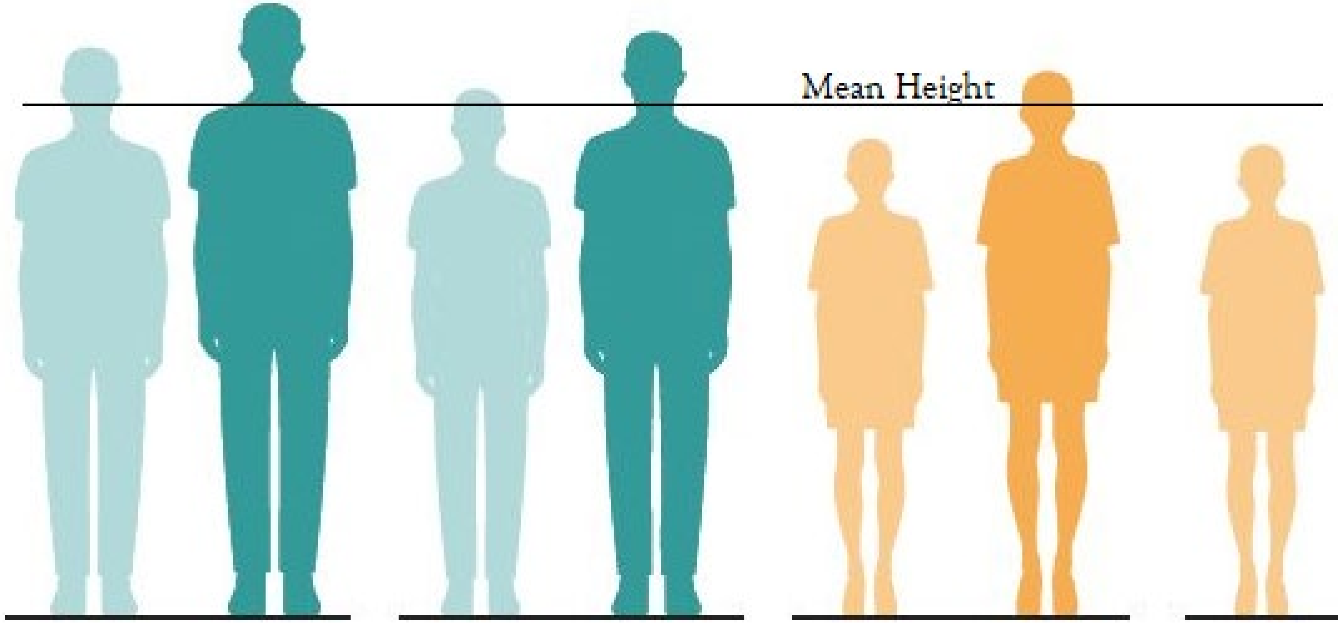
- Heritability, H^2, quantifies the contribution that genetic differences (genetic variation) make to the **variability** of quantitative traits.

- Quantitative traits: Continuous meaurements (e.g. height, weight, blood pressure), and not present/absent.

- Population and time specific.

- Heritability estimates are not measures of the importance of genes in the production of a trait, nor how modifiable a trait is to environmental influence.

The heritability of height in humans is approximately 90%



This <u>does not</u> mean that your height is determined 90% by genetics.

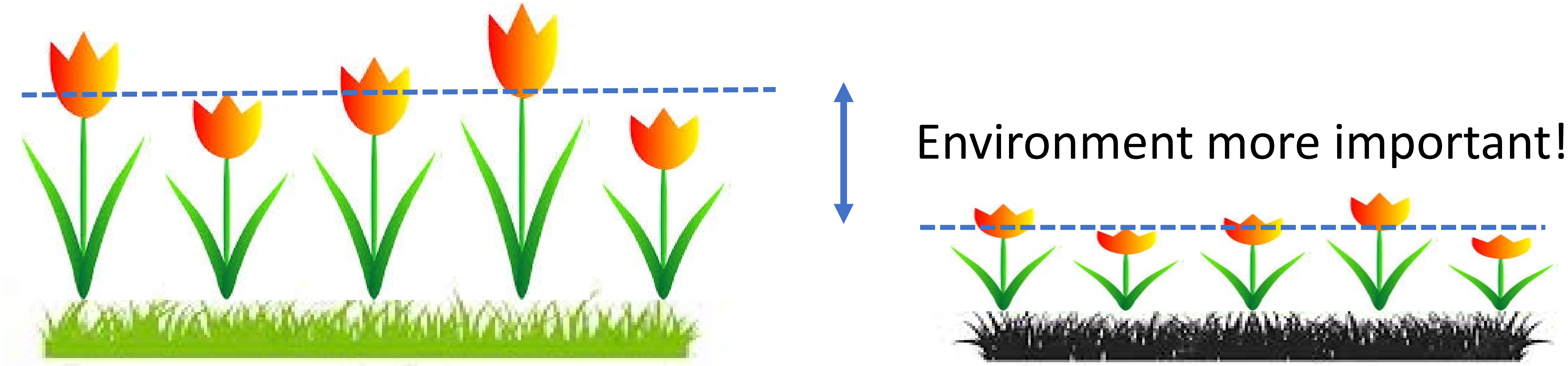
Imagine planting ordinary, genetically diverse seeds into two radically different environments and then allowing them to grow to their full heights.

Ideal amounts of light, water, and nutrients

Just barely enough light, nutrients, and water for survival.

What is the heritability of height?

Imagine planting ordinary, genetically diverse seeds into two radically different environments and then allowing them to grow to their full heights.



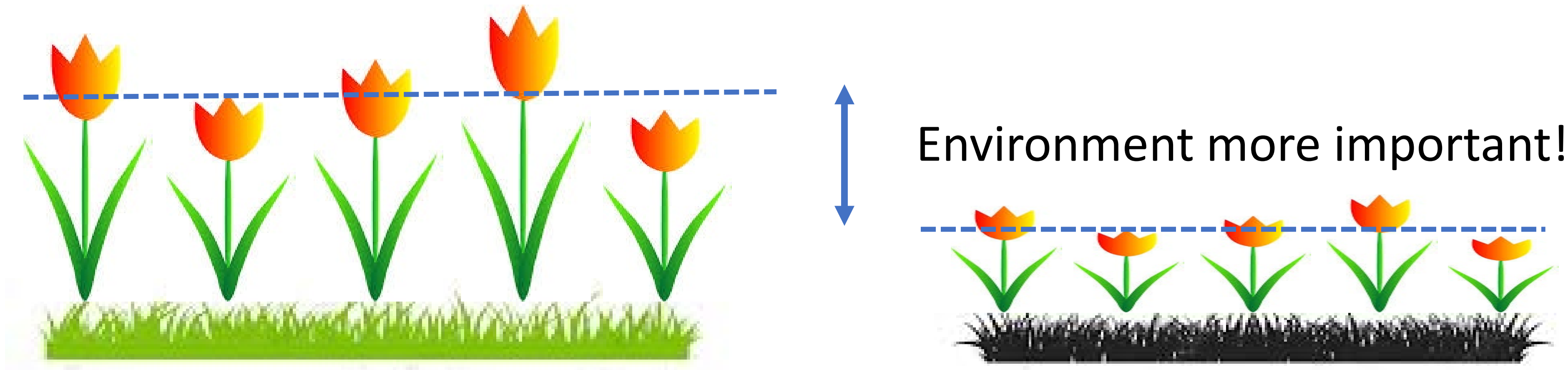
Environment more important!

Ideal amounts of light, water, and nutrients

Just barely enough light, nutrients, and water for survival.

- All of the *variation* in height within each tray must be due to the *genetic diversity* of the seeds, since the seeds developed in identical environments.
- Thus, regardless of the environments in which the plants grew, heritability is 1, or 100 percent, within each tray.

Imagine planting ordinary, genetically diverse seeds into two radically different environments and then allowing them to grow to their full heights.



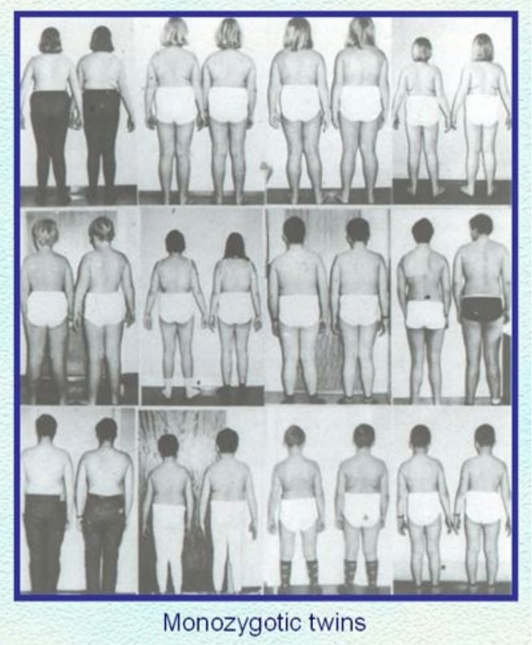
Environment more important!

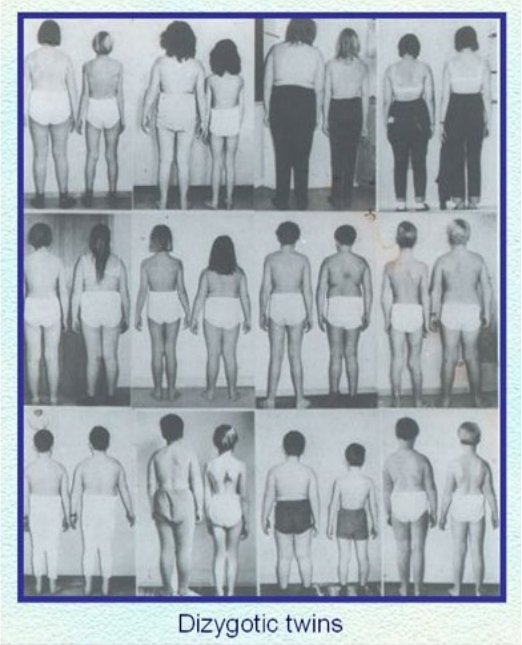
Ideal amounts of light, water, and nutrients

Just barely enough light, nutrients, and water for survival.

- **Heritability** estimates **only** reflect what causes the **variation** in traits in a population; they say nothing about what causes the traits themselves.
- The heritability of a trait within a population may therefore vary over time.

Heritability has historically been estimated from studies of twins.


Monozygotic twins


Dizygotic twins

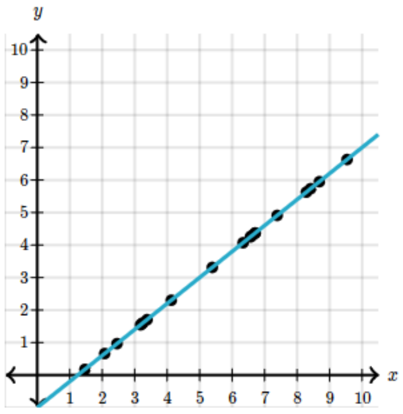
MZ twins share ~**100%** of their alleles (and partly their environment).

DZ twins share on average **50%** of their alleles (and partly their environment).

- If a value of a trait appears to be more similar in MZ twins than in DZ twins, genetic factors likely play an important role in determining the variability of that trait.
- This means, genetic differences between the DZ twins (of a pair) are likely to contribute to the variability of the trait.

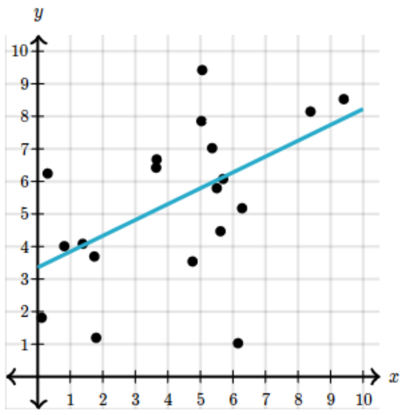
Heritability, H^2

MZ



Example where $r = 1$, which is perfect positive correlation

DZ

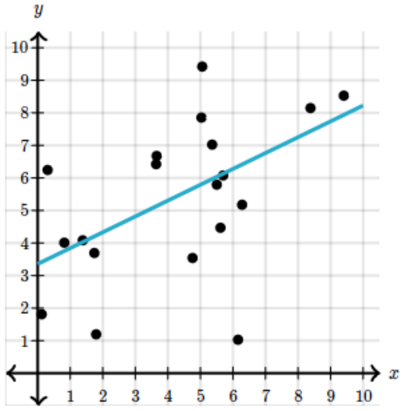


Example where $r = 0.5$, which is weak positive correlation

- We measure pairwise correlation (*r*) in the value of a trait between all twin pairs in both groups (MZ and DZ)
- Heritability: $\boldsymbol{H^2 = 2 \times (r_{MZ} - r_{DZ})}$
- If variation in a trait is influenced <u>mainly by genetic variation</u>, the pairwise correlation in MZ twins would be close to 1, and approximately 0.5 in DZ twins.
- ***H*** $\boldsymbol{H^2}$ **will approach 1**.

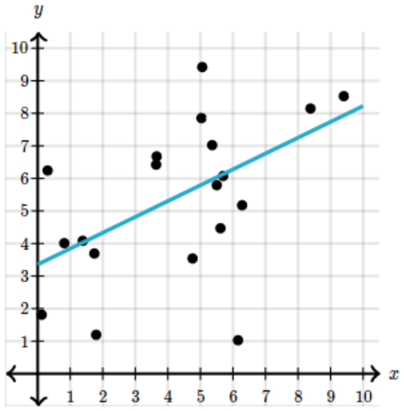
Heritability, H^2

MZ



Example where $r = 0.5$, which is weak positive correlation

DZ



Example where $r = 0.5$, which is weak positive correlation

- We measure pairwise correlation (*r*) in the value of a trait
- Heritability: $\boldsymbol{H^2 = 2 \times (r_{MZ} - r_{DZ})}$

- If trait variation is influenced <u>mainly by environmental factors</u> (and not genetic variation), the pairwise correlation would be similar in MZ and DZ twins.
- ***H*** **$\boldsymbol{H^2}$ will approach 0**.

Pairwise correlation of BMI between MZ and DZ twins brought up together and apart

		Men			Women		
Twin Type	Brought up	No. of pairs	BMI*	Pairwise correl.	No. of pairs	BMI*	Pairwise correl.
MZ	Apart	49	24.8 ± 2.4	0.70	44	24.2 ± 3.4	0.66
	Together	66	24.2 ± 2.9	0.74	88	23.7 ± 3.5	0.66
DZ	Apart	75	25.1 ± 3.0	0.15	143	24.9 ± 4.1	0.25
	Together	89	24.6 ± 2.7	0.33	119	23.9 ± 3.5	0.27

*: mean ± SD

Pairwise correlation of BMI between MZ and DZ twins reared together and apart

		Men			Women		
Twin Type	**Brought up**	**No. of pairs**	**BMI***	**Pairwise correl.**	**No. of pairs**	**BMI***	**Pairwise correl.**
MZ	Apart	49	24.8 ± 2.4	0.70	44	24.2 ± 3.4	0.66
	Together	66	24.2 ± 2.9	0.74	88	23.7 ± 3.5	0.66
DZ	Apart	75	25.1 ± 3.0	0.15	143	24.9 ± 4.1	0.25
	Together	89	24.6 ± 2.7	0.33	119	23.9 ± 3.5	0.27

MZ

DZ

*: mean ± SD

The correlation between pairs of twins was higher for MZ than for DZ twins, regardless of whether they were brought up together or apart. This supports a strong impact of genetic variation in BMI variability.

Pairwise correlation of BMI between MZ and DZ twins reared together and apart

		Men			Women		
Twin Type	**Grown up**	**No. of pairs**	**BMI***	**Pairwise correl.**	**No. of pairs**	**BMI***	**Pairwise correl.**
MZ	Apart	49	24.8 ± 2.4	0.70	44	24.2 ± 3.4	0.66
	Together	66	24.2 ± 2.9	0.74	88	23.7 ± 3.5	0.66
DZ	Apart	75	25.1 ± 3.0	0.15	143	24.9 ± 4.1	0.25
	Together	89	24.6 ± 2.7	0.33	119	23.9 ± 3.5	0.27

MZ

DZ

*: mean ± SD

Grew up together: $H^{2\ \text{men}}$: 2 x (0.74 – 0.33) = 0.82 $H^{2\ \text{women}}$: 2 x (0.66 – 0.27) = 0.78

The usefullness of heritability estimates

- Heritability estimates *can estimate the sources of differences* among people, but only for a particular population, at a particular time, and in particular circumstances.
- Heritability estimates only reflect what causes the variation in traits; they *say nothing about what causes the traits themselves*.
- Heritability estimates say nothing about *which* and *how many* loci in the genome are involved in creating trait variability!

Overview

- The Hardy-Weinberg equilibrium (HWE)
- Linkage disequilibrium (LD)
- Heritability, H^2
- Identity by state & Identity by descent

Identity by State & Identity by Descent

- Relatedness between individuals may cause false signals (type I errors) in genome-wide associaton studies (GWAS), especially if the extent of relatedness differs among cases and controls.
- Number of related pairs grows rapidly with sample size; e.g., in UK Biobank, with 500,000 samples, 30% have close relatives.
- Thus, we need methods to identify close relatedness in sample sets.

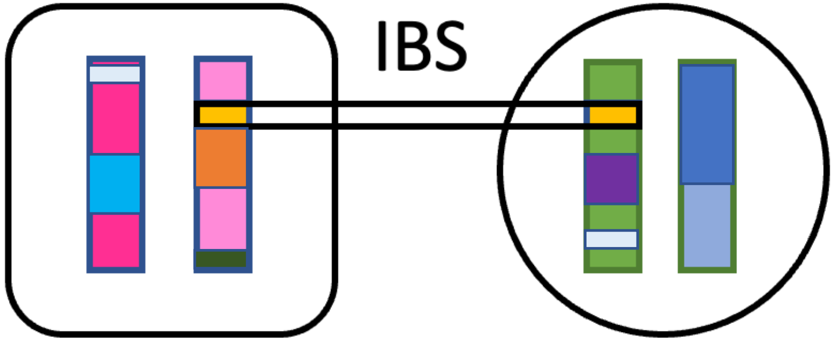
Identity by State & Identity by Descent

- **Identical by state** or **identity by state** (IBS) is a term used in genetics to describe two identical alleles or two identical segments or sequences of DNA.
- A DNA segment is identical by state (IBS) in two or more individuals if they have identical nucleotide sequences in this segment.
- An IBS segment is **identical by descent** (IBD) in two or more individuals if they have inherited it from a common ancestor without recombination, that is, the segment has the same ancestral origin in these individuals.

○ female

□ male

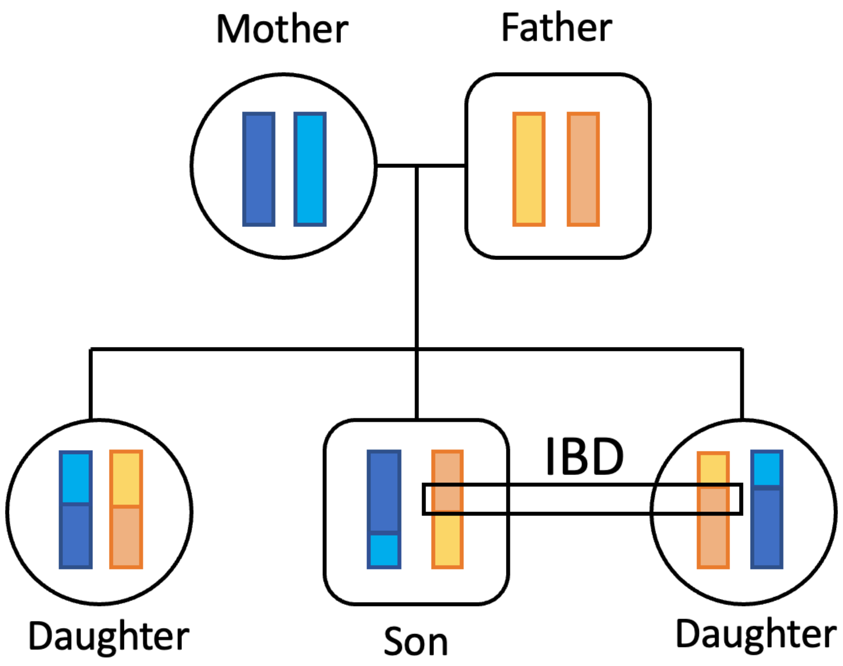
Identity by state (IBS)



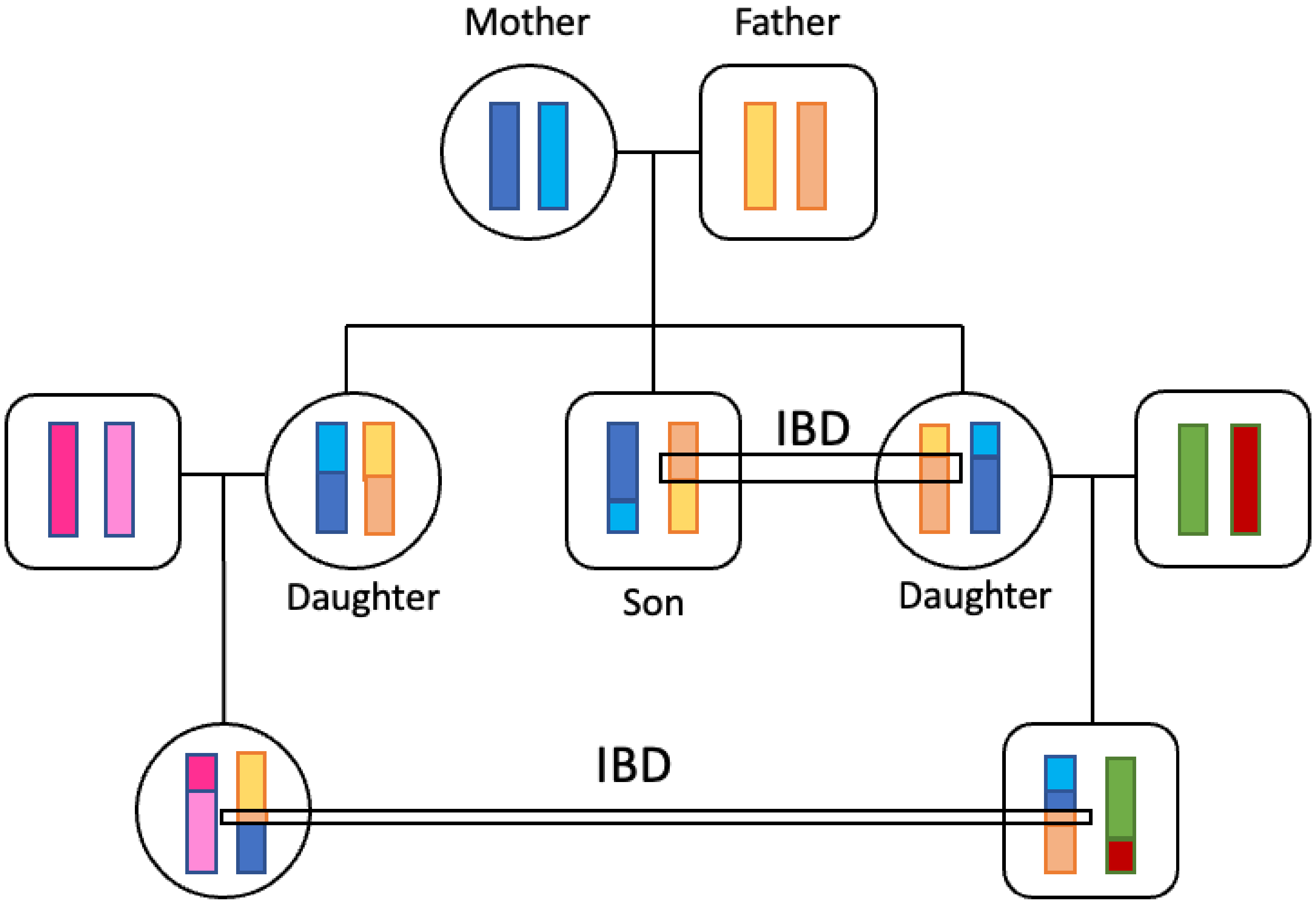
Unrelated individuals

Each box represents an individual with two homologous chromosomes as bars. Two unrelated individuals have a DNA segment in common. These DNA segments are identical by state (**IBS**), but because they are not inherited from a common ancestor, they are not identical by descent (IBD). They could be common in the population and are irrelevant for inferring relatedness.

Identity by descent (IBD)



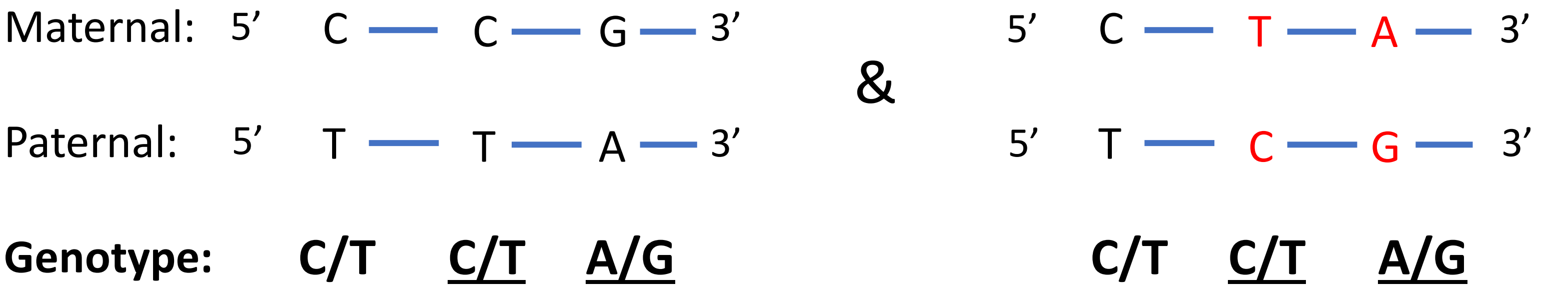
IBD illustrated by a pedigree of 5 individuals: Due to crossing over, the children inherit recombinant chromosomes from their parents. The siblings share several DNA segments that are IBD (inherited from the same parent).



The expected proportion of genome sharing between two individuals varies as a function of their genetic relatedness. The larger the IBD segment, the closer the relationship with the ancestors.

How are IBS and IBD segments identified?

- IBS and IBD segments can be identified from DNA sequencing or SNP array genotyping data
- SNP array data, and to some extent also DNA seq., do not discriminate between which allele is carried on each of the homologous chromosomes



Maternal: 5′ C — C — G — 3′ & 5′ C — T — A — 3′

Paternal: 5′ T — T — A — 3′ 5′ T — C — G — 3′

Genotype: C/T C/T A/G C/T C/T A/G

Different haplotypes may look the same (i.e. not «phased»)

- Complex computational algorithms are needed to identify haplotypes and number and lengths of shared fragments to estimate degree of relatedness.

Brief summary

- The Hardy-Weinberg equilibrium (HWE)
  - Important concept about the relationship between allele frequencies and the expected distribution of genotypes. Used e.g. In genetic counceling or as a quality check in GWAS to detect potential genotyping errors.
- Linkage disequilibrium (LD)
  - Describes the non-random distribution of alleles located close together on a chromosome. Association with LD blocks can be detected by «tag SNPs».
- Heritability, H^2
  - Concept developed to quantify the contribution that genetic differences (genetic variation) make to the variability of quantitative traits.
- Identity by state & Identity by descent
  - The use of genotype data to detect relatedness between individuals in a population.