GENOMICS AND COMPLEX TRAITS

❖ INTRODUCTION

Complex traits are very important. Except for some cases of trauma, it's safe to say that **every human trait has a hereditary component.** Complex traits are important and usually much more different than simple mendelian characters.

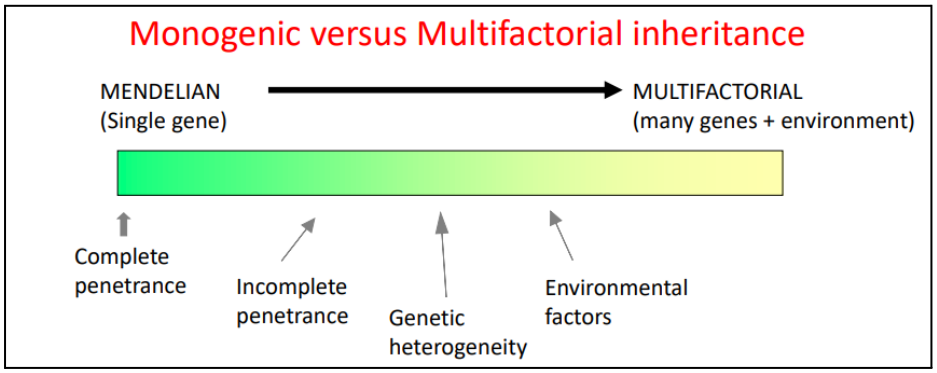
Mendelian Inheritance	Polygenic Inheritance (Fisher)
Genetic characters controlled by discrete factors that exist in pairs **in one locus**	**Large number of independent loci** each having small contribution to the character (e.g. height).
The ***discrete*** factors are the basis of heredity and are passed from generation to generation.	***Continuous*** analogue of Mendelian inheritance.

The presence (or absence) of simple mendelian characters depends on the genotype at a single locus. However most characters around us are not purely mendelian since they depend on genotypes present at more than one locus.

When we consider:

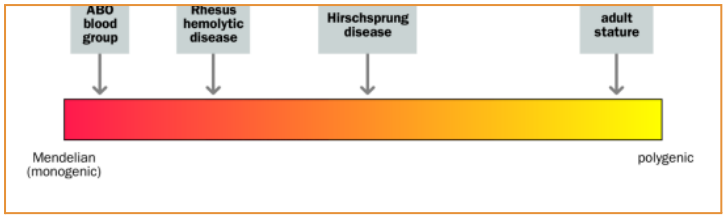
- ★ only one locus: ***monogenic***
- ★ a few loci: ***oligogenic***
- ★ many loci: ***polygenic***
- ★ many loci + environment***: multifactorial or complex***

The initial problem with the coexistence of polygenic and Mendelian (monogenic) traits was that first traits are not explained by Mendelian laws as genes at multiple loci are considered. Fischer reconciled these two theories, saying that **each of the traits that contribute to a complex phenotype acts, on his own, as a mendelian trait.** Moreover, when speaking about complex traits, we must also consider that the **environment plays a role in the phenotype of the trait.**



There is a **SPECTRUM** between mendelian and complex traits.
We can have purely mendelian traits, but also incomplete penetrance, genetic heterogeneity and environmental factors determining the inherited phenotype.

The **environmental effect**, which is not a genetic factor, is also an important characteristic.



During the course of our studies we have already seen examples of all the different kinds of inheritance we can have.

- **Mendelian** traits are **dichotomous**: characters are either present or absent. [1]
- **Complex** traits are either:
  - **dichotomous** (qualitative). The loci characteristic of this type of trait are called ***susceptibility loci***.
  - **continuous** (quantitative). These characters not only have two outcomes so they are not mendelian. However there are still loci of interest, which are called ***quantitative trait loci.*** Their name comes from the fact that each trait contributes only a little to the overall phenotype.

Malfunction of development pathways is likely to involve more factors and thus it is not monogenic. **Many common disorders are multifactorial**: *asthma, diabetes, epilepsy, hypertension, multiple sclerosis, Alzheimer's, manic depression, autism, schizophrenia, etc…*

- Let's take for example ***diabetes***: we have rare cases that are mendelian and are caused at a single locus, moreover, sometimes diabetes is only dictated by environment. However most of the time **both genes and environment play a part.** Hence why we define diabetes as a multifactorial illness.

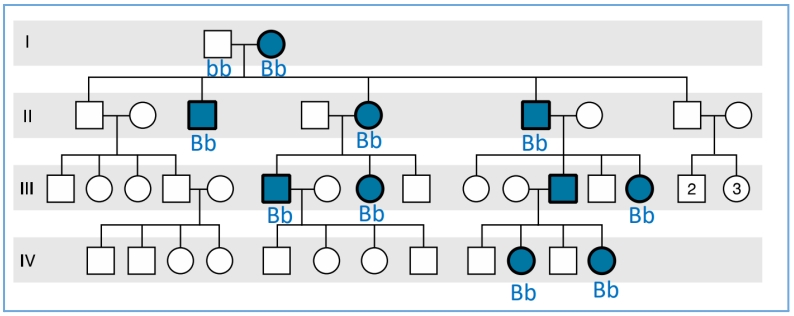
[1] Dichotomous traits are discrete traits that have only two contrasting phenotypic probabilities. Being dichotomous is not restricted to mendelian characters.

❖ SINGLE GENE DISORDERS: MENDELIAN INHERITANCE

Single gene disorders (or monogenic disorders) **are caused by a particular allele being present in only one locus.** More than 4500 Medelian disorders are known, but they constitute only 1% of all diseases

When a monogenic trait is transmitted through the family (hence it's not threatening their life) we can see an inheritance pattern. As we know, there are four inheritance patterns:

★ AUTOSOMAL DOMINANT

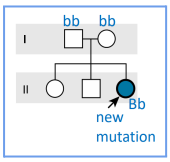


➔ For rare conditions, almost all affected individuals are **heterozygous.**

➔ A child of an affected parent has a 50% risk for inheriting the trait (if other parent is phenotypically normal)

➔ The **phenotype appears in every generation**

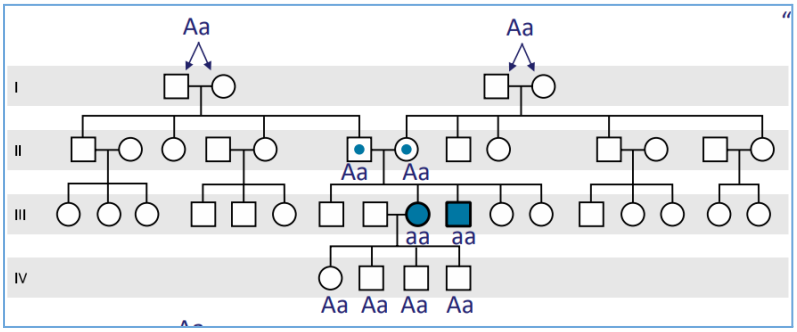
➔ Males and females are equally likely to transmit the phenotype



Isolated cases are usually sporadic cases due to new mutations.

★ AUTOSOMAL RECESSIVE

Autosomal recessive disease occurs only in individuals with two mutant alleles and no wild-type allele. Hence, the affected child must be **homozygous**.



➔ Parents of an affected child are both asymptomatic carriers of mutant alleles.

➔ Males and females are **equally likely to be affected.**

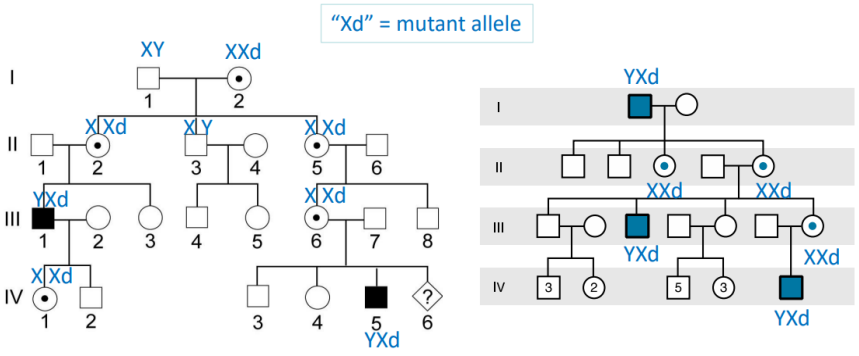
➔ Parents of the affected person may be consanguineous. Especially likely if the condition is rare.

➔ The recurrence risk for each sibling of the proband is ¼

We DO NOT see the disease in every generation as it is rare that both parents are carriers.

★ X LINKED RECESSIVE

X linked recessive disorders are the most popular X linked conditions.



➔ This condition is mainly seen **in males,** who only have one X chromosome that is always expressed.

➔ **Every daughter of an affected male is a carrier.** There is a 50% chance that the daughter will transmit that trait.

➔ The mutant allele is never transmitted directly from father to son. The affected males are related through females

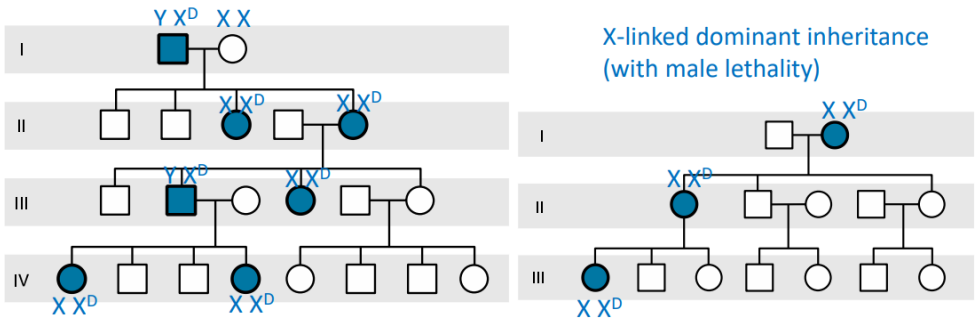
In rare cases, an X linked condition can be seen in **females**:

- ❖ if they are *homozygous*
- ❖ if the female is manifesting *heterozygous*. This can be due to an abnormal pattern of X linked inactivation (we say the pattern is skewed in a favorable manner)

Isolated cases are often due to new mutations.

★ X LINKED DOMINANT

Since the disease is dominant, it is more common **in females**. In general, affected females typically have milder (although variable) expression of the phenotype.



➔ Affected males transmit the phenotype to all the daughters and not to the sons

➔ Both male and female offspring of female carriers have a 50% risk for inheriting the phenotype

---

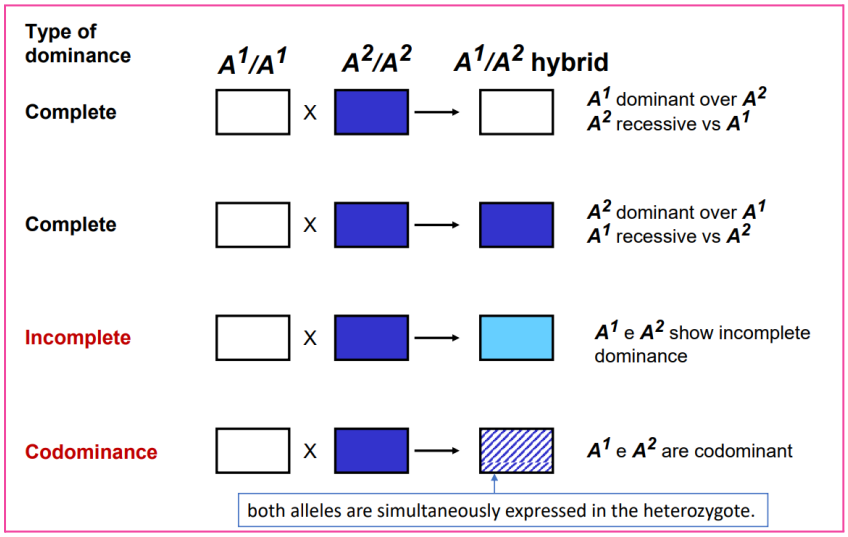
Main features of Mendelian inheritance patterns

	Autosomal dominant	Autosomal recessive	X-linked dominant	X-linked recessive	Mitochondrial
Example	Huntington's disease	Cystic fibrosis	Vitamin D-resistant rickets	Hemophilia	Leber's hereditary optic neuropathy
Multiple generations affected?	YES	**NO** (skips generation)	YES	**NO** (skips generation)	YES
Is a parent always affected?	YES (unless it is a new mutation)	NO	YES	NO	YES
Are both sexes affected?	YES	YES	YES (F > M)	YES (M > F)	YES
Are all affected individuals male?	NO	NO	NO	**YES**[a]	NO
Male-to-male transmission?	POSSIBLE	POSSIBLE	**IMPOSSIBLE**[b]	**IMPOSSIBLE**[c]	IMPOSSIBLE
Is it always transmitted from the mother?	NO	NO	NO	NO	**YES**[d]
Is it always transmitted from the father?	NO	NO	NO	NO	NO
	50% of the children of an affected parent will be affected. Each affected child has an affected parent	One in four children of healthy (carrier) parents will be affected. An affected child usually has unaffected parents	All female children of an affected father are affected	No male children of an affected father are affected	Transmitted only from an affected mother (no transmission from an affected father)

There is also mitochondrial inheritance which depends totally on the mothers genotype.

❖ SINGLE GENE DISORDERS: COMPLEX INHERITANCE

As we said, there is a spectrum that goes from mendelian to complex inheritance.
First of all, we look at **complex inheritance in one locus.**



Mendel talks about complete dominance.

➔ **COMPLETE DOMINANCE**: a trait is completely dominant with respect to the other, which is called a recessive trait.

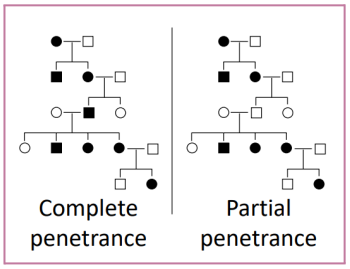
We can have (still in a single locus) other types of interactions:

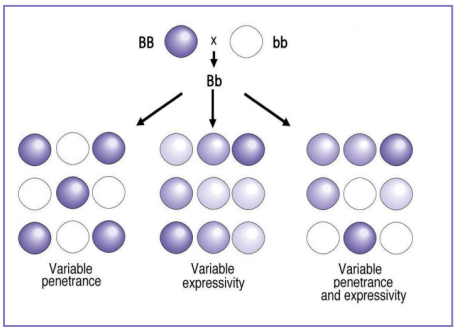
➔ **INCOMPLETE DOMINANCE**: the product of the parents generates an offspring which has a **new, intermediate phenotype.**

➔ **CODOMINANCE**: two alleles at a single loci express at the same time **both alleles.** An example is the existence of the AB blood group.

There can be **several complications** to basic inheritance patterns.

- ★ **EXPRESSIVITY** refers to the **degree of which a character is expressed**. When we have complex traits, all heterozygous individuals express the trait in different ways.
- ★ **PENETRANCE** is the **proportion of people carrying a disease causing allele that are affected.** Penetrance can be either ***complete*** or ***incomplete***, as we can see in the pedigrees on the right.
  - **INCOMPLETE PENETRANCE** refers to the **failure of a genetic traits to be evident** even though the genotype usually producing the phenotype is present





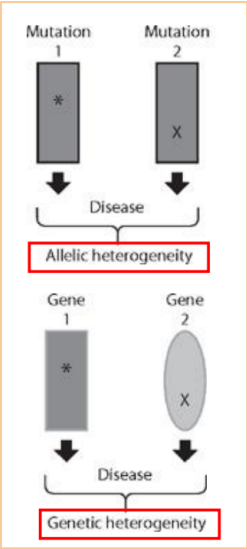
Hence:

➔ **PENETRANCE** is the ***percentage*** **of individuals having a particular genotype that express the expected phenotype**

➔ **EXPRESSIVITY** is **the *degree* to which a character is expressed**

Both are **consequences of how gene expression is affected by environment and genetic background.**

In complex traits we always need to consider the **genomic context**.

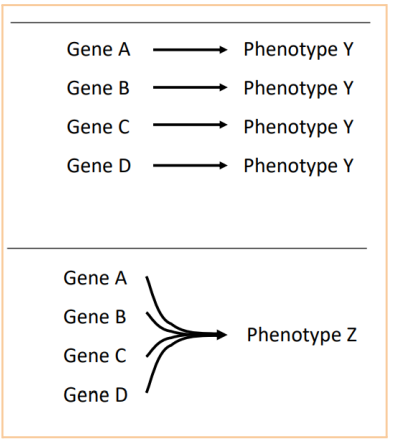


★ **PHENOCOPIES** are conditions which mimic the ones caused by actual genetic diseases. These conditions could be due to:

- **no genetical causes**
- **variants** which may result in the same or indistinguishable phenotypes. These variants are referred to as "**HETEROGENEITY**"

★ we have two main kinds of **HETEROGENEITY** which can contribute to the same disease:

- **ALLELIC HETEROGENEITY:** different variants of the *same* gene increase the risk for the same disease
- **GENETIC HETEROGENEITY:** variants in *different* genes independently increase the risk for the same disease



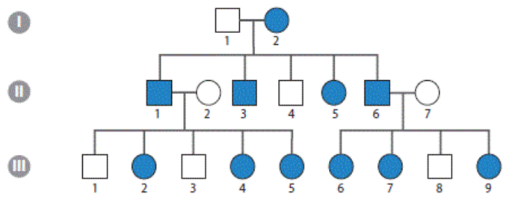
It's very important not to confuse heterogeneity with oligogenicity:

★ **HETEROGENEITY**: variants in **distinct genes** that may result in the **same phenotype**

★ **POLYGENICITY**: variants from several distinct genes acting together to **create a phenotype**

EXERCISES

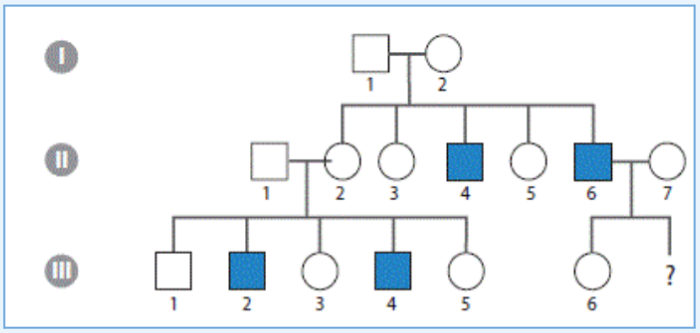
What is the likely inheritance pattern in the pedigree below?



This is clearly not a recessive disorder.

➔ It could be an **autosomal dominant** disorder on the basis that both sexes are affected and affected individuals have an affected parent.

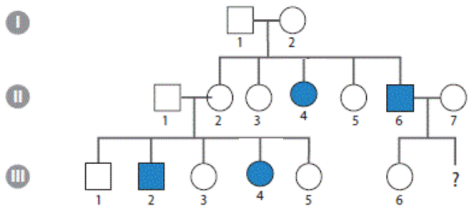
➔ Since there are six affected females and three unaffected males in generation III, each of whom had an affected father, there is also the possibility of an **X-linked dominant** condition.



This could be an autosomal dominant disorder in which there is non-penetrance with two asymptomatic gene carriers (II-2 and one of the grandparents in generation I).
However, given the presence of affected males an **X-linked recessive** is more likely.
Hence II-7's risk of having an affected child would be **zero**:

- ➔ a son could not be affected since it inherits the Y chromosome from the father
- ➔ there is a 50% chance that the affected father, II-6, will transmit an X chromosome with the mutant allele to a daughter, but in that case she would be a carrier but would not be affected.

What is the likely inheritance pattern of a very rare condition in the pedigree below?



II-7 is pregnant again. If the penetrance is of 80%, what is the risk that the child will be affected?

This is an **autosomal dominant disorder.**

1. if it were recessive we should have 3 outside individuals that are sick. This is unlikely since the disease is very rare. It's more possible that it is dominant.
2. we are talking about reduced penetrance: this means that this is a dominant disorder cus if it were homozygous recessive the penetrance would be 100%.

As for the second question, this is a rare disorder so we assume the mother, II 7, is not a carrier with reduced penetrance.

- ➔ There is a 50% chance that the affected father, II-6, will transmit the disease allele to the child, but considering a penetrance of 80%: ½ * 4/5 = 0,4: 40%

❖ MULTIGENIC DISORDERS: COMPLEX INHERITANCE

Traits tend to run in families. We say a trait is "***complex***" when it doesn't depend on a single, dichotomous gene.

➔ For example the habsburg family had a peculiar shape of the jaw. No single genes are responsible for this, but it's sure that the trait runs in the family. Thus, another kind of pattern exists.

Before looking for genes we must be sure that there are genes to be found. Hence **we need to understand if genetic factors play a role in the condition**.

If there is **genetic susceptibility then relatives are more likely to have this trait rather than total strangers.**

DISEASE	RELATIVE RISK FOR SIB (λ_S)
Alzheimer disease (late-onset)	4
Autism spectrum disorder	6.5
Breast cancer, female	2
Crohn's disease	25
Multiple sclerosis	20
Schizophrenia	9
Type 1 diabetes	15
Type 2 diabetes	3

THE RISK RATIO

The Risk Ratio λ_r: is the **risk to a relative (R) of an affected proband compared with the risk in the general population.**

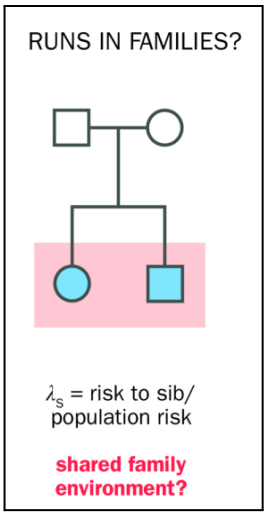
➔ if the risk of the general population is the same as a relative, genetics doesn't have a strong effect on the condition. Otherwise, genetics matters.

➔ λ_r**=1** : no additional risk above that of the general population.

➔ extremely **high λ_s values** are found in monogenic disorders.

➔ **average λ_s values** are found in complex disorders.

In particular, what we calculate is λ_s, which is the **risk to a sibling divided by the risk to the general population**



EXERCISE

We can calculate the risk index for cystic fibrosis with:

1. the probability of a sibling being affected for an autosomal recessive disease like cystic fibrosis)
2. the probability of having cystic fibrosis being 1\2000.

Hence the probability for a sib is ¼ and for a stranger 1/2000. This gives us λ_s = 500, which means that the probability of having cystic fibrosis is strictly connected to the kin.

However **parents** not only **give** genes to their kin, but also **their environment.**

- For example, a family living near a factory increases the possibility of having lung cancer. However this has nothing to do with the genome and nothing to do with the environment.
- The influence of upbringing is very important in psychiatric conditions such as schizophrenia.

This means that the risk ratio could also be influenced by the environment, thus not being definitive about the presence of shared genes.

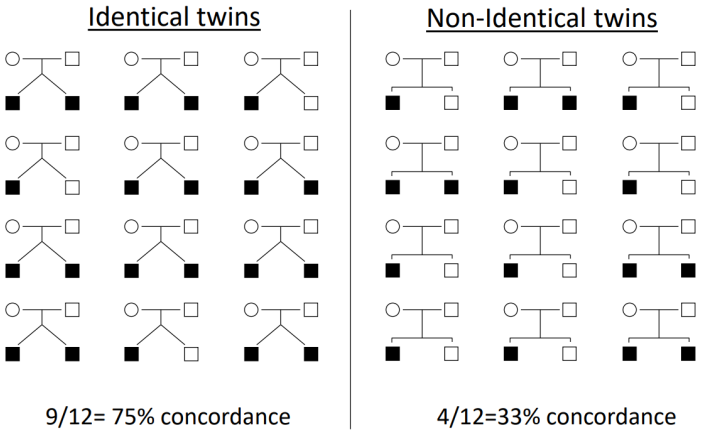
TWIN STUDIES

To overcome this problem we do TWIN STUDIES.

- ★ **MONOZYGOTIC twins** share 100% of their genes because they originate from one egg separating in two
- ★ **DIZYGOTIC twins** share 50% of their genes because they originate from two eggs and two sperms.

Twin studies:

- ➔ **compare the similarity of identical (MZ) and fraternal (DZ) twin pairs** for the trait of interest
- ➔ the **ratio** of MZ:DZ concordance/correlation **can help infer on the genetic component**
- ➔ are usually matched for age and use same-sex pairs of DZ (the MZ are ofc always of the same sex)
- ➔ **include twins raised apart because this removes the effect of shared environment,** allowing better estimation of genetic components.



Genetic characters show higher concordance in monozygotic rather than dizygotic twins.

➔ Concordance is the presence of the same trait in both members of a pair of twins.

Identical twins have a 75% of concordance while in non identical twins the concordance is 33%. This means that **genetics is not the only factor because otherwise there would be a 100% concordance.**

If we don't see a higher concordance in MZ twins it probably means that the trait isn't genetically explainable.

Looking at the table:

- ➔ Genetic factors are more important in schizophrenia since the difference between mono and dizygotic twins is very high
- ➔ Genetic factors are not as important in Parkinson since the concordance percentage is more or less the same

DISEASE	CONCORDANCE (%)	
	In MZ twins	In DZ twins
Type 1 diabetes	42.9	7.4
Type 2 diabetes	34	16
Multiple sclerosis	25.3	5.4
Crohn's disease	37	10
Ulcerative colitis	7	3
Alzheimer disease	32.2	8.7
Parkinson disease	15.5	11.1
Schizophrenia	40.8	5.3

Table 8.4 Genetics and Genomics in Medicine (© Garland Science 2015)

ADOPTION STUDIES

Monozygotic twins raised apart would be better because their environments might be more similar than separated twins.

➔ hence, ***adoption studies*** are more important for disentangling genetic and environmental influences.

However, adoption studies are difficult to perform for lack of subjects and so they haven't been performed a lot. Their main obstacles are:

1. Lack of information about the biological family
2. Selective placement: in the interest of the child, the adoption agency selects a family likely to resemble the biological family. This means that the environmental difference isn't very significant

TABLE 15.3 AN ADOPTION STUDY IN SCHIZOPHRENIA

Case types	Schizophrenia cases among biological relatives	Schizophrenia cases among adoptive relatives
Index cases (47 chronic schizophrenic adoptees)	44/279 (15.8%)	2/111 (1.8%)
Control adoptees (matched for age, sex, social status of adoptive family, and number of years in institutional care before adoption)	5/234 (2.1%)	2/117 (1.7%)

The study involved 14,427 adopted persons aged 20–40 years in Denmark; 47 of them were diagnosed as chronic schizophrenic. The 47 were matched with 47 non-schizophrenic control subjects from the same set of adoptees. [Data from Kety SS, Wender PH, Jacobsen B et al. (1994) *Arch. Gen. Psychiatry* 51, 442–455.]

One of the largest adoption studies has been performed on schizophrenia. As we can see in the table, for biological relatives there's a much higher sharing in comparison to relatives who were just adopted in the same family. This means that schizophrenia has a high genetic base.

Nevertheless, we need to differentiate familiarity and heritability:

★ **FAMILIARITY** indicates the extent to which a 'trait' passed down through generations, considering **both the genetics and the environment**.

★ **HERITABILITY** indicates the proportion of phenotypic variation that is attributable to **genetic variation**

EXERCISE

The table below shows the percentage phenotype concordance in monozygotic (MZ) and dizygotic (DZ) twins in four hypothetical genetic diseases A to D. Which disease would you estimate to have the highest heritability and which the lowest heritability, and why?

DISEASE	CONCORDANCE IN MZ TWINS (%)	CONCORDANCE IN DZ TWINS (%)
A	37.5	16.2
B	15.2	11.7
C	19.2	7.9
D	17.2	1.6

Question 8.3 Genetics and Genomics in Medicine (© Garland Science 2015)

The highest heritability is given by disease D because the ratio (sib/gen) is higher. Furthermore, we can try to see this from a logical point of view: there's a high coordinance between twins who share the same DNA while there's a low coordinance for twins who do not share their DNA. This means that coordinance is given by genetics.

Instead, the smaller heritability is B because there's no difference between monozygotic and dizygotic.

- ➔ A is monogenic
- ➔ B is a complex disorder where the environment plays a big role. However, the ratio isn't one, so some genetic characteristics can be found.

The risk of developing a disease is sometimes expressed as a risk ratio, λ. What is meant by this ratio? Disorder A has a λ_S of 600 and disorder B has a λ_S of 6. What types of disease are A and B?

GENOMICS AND COMPLEX TRAITS

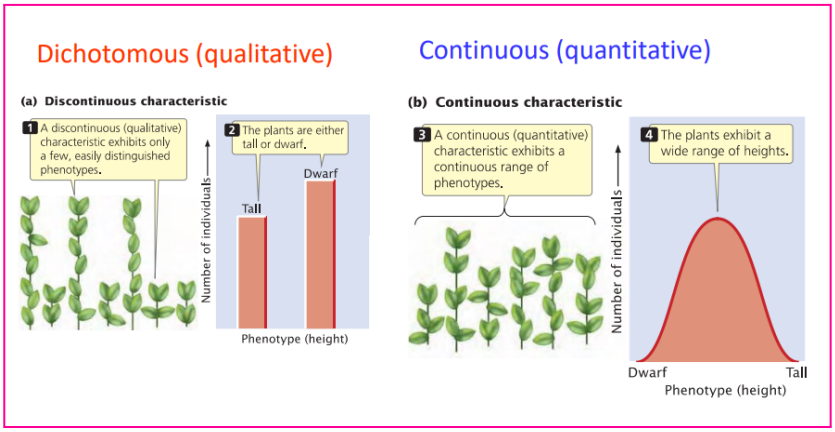
❖ THE DISTRIBUTION OF POLYGENIC TRAITS IS GAUSSIAN

The main idea behind **complex traits** is that there is no clear boundary between them and mendelian traits, rather, **there is a spectrum.**

Here we use complex traits to identify non mendelian traits, making no distinction between polygenic or oligogenic: everything that is not monogenic can be counted as complex.

As we already said:

- ➔ **MENDELIAN** traits are **qualitative** traits, as they come in only two "flavors".
- ➔ **COMPLEX TRAITS** can be either ***dichotomous*** (qualitative) or ***continuous*** (quantitative).
  - ➔ **dichotomous** qualitative traits exhibit only a few, easily distinguishable phenotypes
  - ➔ in **continuous** quantitative traits we don't have clear cut results, rather, we have a spectrum distributed with a gaussian curve.



Mendelian traits can in fact also explain continuous traits. It's just a matter of increasing the number of genes involved.

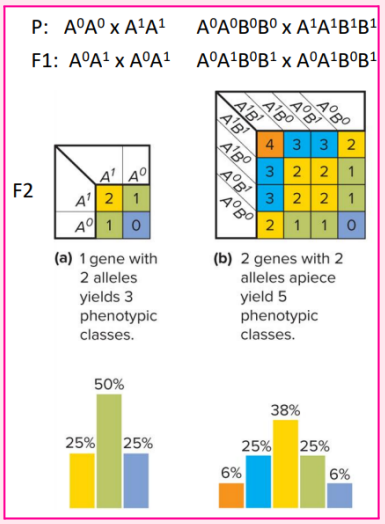
Let's consider A, B, C, … genes, all affecting the height of bean plants. For each gene, two alleles exist:

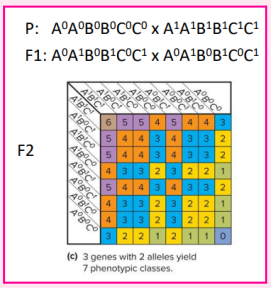
- ➔ 0 allele: contributes nothing to heigh
- ➔ 1 allele: increases the height of a plant of one unit

:

All the alleles are incompletely dominant and have **additive effects**, contributing in different ways to the same trait.

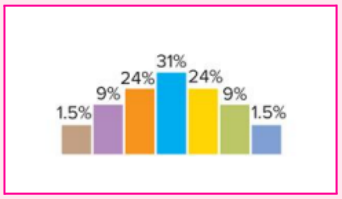
1. In the first case, considering only one gene, we have our usual Punnett square *(mendelian case)*
2. we can input the Punnett square also using two genes, A and B. This still works but we don't have the typical mendelian ratios, as the dominance is incomplete and the effects are additive *(complex case).*





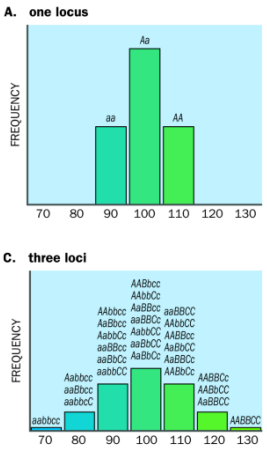
In general **the more genes are involved** the more classes you have and **the more the overall distribution will be similar to a gaussian normal distribution.**

If we add a small contribution from environmental variation, an even smoother curve will appear.



This leads to the **POLYGENIC THEORY by FISHER**. The theory contains the following points:

- A **quantitative trait is influenced by many genes** (polygenes), each behaving in a Mendelian fashion that contribute to the phenotype in a quantitative way (additive effect).
- The **individual effect of each gene on the phenotype is small**
- The additive effects of alleles at several loci produce the quantitative phenotypic variation of the phenotype.
- **Environmental influences may contribute to phenotypic variation.**
  A single genotype produces a range of phenotypes. As a result, the phenotypic differences between genotypic classes become blurred.



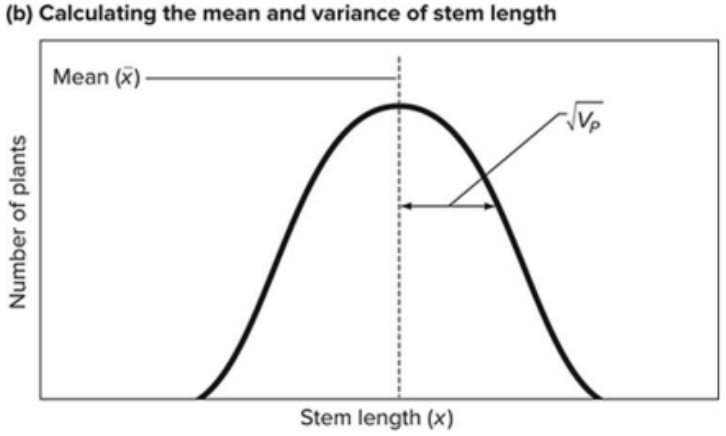
Distribution in the population of an hypothetical continuous character detemined by the additive effects of incompletely dominant alleles

The addition of litte enviromental variation would smooth out the three-locus curve

GAUSSIAN DISTRIBUTION ELEMENTS

The overall **gaussian distribution of the resulting phenotype** is specified by just two parameters:

- ➔ the mean
- ➔ the variance (or the standard deviation). We can also say that variances are additive when they are due to independent causes.



Finding the mean:

Let x_i = the stem length of the plant i in a population of N plants. The mean of stem length, $\bar{x}$, for the population is defined as

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Finding the variance:

The variance V_P of stem length for the population is defined as

$$V_P = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

❖ HERITABILITY: DEFINITION

VARIANCE AND HERITABILITY

The overall **variance of a phenotype V_P** is the sum of the variances due to the individual causes of variation: the **genetic variance** V_G and the **environmental variance** V_E:

➔ **Total phenotypic variance**: $V_P = V_G + V_E$

Heritability (H^2) of a trait is the proportion of **phenotypic variation that is attributable to genetic variation**

➔ **Heritability:** $H^2 = V_G/V_p$

GENETIC VARIANCE AND NARROW SENSE HERITABILITY

Genetic variance can be further partitioned in different components such as: the **additive effect** of genes, the variance of the **dominance** of some genes with respect to others and the variance given by the **interaction** of the genes.

➔ **Genetic variance:** $V_G = V_A + V_D + V_I$

We can also introduce the **Narrow Sense Heritability,** which is the proportion of **phenotypic** variation that results from **additive** genetic variance.

➔ **Narrow sense heritability:** $h^2 = V_A/V_P$

HERITABILITY AND INHERITANCE

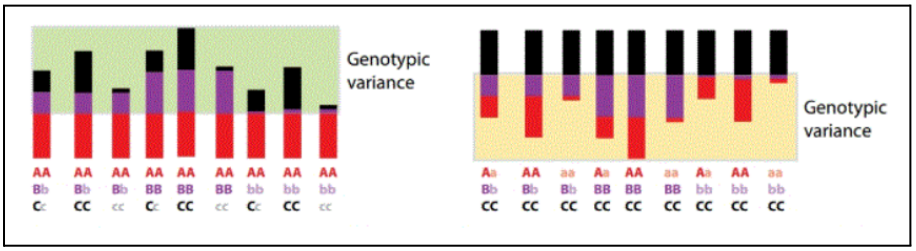
Heritability is not fixed; it always changes depending on the population and the environment. This means that it is different from the mode in which we can inherit traits, as there are different and definite inheritance modes.

Heritability of a trait is always defined for a specific population in a given environment and this **estimation may not be applicable to a different population** with a different genetic substructure and/or different environmental conditions. Instead, **inheritance models are always valid.**

Inheritance models are always valid, whereas heritability depends on the specific dataset that we have.

➔ For example height has different heritability depending on the place: in countries that are fully developed people reach the maximum height they can get, whereas in poor countries many people do not reach their maximum height due to the absence of food or medicines.

In **different populations**, **the same gene** might have **different roles.**



Hence we can go as far as saying that genes that are polymorphic in one population might not be polymorphic in the other.

This extends also with respect to the **environment**, which can have **different roles** depending on the type of population.

- ➔ for example, if we take a population from a first world country and put it in a third world country environment, the effects might be milder.

Hence, we can sum up what **heritability** is and is not:

- ➔ The value of heritability for a character is **specific for a given population in a given environment**. Heritability for a character is not fixed.
- ➔ Heritability **does not indicate the degree to which a character is genetically determined.** It only indicates the degree to which genes determine variation in a character within a defined population.
- ➔ Heritability has no meaning for a specific individual.
- ➔ Even when heritability is high, **environmental factors may influence a characteristic**

❖ HERITABILITY: ESTIMATION

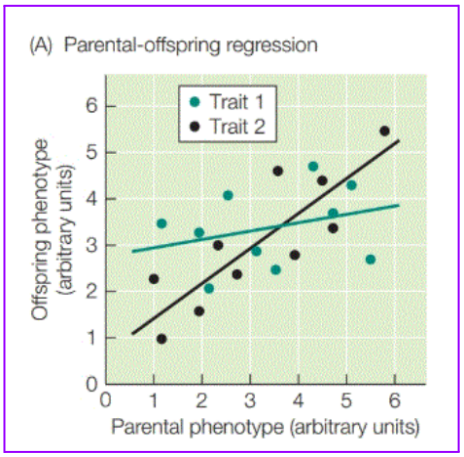
Once we understand what inheritability is, we need to **estimate** it.
We can estimate inheritability looking at two pairs:

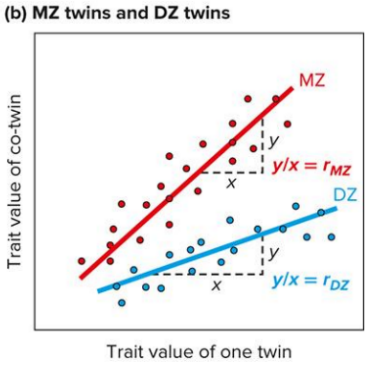
- ➔ parents and children;
- ➔ MZ twins and DZ twins

PARENTS AND CHILDREN

Classically, heritability is measured from the **correlation** between the phenotype of **children** and that of their **parents**.
This correlation is indicated by the **CORRELATION COEFFICIENT**, which is the **slope of the linear regression** that indicates the dependance between parents and children.

- ➔ the more closely the offspring **resembles** its parents, the closer the slope of the line is to **1**;
- ➔ the more **dissimilar** the trait values of parents and offspring, the more the slope of the line moves towards **0**.





TWINS

Phenotypic correlation can be calculated also between **siblings**.

- ➔ If we calculate the phenotypic correlation for the same trait in MZ and DZ sets of twins, we obtain two correlations.

The greater the difference of correlation between the two kinds of twins, the higher the inheritability.

This is because if a trait is environmentally determined, it will have the same effect on children with the same genes and children without the same genes. Instead, if the trait is genetically determined, it'll be more prominent in children who have the same genotype. Hence the correlation coefficients for MZ and DZ will be different.

The advantage of twin studies, is that the total variance can be split up into: genetic, shared or common environmental and unique environmental components, enabling an accurate estimation of heritability.

FALCONER'S EQUATION can be applied only to twin studies, and is able **to reduce the contribution of shared environments to heritability.**

Falconer's equations (removes the contribution of the shared environments to heritability):

$H^2=2(r_{MZ}-r_{DZ})$

$c^2=2r_{DZ}-r_{MZ}$

$e^2=1-r_{MZ}$

c^2: effect of shared environment
e^2: unique environmental effects

Phenotypic variation between twins can be due to:

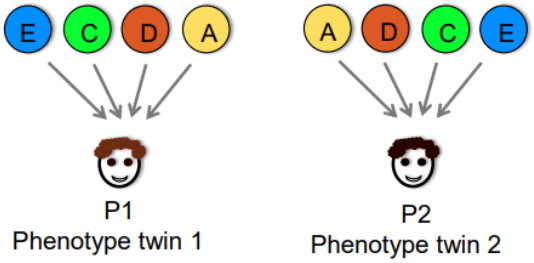
1) Genetic factors
   - A additive genetic effects, V_A or a^2
   - D non-additive genetic effects, (V_D+V_I) or d^2

2) Enviromental factors
   - C effect of shared environment, c^2
   - E unique environmental effects, e^2

$V_E=c^2+e^2$

The total variance is the sum of this four components: A, D, C, E

$$V_P = V_G + V_E = V_A + V_D + V_I + V_E = a^2 + d^2 + c^2 + e^2$$



Let's just consider additive genetic effects

Remembering that:

- A is the additive genetic effect
- D is the non additive genetic effect
- C is the effect of shared environment
- E is the effect of unique environmental effects.

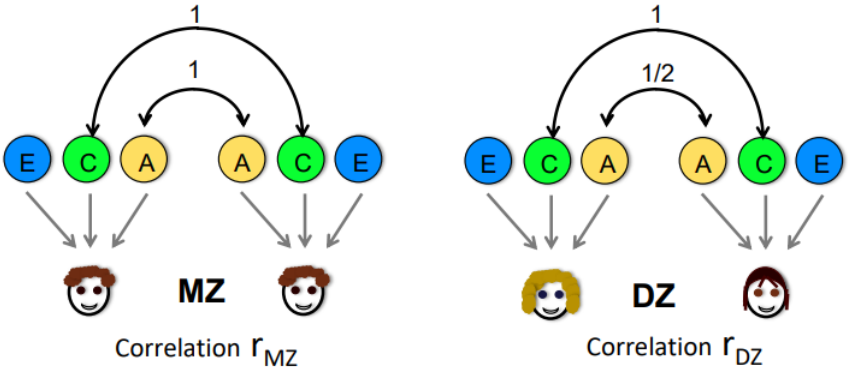
$V_P = a^2 + c^2 + e^2$

A additive genetic effects, V_A or a^2
C effect of shared environment, c^2
E unique environmental effects, e^2

The correlation (r) between MZ and DZ twin is the sum of the additive genetic effects and shared environmental effects:

$r_{MZ} = a^2 + c^2$
$r_{DZ} = 1/2\ a^2 + c^2$
$e^2 = 1 - r_{MZ}$

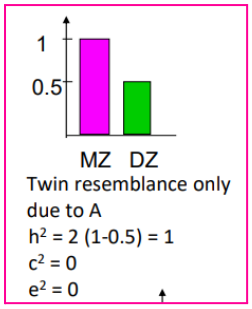
$\mathbf{a^2 = 2\ (r_{MZ} - r_{DZ})}$
$\mathbf{c^2 = 2\ r_{DZ} - r_{MZ}}$
$\mathbf{e^2 = 1 - r_{MZ}}$



Where:

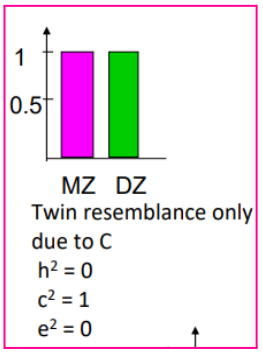
- $r_{MZ} = a^2 + c^2$ is the **concordance for MZ twins**
- $r_{DZ} = 1/2a^2 + c^2$ is the **concordance of DZ twins**, we have ½ because they share 50% of the same genes, and so we have only ½ of additive gene effect, while the shared environment effect remains the same since the DZ twins are raised together like the MZ twins.

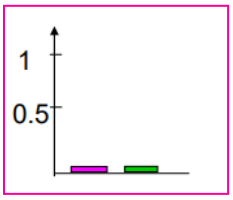
HERITABILITY ESTIMATES FROM TWIN STUDIES



In this first graph**, monozygotic concordance is equal to 1 while for dizygotic concordance is 0.5** so here the trait is only genetically determined. There is no effect on the environment otherwise it would have changed the situation of the monozygotic twins, which is however equal to 1. This result can be obtained also thanks to the falconers equation as we can see in the result.

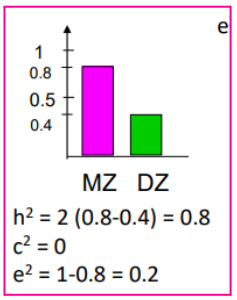
In this second graph, the two **concordances are the same**. This means that no heritability is present which means the actor is the shared environmental condition.

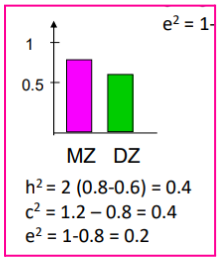




When there is no resemblance at all, the **correlation is absent**. No heritability is involved, only 100% environmental uniqueness.

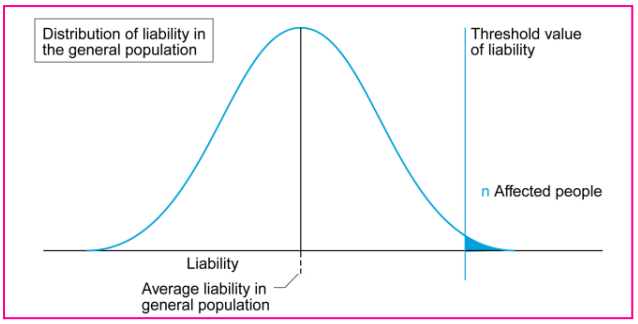
With **0.4 and 0.8** we have an effect of the genes cus the heritability isn't equal and it is different from zero. Since dizygotic **correlation is still half of the monozygotic one, it means that there is no environment playing** at hand. The only important thing is heritability, no environmental changes are present
7





In this case we have correlations 0.8 and 0.6. Since the **concordance of dizygotic twins is not the half of monozygotic twins**, we say we have both heritability and environmental influences.

❖ THRESHOLD MODEL



As we have already said, complex traits can also be dichotomous. In those cases we have susceptibility loci. In this case **a threshold is postulated**.

➔ a threshold is a specific **value of the general population liability.**

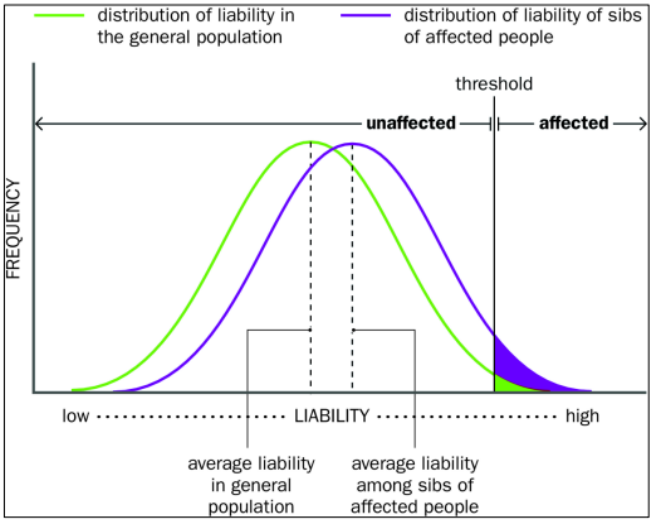
➔ If the susceptibility of an individual is **below the threshold it will not develop the disease.**

We define a **LIABILITY** as the amount of genetic contributors to a determinate risk in the general population. It's polygenic and follows a Gaussian distribution in the population.

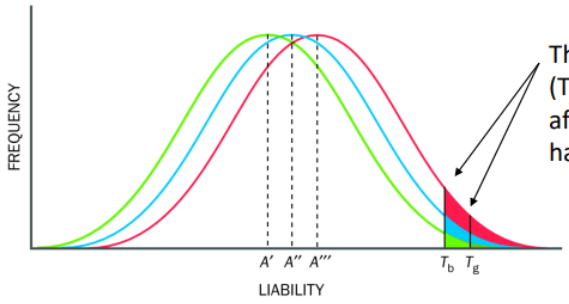
The **distribution of liability among sibs** of an affected person is:

➔ **shifted towards higher liability** because they share genes with their affected sibs.

➔ A greater proportion of them have liability above the threshold: the condition tends to run in families. Hence, the **threshold must go down** since the affected individuals increment.



Congenital pyloric stenosis is 5 time more common in boys than girls.



The threshold is higher for girls (T_g) than for boys (T_b): to be affected a girl must on average have a higher liability than a boy.

Relatives of an affected girl therefore has an higher average liability than relatives of an affected boy.

TABLE 3.2 RECURRENCE RISKS FOR PYLORIC STENOSIS

Relatives of	Sons	Daughters	Brothers	Sisters
Male proband	19/296 (6.42%)	7/274 (2.55%)	5/230 (2.17%)	5/242 (2.07%)
Female proband	14/61 (22.95%)	7/62 (11.48%)	11/101 (10.89%)	9/101 (8.91%)

More boys than girls are affected, but the recurrence risk is higher for relatives of an affected girl. The data fit a polygenic threshold model with sex-specific thresholds (Figure 3.28). [Data from Fuhrmann and Vogel (1976) Genetic Counselling. Springer.]

Thresoholds might also be sex sensitive: some illnesses are more characteristical of males rather than females or vice versa.

EXERCISE 6

Exercise 6

A study published in 1937 examined the average differences between pairs of twins (MZ or DZ) and pairs of siblings for 3 different quantitative traits: height, weight and IQ as measured by the Stanford_bined test.
Some MZ twins were raised together in the same house (RT), while others MZ twins were raised apart in different families (RA).
The results of this study, shown as average differences, are as follows:

	MZ(RT)	MZ(RA)	DZ	Siblings
Height	1.7 cm	1.8 cm	4.4 cm	4.5 cm
Weight	1.86 kg	4.49 kg	4.54 kg	4.72 kg
IQ	5.9	8.2	9.9	9.8

➔ Which of these 3 traits appears to have the highest heritability? The lowest heritability?

The trait that has the highest heritability is the height. This is because it's the trait that has the biggest difference between dizygotic twins and monozygotic ones, which share all genes.

The lowest heritability is the one of the weight, because it's the one where **MZ RA and DZ twins have the smallest difference. Hence, it's where genetics matters the least.**

Hence, in this exercise we should remember to use al of the data we have: MZ RT, MZ RA and DZ.

Question 7

Which of the following statements would be true of a human trait that has high heritability in a population of one country?

a) The phenotypic difference within MZ twin pairs would be about the same as the phenotypic differences among members of DZ twin pairs.

b) Very little phenotypic variation exists between MZ twins but high variability exists between DZ twins.

c) The trait would have the same heritability in a population of another country

a. F, it should be half
b. T
c. F, we know heritability depends on the specific environment and population

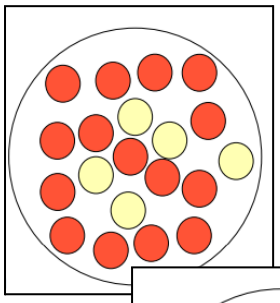
POPULATION GENETICS

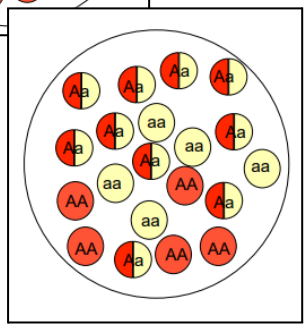
❖ THE DISTRIBUTION OF POLYGENIC TRAITS IS GAUSSIAN

POPULATION GENETICS deals with **changes in the genetics of a population over time** and how these changes occur.

A **POPULATION** is a group of **interbreeding individuals that live in the same time and space**.
For a population we can define:

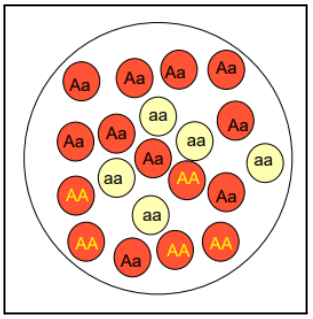
- ➔ a **GENE POOL** is the **total of all alleles** carried in all members of a population
- ➔ a **SAMPLE** is a number of individuals used to make **inference** about the entire population





Other important topics regarding population genetics are:

- **PHENOTYPE FREQUENCY,** which is the proportion of individuals in a population that have a **particular phenotype**.
- **GENOTYPE FREQUENCY**, which is the proportion of individuals in a population that carry a **particular genotype**
- **ALLELE FREQUENCY**: **proportion of a specific allele type in a population**. In this case the denominator is not the population size but the total number of alleles present per allele type.



As we can see, the calculations for allele frequency and genotype frequency are different:

genotypes:

AA	Freq (AA) = N_{AA}/tot
	5/20 = 0.25
Aa	Freq (Aa) = N_{Aa}/tot
	10/20 = 0.5
aa	Freq (aa) = N_{aa}/tot
	5/20 = 0.25

Freq A (p) = N_A/N_{tot}
$p = (2N_{AA} + N_{Aa}) / 2(N_{AA} + N_{Aa} + N_{aa}) =$
$= (2x5 + 10)/40 = 0.5$

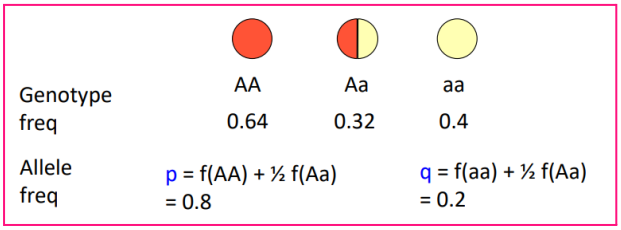
Freq a (q) = N_a/N_{tot}
$q = (2N_{aa} + N_{Aa}) / 2(N_{AA} + N_{Aa} + N_{aa}) =$
$= (2x5 + 10)/40 = 0.5$

or:
$p = f(AA) + ½\ f(Aa)$
$q = f(aa) + ½\ f(Aa)$

$p + q = 1$

❖ HARDY WEINBERG EQUILIBRIUM

The **HARDY WEINBERG EQUILIBRIUM** provides a **relationship between allele frequency and genotype frequency** for an **idealized large population**.



If **no evolutionary pressure** that alters allele frequencies applies to a population, **genotype frequencies** remain stable from one generation to another and they are directly **determined by allele frequencies** (p, q)

As we can gather from the definition of the HWE, in order to use this law the following **assumptions** must be true:

- The population is large
- Individuals **mate at random**
- **No new mutations** (negligible)
- **No migration** (negligible)
- No selection (genotypes have no effect on ability to survive and transmit alleles to the next generation)

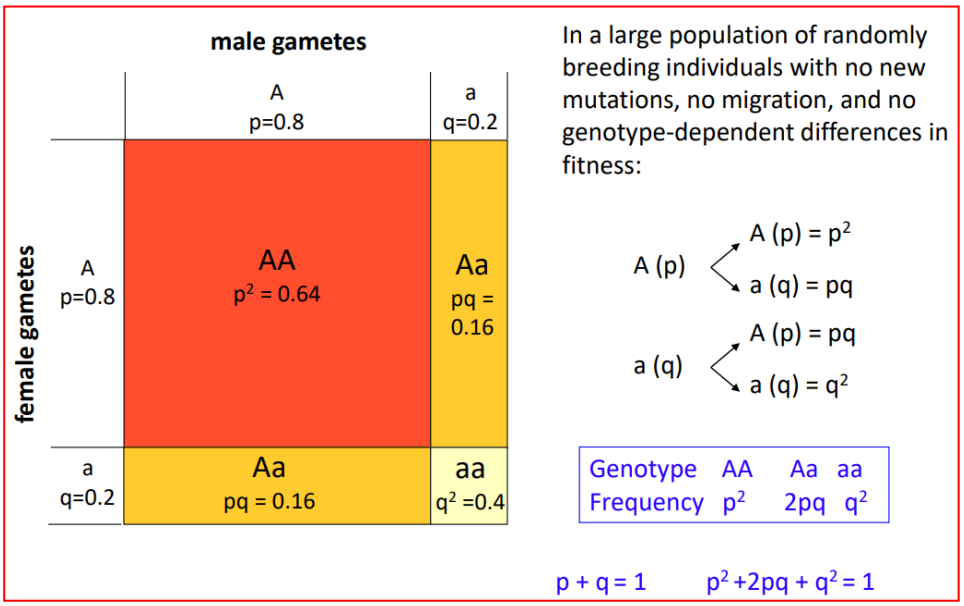
Hence:

- ➔ From the allele sequence we can get genotype sequence
- ➔ Since we are making some assumptions in order for this to work, and
- ➔ no population is a perfect fit for this.
- ➔ If the assumptions are not valid **we can still use this equilibrium law if we use it over a short amount of time** (just a few generations).

In an idea situation the Hardy Weinberg law gives us the following genotype sequences:

We know that, given two allele frequency p and q:

- ➔ $p + q = 1$ so the square of the sum should also be equal to 1:
- ➔ $p^2 + q^2 + 2pq = 1$



❖ CHECKING FOR HARDY-WEINBERG EQUILIBRIUM

Let's see an example where we have to test for HWE.

The exercise begins with giving us the **observed genotype** frequency

- ➔ AA = 15
- ➔ aA = 30
- ➔ aa = 55

We can calculate **allele frequencies**, where n is the population (=100), and n*2 is the total number of alleles, since every individual in the population has two.

- **p(A)** = (AA*2 + Aa) = (15*2 + 30) / n*2 = 0.3
- **q(a)** = (aa*2 + Aa) = (55*2 +30) / n*2 = 0.7

An **allele count** would be given by the allele frequency times n*2.

We can also calculate the **genotype frequency** with HWE from the allele frequency to see if the **expected genotype** is the same as the observed one. Remember that to calculate the genotype frequency you need the allele frequency, not the allele count!

- **[AA]** = $p(A)^2 * n$ = 9
- **[Aa]** = 2*p(A)*q(a)*n = 42
- **[aa]** = $q(a)^2 * n$ =49

What we obtain is the genotype we would have if the HWE were valid. If the observed and the expected genotype (obtained with HWE) are the same, then HWE is valid.

To decide how similar the results are and thus decide if we have the same frequencies or not, we use the **CHI SQUARE VALUE**. Considering that for HWE we only have one degree of freedom:

	AA	Aa	aa
Observed genotypes	15	30	55
Expected genotypes	9	42	49

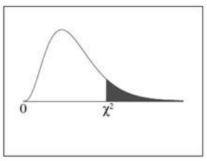
Using the χ^2 **goodness of fit test**

$\chi^2 = \Sigma (obs - exp)^2/exp$

Chi2= 36/9 + 144/42 + 36/49 = 8.16 (1 d.f.) p = 0.0043

Observed and expected values are significantly different
We reject H_0:
The observed genotype frequencies at this locus are NOT in Hardy Weinberg equilibrium

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955

0.05 significance critical values

Given a significance value of 0.05, we can see that the p value p=0.0043 is a lot smaller.

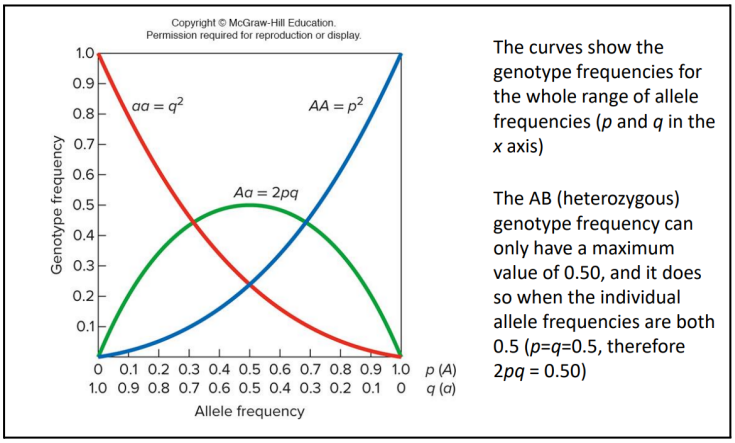
Hence **we reject the null hypothesis.** click here for video to remember how to use it

We can conclude that **HWE is not valid:** observed data does not collide with the theoretical one. This could be because of the small sample size.

In general we can say that the **CHI-SQUARED TEST** is not a very powerful tool, and the **reliability of the result rests on the sample size.**

- Healthy population showing signs of deviance from HWE, the most plausible explanation is not that the population is unhealthy, but that there are genotyping errors. This mostly happens if the sample size is too short.

Hence, **HWE is very powerful**: **if it doesn't work in genome wide association studies (GWAS), it probably means that there were genotyping mistakes.**



The curves show the genotype frequencies for the whole range of allele frequencies (*p* and *q* in the *x* axis)

The AB (heterozygous) genotype frequency can only have a maximum value of 0.50, and it does so when the individual allele frequencies are both 0.5 (p=q=0.5, therefore $2pq$ = 0.50)

10000 unrelated Chinese individuals were typed for the MN blood group

Observed genotypes

MM	MN	NN
3502	4997	1501

Do these frequencies fit HWE proportions?

Sample size (n) = 10 000; Total alleles (2n) = 20 000

Allele frequencies:

- **p** =(3502 x 2+ 4997)/ 20000 = 0.6
- **q** = (1501 x 2 + 4997)/ 20000 = 0.4

Expected genotype frequencies under HWE:

- **[AA]** = $\mathbf{p^2}$ x 10000 = 0.62 x 10000 = 3600
- **[Aa]** = **2pq** x 10000 = 2 x 0.6 x 0.4 x 10000 = 4800
- **[aa]** = $\mathbf{q^2}$ x 10000 = 0.42 x 10000 = 1600

Checking for Hardy-Weinberg equilibrium

	MM	MN	NN
Observed genotypes	3502	4997	1501
Expected genotypes	3600	4800	1600

Using the **χ^2 goodness of fit test**

$\chi^2 = \Sigma (obs - exp)^2/exp = 16.9$

➔ Observed and expected values are **significantly different with the calculated Chi Square value being significantly higher than the expected** one for 0.05

➔ We **reject H_0** : The observed genotype frequencies at this locus are NOT in Hardy Weinberg equilibrium

Application of HWE for calculating genetic risk in single gene disorders

AUTOSOMAL RECESSIVE DISORDER
→ Estimate of carrier frequency

Disease population frequency = 1/2000
What is the carrier frequency?

	Unaffected		Affected
Genotype	AA	Aa	aa
Frequency	p^2	2pq	q^2

all different mutant alleles in the disease gene can be lumped together into a single "disease" allele

assumtions:
random mating
constant allele frequencies

$q^2 = 1/2000$ $q = 0.02$
$p = 1- q = 0.98$
$2pq = 2 \times 0.98 \times 0.02 = 0.039 \approx 4/100$

HWE is the most important law in population genetics but it also has other applications:

➔ In GWAS it's used as a control for genotyping errors.

➔ It is **also used to calculate genetic risk in single autosomal (recessive) disorders.**

For example we have a disease frequency of 1/2000. What is the carrier frequency? The disease frequency corresponds in the slide to q^2, which is the genotype for disease. This gives us q, from which we can get p.

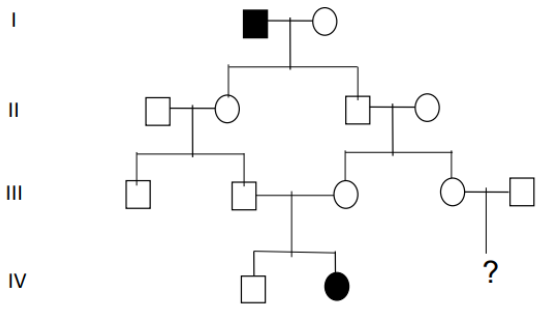
For AUTOSOMAL RECESSIVE DISORDERS, **the frequency of the disease is equal to the frequency of the affected genotype.**

q^2 = population frequency for the disease

EXERCISE 2: HWE CAN EXPLOIT THE POPULATION RISK in AUTOSOMAL DISORDERS

Maple syrup urine disease (MSUD) is autosomal recessive and causes intellectual and physical disability, difficulty feeding and a sweet odor to urine. In Costa Rica, 1 in 8,000 newborns inherit the condition.

a) I-1 and IV-2 are affected by MSUD. What is the probability for III-4 of being a carrier?
b) What is III-4's chance of having an affected child if her partner is an unaffected man from Costa Rica?



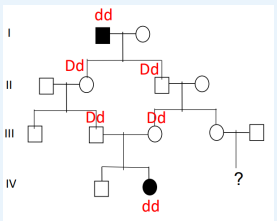
a. III-4: ½

b. **q^2** = (1/8000) is the probability of being affected. Hence, by **assuming that HWE works:**

q = $\sqrt{(1/8000)}$ = 1/89 = 0.011
p = 1- 0.011 ~ 1

possibility of the father being a carrier is given by:
Aa: **2pq** = 2 * 1* 1/89 = 0.2/89 (carrier)

Hence the total probability of IV being affected, considering it must take the disease allele from the mother (½ chance of being a carrier) and from the father, which must be heterozygous::
1/2 * 2/89* 1/4 = 1/356 =0.00275



You can't use 1/8000 directly on the father cus it's the probability of being *affected*, not the probability of being *heterozygous*.

EXERCISE 3: X-LINKED LOCI

The incidence of X-linked hemophilia A is 1 in 10,000 male births.

a. What is the frequency of carrier females?

In this case we don't have q^2 but q because the allele is present only one time for the disease to show in males.

Hence: **q** = 1/10000 (male affected, X^{ill}).

Consider that q stands for the frequency of X with the disease genotype, hence, p will be the allele frequency for X without the disease genotype

p = 1- 1/10000 = 1 (allele frequency for $X^{healthy}$)

Female carriers will have heterozygous genotype:

$\mathbf{X^{ill}X^{healthy}}$ = **2pq** = 2 x 1 x 0.0001 = 0.0002 (female carrier)

b. What is the frequency of affected females?

For females it will be q^2 because you need that allele two times.

Hence: $\mathbf{q^2}$ = 1/100 million (female affected)

❖ DEVIATIONS FROM HARDY WEINBERG EQUILIBRIUM

Here are some factors that are **able to alter the HWE equilibrium:**

- ➔ non random mating
- ➔ mutations
- ➔ genetic drift
- ➔ natural selection

When two populations mix the HWE doesn't work anymore, but it takes just one generation for HWE to be re-established

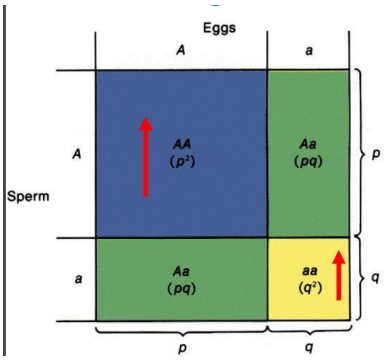
NON RANDOM MATING (INBREEDING)

Inbreeding is the only deviation that **changes the genotype frequency**, the others all change only the allele sequence.

Non random mating is not only composed by incest. It's also made of people mating with the same ethnicity, tribe or cultural group. Sharing the same blood is only the most extreme example.

In general consanguineous mating:

- ➔ results in an increased frequency of homozygous genotypes and a decrease of heterozygous genotype
- ➔ results in an increased frequency of mating between carriers and a correspondingly increased frequency of autosomal recessive disease.



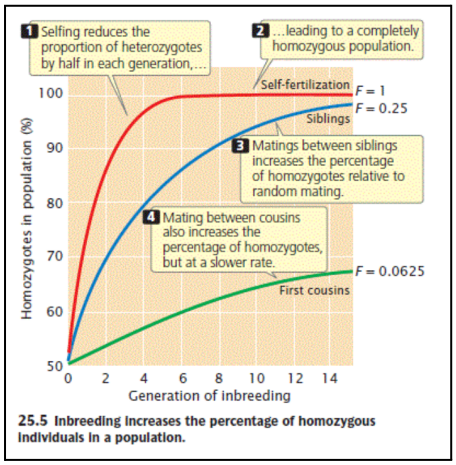
As we can gather from what is written above; genotype frequencies change whereas **allele frequencies remain unchanged.**

As you can see each generation that passes the number of heterozygous diminishes. This extreme would lead to a **completely homozygous population.**

We can see a prove of the "all-homozygous" theory looking at the extreme case of self breeding:

- ➔ with **self breeding** (AA x AA, Aa x Aa, aa x aa), we can see that the number of homozygous increases each time while the number of heterozygous diminishes. This would eventually lead to a population with ½ AA and ½ aa.

Generation	AA	Aa	aa
1	¼	½	¼
2	3/8	1/4	3/8
3	7/16	1/8	7/16
4	15/32	1/16	15/32
N		$1/2^N$	



25.5 Inbreeding increases the percentage of homozygous individuals in a population.

➔ looking at the graph, even if we don't consider something as extreme as self-fertilization, but only **sibling breeding,** the same thing happens at a lower rate. We see that the mating between siblings increases the percentage of homozygotes relative to random mating.

This would also happen, at an even lower rate, with first cousins. In general, it happens every time the breeding is not random.

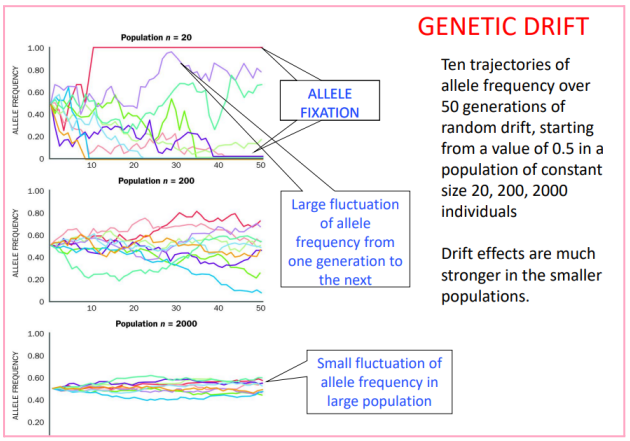
MUTATIONS

Mutations on the other hand are very rare (they **occur at a very low rate**) so they don't change the genotype frequency.

Mutations **create new alleles** they contribute to evolution. However the altering of the allele frequency doesn't happen in only a few generations.

GENETIC DRIFT

Random **variation in allele frequencies due to sampling error** from one generation to the next is called genetic drift. Hence, we say that the genetic drift is a **non-adaptive, random change**. Unlike mutations, genetic drift alters very heavily the allele frequency.

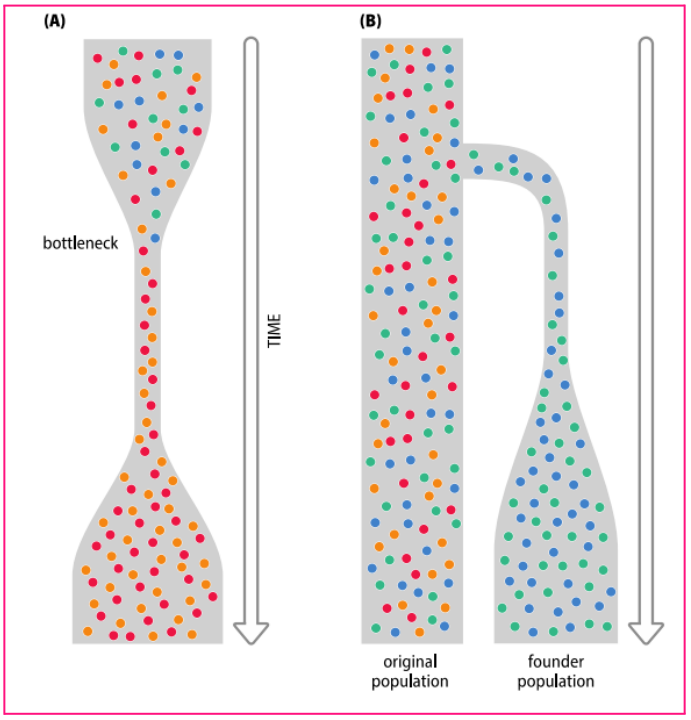


➔ These random effects are **more deleterious for small populations.**
In the graphs we can see that, depending on the size of population. the genetic drift changes. In all cases 50 generations and 10 drifts were considered. As we can see the overall effect of the drift is not really visible in large populations but really small in large populations.

There are two elements that influence genetic drift:

- **population bottlenecks**: **severe reduction in size** before the population starts increasing again. After this effect the genetic variety is reduced by a lot.
- **founder effects**: in this case **a few individuals separate from a larger population** and establish a new colony. The new colony has a subset of the previous genetic variation, hence genetic variety is once again reduced.

Hence, in both cases the problem is that **at some point the population consists of only a small number of individuals.**



A problem that might arise is the increased appearance of s**ome disorders that become prominent because of the lack of genetic variety.** An example of this is the Amish community.

An example of the Founder Effect is the ***Out of Africa Hypothesis.***

- ➔ According to the Out of Africa Hypothesis, **all humans originated in Africa and then expanded from there into the whole world.**

We can understand this thanks to the fact that **genetic variation is higher in Africa** than in other countries. Hence, few people parted from Africa, bringing little genetic variety into their new world.

EXERCISE on MIGRATION

A population of water snakes is found on an island in Lake Erie. Some of the snakes are banded and some are unbanded; **banding is caused by an autosomal allele that is recessive** to an allele for no bands.

- ➔ the frequency of banded snakes on the **island** is 0.4,
- ➔ the frequency of banded snakes on the **mainland** is 0.81.

One summer, a large number of snakes migrate from the mainland to the island. After this migration, **20% of the island population consists of snakes that came from the mainland.**

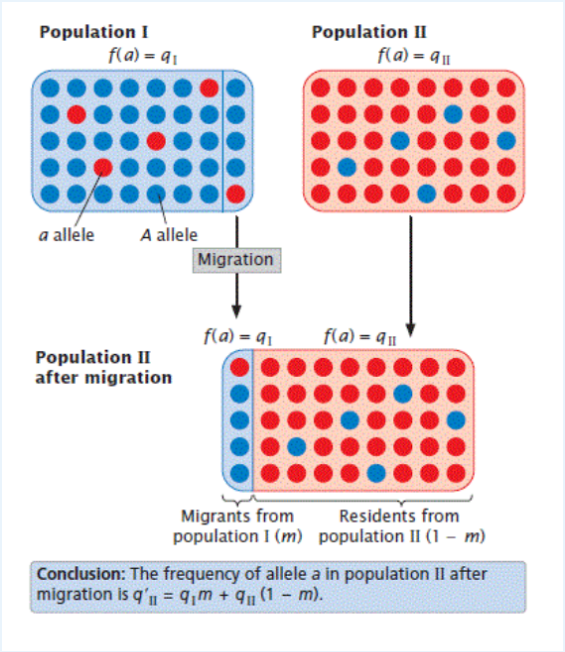
a. If both the mainland population and the island population are assumed to be in Hardy–Weinberg equilibrium for the alleles that affect banding, what is the frequency of the allele for bands (a) on the island and on the mainland before migration?

For the island:

- ➔ $p^2 = 0.4$ (aa)
- ➔ $p(a) = \sqrt{0.4} = 0.63$

For the mainland:

- ➔ $q^2 = 0.81$ (aa)
- ➔ $q(a) = \sqrt{0.81} = 0.9$ (a)



b. After migration has taken place, what is the frequency of the banded allele on the island?

$p_f(a) = 0.8*p(a) + 0.2+q(a) = 0.504 + 0.18 = 0.684$

REMEMBER: the frequency of an allele in a population *after* migration is given by:

$$q_2' = q_1*m + q_2(1-m)$$

where m is the percentage of migrated individuals.

*In this problem the population number is not necessary as we already have the phenotype *frequency* and the genotype frequency is not required.

This chapter is particularly important as in the exam there will be questions about checking the HWE, and about going from allele frequency to genotype frequency. There are **two main types of exercises**:

- ➔ checking for HWE
- ➔ using HWE for autosomal disorders

NATURAL SELECTION

Natural Selection **is a process that progressively eliminates individuals whose fitness is lower than others.**

- ➔ **FITNESS**: an individual's ability to survive and transmit its genes to the next generation.

Because of natural selection, **only individuals with the highest fitness** (highest survival capability) **survive**. That is because **individuals whose fitness is higher are more likely to become the parents** of the next generation.

We say that natural selection is selection ***against*** a certain genotype.

In populations undergoing selection, **each genotype has a relative fitness.**

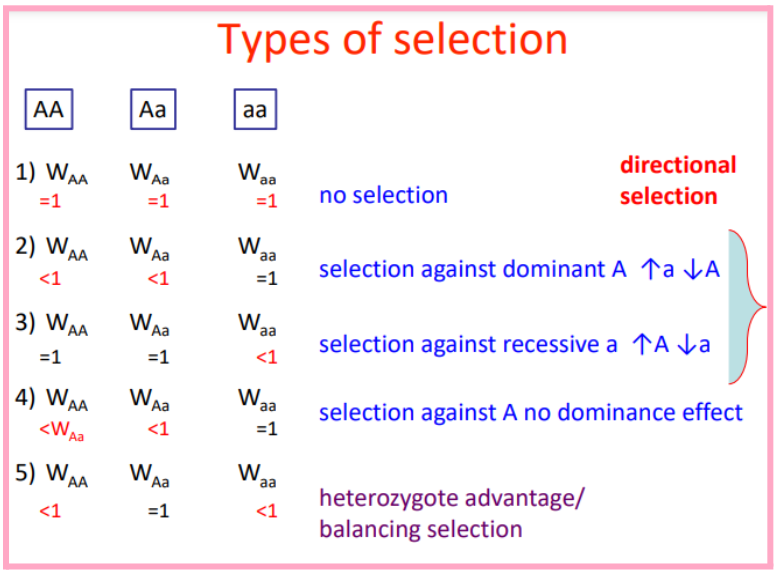
- **W** is the relative **FITNESS OF A GENOTYPE**
  - W=1 → genotype with highest reproductive success
- **s = 1-W** is the **SELECTION COEFFICIENT**
  - s=1 → the genotype is genetically lethal

W_{AA}	W_{Aa}	W_{aa}
1	0.8	0.2
s_{AA}	s_{Aa}	s_{aa}
0	0.2	0.8

The higher the fitness, the lower the selection coefficient against that fitness.

Example on a population with two alleles (A and a)↲

Depending on the values of s and W, we can see different types of selections:



❖ DYNAMIC EQUILIBRIUM

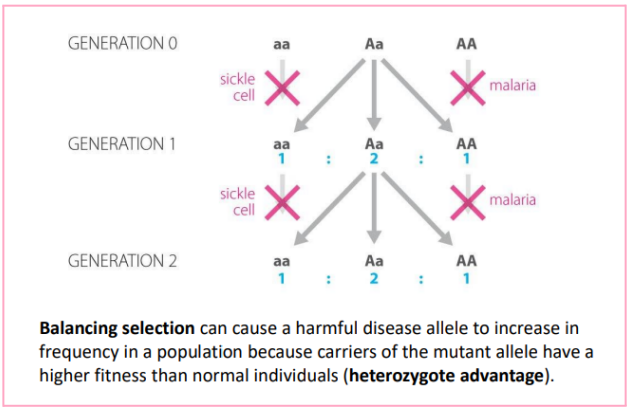
A dynamic equilibrium is created when the population simultaneously tends to **change in opposite directions**. Two opposing evolutionary forces can create a dynamic equilibrium, balancing each other out.

A dynamic equilibrium **can guarantee HWE** even if dynamic.
Two examples:

- **MUTATION SELECTION BALANCE**: **selection eliminates new deleterious alleles** that are introduced in the population by new mutations. When thinking about it, this hypothesis is much more probable than thinking that mutation rate is much lower.

- **BALANCING SELECTION (HETEROZYGOUS ADVANTAGE):** Selection favors the heterozygotes at the **expense of each type of homozygous in the population**.

Heterozygous advantage encourages **MAINTAINING OF DELETERIOUS ALLELES** in a population under the heterozygous genotype. Usually the recessive homozygous s for these types of alleles are lethal, while the dominant homozygous don't have benefits heterozygotes have.



Balancing selection can cause a harmful disease allele to increase in frequency in a population because carriers of the mutant allele have a higher fitness than normal individuals (**heterozygote advantage**).

An example of heterozygous advantage is ***sickle cell anemia***. The allele is very common in areas where malaria was endemic, while it's completely absent in areas where malaria was never present.

- ➔ **recessive homozygous** have a fitness of 0 because they die at an early age.
- ➔ **dominant homozygous** have no mean against malaria, hence they die easily in endemic areas.
- ➔ **heterozygous individuals** have a fitness of 1. This is because sickle cell heterozygous individuals have red blood cells that are inhospitable for malaria parasites.

KEY POINTS

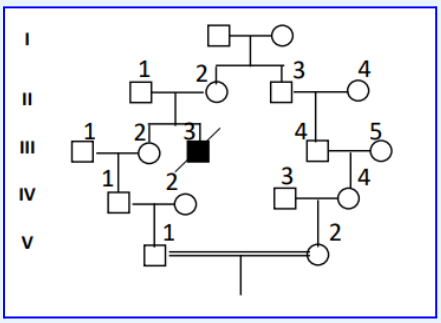
Hardy-Weinberg equilibrium indicates that **no evolutionary forces act on a population.** In such populations, genotype frequencies can be estimated directly from allele frequencies.

- ➔ **allele frequency can always be derived from the genotype,** but the **genotype** can be derived by **allele** frequency **only if** HWE is valid.

- Theoretically, **any deviation** of observed genotype frequencies from expected frequencies (calculated with the allele frequency) **indicates evolutionary forces** or genotyping errors.

- In reality it is very rare that there are no evolutionary forces. It is much more possible that there are many **deviations** creating **a dynamic equilibrium** in which there is no net change in allele frequencies.

FINAL EXERCISE on HWE



III-3 is affected by a **rare autosomal recessive** disease with a population frequency of about 1/40,000.

a. What is the probability for V-2 of having an affected child if her partner is V-1?
We **assume all external people to be dominant homozygous** because it's a rare disease. In practice this means that we do not consider q nor p for this first question.

- II2 = **⅔.** This is because we know its **parents to be both heterozygous**. Its possible genotypes will be AA, Aa, aA. Only two of them are favorable for our calculations since we need to assume that II2 has one recessive allele.
- IV1= **½.** This is because its possible genotypes are Aa or AA since one parent is **homozygous and the other one is heterozygous.**
- V1 = ½
- II3 = ½
- III4 = ½
- IV4 = ½
- V2 = ½
- ¼ probability of the child inheriting

The child has the following probability of being ***affected***:
⅔*½*½*½*½*½*½*¼ = 1/384

b. What is V-2's chance of having an affected child if her partner is an unaffected man from the same population?

- q^2 = 1/40 000
- q =1/200
- p = 1- 1/200 = 199/200 ~1
- 2pq= 2* 1* 1/200= 1/100
- 1/100* ¼*½*½*½*½ = 1/6400

QUESTIONS

1. SELECTION
2. INBREEDING AND FOUNDER EFFECT
3. MIGRATION
4. FOUNDER EFFECT

Exercise

By which mechanisms do the following situations alter HWE?

a) In ovalocytosis, a protein that anchors the red blood cell plasma membrane to the cytoplasm is abnormal, making the membrane so rigid that the parasites that cause malaria cannot enter.

b) In the mid-eighteenth century, a multi-toed male cat from England crossed the sea and settled in Boston, where he left behind many kittens, about half of whom also had extra digits. People loved the odd felines and bred them. Today, in Boston, multi-toed cats are much more common than in other parts of the United States.

c) Many slaves in the United States arrived in groups from Nigeria, which is an area in Africa with many ethnic subgroups. They landed at few sites and settled on widely dispersed plantations. Once emancipated, former slaves in the South were free to travel and disperse.

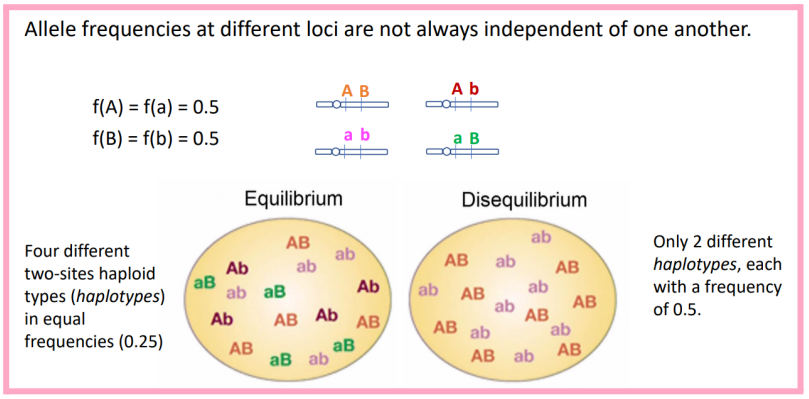
d) About 300,000 people in the United States have Alzheimer disease caused by a mutation in the presenilin-2 gene. They all belong to five families that came from two small villages in Germany and migrated to the United States from 1870 to 1920.

1LINKAGE DISEQUILIBRIUM

❖ LINKAGE DISEQUILIBRIUM & LINKAGE

Allele frequencies at different loci are **not always independent** of one another.

➔ given gene A with certain alleles, do the **alleles of a gene B depend on the alleles of gene A?** If so, it means they are in LD, so we can predict the alleles of gene B depending on the ones from gene A.



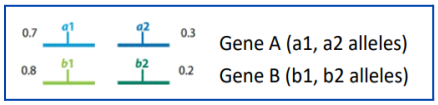
- In an **EQUILIBRIUM** situation we have **no correlation** between the alleles at the two sites, so the variation at these two sites is in linkage equilibrium.
- In a **DISEQUILIBRIUM** situation we have a perfect **positive** correlation between variants at the 2 sites, variation at these two sites is in linkage disequilibrium (LD).

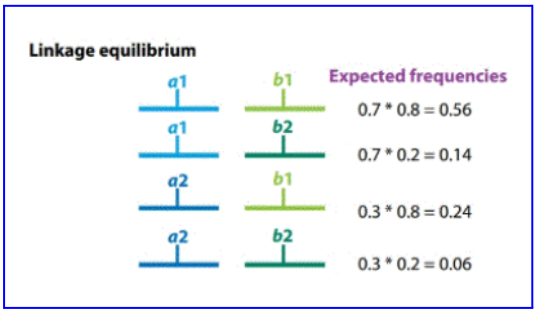
The **equilibrium is the state of total independence**, a random combination of alleles at different loci. Disequilibrium consists of any deviation from the equilibrium state, hence every time there is a correlation between the types of alleles that are present.

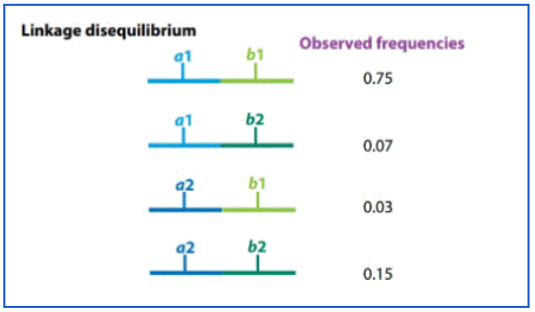
LINKAGE DISEQUILIBRIUM IS THE NON-RANDOM ASSOCIATION OF ALLELES AT TWO OR MORE LOCI.

➔ **HAPLOTYPES** are **combinations of alleles** at two or more loci on the same chromosomal segment, **whose frequencies may be not predictable from the individual allele frequencies**

If the alleles for the 2 genes are in **EQUILIBRIUM**, the frequency of each genotype is the product of the frequencies of the two alleles. If the alleles are linked, hence we are in **DISEQUILIBRIUM**, the frequency of each genotype is NOT equal to the product of the frequencies of the two alleles.



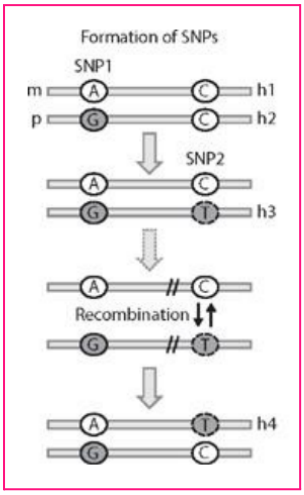




Linkage disequilibrium is a **statistical association** between particular alleles at *separate but linked loci*. This statistical element is based on the genetic effect of **GENETIC LINKAGE.**

➔ genetic linkage is the effect for which, during **meiosis**, **genes that are close to each other on a chromosome will often segregate together**

LINKAGE DISEQUILIBRIUM ARISES BECAUSE GENES ARE LINKED



The T allele of SNP2 **will be always associated** with allele G of SNP1 **until a recombination event breaks this exclusive relationship.**

The closer the two loci are, the harder it will be for them to separate.

Until then, alleles G and T will always be on the same haplotype in this population: this non-independence generates linkage disequilibrium.

Meiosis occurs once per generation, so a recombination event is possible once per generation. LD **decays as a function of the number of generations since the occurrence of a new mutation.**

There can be **various combinations of haplotypes as generations go on.** For example in the picture above we first have GC haplotype, then GT, then GC again because of recombination.

❖ MEASURES OF LINKAGE DISEQUILIBRIUM

LD is a **quantitative phenomenon**: the relationship between alleles at two loci can lie anywhere on a spectrum from completely unrelated to perfectly correlated.

We have 3 measures for LD:

➔ **D value** (Δ value): is the **difference** between the observed and expected value of haplotype frequencies.

$$\Delta \text{ value} = P_{AB} - P_A P_B$$

If there's equilibrium, **Δ=0.** D value is the simplest measure of LD and it makes intuitive sense, but it is not widely used because of its **sensitivity to allele frequency.** → the same magnitude of difference between observed and expected haplotype frequencies will yield a much larger value for common alleles compared to rare alleles.

➔ **D' value** (Δ' value): is the **normalized D value.**

$$\Delta' \text{ value} = \Delta/\Delta\text{max} = (P_{AB} - P_A P_B)/\text{Dmax}$$

The maximum value of **D' is 1**, denoting no recombination between the 2 alleles since the more recent one appeared. **D'< 1** is evidence for historical recombination.

➔ **r^2** (r squared): is the square of the correlation coefficient and it **quantitates the correlation between the presence of two alleles of two SNPs.** Remember that r^2 depicts the **association** between the SNPs rather than saying they are in the same block.

$$r^2 = \Delta^2/(P_A P_B P_a P_b)$$

The maximum value of **r^2 is 1** representing the perfect LD situation: one allele is completely predictable once the other is known. This indicates once again no recombination.

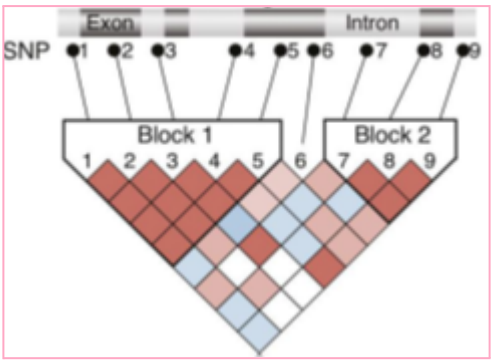
- The threshold for **strong LD is generally accepted as $r^2 \geq 0.80$**, which corresponds to a correlation coefficient of approximately 0.90
- As the intermediate values of r2 are proportional to the degree of correlation, it measures how well one marker can act as a surrogate, or proxy, for another.

Table 3.2 Properties of commonly used bi-allelic LD measures

LD measure	Features
Δ (delta) (no range)	Used in earlier studies when more sophisticated LD measures were not available or could not be computed Strongly influenced by allele frequencies Cannot be used to compare LD quantities among different pairs of alleles when allele frequencies are different
D′ (D-prime) (normalized Δ; range 0 to 1)	Useful in inference of recombinational history. D′ = 1 when there has been no recombination between the two alleles (and only three of four possible gametes are observed) D′ < 1 is evidence for historical recombination Values between 0 and 1 are hard to interpret Small sample size and rare alleles infate D′ D′ may be 1 despite a low r^2 Limited value in comparisons
r^2 (r-squared) (range 0 to 1)	When the two alleles are always on the same chromosome, and are exclusively associated, $r^2 = 1$ Intermediate values are easy to interpret Small sample size does not infate the value Robust for allele frequency differences Most valuable for comparison between studies, for tagSNP selection, and power calculations

❖ VISUALIZATION OF LINKAGE DISEQUILIBRIUM

LD can be easily visualized by **plotting the pairwise LD measures** (or their probability) in a triangular heat map. On the top of the triangle we have the list of SNPs as numbers from *1* to *n*. Each element of the map is a diamond. **Each SNP is connected on a map of the gene.**



The shade of the diamond indicates the strength of the LD.

- dark red means **strong LD**
- the lighter the color the **weaker the LD**

In this heat map D' was used as a measure.

We also have blocks. As we can see, in correspondence of SNP6 we have a **discontinuity, meaning a recombination has occurred**.

➔ **RATES OF RECOMBINATION** in humans average roughly **10^{-8} crossing-over events** per generation **between adjacent base-pairs.**

❖ LINKAGE DISEQUILIBRIUM PATTERNS

LD can be influenced by many factors like:

- genetic linkage
- epistasis
- natural selection
- rate of recombination
- mutation
- genetic drift
- random mating
- gene flow

We have too many players therefore we need to determine LD empirically.

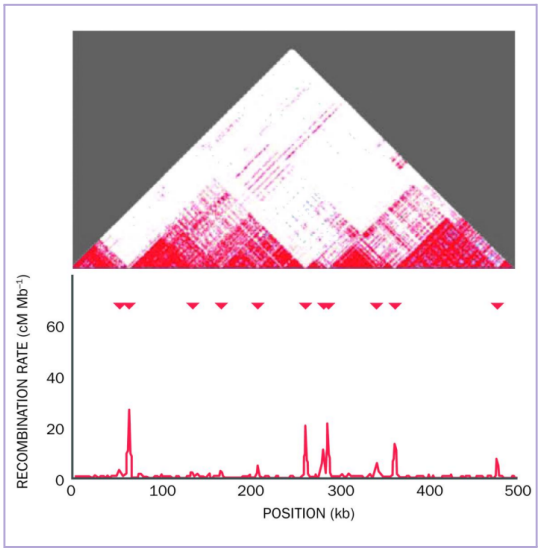
The aim of the International **HapMap Project** is to determine the common patterns of DNA sequence variation in the human genome, by characterizing sequence variants, their frequencies and correlations between them (LD patterns). They genotype one million SNPs from four human populations.

The study takes into consideration DNA samples from populations with ancestry from parts of Africa (YRI), Asia (JPT, CHB) and Europe (CEU).

This project showed that **the genome can be represented as blocks of haplotypes.**
When two people share the same haplotype block, it means that they are both inherited by a common ancestor.

- ➔ in haplotype points recombination is either almost absent or almost always present.

In the graph on the right, block boundaries reflect locations with high recombination rates within the sample



The major discovery by the HapMap project is that there is **much less variation than expected.**

- ➔ The European sample has 70 SNPs *per block.* Each SNPs has two alleles. The power of possible combinations should be 2^{70}. Instead it's less than 5. This accounts for 93 per cent of all chromosomes and it means that loci are dependent between themselves.

TABLE 15.8 HAPLOTYPE BLOCK STRUCTURES IN FOUR HUMAN POPULATIONS AS REPORTED BY PHASE I OF THE HAPMAP PROJECT

Parameter	YRI	CEU	CHB+JPT
Average number of SNPs per block	30.3	70.1	54.4
Average length per block (kb)	7.3	16.3	13.2
Percentage of genome spanned by blocks	67	87	81
Average number of haplotypes per block	5.57	4.66	4.01
Percentage of chromosomes accounted for by these haplotypes	94	93	95

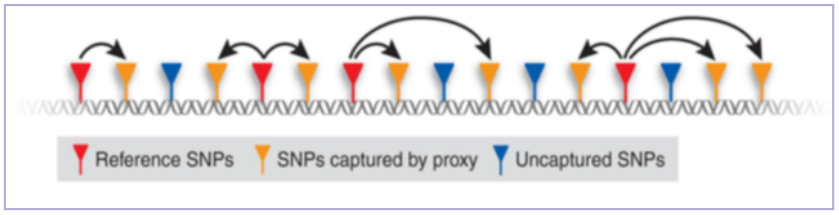
If we compare the European and African database we can see that blocks are shorter in YRI than in CEU. Moreover, the percentage of the genome spanned by blocks is shorter. This is perfectly in line with the Out of Africa theory, in which Africa harbors a much greater diversity in terms of genetics. Europe and Asia have less variability.

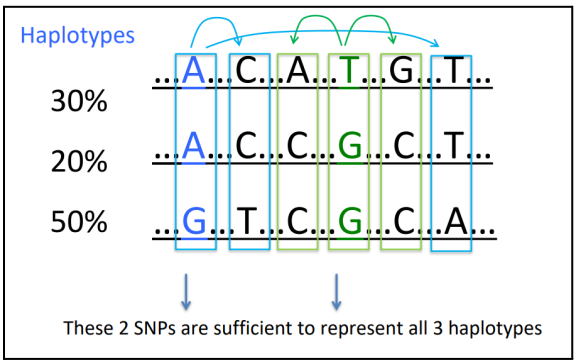
❖ SNP TAGGING IS FACILITATED BY LINKAGE

BLOCKS are defined as **regions of linkage disequilibrium**: within a block genotypes of SNPs are correlated.

If two or more SNPs are highly correlated we don't have to genotype all of them, we can just genotype one and input with high confidence the others.

We can define some **TAG SNPS** that allow us to genotype the majority of the other SNPs, which are captured by proxy. We have also uncaptured SNPs.





Let's look at this example: we have 6 different SNPs, which means that we have 2^6 possible combinations. Or at least, that would be true if all the six loci were independent. Since they aren't, a smaller number of SNPs can be used to map the six loci.

➔ looking at the SNP present in loci 1 and 4, we can understand the SNPs in loci 2,3,5,6.

It has been estimated that **~ 500.000–1 million of (well chosen) tag-SNPs** are sufficient to **capture a significant proportion of all the common genetic variation** in a non-African population (~1,5 million for African populations).

❖ gnomAD

The **HapMap** project was the first public project that represented most genetic variants that have frequencies of **at least 5% in the populations studied.**
However, we are also interested in alleles with **LOWER FREQUENCIES**, since the alleles with frequency of at **least 1%** are the ones with **lethal** or heavy consequences.

Recent improvements in NGS technologies have brought down the cost of sequence.
Now we have newer projects that are more precise and the Hapman data was implemented in the 1000 genome project that included 2535 humans from 26 populations.

After the 1000 genome project, the next biggest project is **gnomAD**.
The Genome Aggregation Database (gnomAD), is a coalition of investigators seeking to aggregate and harmonize exome and genome sequencing data from a variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

It was **released with updates multiple times,** always bringing more specific sequences of a higher number of genomes.

1. **V1 data** (2016), it was called the **Exome Aggregation Consortium (ExAC)** and it included data on exomes (60 000 exomes), so coding regions.
2. The **v2 data set** (GRCh37/hg19) spans 125,748 exome sequences and 15,708 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies.
3. The **v3 data set** (GRCh38) spans 71,702 whole genome sequences as part of various disease-specific and population genetic studies.

gnomADv2.1 reports 241 million small genetic variants (SNPs and InDel) and 335,470 structural variants, compared with 7.4 million small genetic variants identified in ExAC, which did not analyse structural variation.

gnomADv2.1 includes exomes and genomes from European, Latino African and African American, South Asian, East Asian, Ashkenazi Jewish and other populations.

- ★ **RARE VARIANTS** are likely to be younger than the common SNPs that define the haplotypes blocks.
- ★ **HAPLOTYPE BLOCKS** represent ancestral chromosome segments that have been transmitted intact through many generations.
- ★ **MORE RECENT MUTATIONS** will affect particular examples of a given block, so there is a degree of heterogeneity layered on top of the basic structure revealed by the common and ancient SNPs

EXERCISE 1

What is linkage disequilibrium?
Linkage disequilibrium is the non-random association of alleles from different loci in a population. It can be inferred when the population frequency of a haplotype is unexpectedly significantly different from the product of the individual population frequencies of the alleles.

Which of the following haplotypes shows evidence of linkage disequilibrium, given the individual allele frequencies?

- **A.** haplotype A*1-B*3 with a population frequency of 0.101 (frequencies of 0.231 for A*1 and 0.431 for B*3)
- **B.** haplotype C*2-D*1 with a population frequency of 0.071 (frequencies of 0.311 for C*2 and 0.225 for D*1)
- **C.** haplotype E*1-F*1 with a population frequency of 0.205 (frequencies of 0.236 for E*1 and 0.289 for F*1)
- **D.** haplotype X*2-Y*3 with a population frequency of 0.101 (frequencies of 0.532 for X*2 and 0.434 for Y*3)

ANSWER:

A. $D = 0.101 - (0.231 \times 0.431) = 0.101 - 0.0996 = 0.0014 \sim 0$
B. $D = 0.071 - (0.311 \times 0.225) = 0.071 - 0.070 = 0.001 \sim 0$
C. $D = 0.205 - (0.236 \times 0.289) = 0.205 - 0.068 = 0.137$
D. $D = 0.101 - (0.532 \times 0.434) = 0.101 - 0.23 = -0.129$

Hence C and D are potential haplotypes. As we said, delta is sometimes not enough to establish disequilibrium. We thereby calculate D' and R^2.

- **A.** Given the frequency of alleles A and B, we can establish the frequency of alleles a and b for the two loci:

  $a = 1 - A = 0.76$
  $b = 1 - B = 0.569$
  $\Delta_{max} = \min(P_A P_b, P_a P_B) = (P_A P_b) = 0.131$
  $\Delta' = \Delta / \Delta_{max} = 0.011$
  $r^2 = \Delta^2 / P_A P_a P_B P_b = 4.8 * 10^{-5}$

  D_{max} is the :
  $min(P_A P_b , P_a P_B)$ if $D>0$ and
  $min(P_A P_B , P_a P_b)$ if $D<0$

- **B.** Given the frequency of alleles C and D, we can establish the frequency of alleles c and d for the two loci: (consider in this case there is a positive Δ)

  $c = 1-C = 0.689$
  $d=1-D= 0.775$
  $\Delta_{max} = \min(P_C P_d, P_c P_D) = (P_c P_D) = 0.155$
  $\Delta' = 0.0066$
  $r^2 = 2.53 * 10^{-5}$

C. Given the frequency of alleles E and F, we can establish the frequency of alleles e and f for the two loci:

e = 1-E= 0.764
f = 1-F = 0.711
$\Delta_{max} = \min(P_E P_f, P_e P_F) = (P_E P_f) = 0.168$
$\Delta' = 0.81$
$r^2 = 0.486$

D. Given the frequency of alleles X and Y, we can establish the frequency of alleles x and y for the two loci: (consider in this case there is a negative Δ)

x=1-X=0.468
y=1-Y=0.566
$\Delta_{max} = \min(P_X P_Y, P_x P_y) = (P_X P_Y) = 0.230$
$\Delta' = 0.565$
$r^2 = 0.27$

As we can see from the high values of r^2 for cases C and D, we confirm that the alleles taken into consideration are haplotypes.

EXERCISE 2

Calculate D', D and r2 for the following observed haplotypes:

SNP1 & SNP2	Count
CA	162
CT	216
GA	162
GT	0
Total	540

Right off the bat we can see we have only 3 haplotypes out of 4, the SNPs should be in LD: this probably means that **no recombination has occurred** ,and therefore we can **infer that D' is equal to 1.**

However, for higher certainty, we can make all the calculations.

1. Calculate the **GENOTYPE FREQUENCY** for each of the four genotypes:
   - ➔ CA: 162 / 540 = 0.3
   - ➔ CT: 216 / 540 = 0.4
   - ➔ GA: 162 / 540 = 0.3
   - ➔ GT: 0 / 540 = 0

2. Calculate the **ALLELE FREQUENCY** starting from the genotype frequency. Using the following table, all we have to do is sum the probabilities.

	T	A	
C	0.4	0.3	0.7
G	0	0.3	0.3
	0.4	0.6	

As we can see from the initial table, C and G, and T and A are never present together. This means that C and G are two different variations in the same locus, and so are T and A.

3. We can therefore **establish two loci**, a and b, **where we have a dominant and a recessive allele** which we can name, respectively, A and a, B and b.

	T	A		
C	0.4	0.3	0.7	p(A)
G	0	0.3	0.3	p(a)
	0.4	0.6		
	p(b)	p(B)		

4. Then we derive the single allele sequences, and from that write D, D' and r2.

D(AB) = $P_{AB} - P_A P_B$ = 0.3 - 0.42 = -0.12

D'(AB) = D(AB)/min($P_A P_B$, $P_a P_b$) = 0.12/0 = 1

r^2 = $D^2/P_A P_a P_B P_b$ = 0.285714

D' = 1, however the r^2 is very low.

SNP1 and SNP2 belong to the same haplotype block and no recombination has occurred, so they are in LD. But they are not good proxy: D' is very high but it doesn't mean they are good proxy.

LINKAGE ANALYSIS

❖ MODEL FREE LINKAGE ANALYSIS

In the genetics course we have learned **(parametric) linkage analysis**. It is useful to map genes for mendelian diseases.
Inspired by the success of linkage analysis, researchers attempted to apply similar tools to **complex traits.** The aim still being the one of finding genetic variants.

Parametric linkage analysis however requires frequency, penetrance and inheritance models.
These can be named for mendelian characters, but not for complex traits.

- in complex traits we **cannot see a specific inheritance model** even if we see if the trait partially depends on genes
- hence for complex diseases linkage analysis is **NON PARAMETRIC** or MODEL FREE .

Model-free linkage analysis compares the **extent to which** relatives **share alleles** or haplotypes **identical by descent** (IBD) with the **extent to which they share phenotypes.**

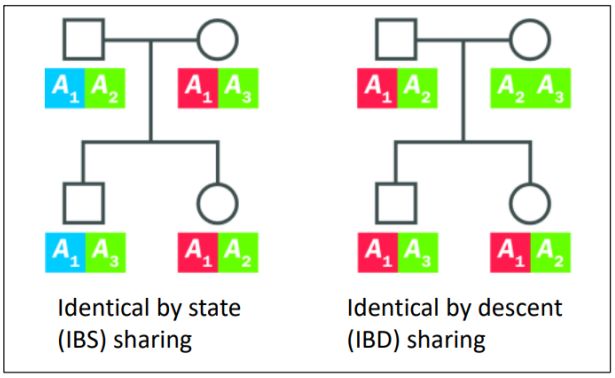
BUT WHAT ARE IDENTICAL BY DESCENT ALLELES?

Two alleles are **identical by state** if they:

- are **identical**
- they **may or may not be identical by descent**.

Two alleles are **identical by descent** if they are known to be:

- **identical**
- both have been inherited from a demonstrable **common ancestor.**



Let's see an example to understand the difference between these two allele types:

1. in the first pedigree, we know the genotype of the children is A_1A_3, A_1A_2 and the one of the parents is A_1A_3, A_1A_2 . In this case we can see that the two alleles A_1 are inherited one from one parent and one from the other, hence, they don't have a common ancestor.
2. In the second case we see that both children have the A_1 allele, but only one of the two parents has it. Hence, the alleles have the same origin.

Free Model linkage **analysis considers only IBD alleles.**

If alleles or haplotypes *identical by descent* are shared by affected relatives more often than would be expected under simple Mendelian principles, that is evidence of linkage.

This is because you expect recombination between alleles at different loci. If it doesn't happen (we know it doesn't because the alleles come from the same origin), it means there is linkage.

Model-free linkage analysis takes two forms:

➔ **RELATIVE PAIR LINKAGE METHODS** are used for **dichotomous characters**.

The genomes of related affected individuals are searched for chromosomal regions in which the **IBD sharing for these pairs differs significantly from what is expected by chance**.

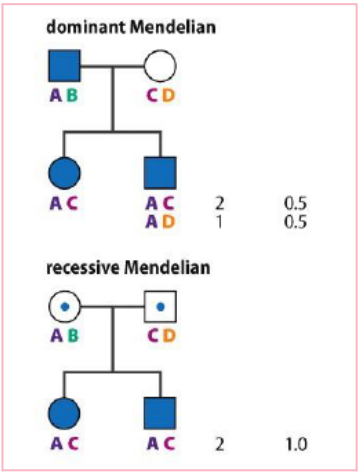
➔ **VARIANCE COMPONENT METHODS** are used for **quantitative traits**.

The variance of quantitative trait loci shared IBD between relatives is **compared to their phenotypic covariance.**

MENDELIAN DISORDERS

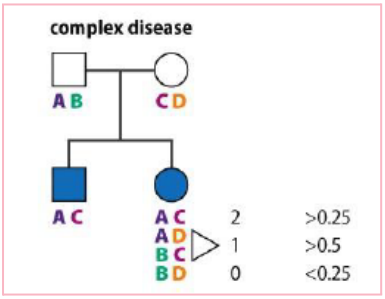
Given a **random chromosomal locus**, siblings will be expected to share 0, 1 or 2 **IBD** haplotypes with frequencies 0.25, 0.5 or 0.25 respectively.

- the overall average across all sets of sibs would be 1 allele in common.



Pairs of sibs that are both affected by a **dominant Mendelian condition** must share the segment that carries the disease allele A and may or may not (a 50:50 chance) share a haplotype from the unaffected parent.

Pairs of sibs that are both affected by a **recessive Mendelian condition** necessarily share the same two parental haplotypes for the relevant chromosomal segment.

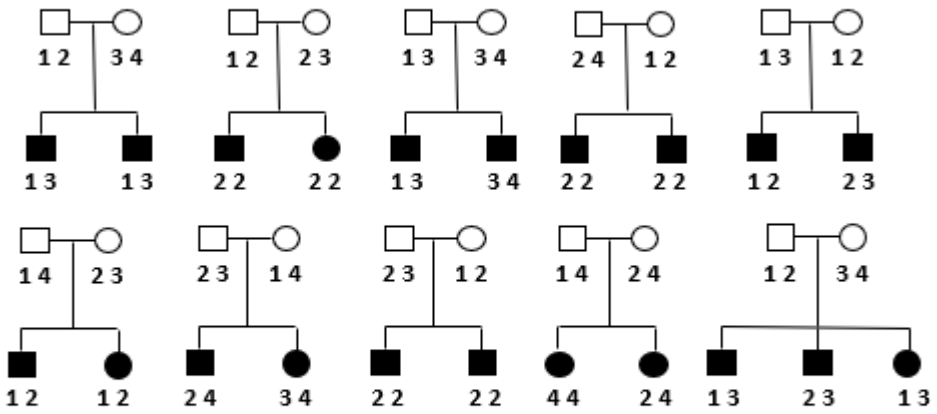


COMPLEX DISORDERS

For complex disorders, **haplotype sharing that is greater than expected by chance** may allow the identification of chromosomal segments containing susceptibility genes. Meaning that the affected children have inherited a similar combination of alleles that could lead to the disease, since they inherited them from their unaffected parents.

EXERCISE:

For the pedigrees below, calculate how many affected sib pairs share 2, 1, 0 alleles identical by descent (IBD) **(count all possible sib-pairs).**



This means that we need to understand when the siblings are sharing an allele, and what type of allele it is, of IBD or IBS.

- Let's take for example the first pedigree: the children are sharing both allele 1, which only comes from the father, and allele 3, which can only come from the mother.
  Since alleles 1 and 3:
  a. are identical
  b. share the same origin
  they are both IBD.
- Let's take now for example the third pedigree: the children share only one allele, which is allele 3. However when we look at it, it is obvious that the first child's allele 3 comes from the mother, while allele 3 of the second child comes from the father. The two alleles are hence identical but they do not share the same origin. We have zero IBD alleles
- Let's take as an example the fifth pedigree: the children share allele 2, which can only come from the mother. This means that we have only one IBD allele.
- Let's take as an example the last pedigree: **we need to consider every possible couple of children.** First couple has only 1 IBD, the second couple also has 1, and the third couple has 2 IBDs

1: 2
2: 2
3: 0
4: 2
5: 1

6: 2
7: 1
8: 2
9: 1
10: 2 1 1

The final results are indicated on the right.

Now we want to see if this disease is in LD. We compare the number of **IBDs** we found with the number **we would expect if there was no linkage**. In the second case we apply frequencies 0.25, 0.5, 0.25 on the total number of IBDs we found (12).

	Alleles IBD		
	2	1	0
Observed	6	5	1
Expected	3	6	3

2-1-0 IBD sharing

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared
O_i = observed value
E_i = expected value

Now that we have the observed and the expected values, all that is left is to apply the **CHI SQUARE TEST.**

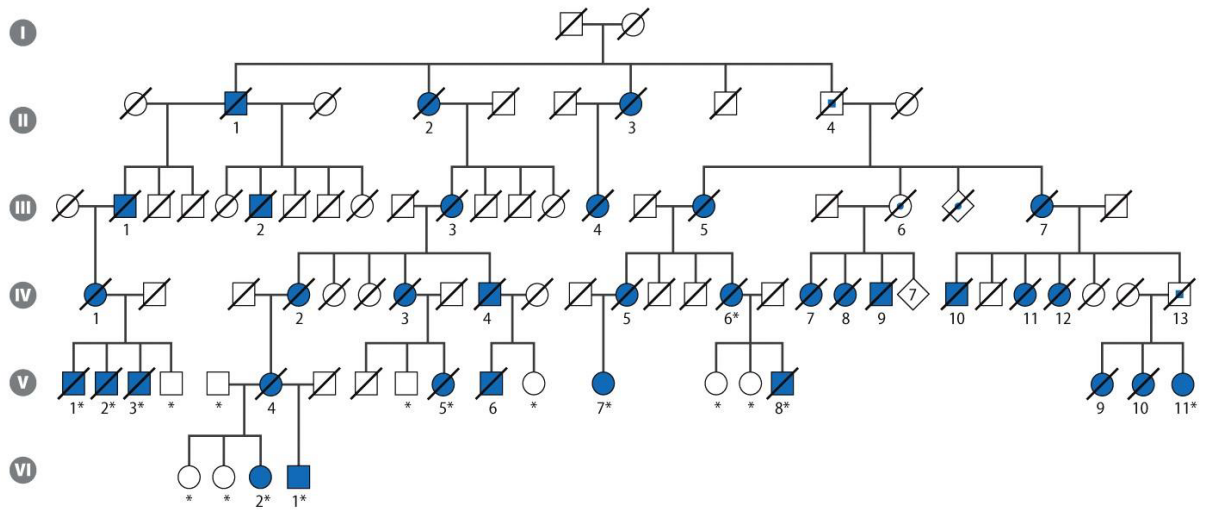
$X^2 = 3+⅙+4/3= 4.5$

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21

Looking at the table we can see that the p value is in between 0.05-0.01.
This means that the p value is very small. We reject the null hypothesis that observed and expected data are the same, hence, **there is disequilibrium.**

This means that the locus we studied is in linkage disequilibrium with the causal variant of the phenotype.

For several complex diseases there are also subsets in which the disorder shows clear **Mendelian inheritance**. That is most obvious when there is dominant or quasi dominant inheritance, as in early onset form of Alzheimer disease, Parkinson disease, diabetes and various type of cancer.



As for **complex traits,** **Model free linkage studies were disappointing in the 90s** when applied to real life scenarios. We have two main problems:

- susceptibility loci are not always necessary or sufficient to cause disease
- low statistical power

You need a lot of samples to have enough statistical power in order to make some statements. The power depends on the frequency of the susceptibility loci, which is not high enough in many cases.

ASSOCIATION ANALYSIS

❖ COMMON DISEASE ALLELES HAVE HIGH FREQUENCY AND LOW EFFECT

Failure of the linkage approach for complex traits implies that genetic mechanisms underlying complex traits are different from the ones at the basis of mendelian disorders.

- ➔ In **MONOGENIC RARE DISORDERS** we assume that there is a **very rare** and **high penetrance** variant causing the disease.
- ➔ In **COMMON DISEASES** instead, variants are more **common** and have a **lower penetrance.**

The **COMMON DISEASE COMMON VARIANT HYPOTHESIS** states that genetic variations with appreciable frequency in the population at large, but relatively low 'penetrance' are the **major contributors to genetic susceptibility to common diseases**

The reasons for which **common variants have LOW PENETRANCE** are the following:

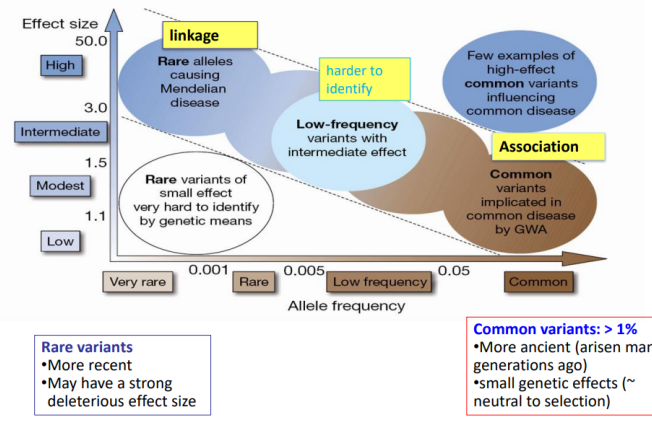
- Variants must have low penetrance otherwise **everyone would have the disease.**
- If the variant is common it's probably from ages ago and it has **passed selection,** so it mustn't have a strong phenotype. Otherwise we wouldn't have it because of natural selection

If common alleles have small genetic effects (low penetrance), but **COMMON DISORDERS SHOW HERITABILITY** (inheritance in families), then **multiple common alleles** must influence disease susceptibility.

We have to assume that there are many common variants that cause the disease.
This leads to another, **POPULATION BASED APPROACH.**

In the graph:

- on the x axis we have the **allele frequency**
- on the y axis the **effect size**



As we can see, **rare variants have big effects.** An example is mendelian disorders, where a single SNP can determine the presence/absence of a disease.

Common variants have low effect because they have already gone through natural selection, hence they are less dangerous. Hence why multiple common variants are needed for disease susceptibility.

There can also be common variates having a large effect and rare variants with small effects. The latter are harder to study and for them we usually rely on sequencing.

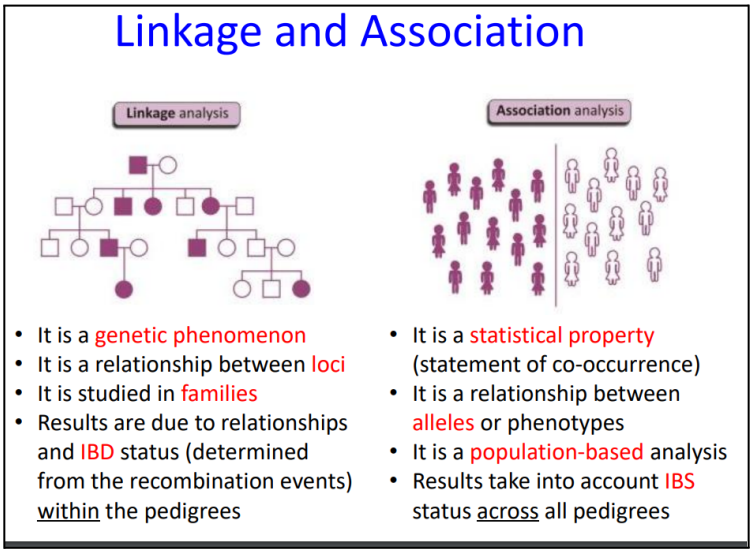
❖ ASSOCIATION STUDIES

➔ **Rare variants (high effect)** can usually be **found by linkage,**
➔ **Common variants (low effect)** can be found by **association studies.**

Linkage analysis has been replaced by association studies for what concerns common variants.

★ **ASSOCIATION** is a **statistical property** that describes the **joint occurrence of alleles** (and/or phenotypes) in individuals within a population.

★ **LINKAGE** is a **genetic phenomena** that describes the relationship **between loci**



Another important difference is that linkage requires IBD alleles, which themselves require the patient's family. Instead association studies only need IBS alleles.

Linkage	Association
Older method	Current method
Analysis within **multi-generation pedigrees**	Analysis across **unrelateds** or pedigrees
Used small data sets	Often uses big data sets
Based on Identity-by-Descent	Based on Identity-by-State
Tracks regions of DNA	Tracks specific values at common markers
Signal spread ~2 **Mb**	Signal spread ~20 **Kb**
Prefers Linkage Equilibrium (LE)	Requires Linkage Disequilibrium (LD)

ASSOCIATION is between alleles and phenotypes at population level. It looks for significant **differences in SNP allele frequency in unrelated cases vs controls** from the same population.

The **GENERAL PIPELINE** of an association study is the following:

1. **Genotyping** from different individuals
2. Find **variations** at a single nucleotide:
   - individuals are **divided** between who has one **allele** and who has the other.
3. Consider the disease you are studying:
   - individuals are **divided** between who is **affected** and unaffected (unrelated individuals)

Look for significance differences in SNP allele frequency in unrelated cases and controls.

❖ TYPE OF EPIDEMIOLOGIC ASSOCIATION STUDIES

There are two main types of association studies:

➔ **COHORT STUDIES** **begin with a large, healthy population** sample. Then, both exposure and outcome are recorded as they occur during **longitudinal follow-up.** This characteristic of cohort studies makes them ***prospective studies.***

Since individuals are taken randomly, you can estimate disease risk directly.

- Determined by **relative risk**

➔ **CASE CONTROL STUDIES** begin with **cases who have already developed the phenotype** of interest (outcome) and collect the exposure information retrospectively. Thus, case-control studies are ***retrospective studies.***

In control studies you select people *because* they have the disease, so the estimate of disease risk is computed in a more indirect way.

- determined by **odd ratio**

The **RELATIVE RISK** is determined in ***cohort studies***, hence subjects are collected independently of their disease status. It is therefore a direct estimation of the risk.
The **relative risk** is **always calculated *relatively* to the exposure to a certain risk factor.**

In the table we have:

- A+ indicates individuals that were exposed
- A- indicated individuals that weren't exposed

"Exposure"		"Outcome" Case	Controls	
	A+	a	b	a+b
	A-	c	d	c+d
		a+c	b+d	N

Since we want to see if the risk is correlated to exposure:

$$\text{Relative Risk} \quad RR = \frac{a/(a+b)}{c/(c+d)} \quad \frac{\text{incidence of disease in individuals WITH exposure}}{\text{incidence of disease in individuals WITHOUT exposure}}$$

The **ODDS RATIO** is determined in ***case-control*** studies, hence the disease risk cannot be estimated directly.
The **odds ratio** is **an *approximation* to Relative Risk.**

$$\text{Odds Ratio} \quad OR = \frac{ODDS\ of\ disease\ given\ exposure}{ODDS\ of\ disease\ given\ NON-exposure}$$

$$ODDS\ of\ disease\ given\ exposure = \frac{a/(a+b)}{b/(a+b)} = \frac{a}{b}$$

$$ODDS\ of\ disease\ given\ NON\text{-}exposure = \frac{c/(c+d)}{d/(c+d)} = \frac{c}{d}$$

$$\text{Odds Ratio} \quad OR = \frac{a/b}{c/d} = \frac{a*d}{c*b} \quad \frac{\text{ODDS of disease given exposure}}{\text{ODDS of disease given non-exposure}}$$

The relative risk requires knowledge of the epidemiology of the disease. It can not be calculated from the results of a case-control study, because the case-control design involves the selection of research subjects on the basis of the ***outcome*** rather than on the basis of the ***exposure***.

If the disease is rare, the odd ratio is a good approximation of the risk ratio.

ADVANTAGES AND DISADVANTAGES OF CASE CONTROL STUDIES

ADVANTAGES	DISADVANTAGES
If a disease is rare, cohort studies would need a large population to get a few affected people in the future. Case control studies have a **smaller sample size.**	**Research of the controls** must be done very well otherwise you can find discrepancies which do not depend on the disease.
Rare diseases with long latency can be studied **without waiting for a long time**, since you already know if the individual is affected or not.	The **risk is approximated**, not directly calculated.
Suitable when exposure is unethical	You can't study multiple variants of the disease because all your patients have already developed the disease.
Multiple exposures can be examined in correlation with the outcome	Causality is hard to establish.

❖ GENOTYPES AS EXPOSURE

SNPs are **BINARY VARIABLES:**

- ➔ If one of the alleles (allele B) is associated with **increased risk of disease**
- ➔ the other allele (allele A) will be a **marker for protection.**

This means that the two **alleles will always be reciprocal.**

- ➔ By default, statistical analysis of associations is done assuming that the **uncommon (minor) allele is the risk marker**

Let's suppose we want to **TEST** whether a given **genetic factor is involved in a disease.**

1. We can sample **N unrelated cases** (who have the disease) and **N unrelated controls** (who do not have the disease)
2. Consider a certain **locus**
3. If the locus is the **disease locus**, what should we expect to observe in the sample is

$$P(\text{disease \& allele}) \neq P(\text{disease}) \times P(\text{allele})$$

In general, what we are about to see are various methods to test if there is association between alleles and the phenotype.

★ ALLELIC TEST

Following HWE, **given a genotype count**, we can get the **ALLELE COUNT.** This can be done through a 2x2 contingency table.

GENOTYPE COUNTS

	CASES	CONTR	
A1-A1	32	75	107
A1-A2	96	150	246
A2-A2	72	75	147
	200	300	500

ALLELE COUNTS

	CASES	CONTR	
A1	160	300	460
A2	240	300	540
	400	600	1000

On the left we can see the calculations necessary to go from a genotype count to an allele count, remembering that the **(observed) count of an allele** is given by: 2*homozygous + heterozygous count.

Moreover, we can get the **frequency of the alleles** in both controls and cases.

- allele frequency of A1 for cases is given by: number of cases with A1/number of cases in total.

	Cases	Controls
Allele frequency of A1	0.4	0.5
Allele frequency of A2	0.6	0.5

Now we need to understand if the **differences** between cases and controls **are significant.** One way to do this is the **CHI SQUARE TEST.**

- ➔ we already have the observed count of alleles thanks to the contingency table
- ➔ now we need the expected count of alleles, which can be calculated from the observed count of alleles.

Observed allele counts

	Case	Control	
A1	n_{11}	n_{12}	n_1
A2	n_{21}	n_{22}	n_2
	n_{Ca}	n_{Co}	N

Expected allele counts

	Case	Control	
A1	$\frac{n_{Ca}\, n_1}{N}$	$\frac{n_{Co}\, n_1}{N}$	n_1
A2	$\frac{n_{Ca}\, n_2}{N}$	$\frac{n_{Co}\, n_2}{N}$	n_2
	n_{Ca}	n_{Co}	N

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

Chi-square distribution with 1 d.f. (for large samples)

In our example, the Chi Square Test is calculated in the following way:

- ➔ We compute on the **hypothesis H_0** that **the alleles** of the disease **are independent.**

Genotypes	Cases	Controls
1-1	32	75
1-2	96	150
2-2	72	75

Observed

	Cases	Controls	
A1	160	300	460
A2	240	300	540
	400	600	1000

Expected

	Cases	Controls	
A1	184	276	460
A2	216	324	540
	400	600	1000

$c^2 = \frac{(160-184)^2}{184} + \frac{(300-276)^2}{276} + \frac{(240-216)^2}{216} + \frac{(300-324)^2}{324} = 9.66$ p = 0.0019

Since the **P value is very small**, this indicates **strong evidence against the null hypothesis**, hence the null hypothesis is rejected. The controls and cases are significantly different, hence alleles are associated with the disease.

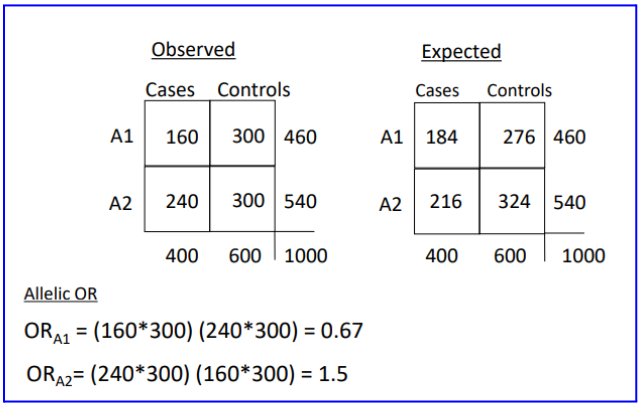
Now that we have established if there is correlation or not with the Chi Square test, we measure the **strength of association** with the **ODDS RATIO.**

"Outcome"

"Exposure"	Case	Controls	
A+	a	b	a+b
A-	c	d	c+d
	a+c	b+d	N

Odds Ratio $OR = \frac{a/b}{c/d} = \frac{a*d}{c*b}$ Odds of disease given exposure / Odds of disease given non-exposure

- ➔ **OR > 1**: increased risk
- ➔ **OR < 1:** protective association, the exposure correlates with a decreased risk
- ➔ **OR = 1**, the SNP allele has no effect on disease risk



We calculate two ORs, one per allele:

The Allelic model is also called the **multiplicative genetic risk model.**
It's very **susceptible to deviations from HWE**: if a SNP is not perfect in HWE, you immediately get results that are not reliable because the allele frequency changes.
This doesn't happen in other tests like genotyping.

- ➔ The multiplicative model is **useful for initial screening** of the data to see whether there is any signal worth pursuing for examination by other genetic risk models.

★ DOMINANT AND RECESSIVE MODELS

We can also use other models, like the dominant and the recessive one. These models **collapse the three different genotypes in a biallelic way** between unaffected and affected.

➔ heterozygous individuals and dominant homozygous are collapsed as one, considering that the **reference genotype is coded as 0** and **risk genotype(s) are coded as 1**

- in a dominant disorder, the risk phenotype is found in heterozygotes and dominant homozygous, which take up value 1.
- in a recessive disorder, the risk phenotype is found only in recessive homozygous, which take up values 1.

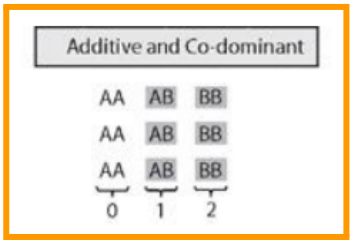
By converting the 3 genotypes to a **BINARY VARIABLE**, coded as 0 or 1, the data can be analyzed by constructing a 2 x 2 contingency table. Afterwards, we can apply either **Chi-squared test** or **Fisher's exact test.**

	Risk genotype (1)	Reference genotype (0)
Cases (1)	a	c
Controls (0)	b	d

★ GENOTYPE (CODOMINANCE) and ADDITIVE MODELS

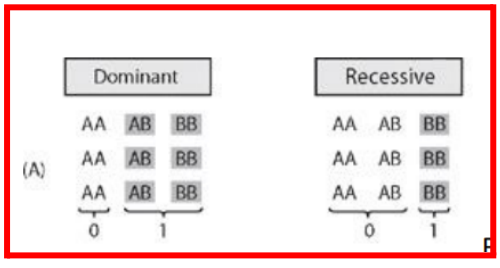
GENOTYPE

In this kind of study, **no genotype groups are collapsed** and all three genotypes are coded. This is because, **if there is codominance, we cannot simplify anything.**
The resulting 2 × 3 table is analyzed differently for the additive and codominant models.



The additive model examines the gene-dosage effect by using all three genotypes and taking into account the **gradual change,** also known as the *trend in risk*, associated with each genotype from AA (referent) to AB to BB.

	Reference genotype (0)	Risk genotype (1)	Risk genotype (2)
Cases (1)	a	c	e
Controls (0)	b	d	f



ADDITIVE MODEL (Cochran-Armitage trend test)

The **additive model** (for B) **assumes that there is a uniform, linear increase** in risk for each copy of the B allele.

The additive model has **greater statistical power** compared with the analysis of the same data after collapsing the three genotypes into two and analyzing them by other models.
Quantitative genetics has shown that it is the most relevant model **for complex phenotypes:** in real life, the additive model is the most biologically plausible model, even though the change may not be perfectly linear.

TABLE 2 | Tests of association using contingency table methods.

Test	Degrees of freedom (d.f.)	Contingency table description	PLINK keyword
Genotypic association	2	2 × 3 table of *N* case-control by genotype (*a/a, a/A, A/A*) counts	GENO
Dominant model	1	2 × 2 table of *N* case-control by dominant genotype pattern of inheritance (*a/a*, not *a/a*) counts	DOM
Recessive model	1	2 × 2 table of *N* case-control by recessive genotype pattern of inheritance (not *A/A, A/A*) counts	REC
Cochran-Armitage trend test	1	2 × 3 table of *N* case-control by genotype (*a/a, a/A, A/A*) counts	TREND
Allelic association	1	2 × 2 table of 2*N* case-control by allele (*a, A*) counts	ALLELIC

d.f. for tests of association based on contingency tables along with associated PLINK keyword are shown for allele and genotype counts in case and control groups, comprising *N* individuals at a bi-allelic locus with alleles *a* and *A*.

EXERCISE: In a case-control study, the genotype counts were as follows:

	AA	AB	BB
cases	28	92	22
controls	58	71	32

To find out if there is association, perform the following tests:allelic, genotype, dominant and recessive model.

1. ALLELIC TEST

Allelic Count - A: CASES: 148 CONTROLS: 187
Allelic Count - B: CASES: 136 CONTROLS: 135

Now we can first input the **OBSERVED ALLELE COUNT TABLE** table.

	Cases	Controls	
A	148	187	148+187=335
B	136	135	136+135=271
	148 + 136 = 284	187+135= 322	Tot: 284+322=606

Now we can input the **EXPECTED ALLELE COUNT TABLE** based on the totals we just calculated:

	Cases	Controls	
A	284*335/606= 157	322*335/606= 178	335
B	284*271/606= 127	332*271/606= 144	271
	284	322	606

Now we can compute the **Chi Square value**:

$$X^2 = \sum \frac{(Observed - Expected)}{Expected}^2 = 2.17$$

Now we can compute the **P value:**
If the Chi Square is of 2.17, we can say looking at the tables that the p value, with one degree is of freedom, is over 0.1
We can also compute the p value on R with the following code:

```
pchisq(q, df, lower.tail=FALSE)
```

$= 0.14$
The p value is big so **we cannot reject the null hypothesis of independence**. Hence, we can say that the alleles are not associated.

Finally, for the allelic test it is also important to compute the **odd ratio.**

"Outcome"

"Exposure"	Case	Controls	
A+	a	b	a+b
A-	c	d	c+d
	a+c	b+d	N

Odds Ratio $OR = \frac{a/b}{c/d} = \frac{a*d}{c*b}$ $\frac{\text{Odds of disease given exposure}}{\text{Odds of disease given non-exposure}}$

In our case:

$OR = \frac{a*d}{b*c} = 0.78$

we can also compute the OR as

$OR = \frac{b*c}{a*d} = 1.27$

as it is the inverse of the previous one.

Just computing one of them is okay, they are one the inverse of the other.

COMPUTE THE ODD RATIOS ON THE OBSERVED VALUES!!!!!!

2. GENOTYPE TEST

Now we can first input the **OBSERVED GENOTYPE COUNT TABLE**.

	Cases	Controls	
AA	28	58	86
BA	92	71	163
BB	22	32	54
tot.	142	161	303

Now we can input the **EXPECTED GENOTYPE COUNT TABLE.**

	Cases	Controls	
AA	40.3	45.7	54
BA	76.38944	86.61056	163
BB	(142*54)/303 = 25.30693	28.7	86
tot.	142	161	

The CHI SQUARE test indicates:

- ➔ **chi-square:** 13.88567
- ➔ **p value**: 0.000966 (TWO DEGREES OF FREEDOM!!!)

The p value for the allelic test is not significant, while for the genotypic test it is. This is not a contradiction, as the two tests test for different things: one looks at the significance of the single alleles, the other looks at the significance of one of the genotypes.

3. DOMINANT.

What we need to do now is compute a table that takes into account the fact that one of the two loci is dominant with respect to the other.
We don't know which one of the two alleles is dominant to the other. However, doing the dominant and the recessive test, we reach the same conclusions. (one could say that the dominant allele is the one with the bigger count, but it's not a specific rule).
Here we assume B to be dominant, but we can also do the opposite.

Now we can first input the **OBSERVED COUNT TABLE**.

	CASES	CONTROLS	
BB+BA	22+92=114	32+71=103	217
AA	28	58	86
	142	161	303

Now we can input the **EXPECTED COUNT TABLE**.
Even if in this case we have a sum, we still just **use the law of marginal probability.**
This is because we are still trying to verify if there is association or not and, as we said at the beginning, a way to do this is seeing if the expected and observed values are the same. If they are, we accept the null hypothesis of independence (that is checked by the law of marginal probabilities).

	CASES	CONTROLS	
BB+BA	(142*217)/303= 101.7	115.30	217
AA	40.30	45.70	86
	142	161	303

CHI SQUARE: 9.8
P VALUE: between 0.01 and 0.001 -> 0.0017
ODDS RATIO: 2.29

4. RECESSIVE.

Now we consider the opposite of the dominant model:

Now we can first input the **OBSERVED COUNT TABLE**.

	CASES	CONTROLS	
AA+BA	120	129	249
BB	22	32	54
	142	161	303

Now we can input the **EXPECTED COUNT TABLE**.

	CASES	CONTROLS	
AA+BA	(142*249)/303 =116.7	132.31	249
BB	25.31	28.7	54
	142	161	303

CHI SQUARE: 0.98
P VALUE: 0.319
ODDS RATIO: 0.73

FOR THE CALCULATIONS CONSIDER THE TABLE:

Chi-square Distribution Tables

1 df							
p	0.1	0.05	0.01	10^{-3}	10^{-5}	10^{-7}	10^{-9}
chi2	2.71	3.84	6.63	10.83	19.51	28.37	37.32
2 df							
p	0.1	0.05	0.01	10^{-3}	10^{-5}	10^{-7}	10^{-9}
chi2	4.61	5.99	9.21	13.82	23.03	32.24	41.45

ASSOCIATION STUDIES

FAMILY BASED STUDIES

❖ TRANSMISSION DISEQUILIBRIUM TEST

CASE-CONTROL STUDIES are the most common type of study in genetic association studies. One of the most crucial aspects of case-control studies is the selection of controls, which can be a major drawback if not done properly. A major complication stems from the possibility of having multiple subsets within a population, called **population substructure.**

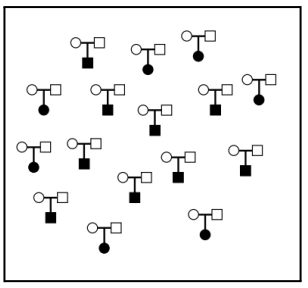
This is why **TRANSMISSION DISEQUILIBRIUM TESTS** exist.

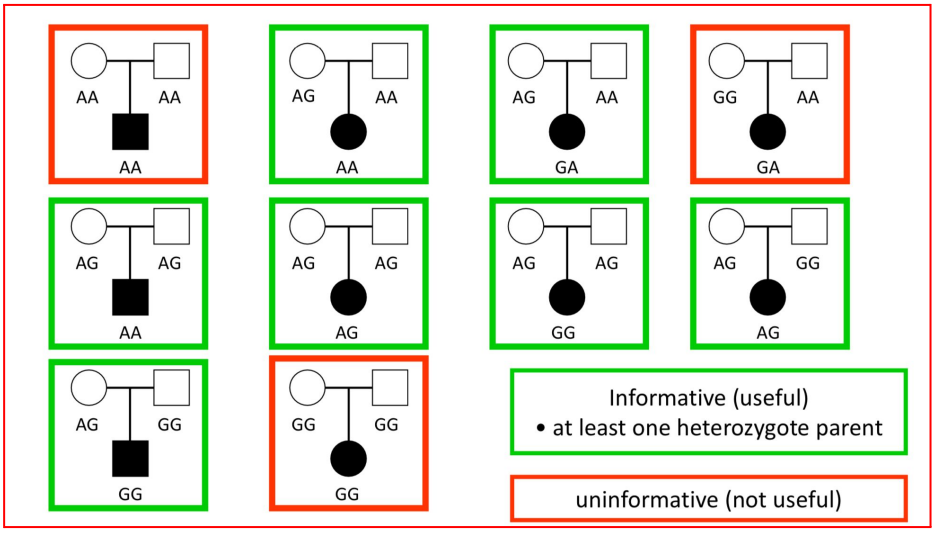
- ➔ The presence of **TDT alleviates the problem of population substructure**.
- ➔ TDT tests are family based. It uses **triads** formed by the parents and affected child.

The information sought in each triad is **whether the affected child inherits any particular allele from a heterozygous parent more than 50% of the time.**

PERFORMANCE of TDTs

1. start collecting sample of trios
2. genotype affected offspring and both parents with SNPS.
3. **To be informative, one heterozygous parent is required.**
4. Test for association with the TDT → we need to look for deviations in the transmission of alleles from parents to affected off-spring.





Looking at one pedigree won't tell us anything about whether an allele is associated with the disease or not. We need to **do statistics basing ourselves on *many* pedigrees.**

➔ Under the **null hypothesis** **of** **no association**, across many pedigrees, we expect 50:50 transmission by Mendelian laws.

➔ Under the **alternative hypothesis**, allele is more likely to have been transmitted to **those who have the disease**. This is the hypothesis TDT wants to assess.

So association is manifested as increased transmission of an allele to all the affected children.

Given a large sample of trios (with varying genotypes) we can form the following table:

Transmitted allele	Untransmitted allele 1	Untransmitted allele 2
1	*A*	*B*
2	*C*	*D*

➔ transmitted alleles in rows

➔ untransmitted alleles in columns.

Under the **null hypothesis** of no association we expect that transmission of **allele 1 is equal to** the number of transmission of **allele 2.**

➔ under the null, B and C are equal.

The test statistics follows a chi square distribution with one degree of freedom according to McNemar's test.

- The test of association is McNemar's test:

$$\chi^2_{TDT,1} = \frac{(B - C)^2}{B + C} = \frac{(n_{1T} - n_{1NT})^2}{n_{1T} + n_{1NT}}$$

Test statistic is asymptotic chi-squared distributions (1df)

- Note that data from only 2 cells (B, C) are used

  Transmissions from homozygous parents (1/1 or 2/2) are ignored in the analysis.

- Simple test – but needs care to avoid bias in the presence of missing data

- Generic table:

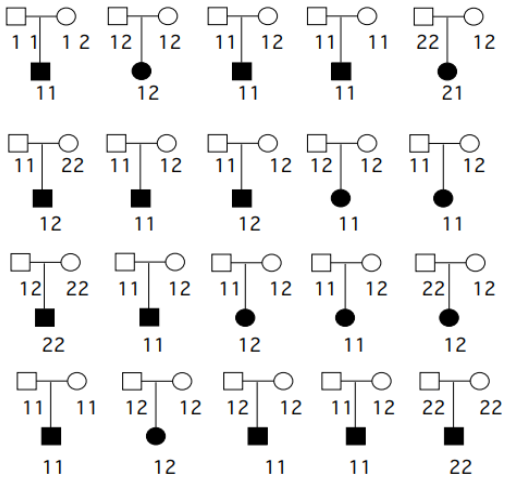
		Non-transmitted allele M_1	Non-transmitted allele M_2	Total
Transmitted allele	M_1	a	b	a+b
	M_2	c	d	c+d
	Total	a+c	b+d	2n

- For *n* affected offspring, there will be 2*n* parents.
- Only data points from heterozygous parents to be considered
  - ie. "b" and "c" in table above
- Assuming no association, or deviation from random patterns of inheritance, then b/(b+c) = c/(b+c) = 0.5
- TDT statistic tests for deviations from the above
  - test statistic is asymptotic chi-squared distributions (1df)

$$\chi^2_{TDT,1} = (b - c)^2 / (b + c)$$

CASE/CONTROL	FAMILY BASED
★ the power is largely equivalent. ★ stratification is indeed a problem, but nowadays we can correct for substructure/relatedness in case/control studies via other approaches (e.g. principal components, linear mixed models)	★ the power is largely equivalent. ★ trios require 3/2 times genotyping (more costly) and may be **harder** (or impossible) **to collect** ★ **powerful against stratifications**, because you do not include any controls. ★ **sensitive to missing data.** ★ allows **examination of more complex effects**, such as maternal genotype and parent-of origin (imprinting) effects

EXERCISE. Calculate the p value for the significance of A1 and A2



		Non-transmitted allele		
		A1	A2	Total
Transmitted allele	A1			
	A2			
	Total			

Considering the number of heterozygous parents, some of them will give the A1 allele (hence not giving A2), others will give the A2 (hence not giving A1). **Eventually, we can fill in spaces B and C.**

Although they are not important, we can fill in tables A and D too: we increase the count in table A every time an homozygous parent 11 gives one A1 allele, but doesn't give the other A1 allele. The opposite is done for table D.

If two parents are heterozygous and you don't know which is transmitting which allele, consider that both cases yield the same result. 1 and 2 or 2 and 1

➔ A1 and not A2

➔ A2 and not A1

is always the same.

The formula we apply to obtain the p value is the one written above. We get the chi square value from McNemar test, and then obtain the p value as usual, looking at the

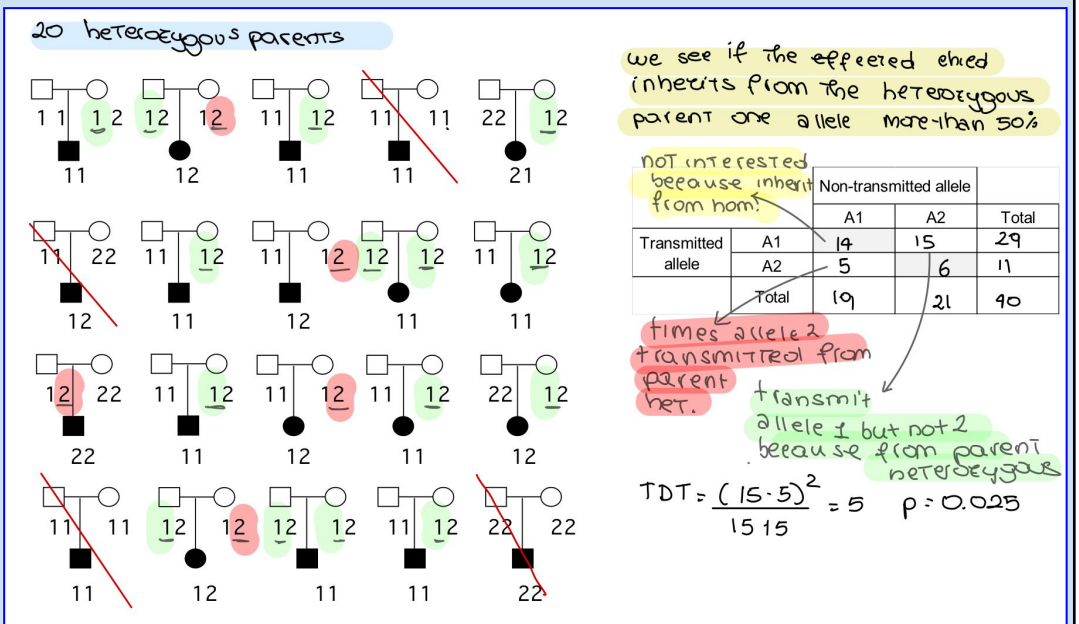


table.

A more sophisticated approach, rather than family studies, is LOGISTIC REGRESSION.

ASSOCIATION STUDIES

LOGISTIC REGRESSION

❖ LINEAR REGRESSION

In statistics, linear regression is a linear approach for **modeling the relationship between a response and one or more explanatory variables.**

$$y=ß_0 +ß_1x$$

- ➔ y is the response variable
- ➔ $ß_1$ is the slope and $ß_0$ the intercept
- ➔ $ß_0 +ß_1x$ is the expected value of y (each yi actually = $ß_0 +ß_1x+\varepsilon_i$), where ε_i is the error
- ➔ **x is a coded genotype variable** taking values (0, 1, 2) for *(dd, dD, DD)* or (aa, aA, AA)
- ➔ $ß_1$ and $ß_0$ can be estimated via least squares or maximum likelihood
- ➔ tests the **null hypothesis that $ß_1 =0$** against the alternative that $ß_1 \neq 0$

❖ LOGISTIC REGRESSION

$$log(\frac{p}{1-p}) = ß_0 + ß_1x$$

- ➔ **p** represents the **probability of being a case** rather than a control
- ➔ we allow the **log odds of disease** $log(\frac{p}{1-p})$ **to vary according to genotype.** This as opposed to allowing the expected value of y to vary according to genotype (*x can be 0, 1 or 2* depending on the genotype as seen below).

The probability of having the disease therefore varies according to genotype.

$$p = \frac{e^{\beta 0+\beta 1\, x}}{1+e^{\beta 0+\beta 1\, x}}$$

The most used regression model is the **ADDITIVE MODEL.**

- ➔ It assumes the log-**odds of disease increase additively with the number of alleles.**

- Our coding scheme assumes two copies of allele A has twice the effect of a single copy on log odds scale
- Corresponds to 'additive' allelic effects on log odds scale, or 'multiplicative' allelic effects on odds scale

Genotype	x	Log odds	Odds	OR
aa	0	β_0	$e^{\beta 0}$	1
aA	1	$\beta_0 + \beta_1$	$e^{\beta 0 + \beta 1}$	$e^{\beta 1}$
AA	2	$\beta_0 + 2\beta_1$	$e^{\beta 0 + 2\beta 1}$	$e^{2\beta 1}=(e^{\beta 1})^2$

- The odds ratio (OR) is the factor by which your odds need to be multiplied if you have 1 (or 2) copies of the risk allele, compared to none
- If genotype has no effect on the odds of disease, then $\beta_1 = 0$ and all ORs=1

The additive model is sometimes called the ***multiplicative*** model.

You **multiply the odds by the odds ratio**, whose value depends on the number of alleles present (which can be, once again, 0 1 or 2)

There is another way of fitting this model (known as a Score Test) and this leads to the test statistic known as the **COCHRAN-ARMITAGE TREND TEST.**

- A more general "genotype" model allows the odds (or probability) of disease to vary arbitrarily in all 3 categories

Genotype	X (factor)	Log odds	Odds	OR
aa	0	β_0	$e^{\beta 0}$	1
aA	1	$\beta_0 + \beta_1$	$e^{\beta 0 + \beta 1}$	$e^{\beta 1}$
AA	2	$\beta_0 + \beta_2$	$e^{\beta 0 + \beta 2}$	$e^{\beta 2}$

- If genotype has no effect on the odds of disease, then β1 = β2 =0 and again all ORs=1

This test allows the odds to vary between three categories.

Genotype	x	Log odds	Odds	OR
aa	0	β_0	$e^{\beta 0}$	1
aA	1	$\beta_0 + \beta_1$	$e^{\beta 0 + \beta 1}$	$e^{\beta 1}$
AA	1	$\beta_0 + \beta_1$	$e^{\beta 0 + \beta 1}$	$e^{\beta 1}$

Genotype	x	Log odds	Odds	OR
aa	0	β_0	$e^{\beta 0}$	1
aA	0	β_0	$e^{\beta 0}$	1
AA	1	$\beta_0 + \beta_1$	$e^{\beta 0 + \beta 1}$	$e^{\beta 1}$

In general, we can say that logistic function **predicts the probability of being a case given a genotype class.**

Logistic regression is often the preferred approach because it **allows for adjustment for clinical covariates** (and other factors), and can provide adjusted odds ratios as a measure of effect size.

Genotype model: $\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 I(G_i=aA) + \beta_2 I(G_i=AA)$

Recessive model: $\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 I(G_i=AA)$

Dominant model: $\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 I(G_i=aA \text{ or } G_i=AA)$

Additive model: $\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_i$

where x_i takes values (0, 1, 2) for genotypes (aa, aA, AA)

For example if we want to carry out tests conditional upon (i.e. allowing for):

- Age
- Sex
- Population of Origin of the individuals

we could compare the following 2 models:

$$\text{Null} : \log\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1\text{Age}_i + b_2\text{Sex}_i + b_3\text{Pop}_i$$

$$\text{Alternative} : \log\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1\text{Age}_i + b_2\text{Sex}_i + b_3\text{Pop}_i + b_4\text{X}_i$$

PRINCIPAL COMPONENT ANALYSIS is able to compute the **covariates** **of a logistic regression analysis as eigenvectors and eigenvalues**.

The results can be:

- ➢ used to **exclude outlier samples** (as we have done in the PLINK practical)
- ➢ **included as covariates** in a logistic regression.

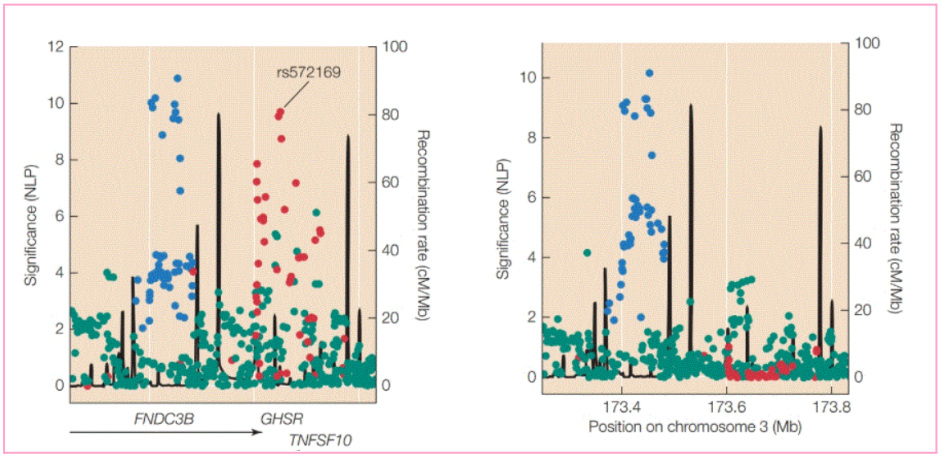
Another useful application of logistic regression is to perform conditional associations.

CONDITIONAL ASSOCIATION is used to distinguish the effects of **different variants** at a **single locus**.

Usually, when coming in contact with different variants in the same haplotype block, we can reach two conclusions. Conditional association helps us understand which one is the true one.

- → **all variants have a small effect so** they contribute to a small proportion and combine to yield a strong signal
- → a **single SNP is fully responsible** and the other SNPs are linked just because in LD with the main SNP.

The idea is to take the **SNP with** the smallest p-value (hence the **highest significance**) **as a term** in the regression model, then evaluate each of the other SNPs conditionally on the already discovered genotype.



In the example here, we do condition associations on red SNPs because it was already found that those are associated with height.

We are trying to see if the blue SNPs are in association with the red ones.

We reduce the red signal in the second picture, however the blue SNP sample remains exactly the same. This means that the blue cluster is independent and hence has independent associations.

Logistic (or linear) regression offers a very convenient way to **model** (statistical) **interactions between variables.**

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

$$\log \frac{p}{1-p} = \beta_0 + \beta_2 x_2$$

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Main effects of factors 1 and 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

Main effects and interaction term

FOR QUANTITATIVE TRAITS USE LINEAR REGRESSION.

ASSOCIATION STUDIES

MULTIPLE TESTING

❖ CORRECTION FOR MULTIPLE TESTING: BONFERRONI

Statistical tests are generally called **SIGNIFICANT** and the null hypothesis is rejected if the **p-value falls below** a predefined alpha value, which is nearly always set to **0.05.**

➔ A p-value is generated for each statistical test.

In the case of **GWAS**, hundreds of thousands to millions of tests are conducted, each one with its own false discovery rate (FDR).

1. Let **α** be the type I error rate for a statistical test (the *probability that a null hypothesis is rejected when it is actually true*).
2. If the test is performed ***n*** times, the **experimental-wise error rate α'** is given by:

$$\alpha' = 1 - (1 - \alpha)^n$$

Because of the multiple testing problem, **the test result may not be that significant** even if its p-value is a significant level α. In order to solve this,
To solve this problem, the **nominal p-value needs to be corrected/adjusted.**

The **BONFERRONI CORRECTION adjusts the alpha value** from a = 0.05 to

$$a = (0.05/k)$$

where k is the number of statistical tests conducted.
This correction is the **most conservative**, as it assumes that each association test of the 1,000,000 is independent of all other tests – an assumption that is generally untrue due to linkage disequilibrium among GWAS markers.

➔ for a typical GWAS using 1,000,000 SNPs, statistical significance of a SNP association would be set at 5x10-8.

❖ CORRECTION FOR MULTIPLE TESTING: FALSE DISCOVERY RATE

FALSE DISCOVERY RATE (FDR) which is an **estimate** of the proportion of significant results (usually at a = 0.05) that are **false positives.**

The Benjamini and Hochberg step-up procedure (BH) and the Storey's q-value are the classical procedures to compute FDR. The q-value was introduced by Storey as a more powerful approach to controlling the FDR.

We can consider the **q value**, which is the **minimum FDR** incurred when calling a test significant. It indicates the proportion of statistically significant results obtained by the original test procedure that are false positives.

The q value is a measure of significance that **can be "attached" to each individual feature** (as each SNP), it can be **derived from ranked p-values** (from smallest to largest).

➔ The q value of a particular feature in a genomewide data set is the expected proportion of false positives among all features as or more extreme than the observed one.

Therefore the q-value provides an **estimate of the number of true results** among those called significant.

➔ Given q, which is the minimum value of FDR, all the features that have **α > FDR** can be considered **true positives.**

❖ CORRECTION FOR MULTIPLE TESTING: PERMUTATION TESTING

It is another method to establish significance through *random assignment.*

A permutation test involves two or more samples. The null hypothesis is that all samples come from the same distribution.

➔ Under the null hypothesis, the distribution of the test statistic is obtained by calculating all possible values of the test statistic under possible rearrangements of the observed data. Permutation tests are, therefore, a form of resampling.

The permutation testing is based on the generation of an empirical distribution of the statistical tests.

1. The standard statistical test is performed and a P value is obtained (**the original P value**).
2. The permutation test **shuffles the case-control status** of each sample randomly and then runs the association analysis.
3. This process is repeated thousands of times (permutations) as determined by the user and the P values estimated in each permutation are retained.
4. The distribution of permutation P values (empirical distribution) is compared with the original P value
5. This comparison leads to an **estimate of how often the original P value would occur by chance if the study was repeated many times**: this estimate is obtained by checking the percentage of permutation P values that are smaller than the original P value.
6. This ***percentage is the P value for the permutation test.***

The resulting p value is more accurate and is preferable.

While the method looks attractive, it is somewhat computationally intensive. Several software packages have been developed to perform permutation testing for GWAS studies, including the PLINK software.

STATISTICAL POWER

❖ INTRODUCTION TO STATISTICAL THINKING

In the following table we see some statistical considerations in a genetic association study.

Phase of the study	Statistical issue	Considerations
Planning	Statistical power	How many subjects should be included Ratio of cases and controls Which variants should be included based on their frequencies and functionality Study design considerations such as collection of data on potential confounders for later statistical adjustments and stratification Instrumental variable choice and assessment for Mendelian randomization studies
Post-genotyping, pre-analysis	Quality control	Identification of cryptic relatedness among subjects to exclude related individuals Systematic error due to batch, study center, or cohort effects Genotyping control tests including missingness assessment and Hardy–Weinberg equilibrium
Analysis	Association tests	Single or multiple variant analysis, haplotype analysis Assessment of confounding and effect modification (interaction) Assessment of population stratification Adjustment for multiple testing
Post-analysis	Validity and utility assessment	Causality assessment Clinical validity analysis Genetic risk profiling Potential biomarker development process

We already saw how quality control should be in the practical part of the course: it includes assessment of missingness and HWE between the others. We have also already seen association testing. Now, we must take a step back and go back to the beginning of any statistical study.

One of the first, main problems, is the **planning of the study.** During planning, one must consider the **STATISTICAL POWER.**

In statistics, the **POWER** of an hypothesis test is the **probability that the test correctly rejects the null hypothesis** H_0 when a specific alternative hypothesis H_1 is true.
This, in brief, means that there is indeed **association**.

However power calculation is not easy as it depends on many different elements.

❖ POWER: USEFUL DEFINITIONS

- H_0 = null hypothesis ; $H_{alt} = H_1$ = alternative hypothesis
- Power (1-β)= P(Reject H_0 | H_{alt} is true)
- Type II error (β)= Fail to reject H_0 when H_0 is false (false negative
- P-value = P(≥ Observed statistic | H_0 is true)
- Type I error (α)= Reject H_0 when H_0 is true (false positive)

	Reject H0	Fail to Reject H0
Reality: H0 is True	Type I error (probability = α)	Probability = 1-α
Reality: H0 is False	Power (1-β)	Type II error (probability = β)

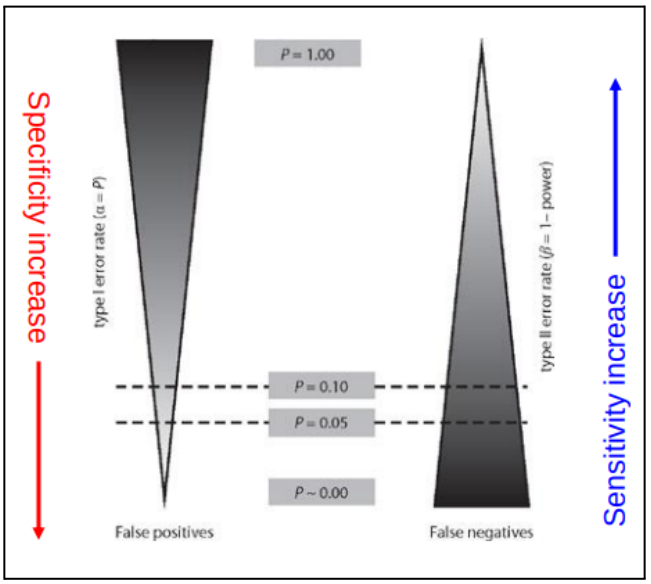
❖ POWER: TRADE OFF BETWEEN SENSITIVITY AND SPECIFICITY

Type I **(false positives)** and type II **(false negatives)** errors have a **relationship** between each other.

In the graph, on the left we have false positives, on the right we have false negatives.

➔ If the statistical significance threshold **(α) is lowered** (*i.e. P=0.01*) **specificity increases** (minimize false positive results), but the **sensitivity decrease** (more false negative, less power to detect associations reaching this significance value)

➔ If the **α is increased** (set for lower stringency i.e. P=0.10) **sensitivity increase**s (more associations will be statistically significant and there will be fewer false negatives), but the **specificity decreases** (more false positives).



Usually:

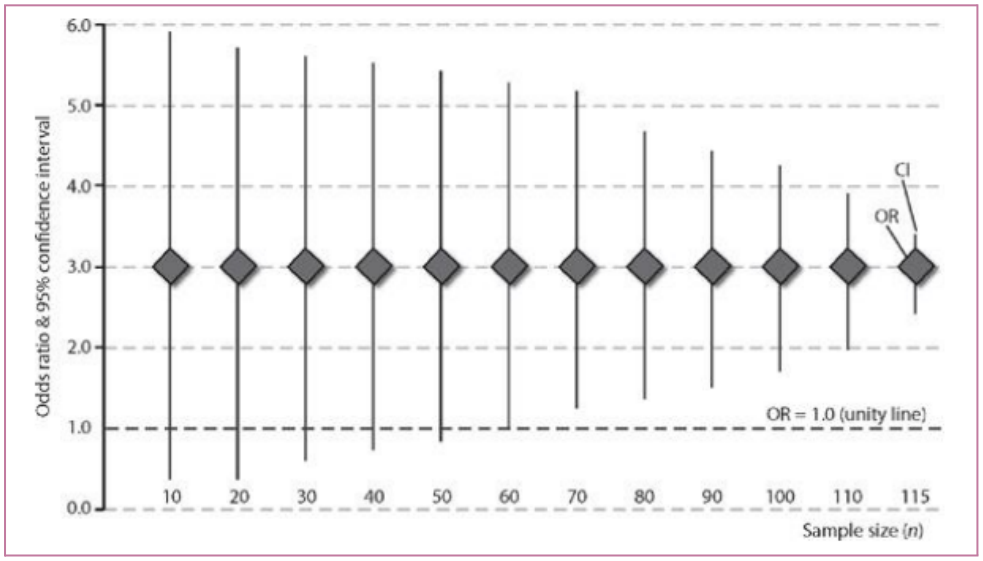
➔ α is placed at **P=0.05**

➔ the statistical power at **0.8**

which is an acceptable trade-off between specificity and sensitivity.

❖ POWER DEPENDS ON VARIOUS ELEMENTS

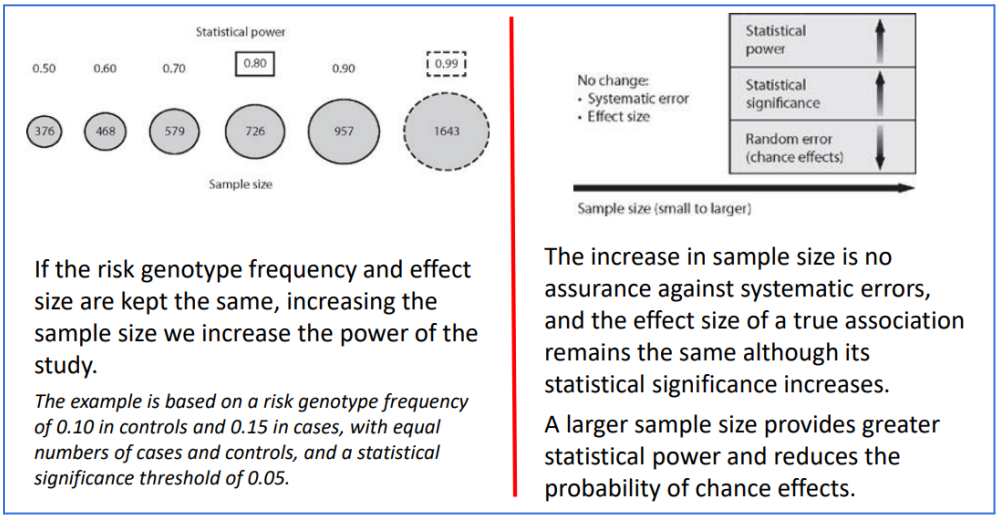
In any association study, **sample size is the major determinant** of statistical power. There are also minor determinants, which will also be described in these pages.



An odd ratio = 1 means no effect. Hence, until we reach a population of size 60, we do not cross the threshold, meaning that we do not have a significant result and a p-value bigger than 0.05.

In practice, like we will see on the statistical power calculator, **statistical power calculations are performed for two scenarios.**

1. For a **set power level** (usually 0.8), what should be the **sample size?**
2. For a **fixed sample size**, what would be the **power** for a certain association?



In any association study, **SAMPLE SIZE** is the major determinant of statistical power.

EFFECT SIZE is a value measuring the strength of the relationship between two variables in a population. In case-control association studies, the **effect size is measured as an OR.** for a complex disease, the effect size will usually be within the range of OR = 1.2 to 1.5 for individual SNPs.

FREQUENCY OF THE RISK GENOTYPE. In genetic association studies, **frequency of exposure equates to frequency of risk genotype**: the higher the risk genotype frequency, the higher the statistical power. For statistical power calculations, **allele frequencies are converted to risk genotype frequencies** using a genetic risk model.

- ➔ common (prevalent) disease, high-risk variants are assumed to be lacking in the control group selected to be disease-free, while cases are expected to be enriched for high-risk variants.
- ➔ For rare diseases, no large divergence is possible to help with statistical power.

This large **difference in the frequency** of alleles between case and control groups **increases statistical power,** hence why common diseases are easier to analyze. This is because it reflects how the cases are associated with the risk genotype.

LINKAGE DISEQUILIBRIUM between the marker being examined and the unknown causal variant must be estimated as it is unknown. In GWAS, when there are so many markers, it is safer to calculate statistical power for a range of potential LD values. As **r^2 gets smaller**, the variant being examined will be **less representative of the causal variant,** and the statistical power will be lower.

- ➔ **Representativeness is an issue in a GWAS when common variants that are presumed to be proxies for usually rare causal variants are used.**

"The situation **MOST FAVORABLE** *for the association test occurs when the* ***disease allele is common*** *and when the* ***marker allele in positive disequilibrium with it has roughly the same frequency."***

Power to detect association is highest if:

- Risk allele is fairly common
- Frequency of associated marker allele is close to the frequency of the risk allele
- Marker is in high LD with risk locus
- Genetic effect is big enough.
- Genetic model is 'simple'
  - no G x G, G x E.

❖ POWER IN REPLICATION STUDIES.

A genetic study is often affected by the **WINNER'S CURSE.**

A genetic **effect size** estimate based on a genomewide scan is **biased upwards**,
This means we should **adjust the effect size estimate** downwards before using it to estimate power. We need high power also in replication studies and not only in association studies. The **replication sample** should be **larger** in order to account for this over-estimation.

Usually a combined analysis (or meta analysis) is more powerful than two stage analysis.

There is no remedy for low power after the study is concluded.

GENOTYPE IMPUTATION

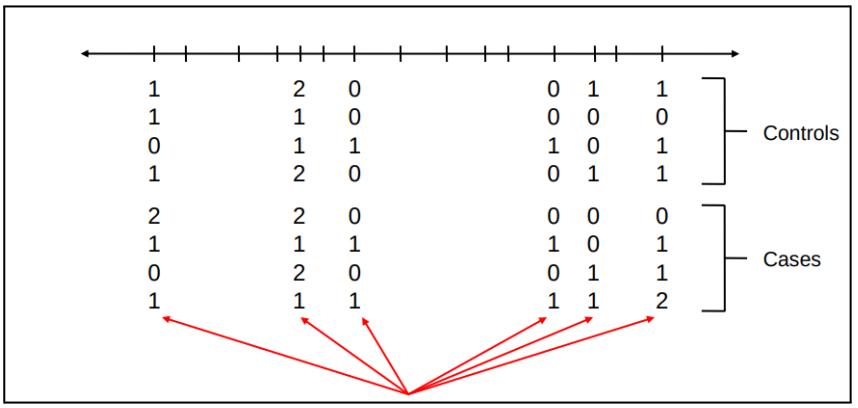
GENOTYPE IMPUTATION has become a standard tool in GWAS, as it enables researchers to **approximate whole genome sequence data** from **genome-wide single-nucleotide polymorphism array data.**

➔ This is possible because of the genotype of **unsequenced SNPs that are in high correlation with sequenced SNPS.** Hence, the method relies on **LINKAGE DISEQUILIBRIUM.**

It therefore increases the power of GWAS.

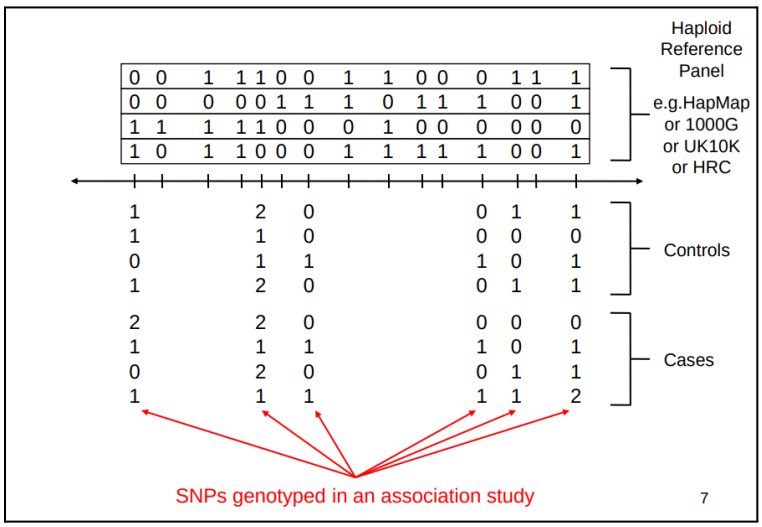
We define **IMPUTATION** as the *process of estimating the most likely genotype at SNPs not directly genotyped in the study*. This is done thanks to the knowledge of LD patterns and haplotype frequencies from reference panels (HapMap, 1000 Genomes).

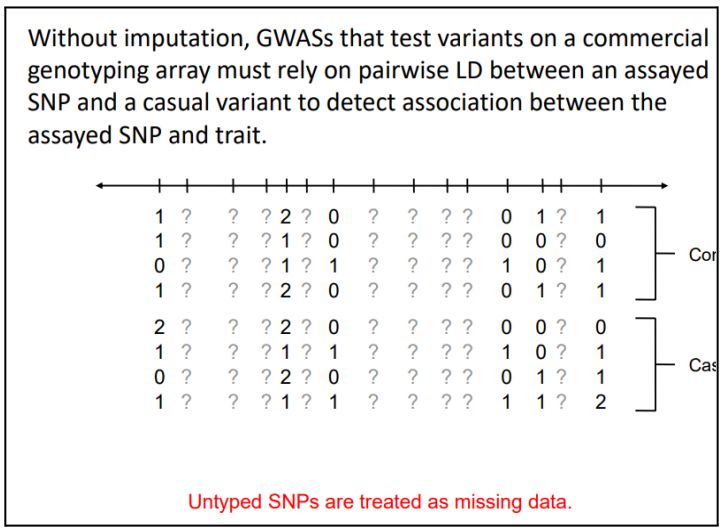
A typical GWAS study consists in genotyping 100.000 SNPs in 1000+ individuals. This is only a small proportion of the actual SNP amount in our genome, but the SNP are selected according to haplotypic blocks.



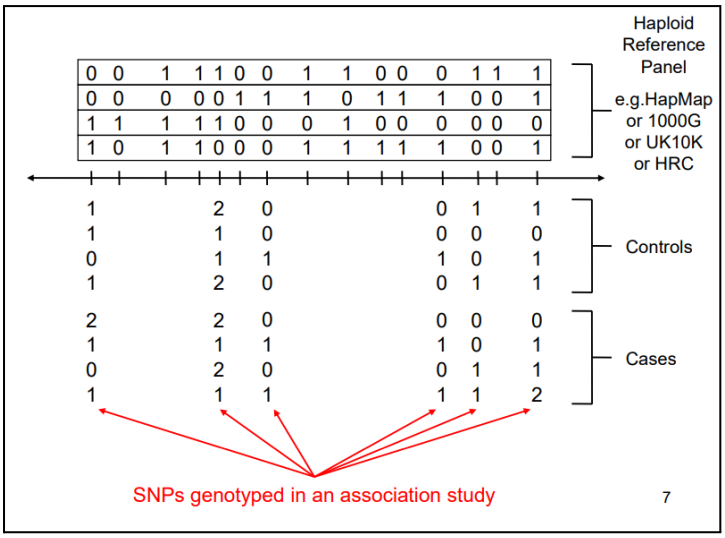
In the image, each row is an individual and each column is a SNP.
As we can see we haven't genotyped each nucleotide, just the tag SNPs.
In this case the numbers 0,1,2 do not indicate three different alleles. Rather, they are indicating the amount of copies of a certain allele *A* present in an individual.
If the individual is homozygous for allele A then he will have 2 copies, if it is heterozygous it will have only 1 copy and, if it's homozygous for the other allele it will have 0 copies.

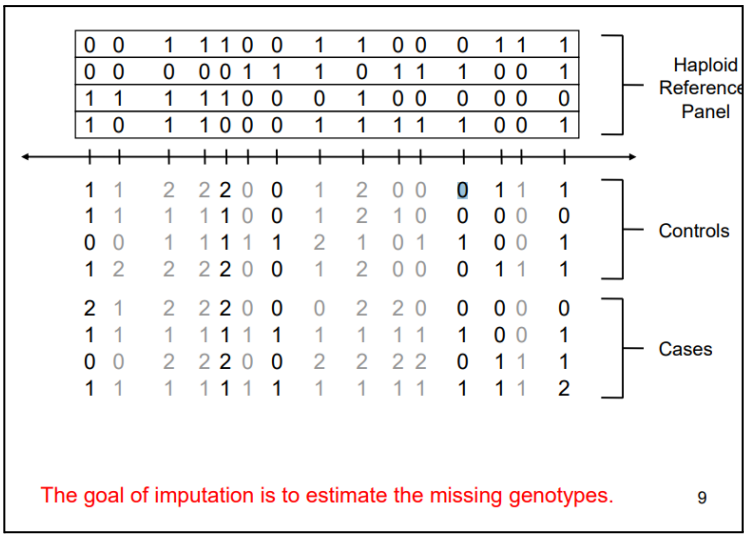
We now take into consideration four haplotypes with different SNPs (in their haploid version). Thanks to the copy number of our tag SNPs, we can understand how many alleles (0 or 1) present in each of the two chromosomes.





Because we know the number of alleles present at every single tag SNP loci, we can understand, thanks to linkage disequilibrium, which non genotyped SNPs are present.





★ Genomic Imputation is **useful for common variants.**

★ **Rare variants** instead are **hard to predict**, as they tend to have low levels of pairwise gene disequilibrium. That is because, to have correlation with the causal variant (the one which represents the block), also the **SNPs we are estimating must have the same frequency as the causal variant.**

After imputation, association analysis is carried out.

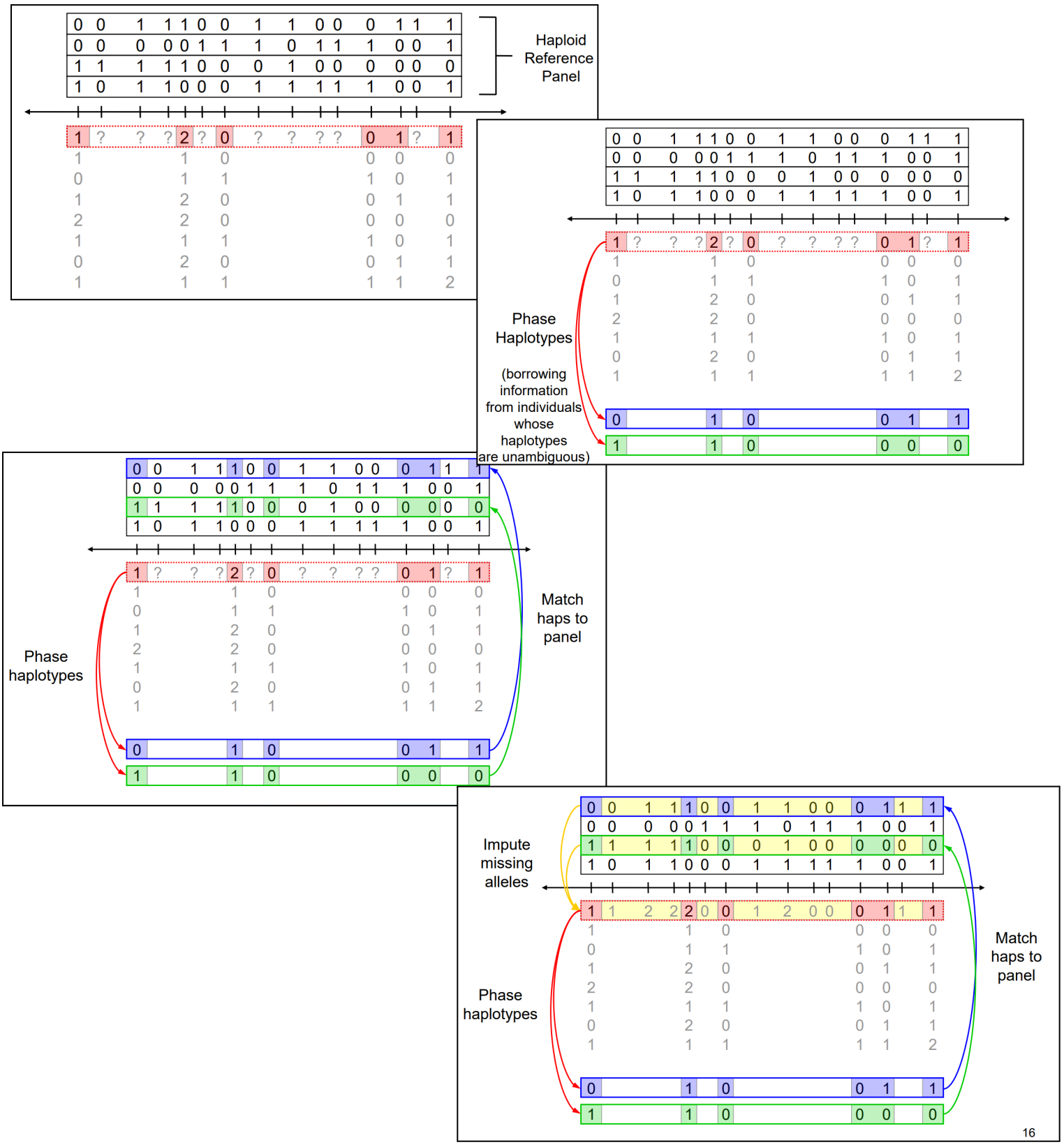
❖ BASIC IDEA AND MODELS

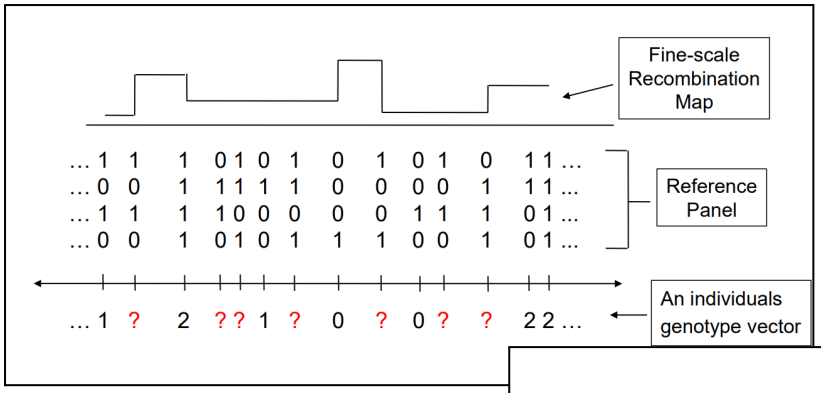
The first methods used for imputation were born in 2007. All the following approaches derive from that one. The point where the genotype changes is the point of recombination.

LI AND STEPHENS MODEL

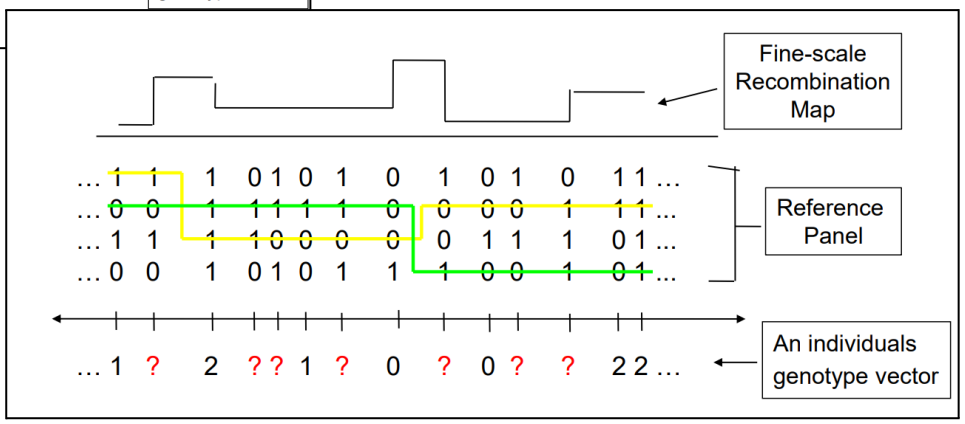
Every individual, even if apparently unrelated, can share short stretches of chromosomes with another individual, deriving from a common ancestor.
The **genotyped SNPs** can be used to find sequences that **match** between the obtained genotype and a **reference set of haplotypes already derived** in the past and present in databases.

Given the **amount of variant alleles** we can understand which haplotypic block we have for **each SMP tag, in each of the two chromosomes.** To do this, we must remember that the haplotypes are in the haploid form, while the genotyping of the tags gives us information about both chromosomes.


Haploid Reference Panel
Phase Haplotypes
(borrowing information from individuals whose haplotypes are unambiguous)
Phase haplotypes
Match haps to panel
Impute missing alleles
Phase haplotypes
Match haps to panel
16



The model says that an individual **genotype is constructed by copying alleles along two paths** (one per chromosome) through the space of haplotypes.

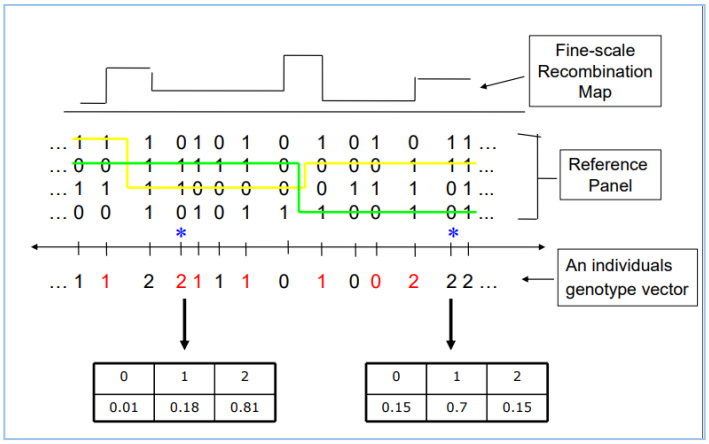


The individual will be a mix of haplotype blocks, some coming from the father and others coming from the mother.

The switch rates of the paths are controlled by the recombination map. Mutation events are also allowed.

Paths are sampled with probabilistic models, each genotype vector has an assigned probability. Meaning that the model produces a **GENOTYPE UNCERTAINTY MEASURE.**

The **score of the associated probabilities** can be used in order to **take into account the uncertainty in the genotype estimation process**. This uncertainty is due to the fact that **we may have different models**



1. Compute a **threshold probability distribution** to give the best guess genotype calls.
2. Rather than assigning a single genotype, the program **assigns multiple alleles, each with their own probability.** From this the ***allele count*** is computed.

In this case instead of using just the most frequent genotype we also consider all the less possible genotypes thanks to the following formula.

AA	AB	BB
0.01	0.18	0.81

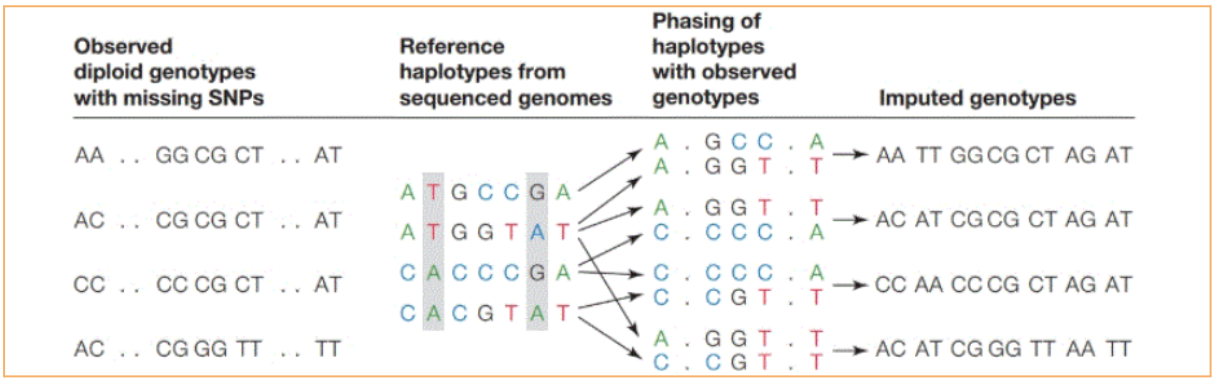
→ $0x0.01 + 1x0.18 + 2x0.81 = 1.8$

3. **Average over the uncertainty** (it can be done in both frequentist and bayesian frameworks).

In general, **imputation is a probabilistic process**.

The computational burden for genotype imputation is very high. To reduce it, the process can be divided into **two phases**:

1. **PHASE** the haplotypes: to estimate the series of haploid genotypes on each of the two contributing chromosomes observed in a diploid individual
2. **IMPUTATION** into the estimated study haplotype: to estimate genotypes for SNPs not directly genotyped in the study



In the picture, each line corresponds to a diploid individual. For each locus we have an allele for each of the two chromosomes, one coming from the mother and one coming from the father. Separately the two chromosomes would be:

1. A…GCC…A
2. A…GGT…T

Now we can compare the tags on our two chromosomes to reference haplotypes. Looking at the reference we can see, in the different cases, how to fill in chromosome gaps. In the last part of the picture we can see the sequences in their diploid form, but with their gaps filled.

Now, from a more theoretical point of view, here are our steps following the table.

1. The first step is to **phase the haplotypes:** given the diploid sequence on which both chromosomes are present, we need to **estimate the series of haploid genotypes** on each of the 2 contributing chromosomes observed in a diploid individual.
   - ➔ if this is homozygous, you will know the haplotype present in both alleles.
   - ➔ if it is heterozygous you must use probability.

2. **Maximum likelihood methods** are used to **optimize the fit of the observed genotypes** to the frequency and identity of haplotypes in a reference population.

The identity of missing genotypes (. .) in the observed sequence is “imputed” by reference to the sequences in the reference panel.

❖ QUALITY CHECKS

Before doing imputation you must do QUALITY CHECK:

- ➔ Check that your **typed SNPs** are mapped to the **same genome build as the reference** panel you are using (the commonly-used genome build is GRCh37.
- ➔ Check your **genotype strand**: strand of study samples needs to match the reference panel strand (+ strand). (Info about chip strand should be found at chip website)
- ➔ Phasing the GWAS samples, to perform a faster imputation.

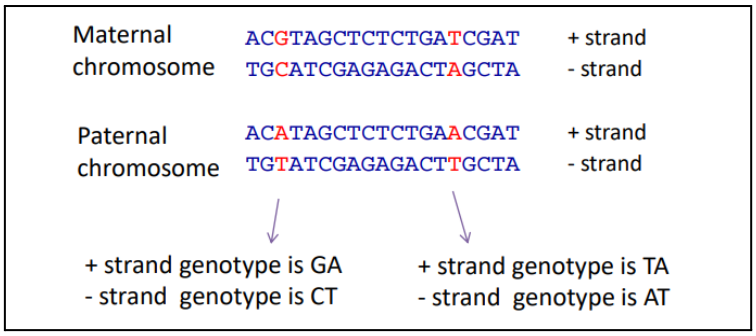
GENOME BUILD

The Human Reference Sequence is updated periodically, each version is referred to as a 'genome build'. **Positions of SNPs can change between builds.**
Almost all imputation programs align SNPs between the reference panels and the GWAS datasets using the position of SNPs. It is very important that the genotypes of your **GWAS are mapped to the same genome build as the reference panel** you are using.

STRAND ISSUE

Genotyping from SNP chips are called relative to either the + or – **(forward or reverse)** strand of the human reference genome.

The brand decides the strand depending on which strand is more suited to sequence that particular SNP. This can lead to mistakes if not done correctly.



Genotype chips can have a mixture of genotypes from + and - strand

This needs to be fixed prior to imputation (strand of study samples needs to match the reference panel). Sometimes you have the genotype data and you don't know which strand we are using: there are **different sites we can use to find the right strand, and we can also use probability.**

QUALITY METRICS

After imputation we need to check **how well imputed the genotypes are at each SNP.**
The most interpretable measures of imputation accuracy are based on the correlation between the imputed and the true dose of an allele (r^2). Statistical methods allow us to estimate the r^2 even if, technically, we don't know the true allele.
We have two types of files:

- ped file
- info file, where for each marker it is recorded how well the imputation was performed

It is common to **exclude poorly imputed markers from downstream analysis by requiring the $r^2 > 0.3$.** Accordingly to the analysis we are carrying out we might be able to exclude/include different markers.

❖ FACTORS AFFECTING THE ACCURACY OF IMPUTATION

1. The genotyping chip we use to genotype our tags can be biased. This is because different populations have different variants, so a sequencing based on **SNPs can be imperfect when it comes to analyzing different populations.**
2. Allele Frequency, as the imputation of **rare alleles** is more difficult.
3. **Increasing the size** leads to an increase accuracy especially for rare variants
4. **Reference panels also affect the accuracy of imputation.** HapMap was the first reference panel, but now we have newer ones as we can see in the table. In particular, **the size of the panel can influence our imputation.** Increasing the size of the sample reference (meaning the ***array is denser***), we increase the accuracy, especially for rare variants. Another factor related to the choice of the reference panel, is the average **sequence coverage.** Ancestry distribution is very important: if we have multiethnic samples, we must use a **multiethnic reference panel.** Moreover, some panels are not publicly available.

Table 2 The most commonly used public reference panels to date

Reference panel	Number of reference samples	Number of sites (autosomes + X chromosome)	Average sequencing coverage	Ancestry distribution	Publicly available	Indels available	Reference
International HapMap Project phase 3	1,011	1.4 million	NA[a]	Multiethnic	Yes	No	47
1000G phase 1	1,092	28.9 million	2–6×	Multiethnic	Yes	Yes	1
1000G phase 3	2,504	81.7 million	7× genomes, 65× exomes	Multiethnic	Yes	Yes	3
UK10K Project	3,781	42.0 million	7× genomes, 80× exomes	European	Yes	Yes	89
HRC	32,470	40.4 million	4–8×[b]	Predominantly European[c]	Partially[d]	No	69
TOPMed	60,039	239.7 million	30×	Multiethnic	Partially[e]	Yes	71

Abbreviations: 1000G, 1000 Genomes Project; HRC, Haplotype Reference Consortium; indel, insertion or deletion; NA, not applicable; TOPMed, Trans-Omics for Precision Medicine.
[a]The International HapMap Project phase 3 data were genotyped on the Illumina Human1M and Affymetrix 6.0 SNP arrays.
[b]The HRC panel was obtained by combining sequencing data across many low-coverage (4–8×) and a few high-coverage sequencing studies.
[c]The only non-European samples in the HRC panel are through the 1000G reference panel (which was a contributing study).
[d]Most of the HRC samples (~27,000) are available for download through controlled access from the European Genome-Phenome Archive.
[e]Some of the TOPMed samples (~18,000) are available for download through controlled access from the Database of Genotypes and Phenotypes (dbGaP).

An example multiethnic panel is obtained by the 1000 genome project.It's size isn't as big as the Haplotype Reference Consortium (HRC), which however focuses on European populations. Hence, HRC is more useful on rare variants, while the 1000 genome project is better for multiethnic populations.

❖ DATABASES

In these servers you upload your sequence, pick a reference panel, then do phasing and imputation. Before this it is important to do quality check, GWAS QC and imputation specific QC. At the end you just download the results, which are potentially really huge files. After imputation, we also carry out post-imputation checks.

HRC IMPUTATION

The HRC data is NOT publicly available, as the HapMap and 1000 GP haplotypes are, due to consent issues. A subset of HRC haplotypes has been made available for the sole purpose of imputation. It is better to not download the reference panel, but use imputation servers, as the reference panel is very big.

TOP MED

The Trans-Omics for Precision Medicine (TOPMed) program, is part of a broader Precision Medicine Initiative, which aims to provide disease treatments tailored to an individual's unique genes and environment.
TOPMed contributes to this Initiative through the integration of whole-genome sequencing (WGS) and other omics (e.g., metabolic profiles, epigenomics, protein and RNA expression patterns) data with molecular, behavioral, imaging, environmental, and clinical data.

- ➔ it includes a variant panel
- ➔ genotyping imputation server includes **a lot of rare variants**
- ➔ reaches a **higher power** in genome association studies.

As of September 2021, TOPMed consists of ~180k participants from >85 **different studies** with varying designs.

- ➔ Prospective ***cohorts*** provide large numbers of disease risk factors, subclinical disease measures, and incident disease cases
- ➔ ***Case-control*** studies provide large numbers of prevalent disease cases;
- ➔ Extended ***family structures*** and population isolates provide improved power to detect rare variant effects.

One of its important features is the **participant diversity,** which fills the lacking diversity there was with non european populations.
Achieving ancestral and ethnic diversity is a priority in selecting contributing studies.
Currently, 60% of the 180k sequenced participants are of predominantly nonEuropean ancestry. This is good, as the patterns of LD are more complex in the African population, rather than European ones.

MICHIGAN IMPUTATION SERVERS.

This server, similar to the other ones, has the following workflow.

1. Prepare your data
2. Register and Login
3. Upload your data
4. Start the Imputation
5. Download Results

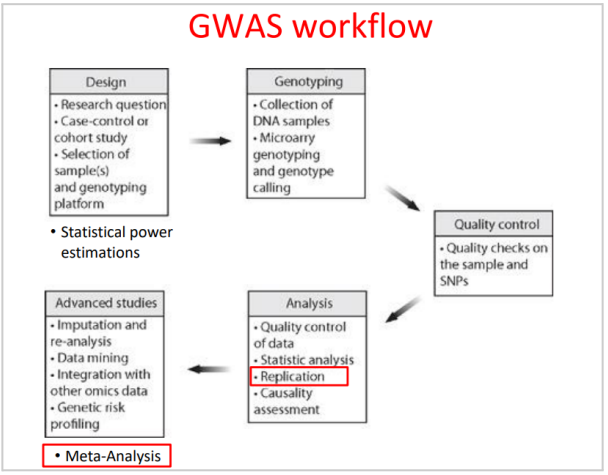
❖ ADVANTAGES OF IMPUTATION:

- ★ **Increase the power** of a study, up to 10%.
- ★ Increase the chance to find the causal variant
- ★ **facilitates meta-analysis**, hence the study of the same variant in different contexts.

META ANALYSIS

In a **GWAS workflow**, we have many different steps:

- **design** (where, as we have learned, we must be careful of the power of the study)
- **genotyping**
- **quality control** of samples and SNPs
- **analysis**: it implies quality control of data, statistic analysis, causality assessment and replication
- **advanced studies** which <u>contain</u> **META-ANALYSIS.**



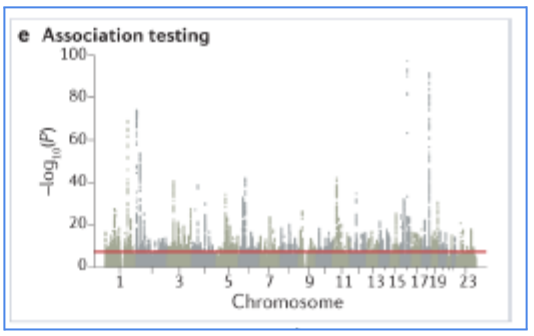
After the quality control and the association testing in particular, we must consider:

- replication
- meta-analysis
- other post GWAS analyses

❖ ASSOCIATION TESTING

Association testing is done through the organization of a **MANHATTAN PLOT.**

- ★ a MANHATTAN plot implies that association analysis is being executed.
- ★ on the Y axis we have a function of the p value, which signals the significance
- ★ on the X axis we have the chromosome, on which the SNPs are marked.



❖ REPLICATION

The gold standard for **validation of any genetic study** is replication, which must be as consistent as possible. Some criteria must be followed to do replication:

1. Replication studies should be **well-powered**: have sufficient sample size to detect the effect of the susceptibility allele. Replication samples should ideally be larger to account for the over-estimation of effect size of the initial GWAS **(winner's curse)**

WINNER'S CURSE refers to the fact that the **detected allele is stronger in the GWAS sample compared to the general population**. The reason behind this is that, when we do association studies, we must correct for multiple testing. This means keeping the threshold quite low.
The positive side is that we get to find the association, however it also means there's inflation. Replication can fix this because you obtain results with different sampling.
This means that a study should have a sample size big enough for resampling to take place and fix the winner's curse.

2. Replication studies should be conducted in **an independent dataset drawn from the same population** as the GWAS. Once an effect is confirmed in the GWAS target population, other populations may be sampled to determine if the SNP has an ethnic-specific effect.
3. **Identical phenotype criteria** should be used in both GWAS and replication studies.
4. **A similar effect** (in magnitude and direction) **should be seen in the replication set from the same SNP,** or a SNP in high LD with the GWAS-identified SNP. The unit of replication for a GWAS should be the genomic region, and all SNPs in high LD are potential replication candidates.

The general strategy for a replication study is to **repeat the ascertainment and design of the GWAS as closely as possible**, but examine only specific genetic effects found significant in the GWAS.

➔ effects that are consistent across the two studies can be labeled replicated effects.

❖ META ANALYSIS

The results of multiple GWAS studies can be pooled together to perform a meta-analysis.

Notice that this doesn't mean various replications of the study are put together, rather, ***various studies*** **are put together.** All studies included must have the same hypothesis.

- ➔ ***META*-ANALYSIS** refers to the practice of **combining summary statistics** from multiple studies
- ➔ ***MEGA*-ANALYSIS** is the joint-**analysis on all the individuals** from multiple studies (which is usually logistically very difficult (if not impossible) to do).

Combining summary statistics means using summary data rather than all data. This eliminates privacy problems.
By **increasing the effective sample-size**, meta-analysis **increase the statistical power** of the analysis and **reduces false-positive findings.**

Each study included in meta-analysis should:

1. have a **similar general design**
2. follow **near-identical procedures** for the study-level SNP analysis
3. allow ***standardized*** **SNPs quality control procedures** (along with any covariate adjustments, and the measurement of clinical covariates and phenotypes should be consistent)
4. involve an ***independent sample set.*** We cannot have the same sample (or individual) in different studies, because it would lead to inflation.
5. report results **relative to a common genomic build** and reference allele.

It's **rare to find studies that match on everything.** Hence why we need to do studies on heterogeneity.

- ➔ a **HETEROGENEITY STUDY** statistically quantifies the degree to which studies differ.

Different studies often used **different genotyping platforms that genotype different SNPs** within a haplotype block. To solve this:

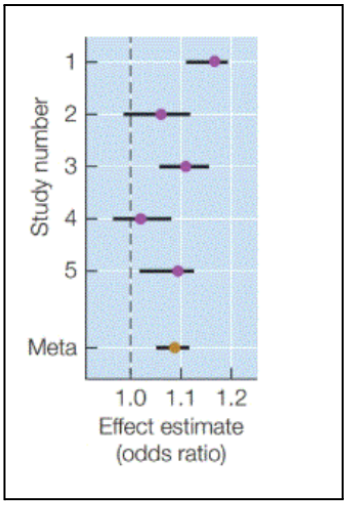
- ➔ To conduct a meta-analysis properly, it is necessary to **generate a common set of SNPs** so that the separate datasets can be combined and the effect of the same allele across multiple distinct studies must be assessed.

Meta-analysis of genome-wide association summary statistics:

- is useful as it doesn't cause problems for privacy
- it does not involve the integration of genotypes, only of the summary statistics
- it is less affected than usual by the Winner's Course. That is because one study may be affected by it, but not all of them are.

❖ FOREST PLOTS

Results of meta analysis can be seen on forest plots.



- ***x axis*** we have the *effect estimate.* In particular, for each study we have a **95% confidence interval for the effect of each allele.**
- **y axis** we have *different studies*

All the **studies are aligned with one another.**
What we get at the end is the meta analysis effect **odd ratio** and the meta analysis **p value.** We have a greater power from the combined sample sizes.

➔ The meta-analysis p-value can be thought of as an average p-value weighted by the study sizes and boosted by the combined sample size.

In the plot:

➔ In studies 2 and 4 do not have SNPs which contribute to the disease (the **OR is under 1**). This might be due to the fact that those studies do **not have enough power**, having a smaller sample size.

➔ In study 1 instead we have a high odd ratio (**OR is over 1**), which indicates contribution to the disease. However it might be just because of the **winner's curse**.

This is **why we need to consider different studies.**

The p value in data analysis is smaller, hence being more precise.
This is because according to the CENTRAL LIMIT THEOREM, the more studies we use the better the estimate.

We can have **DIFFERENT MODELS FOR META- ANALYSIS,** based on different things.

1. **P-VALUE BASED** (Z score approach): It converts the direction (positive/negative) of the effect and P-value observed in each study into a new statistics, which is a signed **Z-score.**

Z-scores are **combined across samples in a weighted sum**, with weights proportional to the square-root of the sample size for each study. This is done in order to consider the sample size.

- **Pros:** very flexible and allows results to be combined even when effect size estimates are not available or the β coefficients and the standard errors from individual studies are **in different units**
- **Cons**: it cannot provide an overall estimate of effect size, because the p value only gives us the significance of the SNP.

2. **EFFECT SIZE ESTIMATE,** which can be either fixed or random. In any case, it makes effect size estimates weighting on the standard error.
    a. **Fixed effects** (inverse variance based): It weighs the effect size estimates, or β-coefficients, by their estimated standard errors. That is because our confidence intervals are based on it. As we can notice, the effect is not based on the sample size. This method assumes that the true effect at each risk allele is the *same* in each data set.
        - Pros: very powerful and it gives an idea on the effect size
        - Cons: It requires effect size estimates and their standard errors to be in **consistent units** across all studies.
    b. **Random effects:** it's another synthesis of effect sizes. It assumes that each study estimates different effects.
        - Pros: more limited power compared to fixed effect
        - Cons: useful for predictive purpose

FIXED EFFECT	RANDOM EFFECT
All SNPs have the same effect in all of the studies, so we can just use them and make an average	This method considers the fact that **the same risk allele can have different effects** on different populations. Hence, we make an average considering also the population.

Therefore effect size estimate:

- it is not robust
- needs effect size estimate in order to work
- it is useful for predictive purpose not only under a fixed effects model, however the latter is the most used
- it provides an overall estimate of effect size

BENEFITS AND LIMITATIONS OF GWAS

As for 2016, approximately 10.000 strong associations were reported between genetic variants and complex traits.

The discovery of new association signals has increased overtime, so GWAS has demonstrated to be very powerful when the sample size is big enough.

➔ we have an **increase in association studies,** but more importantly **sample size has increased** overtime. This is of paramount importance as the sample size is the main determinant of the power.

In 2019 the number of significant associations was >50.000, they are all stored in the ***GWAS Catalog***, where you can plot, visualize and download data.

❖ BRIEF LIST OF GWAS BENEFITS

Two main benefits:

1. led into **insights on disease architecture**: GWAS was able to cover **identification of novel disease causing genes and mechanisms thanks to the finding of variants.** That said, finding an association and understanding the real causal variant is different and takes a lot of time

2. led to **advances in clinical care** and personalised medicine. GWAS favored identification of **new drug targets** and disease biomarkers. This also led to a **better risk prediction** (because of the presence of genotype markers) and optimization of therapies based on genotype.

Moreover:

- Some risk loci might have different frequencies in different populations. Performing GWAS on different populations **highlights heterogeneity between populations** with regards to risk factors.
- GWAS has **multiple applications**: determination of polygenic scores, determination of ancestry, assessment of population substructure, assessment of LD in genome wide study, estimation of genetic correlation between traits.
- Data generation and analysis is very straightforward, with techniques for data generation that are up to date
- Thanks to meta-analysis, GWAS data is **easily shared** and **publicly available**.

❖ BRIEF LIST OF GWAS LIMITS

We have also some important limitations in GWAS:

- ➔ usually associated **variants have a small effect** and correspond to only a small fraction of the truly associated variants, hence predictive power is low.
- ➔ variants tend to **explain only a small fraction of heritability** of complex traits
- ➔ **association variants do not necessarily point to causal variants** and genes. Going from association to causality is a very lengthy process.
- ➔ GWAS yields too many loci. If we imply too many different genes, then the study is not informative.

❖ HERITABILITY

GWAS have facilitated estimation of SNP heritability.

SNP heritability is the **proportion** of additive variance which **can be explained by linkage disequilibrium. The linkage disequilibrium is between the imputed SNP and the unknown causal variants.**

The heritability of a character can be estimated in two ways:

- ➔ **family based studies**
- ➔ **identifying individual susceptibility factors** (SNP heritability).

Almost always the combined effects of all known susceptibility factors identified by GWAS studies account for less than half the heritability estimated from family-studies

Hence **there is some missing heritability.**

There are six hypotheses that explain missing heritability, they are not mutually exclusive so they can explain different cases.

1. **RARE VARIANTS HAVE LARGE EFFECTS**: as we have studied, at the beginning complex traits were studied similarly to mendelian traits. This led to the study of large effect variants, because they are the most similar to mendelian variants. However, large-effect variants are too rare to be assayed by the currently available commercial SNP arrays, and they are not well tagged by the SNPs on the arrays. This means that there **can be missing heritability due to the variants being too rare.**
2. **GENE INTERACTIONS:** missing heritability can come from gene-gene or gene-environment interactions. **GWAS can only study what single variants are doing, so they overlook interactions.** However, even if genetic interactions are important, they do not seem to have any great effect from a biological point of view. Hence this point doesn't have that much of an impact. They're related to another type of the heritability that is not the additive one.
3. **EPIGENETIC EFFECTS** also play a part in heritability and gene expression but they do not depend on the genomic sequence.
4. **FAMILY BASED STUDIES OVERESTIMATE HERITABILITY** which means that, in fact, there is no missing heritability. Instead of shared genes, **family based studies might mistakenly consider the shared environment.** Nevertheless, in recent times a lot of

effort has been done to separate genetics and environments, so this hypothesis is not true anymore.

5. **CAUSATIVE VARIANTS MIGHT NOT BE RELIABLY TAGGED BY SNP CHIPS: Causative variants are likely to have a lower minor allele frequency (MAF) than tag SNPs**. Using the p value of a snp which is not the causal one, means taking into consideration a higher frequency for the causal one
6. **COMMON VARIANTS HAVE SMALL EFFECTS**: **heritability might just be hidden below the threshold of significance because the effect of the variant is too small to be seen.** Hence, a part of missing heritability might be hidden heritability, which can increase just by increasing the sample size. Looking at GWAS, increasing the sample size we manage to recover heritability: hence heritability was really hidden.

In general, we can say **PHENOTYPE** can be explained by three main elements:

- ➔ **environment**
- ➔ **discovered genetics**
- ➔ **undiscovered genetics** (missing heritability)

Increasing the sample size of GWAS should continue to yield new loci of smaller effect.

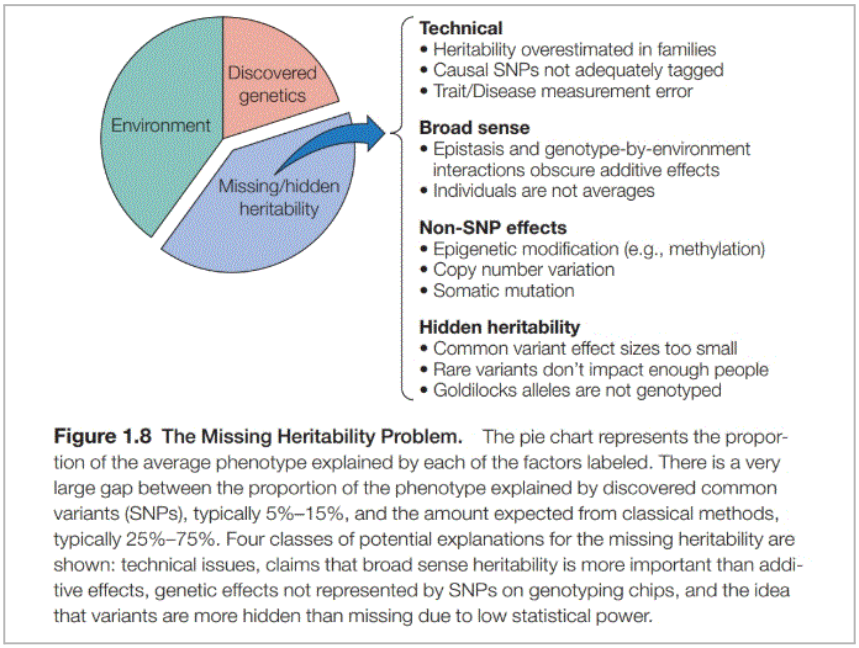


Figure 1.8 The Missing Heritability Problem. The pie chart represents the proportion of the average phenotype explained by each of the factors labeled. There is a very large gap between the proportion of the phenotype explained by discovered common variants (SNPs), typically 5%–15%, and the amount expected from classical methods, typically 25%–75%. Four classes of potential explanations for the missing heritability are shown: technical issues, claims that broad sense heritability is more important than additive effects, genetic effects not represented by SNPs on genotyping chips, and the idea that variants are more hidden than missing due to low statistical power.

Empirical evidence demonstrates that for each complex trait **there is a threshold sample size above which the rate of locus discovery accelerates** in GWAS. **At a certain point, the identification of risk loci plateaus for any trait.**

The missing heritability has been found in some populations, for example the European sample has reached ***saturation***, so 100% of heritability, while other populations need further studies and data. **SATURATION CAN ONLY BE REACHED WITH AN INCREASED SAMPLE SIZE.**

- ➔ So the majority of heritability is actually hidden.

❖ POLYGENICITY

A trait is called **POLYGENIC** when it's influenced by many alleles at many different loci.

- ➔ **GWAS provided evidence that complex traits are highly polygenic.**

For almost any complex trait that has been studied, many loci (polymorphisms in many genes) contribute to genetic variation in the population. The proportion of variance explained by the

individual variants is small. This means that we need large sized experiments, as most of them end up being underpowered for now.

From the individual point of view:

- ➔ each individual will carry a number of **alleles that either increase or decrease the trait or disease risk.**
- ➔ there are so many possible combinations of these sets of alleles that each individual is likely to have a **unique combination**, and the effect size of each locus is found to be small.

The meta-analysis confirms this hypothesis.

The only way to analyze complex polygenic traits is to increase sample size.

❖ OMNIGENICITY

The **OMNIGENIC MODEL** is a hypothesis which proposes that gene regulatory networks are sufficiently interconnected that **all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes and that most heritability can be explained by effects on genes outside core pathways.**

The authors propose an omnigenic model in which the genetic architecture of regulatory networks is composed of:

- ➔ a small number of **core genes** that directly affect a trait
- ➔ a large number of **peripheral genes** that indirectly affect the trait

However, the **majority of heritability is explained by *peripheral* genetic effects** that are propagated through genetic regulatory networks and genetic variation in core genes explains a small fraction of the overall heritability.

This model opposes the polygenic one, which assumes that *some* genes are responsible for a complex trait.

GWAS provided evidence for a **WIDESPREAD PLEIOTROPY FOR COMPLEX TRAITS.** The number of segregating variants in the human population is large but finite.

- **PLEIOTROPY** occurs when one gene influences two or more seemingly unrelated phenotypic traits.

The number of segregating variants in the human population is large but finite.
The fact that each of the many studied traits is associated with variants at hundreds to thousands of loci in the genome strongly suggests that **some of the underlying causal variants are the same for different diseases.**

❖ POLYGENIC RISK SCORE

GWAS data can be used to create genetic predictors for complex traits by **estimating the effect size at multiple loci** in a discovery sample and using those estimated SNP effects in independent samples **to generate** a **POLYGENIC RISK SCORE** (PRS) per individual.

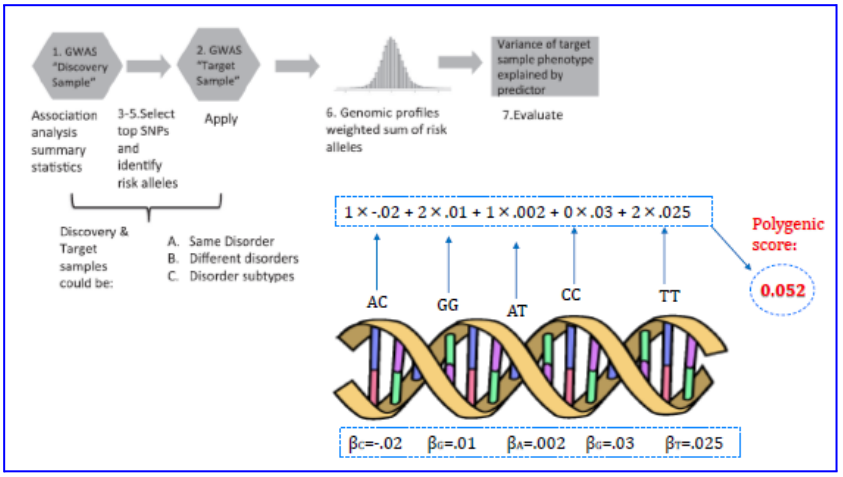
A. In the discovery sample: **select all the variants** with a p value below a certain thresholds and estimate the effect size;
B. In the test sample: compute risk score for each individual based on **how many of the variants a person has,** weighted by the effect size.

GWAS summary statistics can be used to estimate the effect size in the discovery sample, then we use it in the test sample in order to compute the risk score for an individual.

PRS is a **quantitative measure** of the cumulative genetic risk or vulnerability that an individual possesses for a trait.

The traditional approach to **CALCULATING PRS** is to construct a weighted sum of the ßs (or other effect size measure) for a set of independent loci thresholded at different significance levels. Typically, this **independence is LD based** (LD $R^2 \leq 0.2$), while the **threshold has to be defined:** the optimal p-value threshold is unknown a priori.

1. First of all we perform GWAS in the discovery sample
2. Then we apply the summary statistics to the GWAS target sample, that must be independent
3. We must select the top common SNPs between samples and identify the risk alleles, we do not use the original genotype or phenotype, only the odds ratio.
4. We compute the PRS for each individual, we sum the risk factor for each SNP, we can do it because the PRS is a linear combination;
5. Compute the polygenic score;
6. Then we compute the genomic profiles weighted sum of risk alleles
7. Evaluate the variance of target sample phenotype explained by predictor



PRS for an individual *i* can be calculated as the sum of the allele counts (dosage, which can be 0, 1 or 2) for each SNP, multiplied by a weight (ß):

$$PRS_i = \sum_j^M \hat{B}_j \cdot dosage_{ij}$$

Once we decide the SNPs, we just multiply the dosage of said SNP with the risk factor. Then we sum all of them.

Discovery GWAS

	Weight*	Risk Allele
SNP1	0.2	A
SNP2	-0.3	C
SNP3	0.1	G

Individual	Alleles SNP1	Alleles SNP2	Alleles SNP3
1	AT	AA	CG
2	AA	CA	GG
3	TT	AC	CG
4	TT	AA	GG
5	TA	CA	GC
6	AT	CA	CG
7	AA	AA	GG
8	AA	CC	CG
9	TA	CC	GC
10	AT	AA	CG

PRS:

Individual	SNP 1	SNP 2	SNP 3	PRS
1	0.2+0.0	0.0+0.0	0.0+0.1	**0.3**
2	0.2+0.2	-0.3+0.0	0.1+0.1	**0.3**
3	0.0+0.0	0.0-0.3	0.0+0.1	**-0.2**
4	0.0+0.0	0.0+0.0	0.1+0.1	**0.2**
5	0.0+0.2	-0.3+0.0	0.1+0.0	**0.0**
6	0.2+0.0	-0.3+0.0	0.0+0.1	**0.0**
7	0.2+0.2	0.0+0.0	+0.1+0.1	**0.6**
8	0.2+0.2	-0.3-0.3	0.0+0.1	**-0.1**
9	0.0+0.2	-0.3-0.3	0.1+0.0	**-0.3**
10	0.2+0.0	0.0+0.0	0.0+0.1	**0.3**

We can therefore resume all the PRS analysis process:

A. A **GWAS is carried out** in the discovery sample.

B. The **risk alleles** and their **effect sizes** are used to **generate genomic polygenic risk scores** in an independent 'target' sample, using partially independent SNPs whose p values in the discovery sample are below some threshold.

C. Both the discovery and target sample must undergo some **quality checks**. These, as usual, are: L*D adjustment, beta shrinkage, p-value thresholding, missingness rate for both markers and samples, HWE, sample heterozygosity, sex chromosomes, MAF, imputation score.*

After this, we apply some **QC specific for PRS analysis:**

- Genome build check
- ambiguous SNPs (C/G or A/T): the best solution is to remove them in order to avoid systematic errors
- mismatching SNPs
- duplicate SNPs
- sample overlap
- relatedness

D. A **PRS is calculated for each individual** in the target sample as the sum of the count of risk alleles weighted by the effect size (e.g. log(OR) for case control) in the discovery sample.

The **key factors in the development** of methods for calculating PRS are:

1. **LD adjustment:**
   a. **SNPs are clumped** so that the retained SNPs are largely independent of each other, and, thus, their effects can be summed, assuming additivity
   b. **All SNPs are included**, accounting for the LD between them
2. GWAS estimated **effect sizes adjustment:**
   a. **Shrinkage**: statistical technique applied to reduce estimated effect sizes
   b. **P-value thresholding:** only those SNPs with a GWAS association P value below a certain threshold are included in the calculation of the PRS

E. Then we proceed with association testing, we **test the accuracy of the PRS:**

➔ For continuous traits, the phenotypic variance explained by the PRS is typically quantified as a **coefficient of determination** (R^2).

➔ For binary traits, pseudo-R 2 values are typically computed using logistic **regression models.**

The profile score is evaluated through regression of the target sample phenotype on the PRS after accounting for other known covariates.

$$H_0 : \text{Phenotype} \sim \text{covariates} + e$$

$$H_1 : \text{Phenotype} \sim \text{PRS} + \text{covariates} + e$$

A commonly used metric for assessing PRS accuracy is the **area under the receiver operating characteristic curve (AUC).** The AUC quantifies the performance of the models when the aim is to discriminate between two groups.

POLYGENIC RISK SCORE (PRS): A weighted sum of the number of risk alleles carried by an individual, where the risk alleles and their weights are defined by the loci and their measured effects as detected by GWAS.

Polygenic predictions are **not particularly informative for an individual** (hence at the end the risk scores for the individuals are put together) but they do explain a sufficient proportion of variation to separate groups, for example, samples with the highest and lowest risk.
PRS is also useful to establish a common genetic basis for related disorders.

MAIN USES OF PRS:

As GWAS sample sizes increase and PRSs become more powerful, PRSs are set to play a key role in research and stratified medicine, for:

➔ Single disorder analyses
➔ Cross-disorder analysis
➔ Subtype analysis

PRSs in particular are exploited:

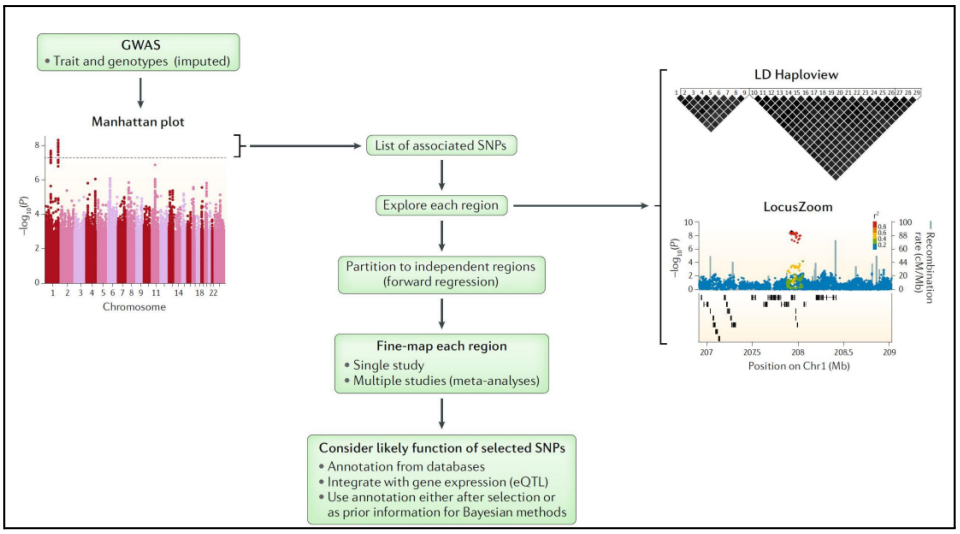
1. to assess shared etiology between phenotypes
2. to **evaluate the clinical utility of genetic data** for complex disease
3. as part of experimental studies in which, for example, experiments are performed that **compare outcomes** (e.g., gene expression and cellular response to treatment) **between individuals with low and high PRS values.**

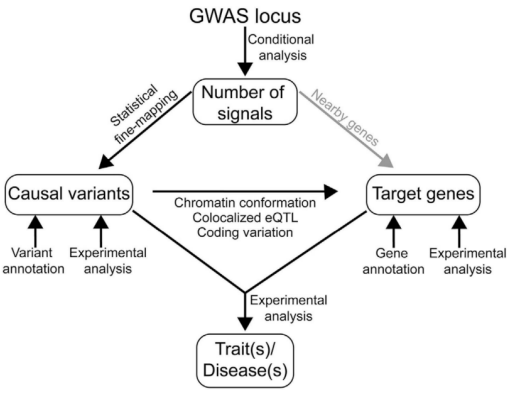
GWAS have been extremely successful in **identifying the variants with reasonably large effect** sizes that are the most likely to be biologically interesting.

However, what we want is to **move from tag SNPs to the causative variants** and identify the biology underlying the associations, so that we have the full picture of a disease.

In order to go from GWAS to causative variants we use **GENOMIC ANNOTATION.**

- ➔ Genomic annotation assigns biological function to DNA sequences and therefore it can be informative about the likely function of SNPs selected by fine-mapping analysis and can aid prioritization of follow-up functional studies.





The maps of regulatory annotations and connections in disease-relevant tissues, generated by projects such as ENCODE, Epigenome RoadMap, and GTEx have been crucial to interpretation of the non-coding variants that account for the majority of GWAS-identified risk alleles

When we have a GWAS we have **two steps**:

- ➔ **FINE-MAPPING**: we try to identify the most likely regions that can explain the result, hence containing the causal variant.
- ➔ **ANNOTATION**: we have to try to understand the biological function of the association variant.

In parallel to GWAS it's important to move and **study rare variants** because they provide a much direct link between variants and diseases.

RARE VARIANTS

A part of heritability is explained by rare variants.

- ➔ **rare variants** usually have a **strong effect** in fewer individuals. They are able to directly link variants and traits.

According to the ***common disease-rare variant hypothesis***:

- ➔ Since a disease is deleterious to fitness, the **variant** which causes it **should be selected against,** hence being rare.
- ➔ **Mutation rate is high** so variations are a possibility, but thanks to selection they are a pretty rare one.

There can be **DIFFERENT TYPES OF RARENESS.** At the two extremes we see:

- ***moderately rare:*** their frequency is just below the GWAS threshold of 0.01%. This threshold is the one we identify with "MAF" in plink association studies.
- ***private variants*** are identified in only one family, usually being de novo and explaining sporadic cases.

This means that **most family disorders are due to rare variants.** This doesn't only apply to mendelian conditions, but also to complex disorders.

The role of **rare variants in complex diseases** was first highlighted from the analysis of **copy number varianìtion** (rare copy number variants are a kind of rare variant).

❖ COPY NUMBER VARIANTS

COPY NUMBER VARIATION (abbreviated CNV) refers to a circumstance in which the number of copies of a specific segment of DNA varies among different individuals' genomes.

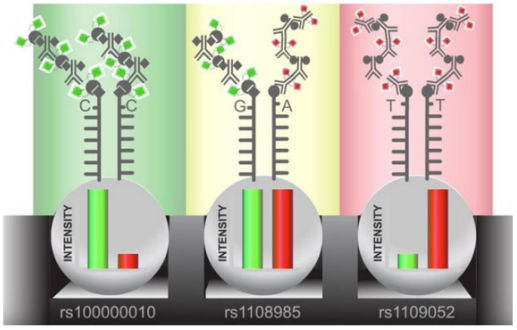
1. FINDING MUTATIONS THROUGH SEQUENCING.

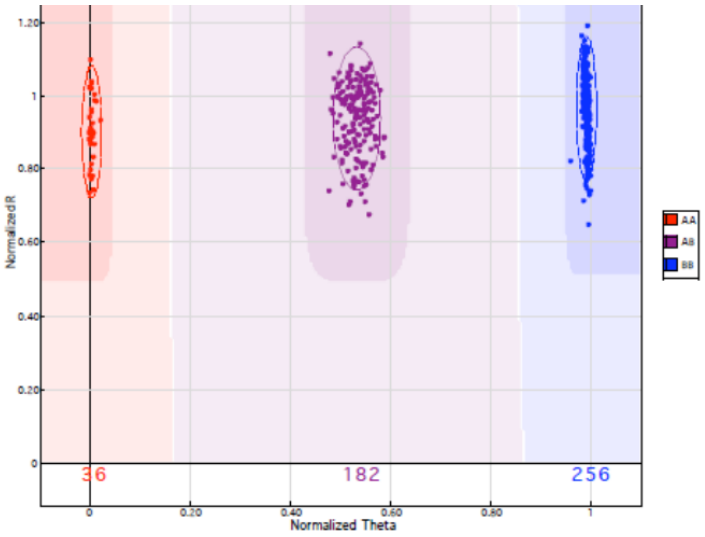
The SNP array data generated to perform GWAS can also be used for the analysis of copy number variants. This analysis is done by ***signal intensity.***

Each Illumina assay genotypes a single locus and uses two colors, one per allele. The probe binds to a complementary DNA sequence, but the annealing stops before the variation site. Then we have the incorporation of one nucleotide in the SNP locus: when the nucleotide binds, it emits a signal.

Depending on the color of the signal we have information on the ratio of the two alleles:

- ➔ ***homozygous*** for the SNPs will be one of the two colors
- ➔ ***heterozygous*** will have an average of the two colors.



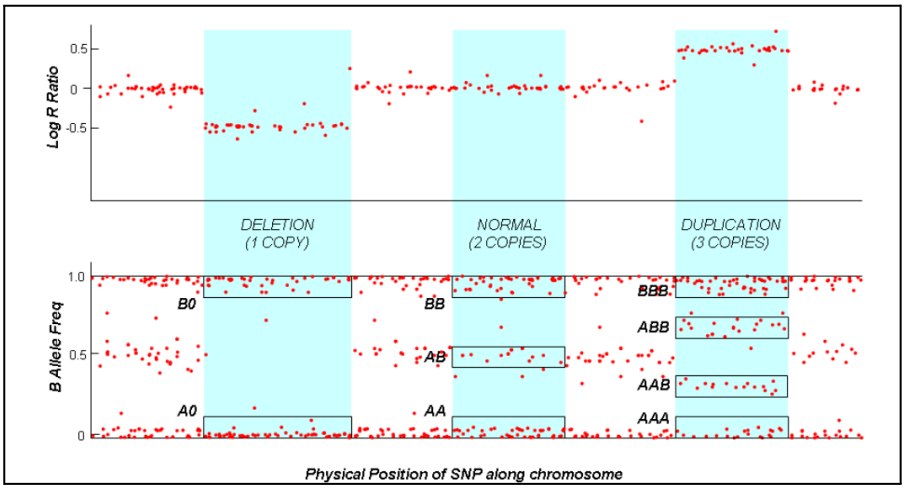


Each point in the graph represents one individual. Individuals are divided into three clusters: homozygous AA, homozygous BB on the sites, and in the middle we have heterozygous AB.
The plot allows us to know the genotype of each individual for a SNP.

If you have SNPs you expect heterozygotes, since the trait isn't common.

2. FIND COPY NUMBER DATA

From the data we can **derive two measures:**



- **LOG R RATIO**. In autosomal regions without copy number variants the LRR should be 0, if you have a deletion LRR<0 and if you have an insertion LRR>0
  - The LRR is relative to a single individual which has many SNPs.
- **B ALLELE FREQUENCY.** It ranges from 0 to 1. If the allele frequency is close to 1, we have an individual homozygous for allele A, if it's close to 0, we have an individual homozygous for allele B. 0.5 means heterozygous for A and B.

Exploiting these two measures we can predict the presence of copy number variants.

Exploiting GWAS data to find copy number variants, many studies have suggested that structural variants have an important role in disease susceptibility and gene expression.

GENOMICS DISORDERS are **sites** of rare **copy number variants.** These sites are involved in many different disorders, **exhibiting variable phenotype** which goes from incomplete to complete penetrance. These genomic disorders show a sort of mirror phenotype, where increased dosage leads to an opposite phenotype. They are one of the most common causes for abnormal neural development.

- ➔ the expression of some **genes** is not dependent on the **dosage** in some cases, while in others it is.
- ➔ This means that the **dosage of the protein can either interfere or not with the phenotype.**

In some cases having 1/2/3 copies doesn't matter at all until the protein is present, in other cases it matters. Moreover, with complex traits, different genes can be sensitive or non sensitive while contributing to the same disease.

❖ SEQUENCING OF COPY VARIANTS

NGS technologies allow large scale sequencing studies for many complex diseases. So in the late 2000s researchers started to do large scale exome sequencing and genome sequencing to study rare variants in case-control studies.

In **AUTISM**, the number of copy number variations is increased.

- ➔ 27% of families with syndromic autism have ***de novo mutation,***
- ➔ some individuals with ASD have two or more de novo CNVs. The ***CNVs*** are ***more frequent in cases*** compared to controls.
- ➔ low frequency of of microdeletions and microduplications correspond to ***"genomic disorders".***

In **SCHIZOPHRENIA,** ***11 CNVs*** were convincingly shown to increase the risk of developing schizophrenia. They all increase risk for other neurodevelopmental disorders (ASD, BD, DD).

Schizophrenia studies were successful and they managed to identify many loci. For autism however, only with the biggest meta-analysis ever in 2019 5 variants were found.
Even if in autism the rare variants were identified in sequencing in **different genes,** these genes tend to **converge in the same pathways.**

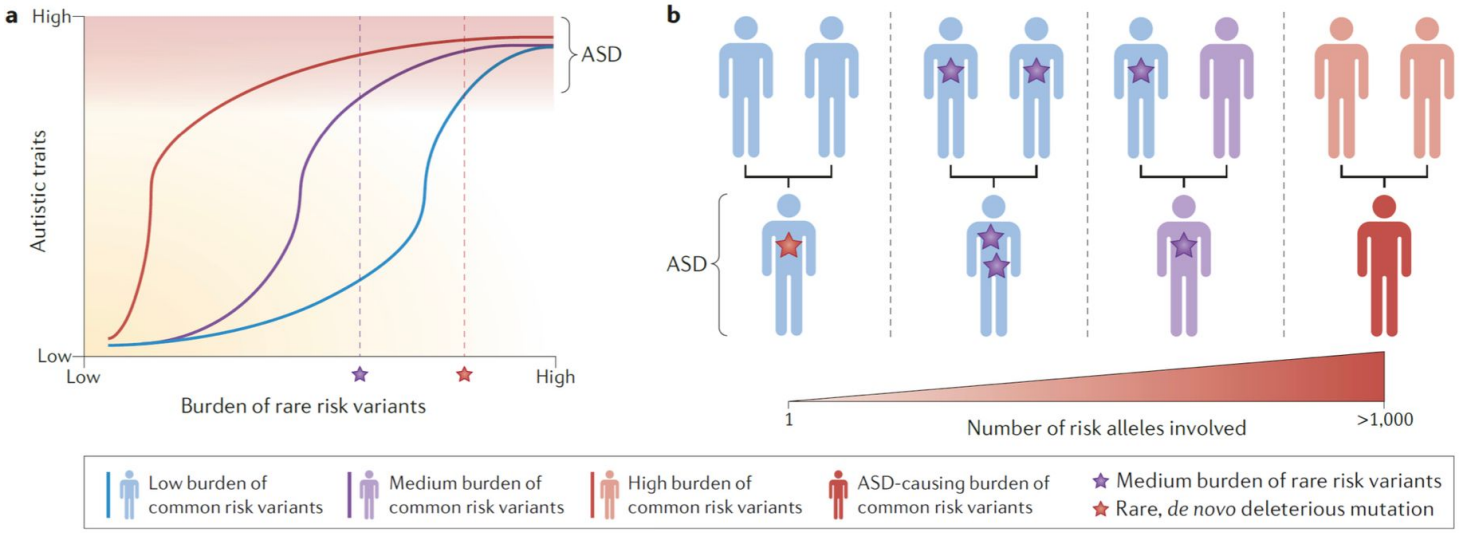
To sum up, ***NGS variant studies*** provide insight in:

- **Identifying specific variants that contribute to disease risk;**
- Evaluating the relative contributions of individual disease genes to overall disease burden;
- Assessing aspects of genetic architecture, including comparing the contributions of different allele frequencies and variant effect classes.

❖ VARIANT INTERPLAY: RARE AND COMMON

Complex traits are very complex**. Comparing genetic architecture is difficult,** even with individuals which have the same disorders. An example of the latter situation (same trait, different genetic architecture), is shown in autism.

In **AUTISM**, the **genetic mutations** that contribute to ASD **are a combination of rare deleterious variants and a myriad of low risk alleles.**



The colors of the individuals represent the burden of the common variant (blue=low, red=high), so they represent the genetic background that gives you a high or low risk of ASD. The stars are the rare deleterious variants. All of these individuals have autism but their genetic architecture is different. For example the first individual has a low-risk background, but a highly penetrant risk mutation.

Power for rare variants is very low: effects of rare variants only detectable if we pool across multiple variants within a region of interest.
Each causal allele is expected to explain only a very small fraction of the cases under study, but **different variants** in the same gene **may have a larger cumulative contribution.**

❖ VARIANT FINDING

How do we define a variant? And in particular, how can we find rare variants?
The most conservative approach is to focus only on **loss of function variant.**

1. GENE BASED COLLAPSING APPROACH

According to this approach, each individual causal allele is expected to explain only a very small fraction of the cases under study, but **different variants in the same gene may have a larger cumulative contribution.** Collapsing models focused on ***PTVs*** (protein-truncating variants) explained 80% of binary and 55% of quantitative associations.

Example: haploinsufficiency-mediated disorders.
In haploinsufficient disease genes, the number of different alleles that confer equivalent risk is expected to be large and generally recognizable: any loss of function (LOF) allele, whether in an essential splice site, a frameshift or a stop mutation, will result in haplo-insufficiency and, hence, disease.

It is reasonable to **flag the presence (or absence) of any LOF variant** and simply to test whether an **increased number of cases have an LOF variant**, relative to controls.

- ★ Qualifying variant: the successful application of collapsing analyses depends on optimizing parameters in order to focus on the class of variation that will **enrich for variants that confer risk** and to reduce the impact of neutral background variation.
- ★ Protein coding boundaries of genes to **evaluate whether there is a significant difference in the counts of cases versus controls** who carry at least one qualifying variant.

Given the approximately 19,000 protein- coding genes, the conventional exome- wide multiplicity- adjusted significance threshold is assigned as $\alpha = (0.05/19{,}000) \approx 2.6 \times 10{-}6$

2. RARE VARIANT BURDEN TEST

RARE-VARIANT BURDEN METHODS **aggregate the information found within a defined genetic region into a summary** dose variable.

- In **WEIGHTED BURDEN TESTS,** variants are additionally weighted according to their frequency or functional impact.
- **ADAPTIVE BURDEN TESTS** try to account for bidirectional effects by selecting appropriate weights.
- **VARIANCE COMPONENT** (kernel) tests (such as C- alpha or SKAT) allow for bidirectional effects, but they are underpowered compared to collapsing or burden tests if many variants are causal and/or if effects are mostly unidirectional within a gene. In this case we don't focus on the mean, we calculate the variance of the distribution of the assignment of alleles. The variance is higher because there are many variants. They test the variability of the assignment of the allele rather than testing the allele itself.
- **OMNIBUS TESTS** such as SKAT- O use a combination of burden and variance component tests, which helps for settings with limited prior knowledge of the underlying disease architecture.

! When we talk about ***direction***, we talk about the fact that SNPs can increase or decrease the risk of having a trait.

Here is a summary table:

Table 2. Summary of Statistical Methods for Rare-Variant Association Testing

	Description	Methods	Advantage	Disadvantage	Software Packages[a]
Burden tests	collapse rare variants into genetic scores	ARIEL test,[50] CAST,[51] CMC method,[52] MZ test,[53] WSS[54]	are powerful when a large proportion of variants are causal and effects are in the same direction	lose power in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	EPACTS, GRANVIL, PLINK/SEQ, Rvtests, SCORE-Seq, SKAT, VAT
Adaptive burden tests	use data-adaptive weights or thresholds	aSum,[55] Step-up,[56] EREC test,[57] VT,[58] KBAC method,[59] RBT[60]	are more robust than burden tests using fixed weights or thresholds; some tests can improve result interpretation	are often computationally intensive; VT requires the same assumptions as burden tests	EPACTS, KBAC, PLINK/SEQ, Rvtests, SCORE-Seq, VAT
Variance-component tests	test variance of genetic effects	SKAT,[61] SSU test,[62] C-alpha test[63]	are powerful in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	are less powerful than burden tests when most variants are causal and effects are in the same direction	EPACTS, PLINK/SEQ, SCORE-Seq, SKAT, VAT
Combined tests	combine burden and variance-component tests	SKAT-O,[64] Fisher method,[65] MiST[66]	are more robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait-decreasing variants	can be slightly less powerful than burden or variance-component tests if their assumptions are largely held; some methods (e.g., the Fisher method) are computationally intensive	EPACTS, PLINK/SEQ, MiST, SKAT

A paper we have read during the lectures considers 80000 medical conditions and identifies thousands of new gene-phenotype relationships. The **gene collapsing analysis reveals many associations for binary traits.** 83% of this risk association was undetectable via single gene testing: this signifies the importance of collapsing testing.
The statistical significance for the traits is much higher in the collapsing method.

Including missense and deleterious effects can increase the power:

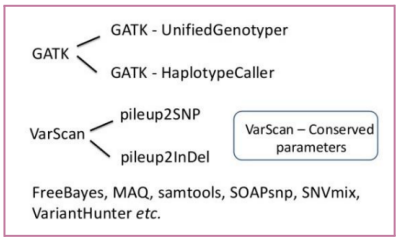
- to include missense variants we have to pay attention to the direction of the effect, hence we need to use bidirectional testing.
- rare variants apply also to common diseases. To find these variants you need specific approaches that go beyond simple SNP genotyping of GWAS.

GWAS can be successfully applied to copy number variants and they lead to very important information.

ANNOTATION

When reads are aligned into a reference genome, comparison can start and mismatches are highlighted. This process, known as **VARIANT CALLING**, leads to the formation of a variant call format (**VCF**).

➔ VCF is the standardized generic format for storing sequence variation including SNPs, indels, larger structural variants and annotations



The most useful tools for variant detections are GATKs, which have two different algorithms. Other softwares also exist.

The **difficulty of variant calling** is given by

1. the presence of indels, which represent a major source of **false positive** SNV identifications, especially if alignment algorithms do not perform gapped alignments;
2. errors from library preparation due to PCR artifacts and variable GC content in the short reads unless paired-end sequencing is utilized;
3. variable quality scores, with higher error rates generally found at bases at the ends of reads

UnifiedGenotyper was the first program in this field, but now ***HaplotypeCaller*** is the most used, as it also allows de-novo assembly.

- **UnifiedGenotyper**

  Call SNPs and indels separately by considering each variant locus independently

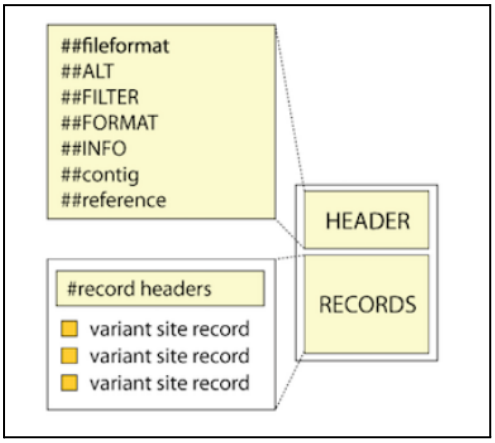
- **HaplotypeCaller**

  Call SNPs, indels, and some SVs simultaneously by performing a local *de-novo* assembly

Developed by the Broad Institute, the Genome Analysis ToolKit (GATK) is one of the most popular methods for variant calling using aligned reads.
The package has been used for projects such as the Cancer Genome Atlas and the 1000 Genomes Project that have covered analyses of HLA typing, multiple-sequence realignment, quality score recalibration, multiple-sample SNP genotyping and indel discovery and genotyping.

❖ VCF FORMAT

It is characterized by a **header** which contains information about the dataset.



The header contains information about the dataset and relevant reference sources (e.g. the organism, genome build version etc.), as well as definitions of all the annotations used to qualify and quantify the properties of the variant calls contained in the VCF file. The header of VCFs generated by GATK tools also include the command line that was used to generate them.

After the header we have the **variant call records,** which has lines for each single variant, with certain fields that are compulsorily associated to it.

- CHROM: chromosome
- POS: one based position
- ID
- REF: wild type base
- ALT: variant
- QUAL: Phred score for variant
- FILTER: pass / not pass of the filter
- INFO: Various site-level annotations
- FORMAT : GT: AD:DP:QC:PL
- genotype call

PHRED SCORE

The QUAL field is a **Phred-scaled quality score for the assertion made in ALT.** If the ALT field contains a variant call, then **QUAL reflects the estimated probability that the variant call is wrong. The Phred scale is -10 * log(1-p).**
A value of 10 indicates a 1 in 10 chance of error, while a 100 indicates a 1 in 10^{10} chance of error.

FORMAT

GT: The genotype of this sample at this site.
When there's a single ALT allele (by far the more common case), GT will be either:
- 0/0 : the sample is homozygous reference
- 0/1 : the sample is heterozygous, carrying 1 copy of each of the REF and ALT alleles
- 1/1 : the sample is homozygous alternate

AD and DP: Allele depth (AD) and depth of coverage (DP)
AD is the unfiltered allele depth, all reads at the position (including reads that did not pass the variant caller's filters) are included in this number, except reads that were considered uninformative. Reads are considered uninformative when they do not provide enough statistical evidence to support one allele over another.
DP is the filtered depth, at the sample level. This gives you the number of filtered reads that support each of the reported alleles.

PL: "Normalized" Phred-scaled likelihoods of the three possible genotypes 0/0, 0/1, 1/1.
Value of 0 for the most likely genotype

GQ: The Genotype Quality represents the Phred-scaled confidence that the genotype assignment (GT) is correct, derived from the genotype PLs. Specifically, the GQ is the difference between the PL of the second most likely genotype, and the PL of the most likely genotype. The value of GQ is capped at 99. So if the second most likely PL is greater than 99, we still assign a GQ of 99.

GENOME ANNOTATION & TOOLS

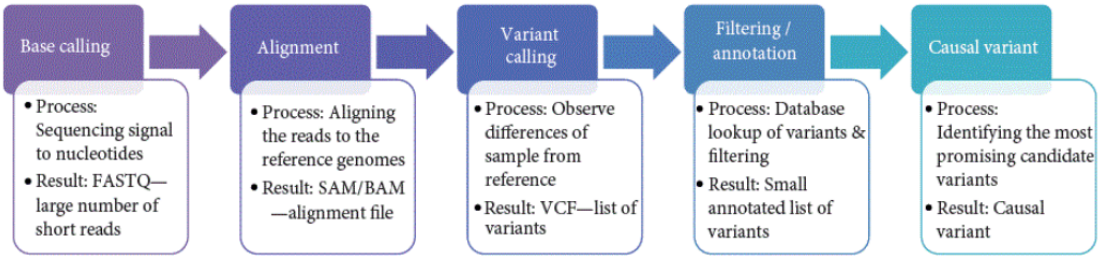


FIGURE 1: Next-generation sequencing bioinformatics workflow.

After alignment and variant calling, a list of thousands of potential differences between the genome under study and the reference genome is generated.

We have many different variants. Among them, we have to highlight the ones that are most likely to cause the trait. This means we must filter the variants and, in order to do that, **we need to know information about the variants themselves.**

➔ information on variants is given by **GENOME ANNOTATION**

There exist many tools to examine relevant variants. They work by referencing previously known information about their biological function and inferring potential affects based on their genomic context. These tools can be found online.

Table 2 | A list of tools for variant annotation.

Name	Website	Reference
ANNOVAR	http://www.openbioinformatics.org/annovar/	Wang et al. (2010)
HaploReg	http://www.broadinstitute.org/mammals/haploreg/haploreg.php	Ward and Kellis (2012a)
RegulomeDB	http://regulome.stanford.edu/	Boyle et al. (2012)
SeattleSeq	http://snp.gs.washington.edu/SeattleSeqAnnotation137/	Ng et al. (2009)
Snpeff	http://snpeff.sourceforge.net	Cingolani et al. (2012)
VEP	http://useast.ensembl.org/info/docs/variation/vep/index.html	McLaren et al. (2010)

ANNOVAR

ANNOVAR provides a lot of **information that can be based on genes, regions and filters.**
It can utilize annotation databases from the UCSC Genome Browser or any annotation data set.
In particular, it provides:

- ➔ **gene based annotation**
- ➔ **region based annotation:** in which region the variant is present, if it's in a conserved region, or if the region is expressed, particularly sensitive, etc..
- ➔ **filter based annotation:** identifies variants that are already written in the databases.

ANNOVAR can be used on computers, by downloading information which is a very lengthy process. wANNOVAR is the online version of this tool: it's easier to use but it is less customizable.

dbNSFP

It is another important database. It gives us the prediction score of many different variants.
This is done in order to avoid the the loading of a heavier kind of data.

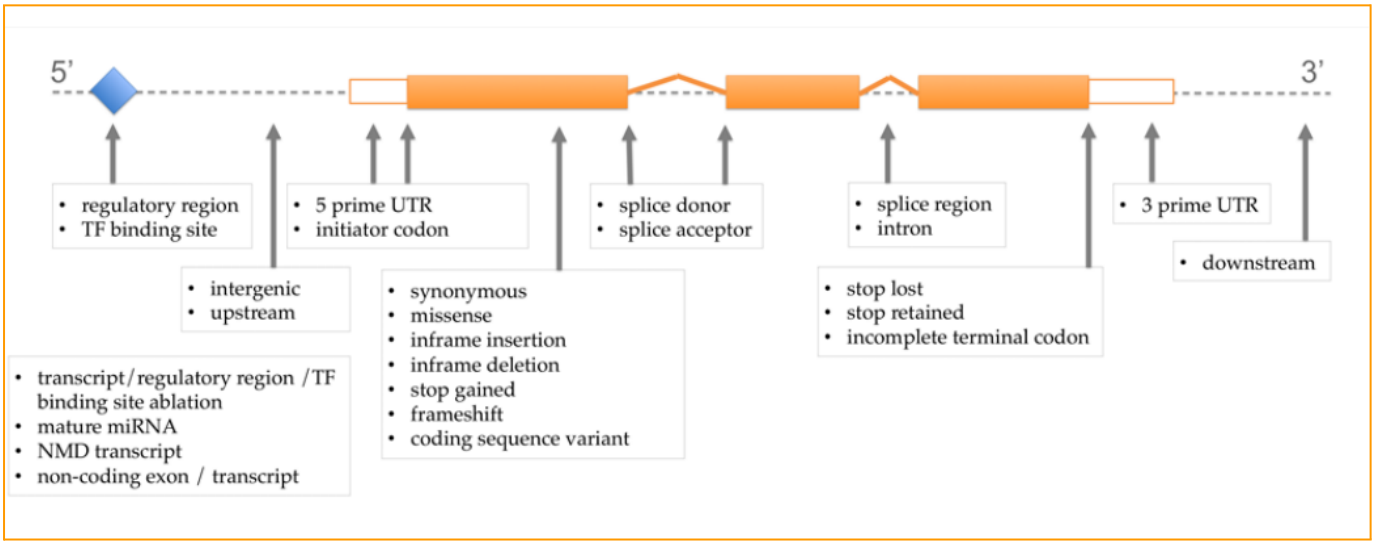
Pros

- Easy usage
- Prediction scores from 20 prediction algorithms
  - SIFT, Polyphen2-HDIV, Polyphen2-HVAR, LRT, MutationTaster2, MutationAssessor, FATHMM, MetaSVM, MetaLR, CADD, VEST3, PROVEAN, FATHMM-MKL coding, fitCons, DANN, GenoCanyon, Eigen coding, Eigen-PC, M-CAP, REVEL, MutPred
- 6 conservation scores
  - PhyloP x 2, phastCons x 2, GERP++ and SiPhy
- Allele frequencies from several databases
  - 1000 Genomes Project phase 3 data, UK10K cohorts data, ExAC consortium data and the NHLBI Exome Sequencing Project ESP6500

Cons

- Provides annotation for only coding variants

GENOMIC VARIANTS CLASSIFICATION



Variants can be classified in:

- **INTRONIC:** the variant is inside a gene, in one of the introns
- **INTERGENIC:** the variant is between two genes. Remember that DNA is made up of non gene sequences.
- **CODING:** the variant is found in one of the gene eons. They are divided in:
    - Synonymous
    - Non synonymous (25% protein disrupting)
    - Nonsense (protein disrupting)
    - Frameshift (protein disrupting)
    - Splice site (protein disrupting)

There are many softwares that try to **predict the deleteriousness** (=danger) **of a variant** in a sequence.

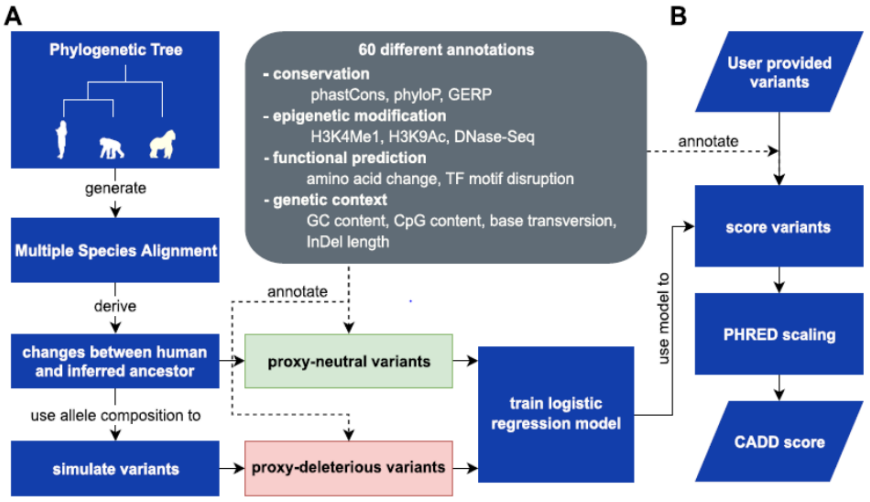
➔ They are based on the idea that **mutations in conserved sites** (which are shared among species) are the most deleterious, as they have a more important role.

➔ Other modalities to predict disease-causing variants include protein biochemistry, such as amino acid charge, the presence of a binding site, and structure information of protein. SNVs that are predicted to **alter protein features** (such as polarity and hydropathy) and structure (binding ability and alteration of secondary/tertiary structure) have a higher probability of being deleterious.

➔ Other exploit **machine learning approaches** (differences in training datasets and algorithms)

CADD

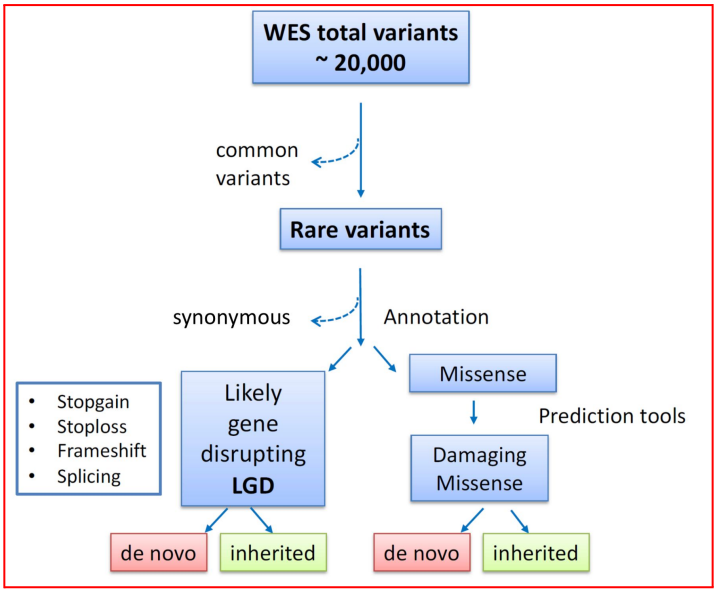
Another score is the **Combined Annotation Dependent Depletion** (CADD) score, which **combines various annotations into ONE metric.** It contrasts variants that survived natural selection with simulated mutations.
CADD can quantitatively prioritize functional, deleterious, and disease causal variants across a wide range of functional categories, effect sizes and genetic architectures and can be used to prioritize causal variation in both research and clinical settings.

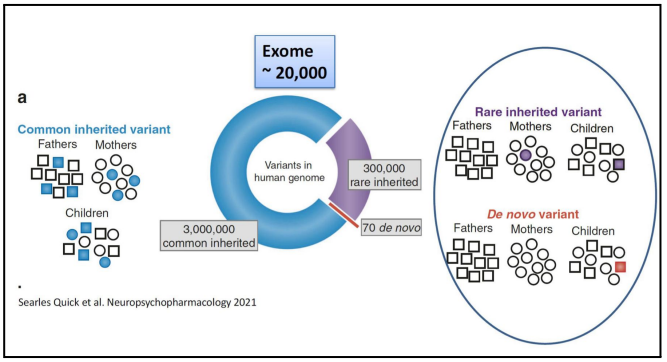


There are around **20.000 variants.** We need to **FILTER** them.

1. remove **non coding variants**
2. remove **common variants**
3. remove coding variants that are synonymous or **bening** according to he CADD score
4. **prioritize gene disrupting mutations and missense mutations,** as those are the most important ones.
5. look also for ***de novo variants***

Moreover, look for shared candidate genes/variants among affected.
Remove variants that are present in unaffected relatives/controls if available.





META ANALYSIS

The input files for our protocol are:

- ➔ rm1.csv and its txt version.
- ➔ rm2.csv
- ➔ r_total.csv
- ➔ gwadata.map

.txt files are for PLINK, .csv are for METAL. ***rm1*** and ***rm2*** are two different association studies.

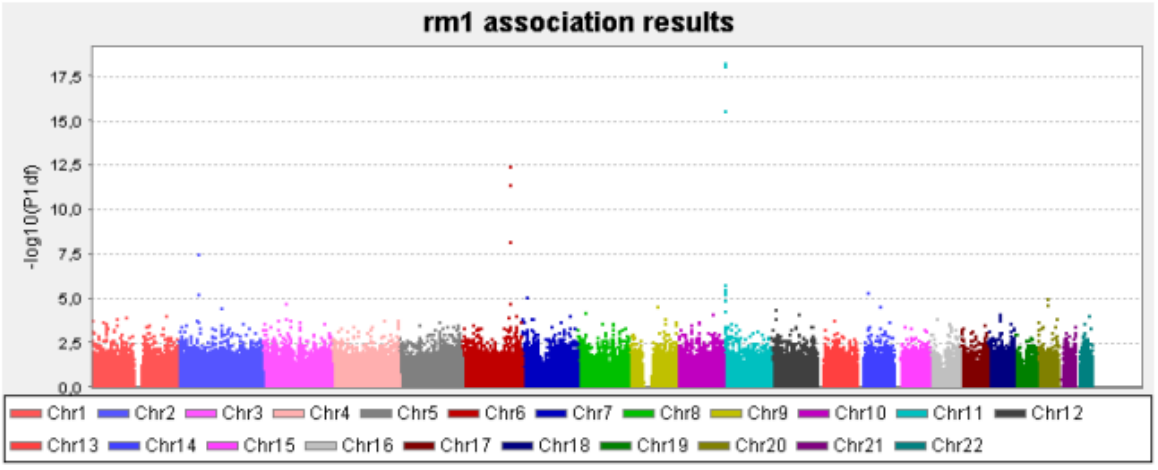
1. GWAS RESULTS

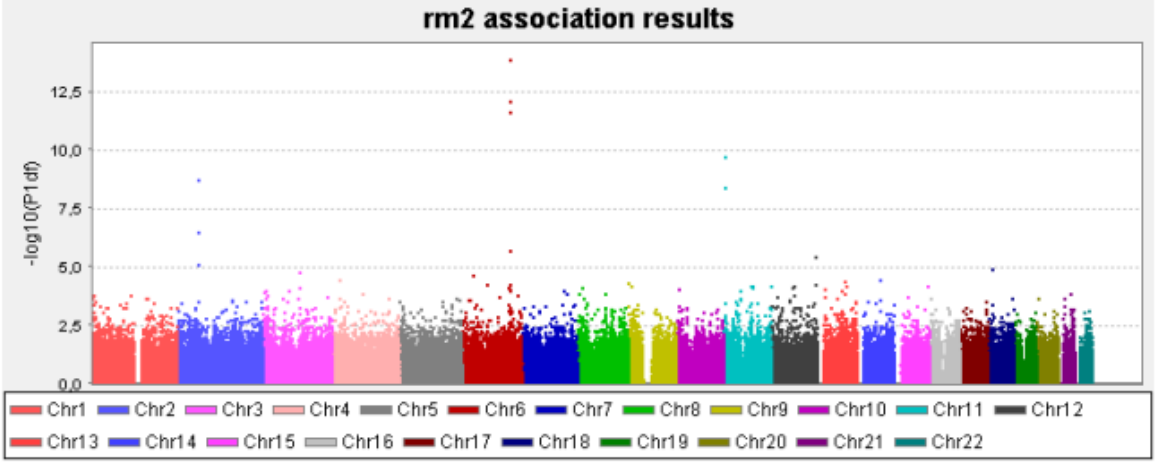
Visualize the data in a **MANHATTAN PLOT.** Manhattan plots can be easily done with ***Haploview***. (Use rm1.***txt***, rm2.***txt*** and r_total.***txt*** files and the map file gwadata.map).

```
java -jar /opt/Haploview/Haploview.jar
```

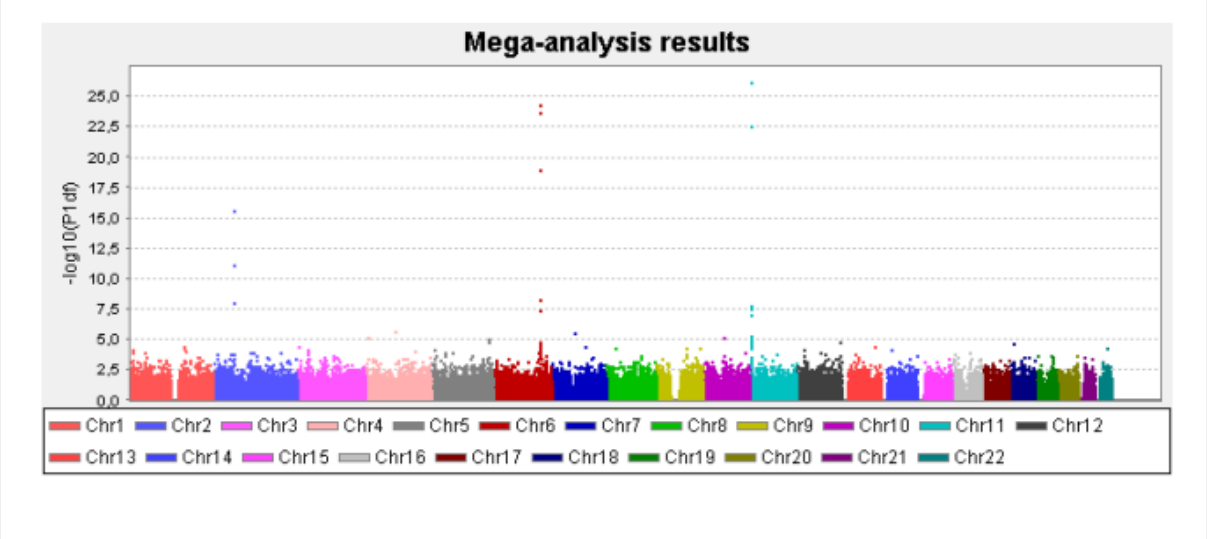
This command will open a new page.

1. In the 'Welcome to Haploview' window, **select the 'PLINK Format' tab.** In 'Results File' click 'Browse' and import the result file (**rm1** and **rm2**). In 'Map file' click 'Browse' and choose '**gwadata.map**'. Leave other options as they are (ignore pairwise comparison of markers > 500 kb apart and exclude individuals with > 50% missing genotypes). Click 'OK'.
2. Click 'Plot', insert the Title of the graph (trend chr9 result), in the 'X-Axis' select **Chromosome**, in 'YAxis' select **P1df**, in 'Scale' select **–log10.** Leave other options as they are.





Compare this with the mega-analysis results on the combined data set 'r_total':



Do the two GWAS have signals on the same chromosomes or not?

Now sort by **P1df** (which is the *p-value*) the top 15 results of each file: here are the top hits for the three different scans. Notice that while there is overlap between the 'rm1' and 'rm2' top hits, their relative rankings differ, as do their estimated effect sizes.

$ sort -g -k11,11 rm1.txt | head -15

rm1 results
Summary for top 15 results, sorted by P1df

SNP	Chrom	Position	Strand	A1	A2	N	effB	se_effB	chi2.1df	P1df
rs2855039	11	5228247	u	1	2	85	0.023	0.003	79.242	5.50E-19
rs2071348	11	5220722	u	1	2	86	37.333	4.215	78.457	8.17E-19
rs16912979	11	5266271	u	1	2	86	21.467	2.628	66.747	3.09E-16
rs9399137	6	135460711	u	1	2	87	13.800	1.898	52.847	3.61E-13
rs9376092	6	135468837	u	1	2	87	0.086	0.012	48.225	3.80E-12
rs9494145	6	135474245	u	1	2	87	8.455	1.462	33.436	7.36E-09
rs766432	2	60631621	u	1	2	87	13.889	2.511	30.593	3.18E-08
rs3813727	11	5212488	u	1	2	87	0.201	0.042	22.857	1.74E-06
rs968856	11	5217152	u	1	2	85	0.214	0.046	21.501	3.54E-06
rs2255519	11	5230117	u	1	2	87	4.807	1.041	21.333	3.86E-06
rs1955490	14	32408251	u	1	2	87	4.474	0.977	20.947	4.72E-06
rs2855122	11	5233812	u	1	2	87	4.602	1.013	20.638	5.55E-06
rs6732518	2	60620248	u	1	2	87	4.426	0.982	20.329	6.52E-06

rm2 results
Summary for top 15 results, sorted by P1df

SNP	Chr	Position	Strand	A1	A2	N	effB	se_effB	chi2.1df	P1df
rs9376092	6	135468837	u	1	2	87	0.052	0.007	59.314	1.34E-14
rs9399137	6	135460711	u	1	2	82	16.704	2.332	51.296	7.94E-13
rs9494145	6	135474245	u	1	2	87	16.059	2.294	48.994	2.57E-12
rs2071348	11	5220722	u	1	2	87	8.786	1.383	40.373	2.10E-10
rs2855039	11	5228247	u	1	2	87	0.114	0.018	40.373	2.10E-10
rs766432	2	60631621	u	1	2	87	16.189	2.702	35.894	2.08E-09
rs16912979	11	5266271	u	1	2	85	7.407	1.258	34.658	3.93E-09
rs6732518	2	60620248	u	1	2	87	5.297	1.037	26.068	3.30E-07
rs1547247	6	135432529	u	1	2	87	0.203	0.043	22.614	1.98E-06
rs7135277	12	127262481	u	1	2	87	0.226	0.049	21.393	3.74E-06
rs243081	2	60525427	u	1	2	87	4.236	0.951	19.831	8.46E-06
rs7243066	18	11272319	u	1	2	87	5.158	1.184	18.966	1.33E-05
rs2398912	3	105426470	u	1	2	87	0.217	0.051	18.430	1.76E-05
rs2248372	6	31554445	u	1	2	84	0.229	0.054	17.952	2.27E-05

$ sort -g -k11,11 r_total.txt | head -15

total results
Summary for top 15 results, sorted by P1df

SNP	Chr	Position	Strand	A1	A2	N	effB	se_effB	chi2.1df	P1df
rs2855039	11	5228247	u	1	2	172	0.061	0.006	115.268	6.87E-27
rs2071348	11	5220722	u	1	2	173	16.017	1.494	114.949	8.07E-27
rs9376092	6	135468837	u	1	2	174	0.070	0.007	106.812	4.89E-25
rs9399137	6	135460711	u	1	2	169	14.885	1.461	103.836	2.20E-24
rs16912979	11	5266271	u	1	2	171	12.030	1.212	98.595	3.10E-23
rs9494145	6	135474245	u	1	2	174	11.407	1.260	82.027	1.34E-19
rs766432	2	60631621	u	1	2	174	15.089	1.846	66.803	3.00E-16
rs6732518	2	60620248	u	1	2	174	4.865	0.713	46.586	8.77E-12
rs1320963	6	135484905	u	1	2	174	0.165	0.028	33.838	5.99E-09
rs243081	2	60525427	u	1	2	174	3.757	0.655	32.852	9.94E-09
rs5010981	11	5235931	u	1	2	173	0.174	0.031	31.556	1.94E-08
rs2255519	11	5230117	u	1	2	174	3.653	0.657	30.899	2.72E-08
rs3813727	11	5212488	u	1	2	174	0.277	0.050	30.872	2.76E-08
rs2855122	11	5233812	u	1	2	174	3.593	0.652	30.396	3.52E-08

2. P-VALUE BASED META ANALYSIS

As we know, meta analysis can be done on the basis of two different things: p value or effect size. Let's start from p value based analysis. In order to do meta-analysis we use the program **METAL.** METAL takes its instructions from a control file, which we need to create.

Set up a text file named *'metal_p.script'* containing the appropriate commands for reading our two csv files and carrying out a sample-size weighted analysis (based on p-values).

If you are carrying out a sample size weighted analysis (**based on p-values**), you need:

- ➔ A column indicating the **direction of effect for the tested allele**
- ➔ A column indicating the **corresponding p-value**
- ➔ A column indicating the **sample size** (if the sample size varies by marker)

These are the step-by-step instructions:

A. The first thing you should specify is the **Column Separator.** By default, METAL assumes columns are separated by whitespace (which consists of any combination of space and tab characters). However, since our files are .csv, we need to indicate another column separator. The choices are the following:

```
SEPARATOR  WHITESPACE - the default
SEPARATOR  COMMA - for comma delimited files
SEPARATOR  TAB - columns separated by a single tab, so that consecutive tabs
indicate an empty column
```

Our two csv files are **comma delimited** (rm1.csv and rm2.csv), so we choose the second option.

B. Then you should select an **Analysis Scheme.**

By default, METAL combines p-values across studies taking into account a study specific weight (typically, the sample size) and direction of effect.

- ➔ This behavior can be requested explicitly with the `SCHEME SAMPLESIZE` command.

There are also other options, although in our case it is best to choose this first one.

```
SCHEME SAMPLESIZE - default approach, uses p-value and direction of effect, weighted
according to sample size
SCHEME STDERR - classical approach, uses effect size estimates and standard errors
STDERR SE - specify the label for the standard error column.
```

C. Then you'll need to specify the **Input File Columns:**

The columns that must be present are the following:

- ➔ The name of the column with the **marker names,** which should be consistent across studies
- ➔ The name of the column indicating the tested allele
- ➔ The name of the column indicating the other allele
- ➔ The name of the column indicating the corresponding p-value
- ➔ The name of the column indicating the direction of effect for the tested allele

The commands we need to write are as follows. However, one must consider that we also need to indicate the name of the columns from our results file that correspond to the columns of this new txt. Hence, the names of the column must be followed by the names of the column in our file

```
MARKERLABEL   SNP
ALLELELABELS  RefAllele NonRefAllele
PVALUELABEL   P-value
EFFECTLABEL   Effect
```

These can be abbreviated as:

```
MARKER        SNP
ALLELE        RefAllele NonRefAllele
PVALUE        P-value
EFFECT        Effect
```

Metal checks whether the direction of effect is positive or negative when combining information across samples. This means checking if each allele increases or decreases the chance of having the trait we study. The direction is either positive or negative, however, most of the time, the **direction is indicated by the Odds Ratio, which is always positive.**

Typically, METAL checks whether the value in the **`EFFECT`** column is positive or negative. If the EFFECT column includes an odds-ratio or relative risk, you actually need METAL to **check whether the *log-odds ratio* is positive or negative**. You have to modify your EFFECTLABEL command to tell METAL it should take the log of whatever value is listed. For example, if the EFFECT column is labeled ODDS, use the syntax **`EFFECTLABEL log(ODDS)`**

```
EFFECT      log(effB)
```

D. Finally, you need to specify **`WEIGHTS IN P-VALUE`** **based analysis**

```
WEIGHTLABEL     N
```

E. Then you have to specify the input files with the **`PROCESS`** command.

```
PROCESS      rm1.csv
```

If you have multiple files, repeat the process command for each file.

F. To add a specific prefix to the output file, use the **`OUTFILE`** command

```
OUTFILE   METAANALYSIS_P_ .TBL            (default = 'METAANALYSIS','.TBL')
```

G. To perform the final analysis, once all input files have been processed, simply issue the ANALYZE command to execute a meta-analysis.

```
ANALYZE
```

H. To end the script, type:

```
QUIT
```

At the end, the script in our file should be:

```
SEPARATOR COMMA
SCHEME SAMPLESIZE
MARKER SNP
ALLELE A1 A2
PVALUE P1df
EFFECT log(effB)
WEIGHT N
PROCESS rm1.csv
PROCESS rm2.csv
OUTFILE METAANALYSIS_P_ .TBL
ANALYZE
QUIT
```

Now we can RUN METAL, remembering to go in the correct folder and to keep all the files there.

```
STUDENTI^bianca.solazzo@BNFAB-L002040DL:~$ cd Desktop
STUDENTI^bianca.solazzo@BNFAB-L002040DL:~/Desktop$ cd meta
STUDENTI^bianca.solazzo@BNFAB-L002040DL:~/Desktop/meta$ metal_p.script
metal_p.script: command not found
STUDENTI^bianca.solazzo@BNFAB-L002040DL:~/Desktop/meta$ metal metal_p.script
MetaAnalysis Helper - (c) 2007 - 2009 Goncalo Abecasis
```

You can have a look at the beginning of the file, typing:

```
head METAANALYSIS_P_1.TBL
```

MarkerName	Allele1	Allele2	Weight	Zscore	P-value	Direction
rs2720934	a	c	174	0.806	0.4202	++
rs908551	a	c	174	-0.198	0.8427	+-
rs982887	a	c	174	-0.421	0.6741	-+
rs1996182	a	c	174	-1.986	0.04698	--
rs7521783	a	c	174	-0.635	0.5252	-+
rs11004591	a	c	174	0.539	0.5902	-+
rs1194463	a	c	174	2.241	0.02505	++
rs10486503	a	c	174	-0.177	0.8595	+-
rs1289744	a	c	173	-1.279	0.2009	--

You can sort the file and have a look at the best results:

```
$ sort -g -k6,6 METAANALYSIS_P_1.TBL | head -20
```

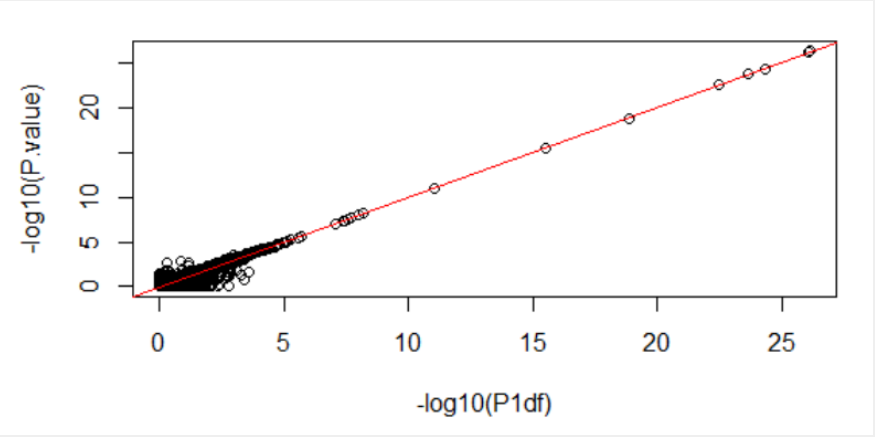
MarkerName	Allele1	Allele2	Weight	Zscore	P-value	Direction
rs2855039	a	c	172	-10.777	4.44E-27	--
rs2071348	a	c	173	10.751	5.86E-27	++
rs9376092	a	c	174	-10.357	3.91E-25	--
rs9399137	a	c	169	10.205	1.89E-24	++
rs16912979	a	c	171	9.944	2.67E-23	++
rs9494145	a	c	174	9.038	1.59E-19	++
rs766432	a	c	174	8.148	3.71E-16	++
rs6732518	a	c	174	6.798	1.06E-11	++
rs1320963	a	c	174	-5.785	7.26E-09	--
rs243081	a	c	174	5.751	8.85E-09	++
rs5010981	a	c	173	-5.598	2.17E-08	--
rs2255519	a	c	174	5.559	2.72E-08	++
rs3813727	a	c	174	-5.546	2.92E-08	--
rs2855122	a	c	174	5.505	3.69E-08	++
rs1547247	a	c	174	-5.475	4.38E-08	--
rs968856	a	c	165	-5.333	9.67E-08	--
rs6532013	a	c	174	4.706	2.52E-06	++
rs12667374	a	c	173	-4.657	3.22E-06	--
rs4601817	a	c	172	-4.575	4.75E-06	--

Now we can compare METAL results to the original analyses.

➔ we want to **compare the results of the meta-analysis to the mega analysis**. You can do this in **R**. In this case we consider the mega analysis the analysis of the two GWAS studies in r_total.

To properly create a graph of the 'mega' vs. the 'meta' -log10(P-values), we first need to merge the 'mega' and 'meta' results into a single data frame, matching on SNP name

```
mt1 <- read.table('METAANALYSIS_P_1.TBL',header=T)
mega <- read.csv('r_total.csv',header=T)
b <- merge(mt1,mega,by.x='MarkerName',by.y='SNP')
with(b,plot(-log10(P1df),-log10(P.value)))
abline(a=0,b=1,col='red')
```



Here's a graph comparing the mega p-values ('P1df') vs. the meta p-values ('P.value'):
While there is not a perfect agreement among some of the smaller P-values, the most significant p-values do agree.

3. ERROR-WEIGHTED BASED META ANALYSIS

Set up a text file named *metal_se.script* containing the appropriate commands for reading our two csv files and carrying **analysis on effect size estimates** and the **standard errors.**

If you are carrying out a sample size weighted analysis (**based on standard error**), you need:

- ➔ A column indicating the **estimated effect size** for each marker
- ➔ - A column indicating the **standard error of this effect size estimate**

The way in which we have to write our file is very similar to the previous case, however the **scheme is different**, as we are not **using** the p value approach, but the **effect size estimate approach.**

```
SCHEME SAMPLESIZE - default approach, uses p-value and direction of effect, weighted
according to sample size
SCHEME STDERR - classical approach, uses effect size estimates and standard errors
STDERR SE - specify the label for the standard error column.
```

The final script should be equal to:

```
SEPARATOR COMMA
SCHEME STDERR
STDERR se_effB
MARKER SNP
ALLELE A1 A2
PVALUE P1df
EFFECT log(effB)
WEIGHT N
PROCESS rm1.csv
PROCESS rm2.csv
OUTFILE METAANALYSIS_SE_ .TBL
ANALYZE
QUIT
```

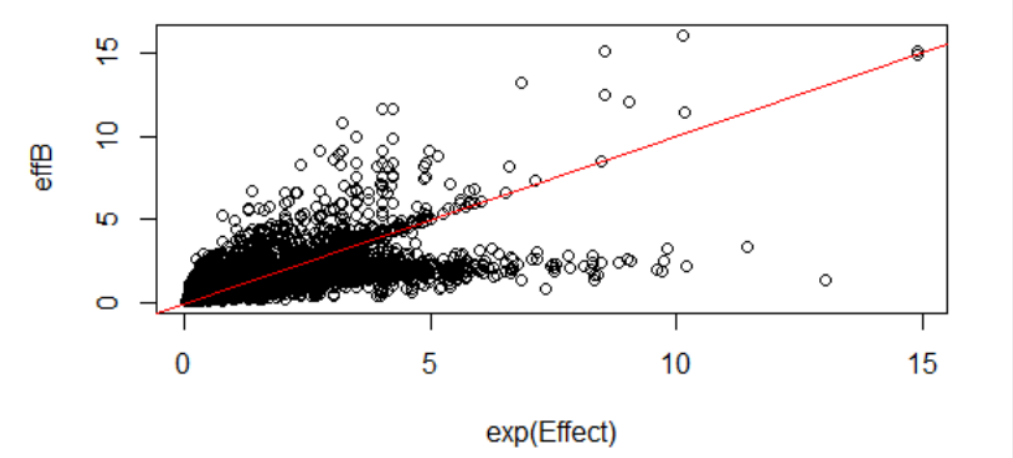
Once again, we want to compare meta and mega analysis. In this case however **we will not compute the p values against each other, but the effect size estimates.**

This is once again done on **R**:

```
mt1 <- read.table('METAANALYSIS_SE_1.TBL',header=T)
mega <- read.csv('r_total.csv',header=T)
b <- merge(mt1,mega,by.x='MarkerName',by.y='SNP')
with(b[exp(b$Effect)<100,],plot(exp(Effect),effB))
abline(a=0,b=1,col='red')
```

- ➔ We have filtered out the excessively large effect values, which are probably erroneous, with the command: with(b[exp(b$Effect)<100,],plot(exp(Effect),effB))

Here's a graph of the mega OR ('effB' on the Y axis) vs. the meta OR ('exp(Effect)' on the X axis):



As we can see from the plot above, the standard error-based meta-analysis results do not agree as well with the mega analysis results as the Pvalue meta-analysis results. This may be because the data set is so small, containing only 87 samples in each half of the data.

Such small data sets would not estimate effect sizes very well, and the uncertainty in the estimates of the effect size would propagate into the meta-analysis results, as we observe here.

META ANALYSIS WITH PLINK (effect size based)

We can perform a meta-analysis based on the effect size estimates and the standard errors also using PLINK. You just need input files in plink format (rm1_plink.txt and rm2_plink.txt in "metaanalysis files 4 plink" folder).

```
plink --meta-analysis rm1_plink.txt rm2_plink.txt
```

A ***plink.meta*** file will be created. We can have a look at the beginning of the file using the command `head.`

head plink.meta

CHR	BP	SNP	A1	A2	N	P	P(R)	OR	OR(R)	Q	I
1	1045729	rs3934834	1	2	2	0.7449	0.7449	1.4182	1.4182	0.7324	0.00
1	1061338	rs3737728	1	2	2	0.5503	0.5503	0.6963	0.6963	0.9168	0.00
1	1070488	rs6687776	1	2	2	0.8082	0.8082	1.4499	1.4499	0.9973	0.00
1	1071463	rs9651273	1	2	2	0.9998	0.9998	0.9972	0.9972	0.9982	0.00
1	1088878	rs4970405	1	2	2	0.0007785	0.0007785	0.4395	0.4395	0.9161	0.00
1	1089873	rs12726255	1	2	2	0.05146	0.05146	0.5196	0.5196	0.9892	0.00
1	1104902	rs2298217	1	2	2	0.8381	0.8381	1.2435	1.2435	0.7574	0.00
1	1116987	rs4970357	1	2	2	0.9804	0.9804	0.8999	0.8999	0.9982	0.00
1	1134661	rs4970362	1	2	2	0.8857	0.8857	1.2232	1.2232	0.9925	0.00

The plink.meta file has the following columns:

CHR	Chromosome code
BP	Basepair position
SNP	SNP identifier
A1	First allele code
A2	Second allele code
N	Number of valid studies for this SNP
P	Fixed-effects meta-analysis p-value
P(R)	Random-effects meta-analysis p-value
OR	Fixed-effects OR estimate
OR(R)	Random-effects OR estimate
Q	p-value for Cochrane's Q statistic
I	I^2 heterogeneity index (0-100)

The effect (**OR**, or BETA in case of quantitative trait) i**s with respect to the A1 allele**.

➔ i.e. **if OR is greater than 1**, implies A1 **increases risk relative to A2.**

Since the ***plink.meta*** and the ***METAANALYSIS_SE_1.TBL files*** are sorted by default in different orders, you can quickly compare the results, just **searching for specific SNPs**. You can use the grep command:

```
grep -E "rs3934834|Effect" METAANALYSIS_SE_1.TBL
grep -E "rs3934834|OR" plink.meta

grep -E "rs3737728|Effect" METAANALYSIS_SE_1.TBL
grep -E "rs3737728|OR" plink.meta

grep -E "rs6687776|Effect" METAANALYSIS_SE_1.TBL
grep -E "rs6687776|OR" plink.meta
```

```
grep -E "rs3934834|Effect" METAANALYSIS_SE_1.TBL

MarkerName  Allele1  Allele2  Effect  StdErr  P-value  Direction
rs3934834   a        c        0.3494  1.0737  0.7449   +-

grep -E "rs3934834|OR" plink.meta

CHR  BP       SNP        A1  A2  N  P       P(R)    OR      OR(R)   Q       I
1    1045729  rs3934834  1   2   2  0.7449  0.7449  1.4182  1.4182  0.7324  0.00

grep -E "rs3737728|Effect" METAANALYSIS_SE_1.TBL

MarkerName  Allele1  Allele2  Effect   StdErr  P-value  Direction
rs3737728   a        c        -0.3620  0.6060  0.5503   +-

grep -E "rs3737728|OR" plink.meta

CHR  BP       SNP        A1  A2  N  P       P(R)    OR      OR(R)   Q       I
1    1061338  rs3737728  1   2   2  0.5503  0.5503  0.6963  0.6963  0.9168  0.00

grep -E "rs6687776|Effect" METAANALYSIS_SE_1.TBL

MarkerName  Allele1  Allele2  Effect  StdErr  P-value  Direction
rs6687776   a        c        0.3715  1.5301  0.8082   +-

grep -E "rs6687776|OR" plink.meta

CHR  BP       SNP        A1  A2  N  P       P(R)    OR      OR(R)   Q       I
1    1070488  rs6687776  1   2   2  0.8082  0.8082  1.4499  1.4499  0.9973  0.00
```

To **COMPARE THE RESULTS,** remember that:

➔ ***METAANALYSIS_SE_1.TBL*** the "Effect" is the **log(OR).**

➔ ***plink.meta*** has its Effect simply as an **OR**