UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Statistical Genetics and Epidemiology

MDS/MIRI

Fall 2025/2026

Marta Castellano

Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

marta.castellano@upc.edu

Syllabus

Tuesdays 5-8pm

Mid-term exam:
04/11/2025 13:00 - 15:00

Final exam:
14/01/2026 11:30 - 14:30

September	**9**
	1. Introduction to Statistical Genetics (Marta Castellano)
	16
	2. Hardy Weinberg equilibrium (Marta Castellano)
	23
	3. Linkage disequilibrium and Phase estimation (Marta Castellano)
	30
	4. Population substructure (Marta Castellano)
October	**7**
	5. Family relationships and allele sharing (Marta Castellano)
	14
	6. Genetic Association Analysis (Marta Castellano)
	21
	7. State of the ART (Marta Castellano)
	28
	8. Introduction to Epidemiology (Cristian Tebe)
November	**4**
	Midterm week
	11
	9. Measures of Disease Frequency (Cristian Tebe)
	18
	10. Analytical Study Designs and Their Core Measures (I) (Cristian Tebe)
	25
	11. Analytical Study Designs and Their Core Measures (II) (Cristian Tebe)
December	**2**
	12. Bias, Confounding and Causality (Cristian Tebe)
	9
	13. Introduction to Risk Assessment (Cristian Tebe)
	16
	14. Applications and Future Directions (Cristian Tebe)

Syllabus

Bioinformatics and **Statistical Genetics**

1. Introduction to statistical genetics
2. Hardy-Weinberg equilibrium
3. Linkage disequilibrium and haplotype estimation
4. Population substructure
5. Relatedness analysis (allele sharing)
6. Genetic association analysis

Tuesdays 5-8pm

Content

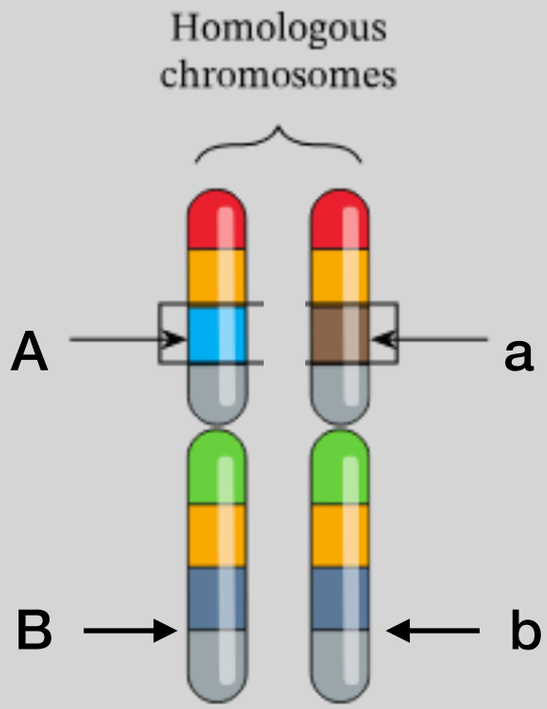
Genetic Association Studies

1. Introduction
2. Allele based tests
3. Genotype based tests
4. Quantitative traits and multiple polymorphisms
5. Computer exercise

Genetic association studies

Explain Like I'm 5 - RECALL

The **Hardy-Weinberg Equilibrium** is fundamental in population genetics: **Linkage disequilibrium**
Haplotype estimation
Population substructure
Family-relationships

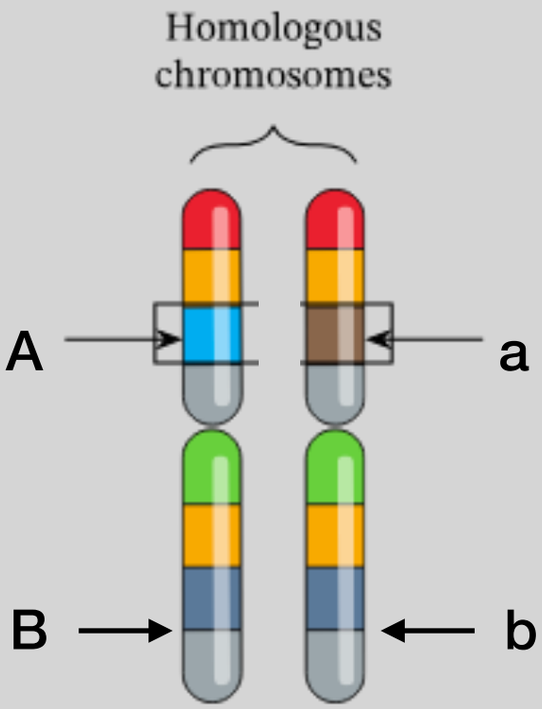


Finally….phenotype-marker association studies!

Genetic association studies

Explain Like I'm 5 - RECALL

The **Hardy-Weinberg Equilibrium** is fundamental in population genetics: **Linkage disequilibrium**
Haplotype estimation
Population substructure
Family-relationships



Finally….phenotype-marker association studies!

RECALL:

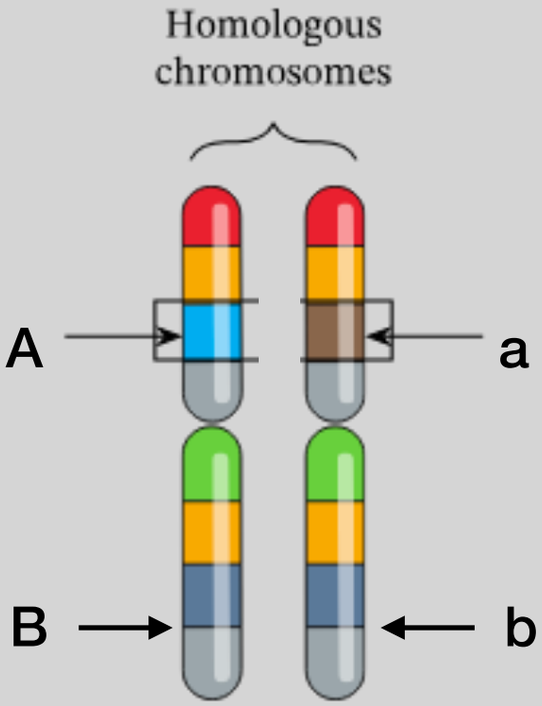
A **trait** (**phenotype**) is a specific characteristic of an individual. Traits can be determined by genes, environmental factors or by a combination of both.

- Traits can be qualitative (such as eye color) or quantitative (such as height or blood pressure).
- In many studies in statistical genetics, some trait (e.g. yield or disease status) of an organism is considered to depend on one or more genetic variables.
- The position of genetic factors determining a trait is often unknown.

Genetic association studies

Explain Like I'm 5 - RECALL

The **Hardy-Weinberg Equilibrium** is fundamental in population genetics: **Linkage disequilibrium**
Haplotype estimation
Population substructure
Family-relationships

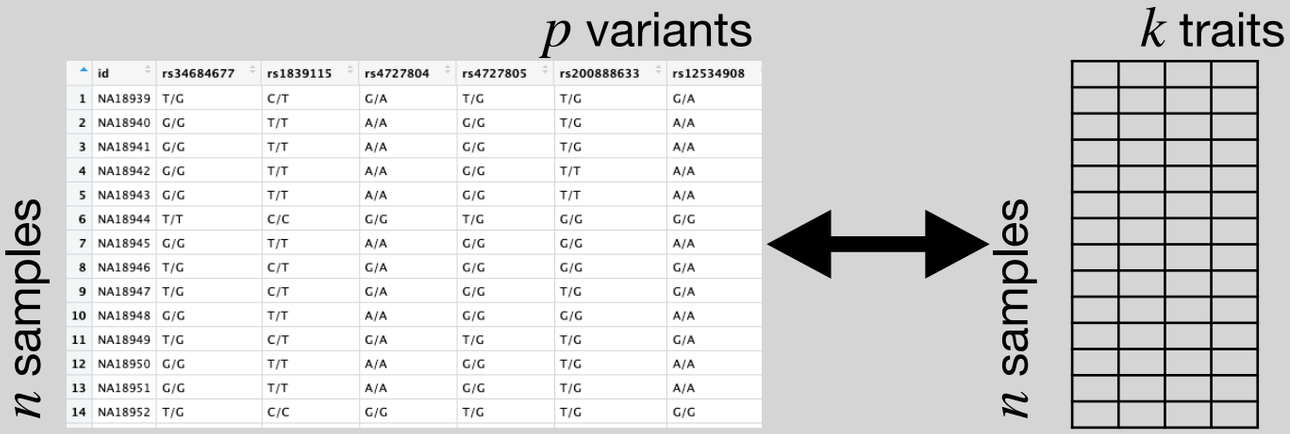


Finally….phenotype-marker association studies!

Genetic association studies

Explain Like I'm 5 - RECALL

The **Hardy-Weinberg Equilibrium** is fundamental in population genetics: **Linkage disequilibrium**
Haplotype estimation
Population substructure
Family-relationships



	id	rs34684677	rs1839115	rs4727804	rs4727805	rs200888633	rs12534908
1	NA18939	T/G	C/T	G/A	T/G	T/G	G/A
2	NA18940	G/G	T/T	A/A	G/G	T/G	A/A
3	NA18941	G/G	T/T	A/A	G/G	T/G	A/A
4	NA18942	G/G	T/T	A/A	G/G	T/T	A/A
5	NA18943	G/G	T/T	A/A	G/G	T/T	A/A
6	NA18944	T/T	C/C	G/G	T/G	G/G	G/G
7	NA18945	G/G	T/T	A/A	G/G	G/G	A/A
8	NA18946	T/G	C/T	G/A	G/G	G/G	G/A
9	NA18947	T/G	C/T	G/A	G/G	T/G	G/A
10	NA18948	G/G	T/T	A/A	G/G	G/G	A/A
11	NA18949	T/G	C/T	G/A	T/G	T/G	G/A
12	NA18950	G/G	T/T	A/A	G/G	T/G	A/A
13	NA18951	G/G	T/T	A/A	G/G	T/G	A/A
14	NA18952	T/G	C/C	G/G	T/G	T/G	G/G

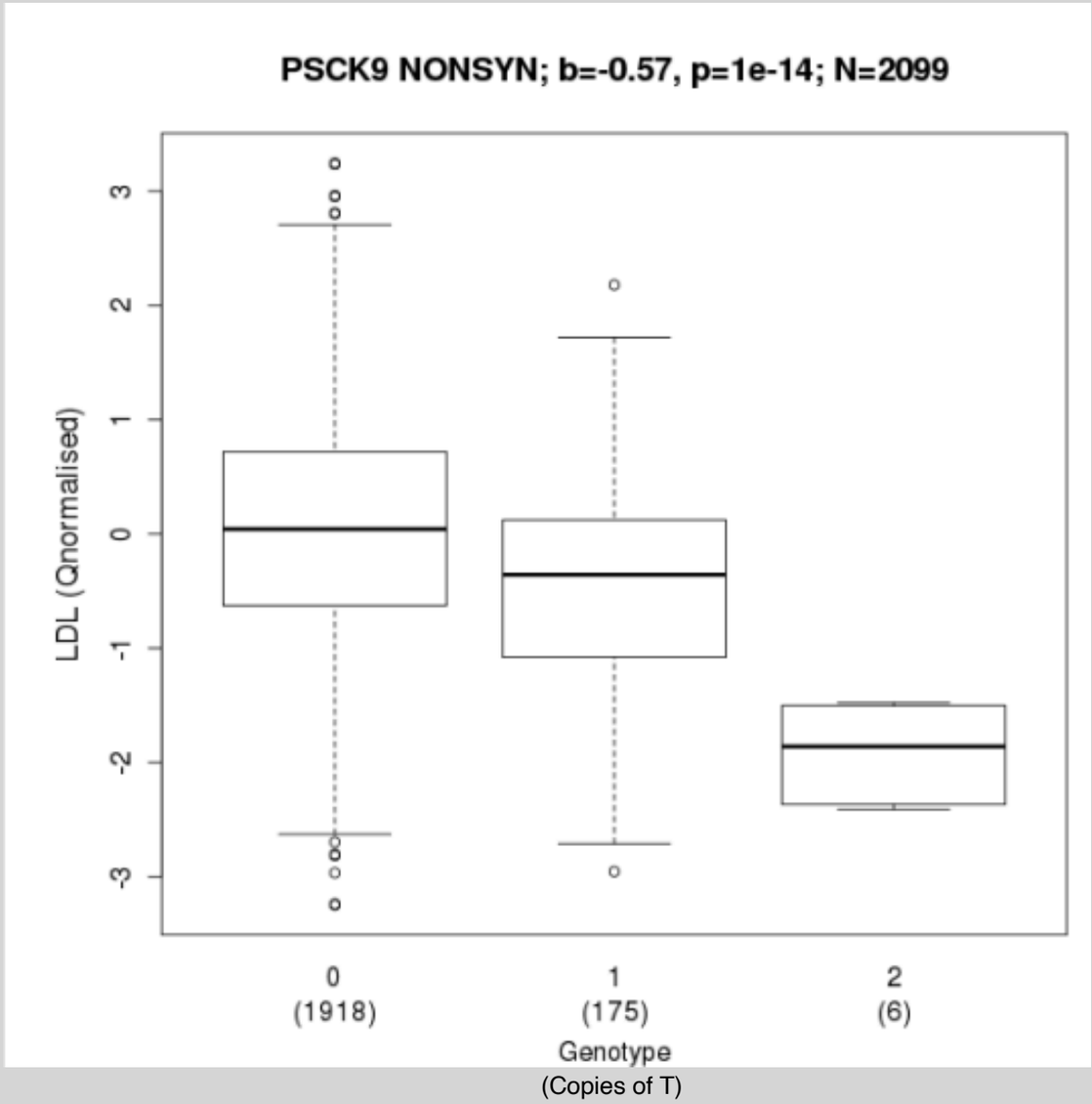
Finally….phenotype-marker association studies!

Genetic association studies

Genetic markers

- A **genetic marker** is a genetic variable that has a known variation over individuals, has a known locus.

- Example: PCSK9 is associated with cholesterol levels
  - PCSK9 is a protein that helps regulate blood cholesterol levels by promoting the breakdown of LDL receptors on liver cells, which are crucial for removing "bad" LDL cholesterol from the blood.
  - Increased PCSK9 activity leads to high LDL levels in blood, which leads to a higher risk for heart disease.
  - Carriers of T variant have lower levels of LDL cholesterol than carriers of G variant.



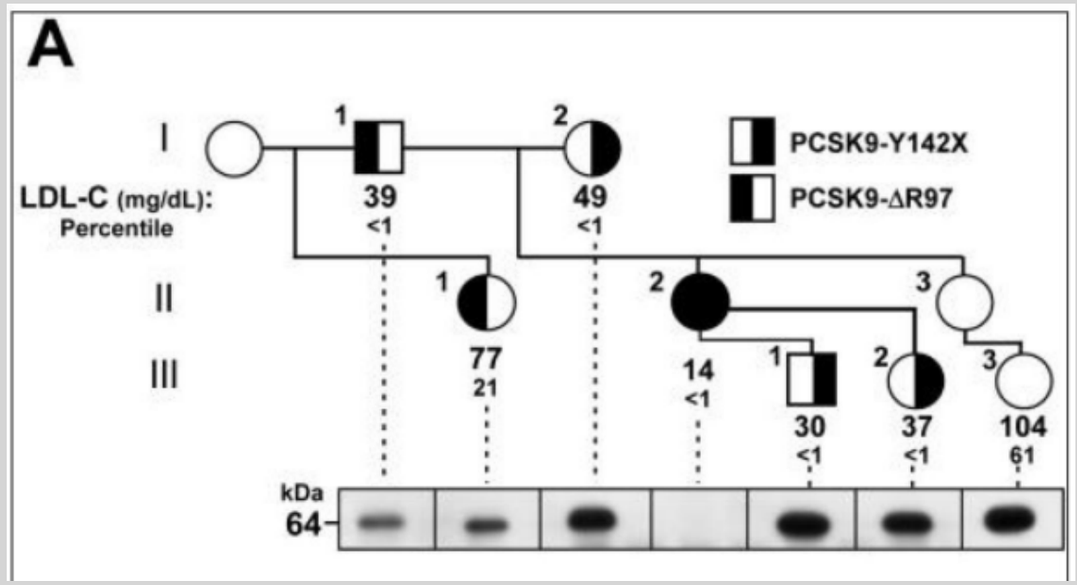
Grejtakova, D., Boronova, I., Bernasovska, J., & Bellosta, S. (2024). PCSK9 and lipid metabolism: Genetic variants, current therapies, and cardiovascular outcomes. *Cardiovascular Drugs and Therapy*, 1-13.

Genetic association studies

Genetic markers

- A **genetic marker** is a genetic variable that has a known variation over individuals, has a known locus.

- Example: PCSK9 is associated with cholesterol levels



Zhao, C., Bellur, D. L., Lu, S., Zhao, F., Grassi, M. A., Bowne, S. J., ... & Larsson, C. (**2009**). Autosomal-dominant retinitis pigmentosa caused by a mutation in SNRNP200, a gene required for unwinding of U4/U6 snRNAs. The American Journal of Human Genetics, 85(5), 617-627.

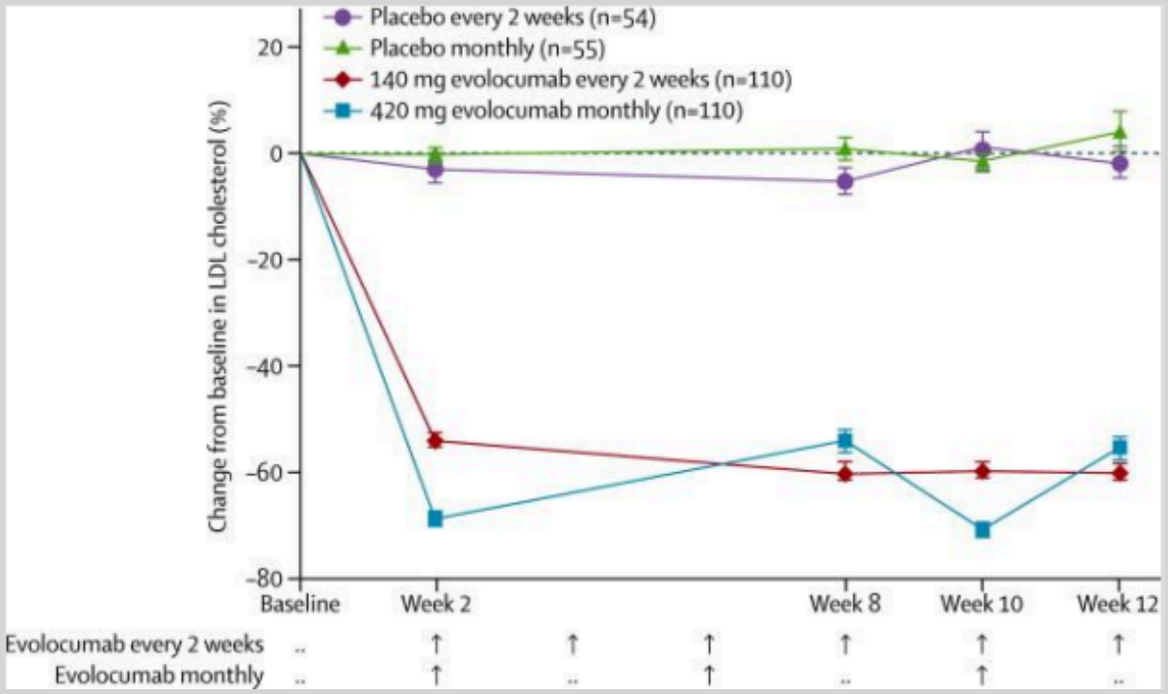
- Can we inhibit PCSK9 safely to reduce LDL?

- Human knock-out pf PCSK9 observed, with very low levels of LDL in blood

FDA Approves Amgen's Repatha (evolocumab) to Prevent Heart Attack and Stroke

Dec 1 2017

In the Repatha cardiovascular outcomes study (FOURIER), Repatha reduced the risk of heart attack by 27%, the risk of stroke by 21% and the risk of coronary revascularization by 22%.



Raal, F. J., Stein, E. A., Dufour, R., Turner, T., Civeira, F., Burgess, L., ... & Gaudet, D. (**2015**). PCSK9 inhibition with evolocumab (AMG 145) in heterozygous familial hypercholesterolaemia (RUTHERFORD-2): a randomised, double-blind, placebo-controlled trial. *The Lancet*, *385*(9965), 331-340.

Genetic association studies

Why? (Still ELI5)

Genetic association studies

Why? (Still ELI5)

To what extent certain traits are genetic?

What are the genetic mechanisms of that trait?

To find gene-trait association that allows for:

- **Disease diagnosis**: determines if a particular person has a disease.
- **Prognosis**: determines if a particular person will develop the disease in the future.
- **Conversion**: determines if a particular person converts from one stage to another stage of a disease.
- **Therapy**: determines if a therapy is successful.
- **Aging**: determines general processes of aging but not the disease, determines the biological age of a person.
- **Individualised therapy**: determines the likelihood of benefiting from a specific therapy.
- …

Genetic association studies

Introduction

Goal:
Investigate associations between markers and a trait (disease).

Designs:

- Unrelated subjects (population-based)
- Related subjects from pedigrees (family-based)

Genetic association studies

Introduction

Goal:
Investigate associations between markers and a trait (disease).

NOTE:

Genetic association studies test for a **correlation** between a trait (usually disease status) and genetic variation to identify candidate genes or genome regions that **contribute** to that specific trait/disease.

Designs:

- Unrelated subjects (population-based)
- Related subjects from pedigrees (family-based)

Genetic association studies

Introduction

Goal:
Investigate associations between markers and a trait (disease).

NOTE:

Genetic association studies test for a **correlation** between a trait (usually disease status) and genetic variation to identify candidate genes or genome regions that **contribute** to that specific trait/disease.

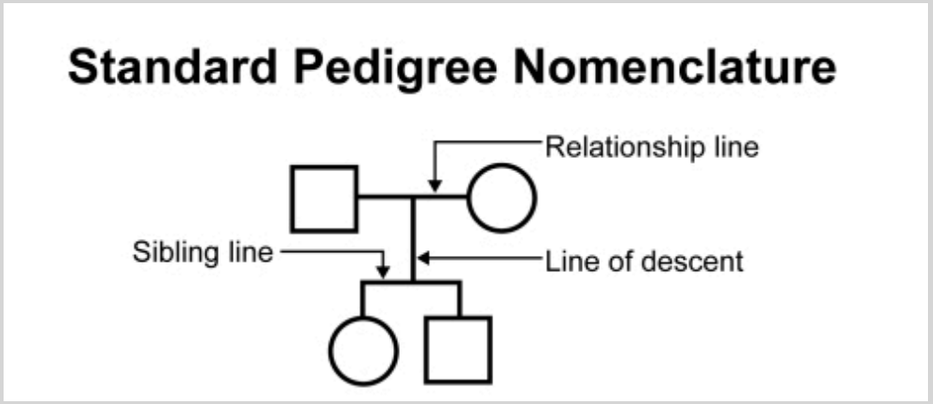
Designs:

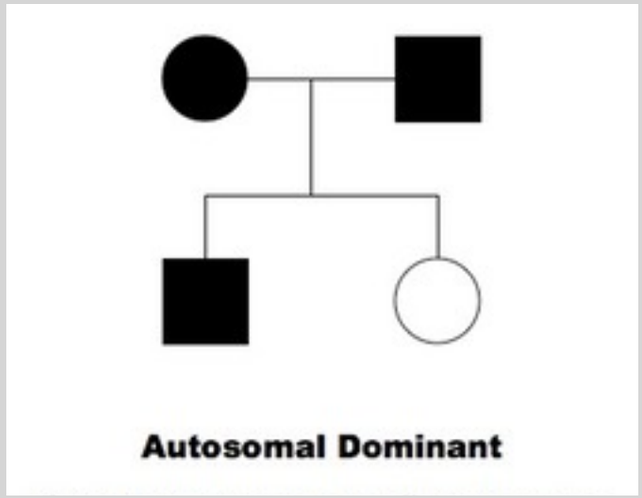
- Unrelated subjects (population-based)
- Related subjects from pedigrees (family-based)

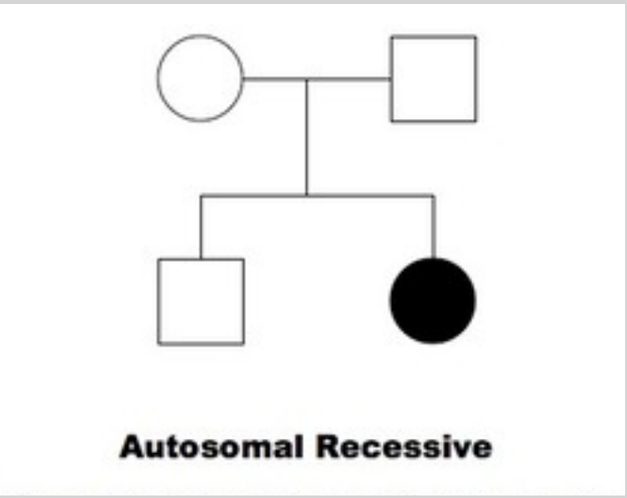
Family-based association studies are often aimed at finding rare variants underlying rare conditions or rare sub-phenotypes of a common condition.

Genetic association studies

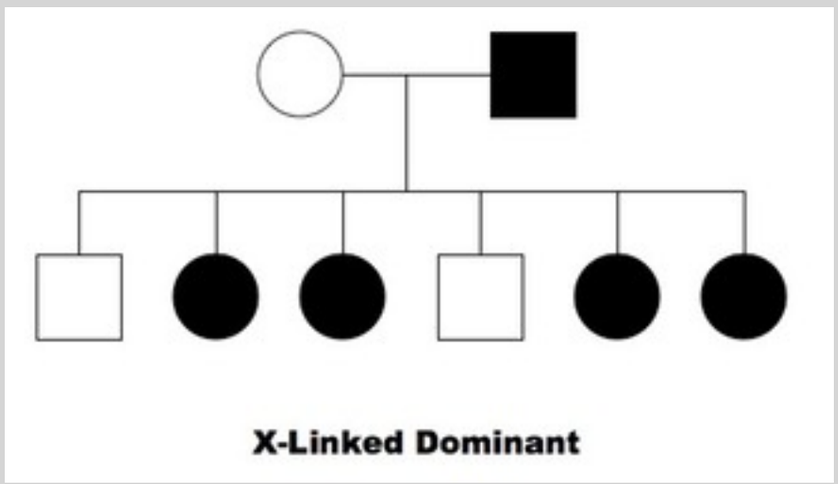
Introduction

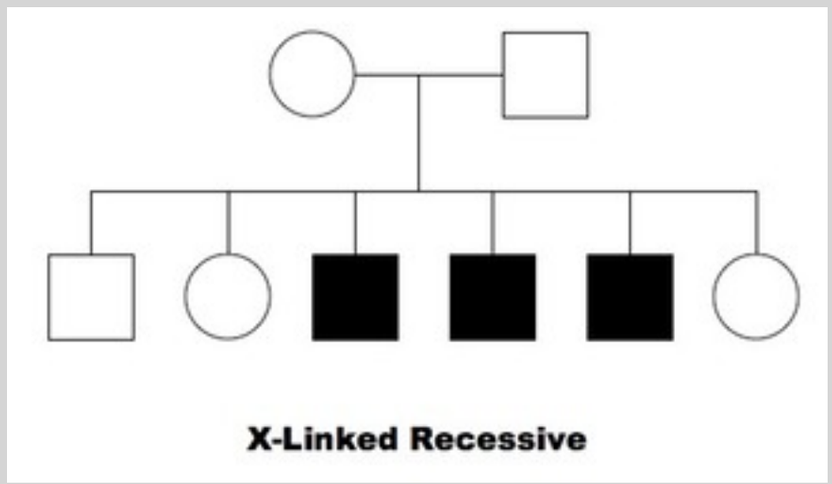






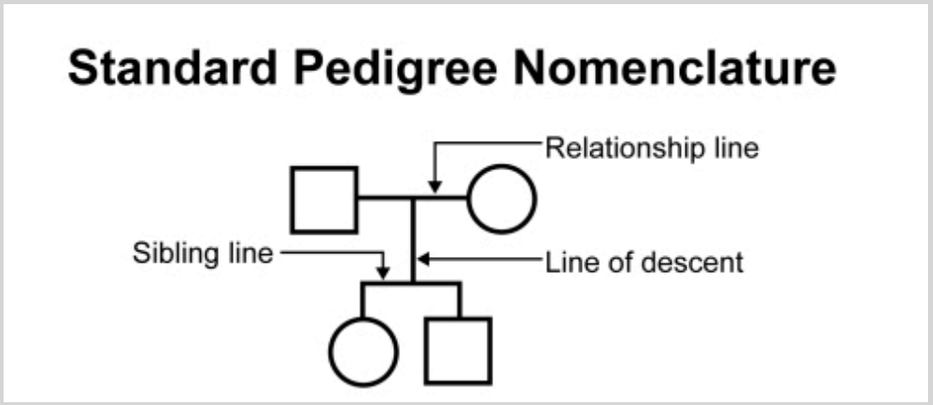
Some examples of different modes of inheritance inferred from pedigree information

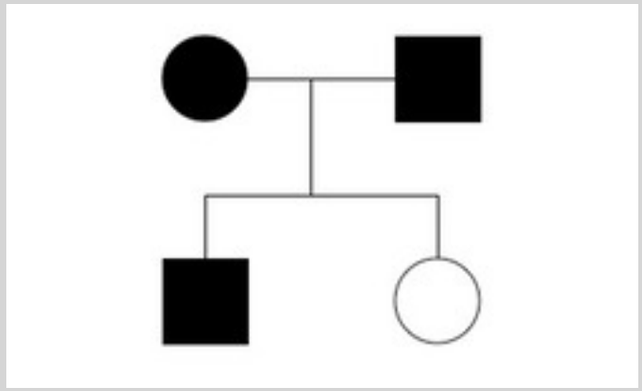


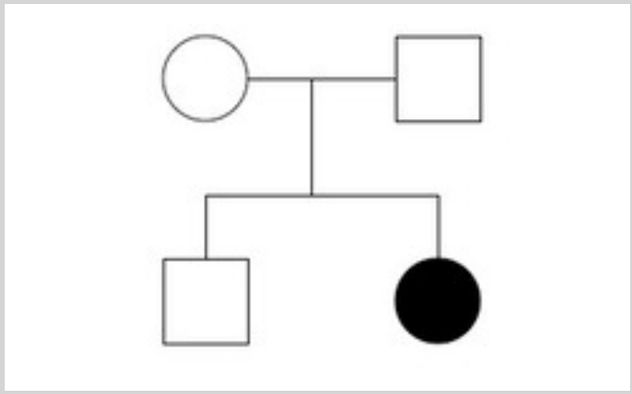


Genetic association studies

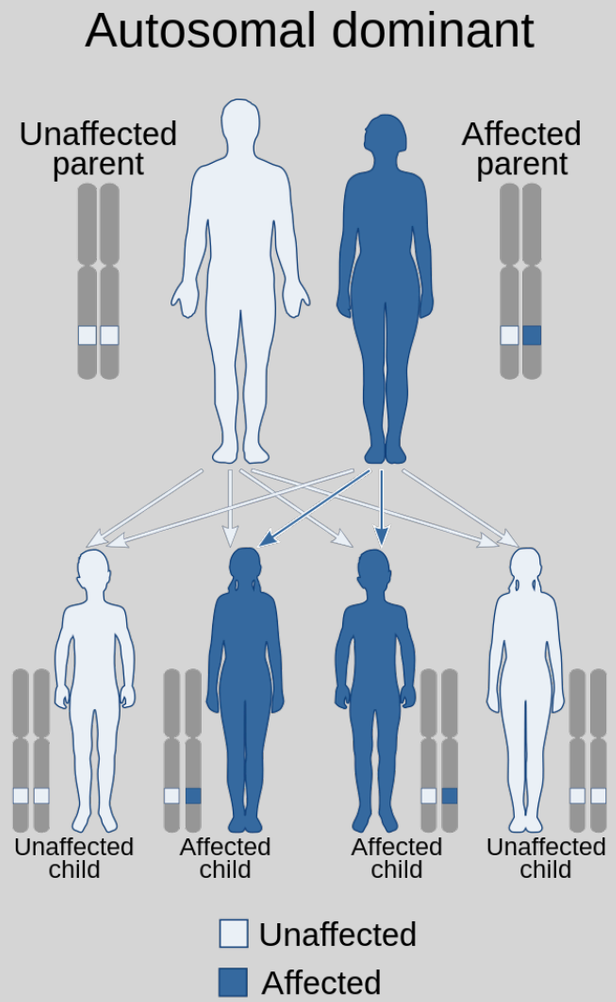
Introduction

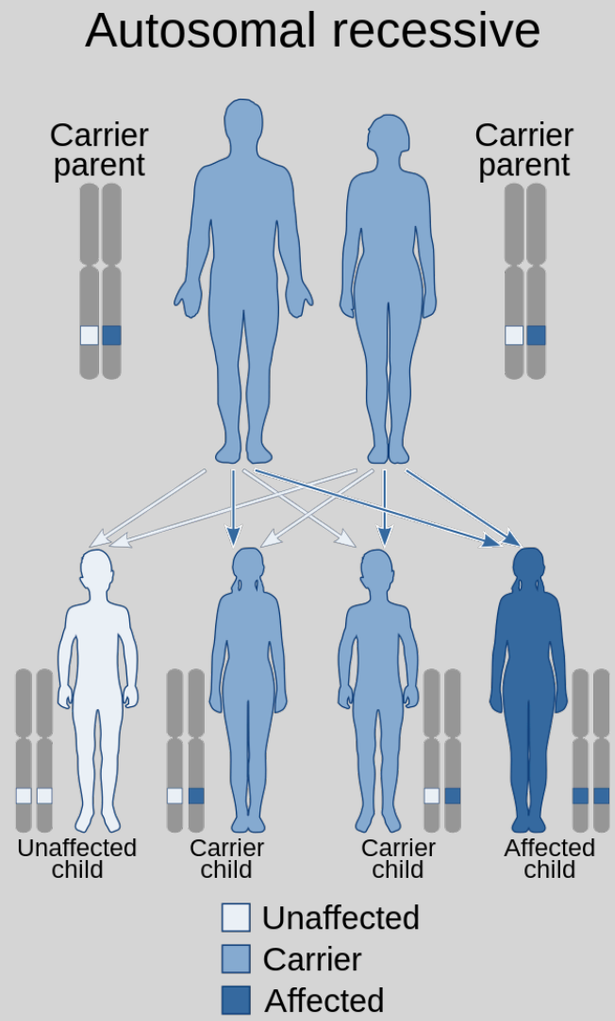






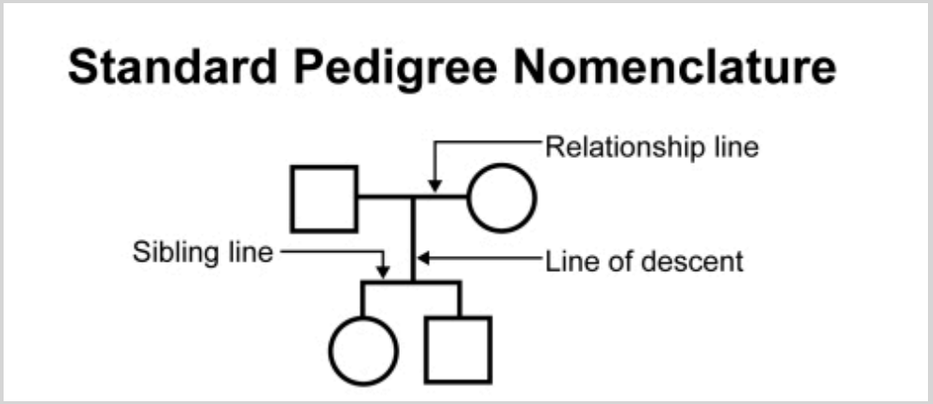
Some examples of different modes of inheritance inferred from pedigree information

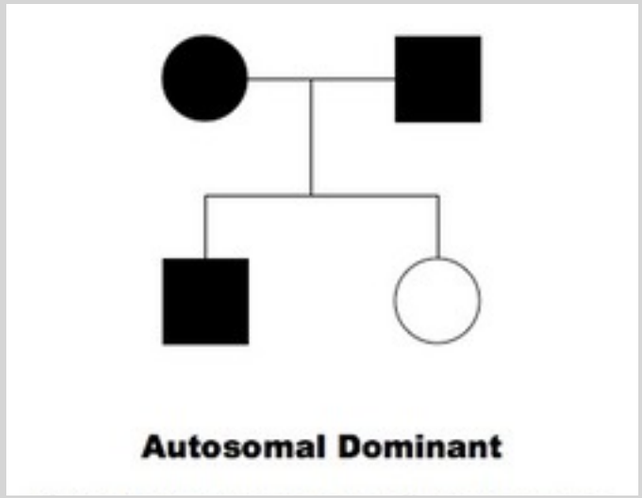


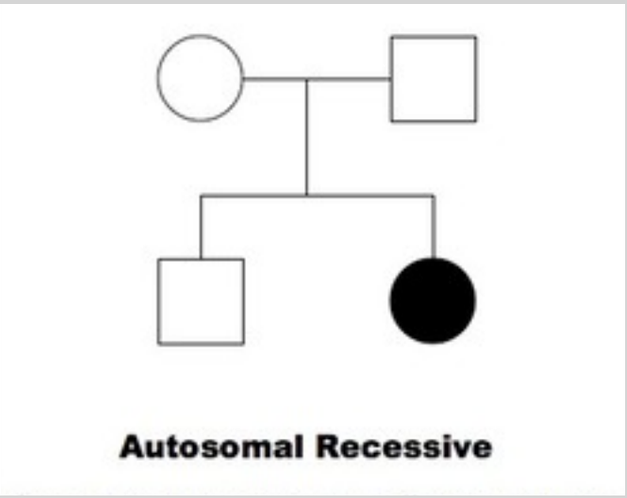


Genetic association studies

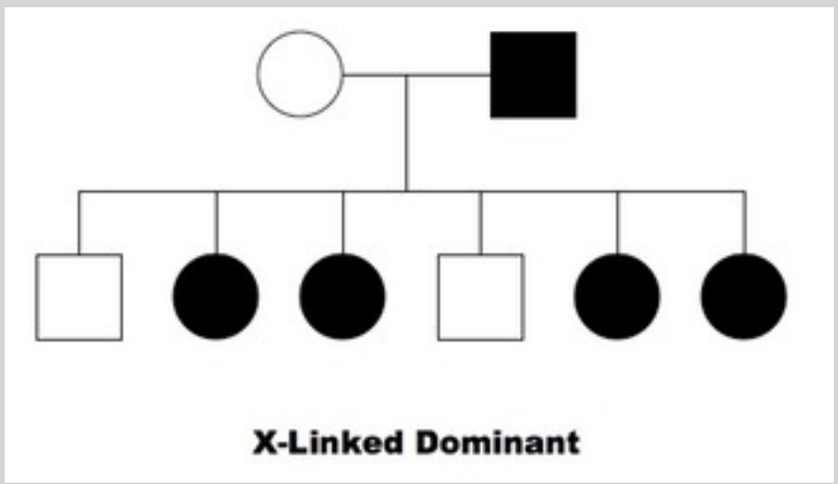
Introduction

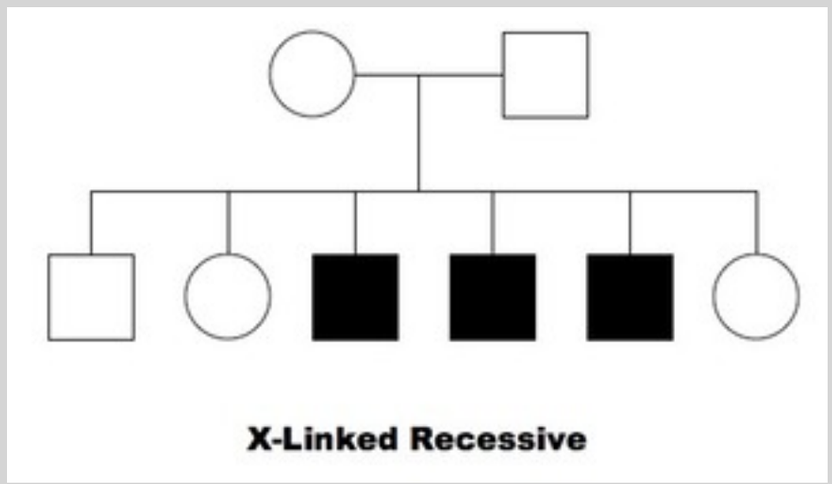






Some examples of different modes of inheritance inferred from pedigree information





Genetic association studies

Introduction

Goal:
Investigate associations between markers and a trait (disease).

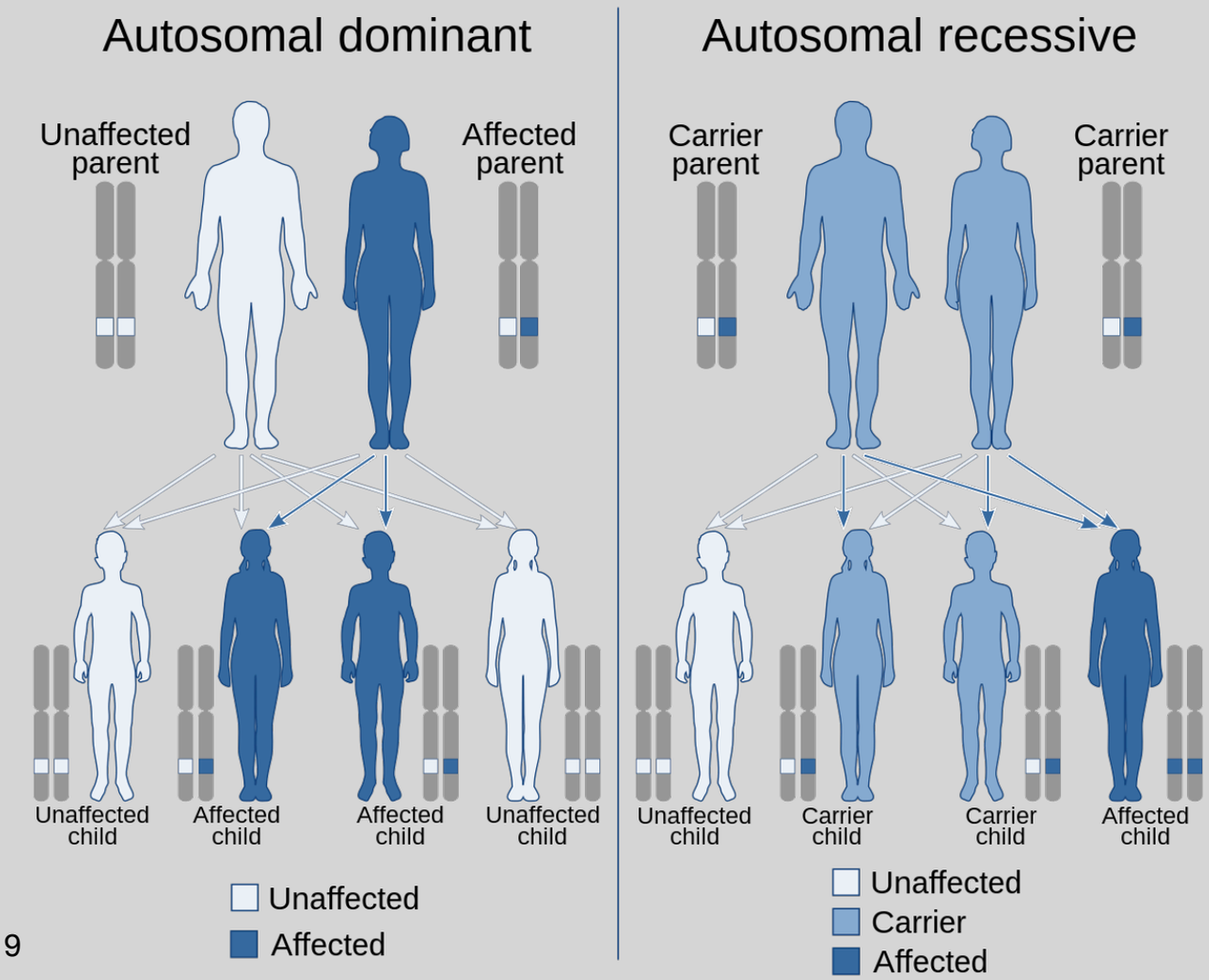
Early studies investigated rare conditions that show clear Mendelian segregation through families, and ver successfully located those genetic variations because they carry 100% of the risk

Examples:

Huntington's disease is a neurodegenerative disease where patients show variety of movement disorders, with involuntary and uncoordinated body movements.

Common autosomal dominant inheritance disorders
- Huntington's disease
- Marfan syndrome (cardiovascular)
- Acondroplasia
- …about 200 disorders

Common autosomal recessive inheritance disorders
- Spinal muscular atrophy
- Cystic fibrosis
- ….about 450 disorders
-



Genetic association studies

Introduction

Goal:
Investigate associations between markers and a trait (disease).

Complex diseases are caused by a combination of genetic, environmental, and lifestyle factors, ….and determining the heritability (genetic contribution to the trait) its difficult as there is no clear mendelian inheritance patterns.

- In almost any complex trait that has been studied, **many loci contribute to standing genetic variation**…so that mutations in many genes contribute to genetic variation in the population.
- On average, the proportion of variance explained at the individual variants is small.
- Each variant is only one of the many genetic and environmental causal factors, each of which are neither necessary nor sufficient to individually cause the disease. Thus, they **predispose** to—rather than directly result in—its development.

Examples:

Complex diseases are also referred as multi-factorial traits, non-communicable diseases or chronic diseases

- Diabetes
- Stroke
- Asthma
- Obesity
- Hypothyroidism
- Cancer
- Schizophrenia
- Depression
- Epilepsy
- …

Genetic association studies

Introduction

Goal:
Investigate associations between markers and a trait (disease).

Designs:

- Unrelated subjects (population-based)
- Related subjects from pedigrees (family-based)

Genetic association studies

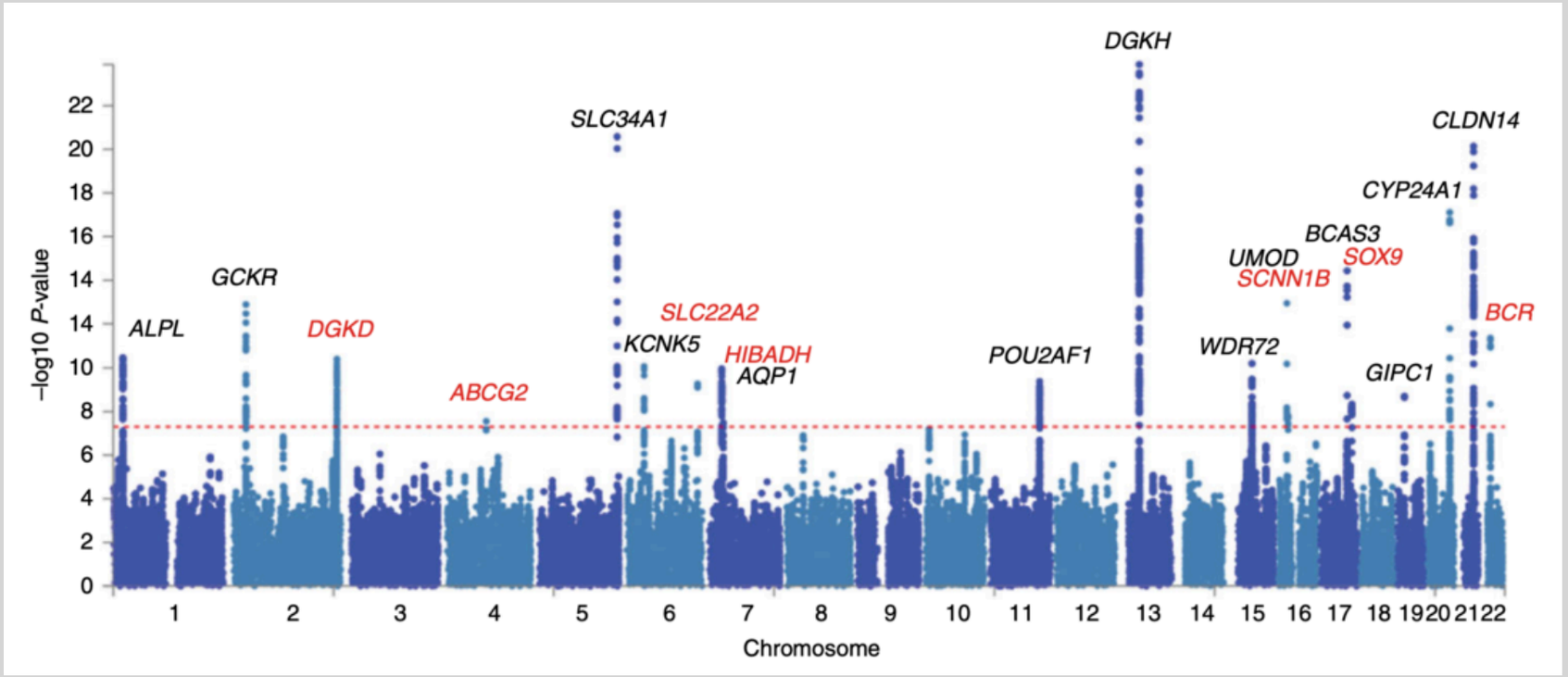
Introduction - population-based studies

- We will focus on population-based association studies, executed on unrelated subjects. Two types:
  - **Allele-based tests and genotype-based tests**: are hypothesis driven, where a candidate locus is being tested.
  - **Genome-wide association studies (GWAS)**: involve the analysis of multiple polymorphisms conducted without prior hypothesis.

Genetic association studies

Introduction - population-based studies

- We will focus on population-based association studies, executed on unrelated subjects. Two types:
  - **Allele-based tests and genotype-based tests**: are hypothesis driven, where a candidate locus is being tested.
  - **Genome-wide association studies (GWAS)**: involve the analysis of multiple polymorphisms conducted without prior hypothesis.



Typically, a single test statistic (for case–control studies, a chi-squared ($\chi 2$) comparison of absolute genotype counts) is calculated for each variant passing quality control.

Increasing number of studies are being extended from case-control studies to population-based cohorts.

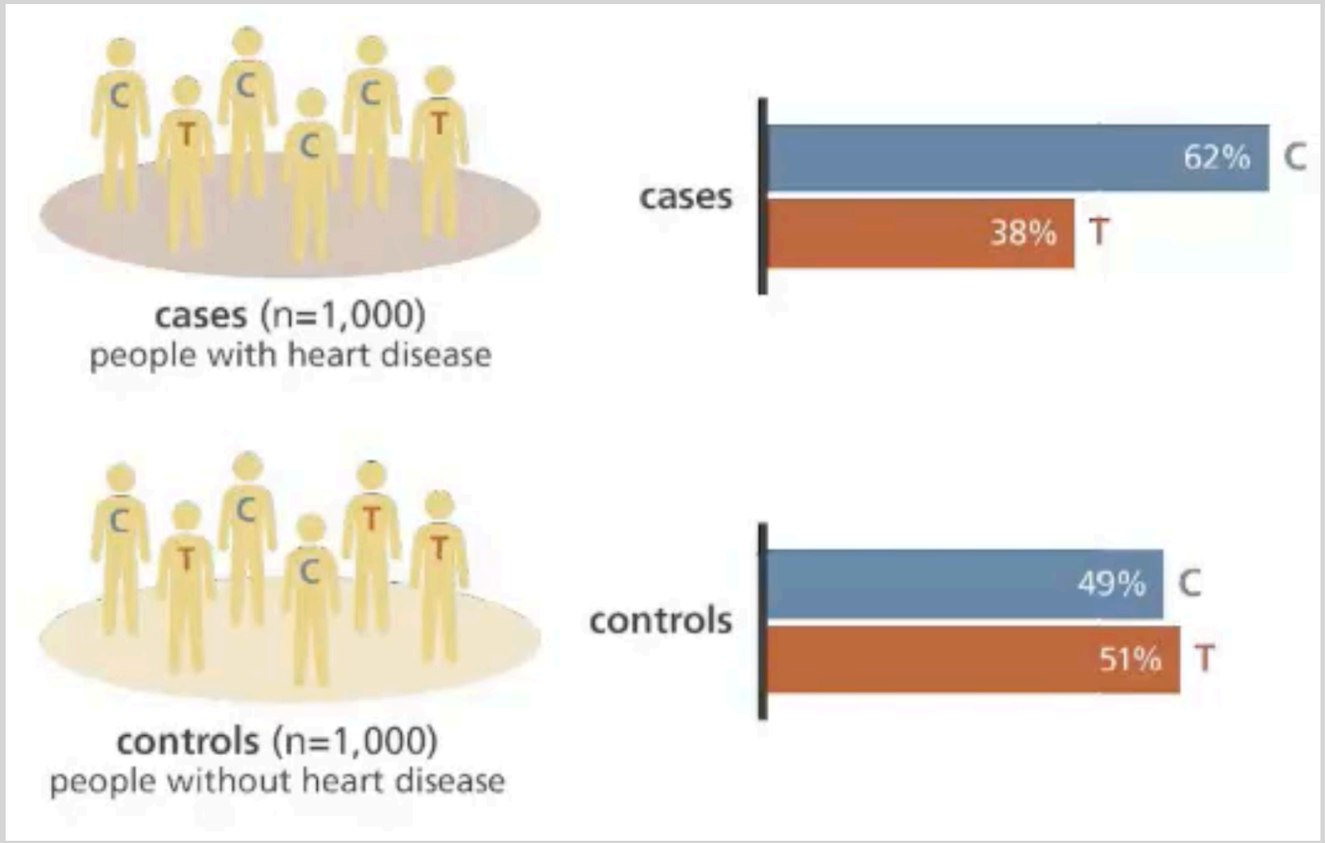
Manhattan plot depicting several strongly associated risk loci. Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level. The peaks indicate genetic variants that are found more often in individuals with kidney stones.

When we are looking for regions of the genome or SNP that is causal for a gene, we often find that a whole bunch of SNPs are associated with the disease. Its not that they all cause disease, it is just that a whole bunch are correlated with the causal SNP (passenger mutations). Thus it is our job to identify the causal needle in the haystack.

Genetic association studies

Introduction - population-based studies

- We will focus on population-based association studies, executed on unrelated subjects. Two types:
  - **Allele-based tests and genotype-based tests**: are hypothesis driven, where a candidate locus is being tested.



Requires a case-control study.

The success rate of candidate gene case-control studies has been very poor. In 2002:

- 603 published disease-genetic variant associations found that only 6 appeared to be independently replicated

Complex traits are not caused by common genetic polymorphisms but by multiple rare ones

Zondervan, K.T. and Cardon, L.R., 2007. Designing candidate gene and genome-wide case–control association studies. Nature protocols, 2(10), pp.2492-2501.

Genetic association studies

Introduction - some notes on study designs

Observational (descriptive):

- **Case report** (single cases) **and case series**
- **Cross-sectional studies:** involves the analysis of a trait and a set of markers in a particular point in time given a random sample of a population. Can't inform about causal relationship with a trait nor the risk of a disease but can inform about potential relationships.

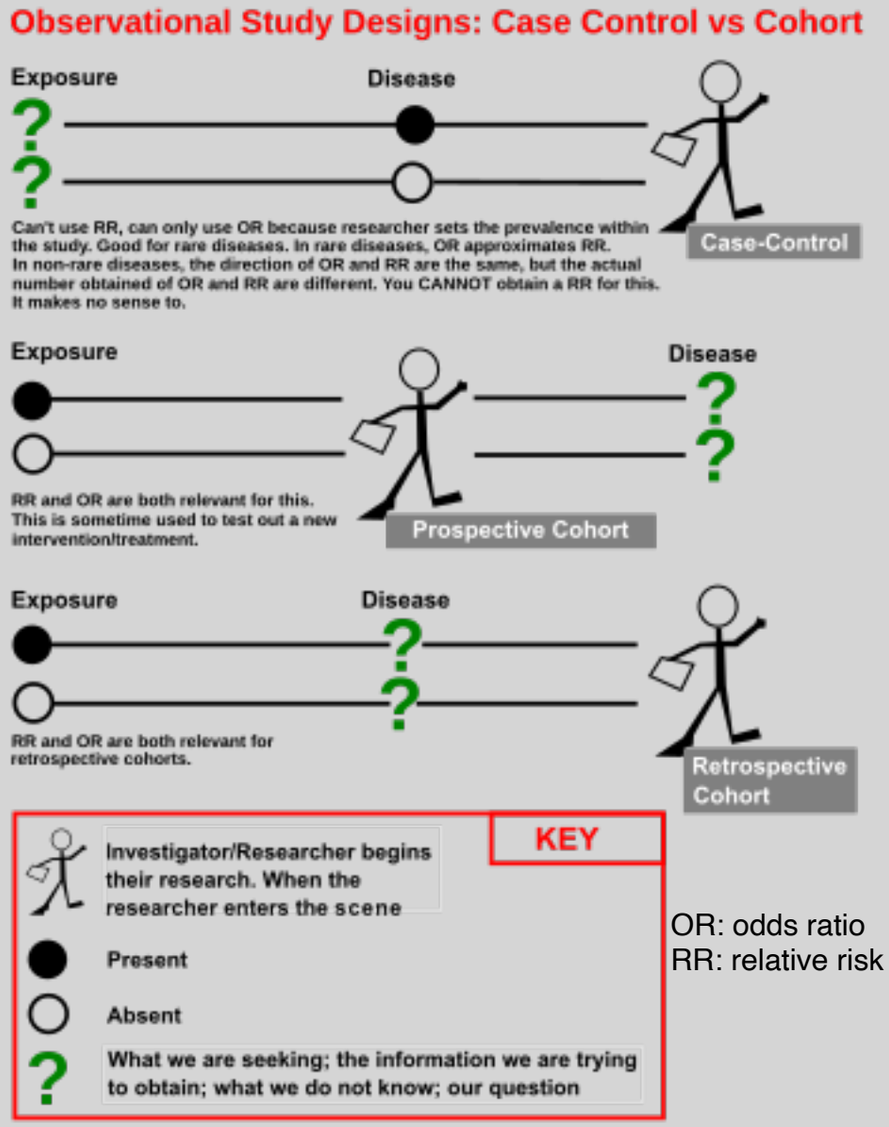
Observational (and analytical):

- **Case-control studies**: compares individuals with the condition (cases) to those without (controls) to investigate causes of a trait. Given a trait we aim to infer what are the causes (exposures) that caused it.
- **Cohort studies**: a type of longitudinal study that recruit participants of a specific population (given a risk factor) and observe them over time. Given a set of risks factors we aim to estimate if participants have developed the trait.

Experimental:

- **Randomized Controlled Trials:** random assignment to treatment or placebo (control) groups. Possible risk factors (exposures) are randomized and thus, strongest inference of causality.
- **Non-randomized Trials:** mostly used when RCT is not applicable due to practical or ethical limitations.

- Within each class, we classify the study by timeline and randomization structure.
- The choice of study design or data resource depends on the required sample size, the experimental question, the availability of existing data or…the ease by which new data can be obtained.



OR: odds ratio
RR: relative risk

Content

Genetic Association Studies

1. Introduction
2. Allele based tests
3. Genotype based tests
4. Quantitative traits and multiple polymorphisms
5. Computer exercise

Genetic association studies

The data

The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

r_0, r_1 and r_2 refer to the number of observed individuals with genotype aa, aA and AA respectively, in the case group

s_0, s_1 and s_2 refer to the number of observed individuals with genotype aa, aA and AA respectively, in the control group

n_0, n_1 and n_2 refer to the total number of observed individuals with genotype aa, aA and AA respectively

Genetic association studies

The data

The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

r_0, r_1 and r_2 refer to the number of observed individuals with genotype aa, aA and AA respectively, in the case group

s_0, s_1 and s_2 refer to the number of observed individuals with genotype aa, aA and AA respectively, in the control group

n_0, n_1 and n_2 refer to the total number of observed individuals with genotype aa, aA and AA respectively

Risk of developing the disease for each genotype : $R_{aa} = \frac{r_0}{n_0}$, $R_{aA} = \frac{r_1}{n_1}$ and $R_{AA} = \frac{r_2}{n_2}$

Genetic association studies

The data

- Let p be the allele frequency of the A allele and $P_{cases}(A)$ the frequency of allele A in the case group and $P_{control}(A)$ the frequency of allele A in the control group.
- Hypothesis:

$$H_0 : P_{cases}(A) = P_{control}(A)$$

$$H_0 : P_{cases}(A) \neq P_{control}(A)$$

- The test assumes Hardy-Weinberg equilibrium.

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

- Contingency table:

	a	A	Total	$\hat{p}$
Cases	$r_a = 2r_0 + r_1$	$r_A = 2r_2 + r_1$	$2r$	$r_A/(2r)$
Controls	$s_a = 2s_0 + s_1$	$s_A = 2s_2 + s_1$	$2s$	$s_A/(2s)$
Total	$n_a = 2n_0 + n_1$	$n_A = 2n_2 + n_1$	$2n$	$n_A/(2n)$

Risk of developing the disease for each allele : $R_a = \frac{r_a}{n_a}$ and $R_A = \frac{r_A}{n_A}$

Genetic association studies

Allele based tests

- Statistical tests:
  - Chi square test for independence
  - Fisher's exact test
- Odds ratio for the effect size

Genetic association studies

Allele based tests

- Let p be the allele frequency of the A allele and $P_{cases}(A)$ the frequency of allele A in the case group and $P_{control}(A)$ the frequency of allele A in the control group.
- Hypothesis:

$$H_0 : P_{cases}(A) = P_{control}(A)$$

$$H_0 : P_{cases}(A) \neq P_{control}(A)$$

- The test assumes Hardy-Weinberg equilibrium.

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

- Contingency table:

	a	A	Total	$\hat{p}$
Cases	$r_a = 2r_0 + r_1$	$r_A = 2r_2 + r_1$	$2r$	$r_A/(2r)$
Controls	$s_a = 2s_0 + s_1$	$s_A = 2s_2 + s_1$	$2s$	$s_A/(2s)$
Total	$n_a = 2n_0 + n_1$	$n_A = 2n_2 + n_1$	$2n$	$n_A/(2n)$

Genetic association studies

Allele based tests - Pearson's χ^2 test

- Let p be the allele frequency of the A allele and $P_{cases}(A)$ the frequency of allele A in the case group and $P_{control}(A)$ the frequency of allele A in the control group.
- Hypothesis:

$$H_0 : P_{cases}(A) = P_{control}(A)$$

$$H_0 : P_{cases}(A) \neq P_{control}(A)$$

- The test assumes Hardy-Weinberg equilibrium.

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

- Contingency table:

	a	A	Total	$\hat{p}$
Cases	$r_a = 2r_0 + r_1$	$r_A = 2r_2 + r_1$	$2r$	$r_A/(2r)$
Controls	$s_a = 2s_0 + s_1$	$s_A = 2s_2 + s_1$	$2s$	$s_A/(2s)$
Total	$n_a = 2n_0 + n_1$	$n_A = 2n_2 + n_1$	$2n$	$n_A/(2n)$

The test is a χ^2 test for independence in a 2 × 2 contingency table of alleles, which can be obtained from the genotype count table.

- χ^2 square test for independence

$$X^2 = \sum_{i,j} \frac{(observed - expected)^2}{expected}$$

Expected count:
e_{ij} = total row i × total colum j / total of table

- If H_0 true, then $X^2 \sim \chi^2_1$

Genetic association studies

Allele based tests - Pearson's χ^2 test

Example

- A polymorphism in the Dopamine receptor is supposed to be involved in Schizophrenia. In a case-control study, the following data were obtained:

	11	12	22	Total
Cases	7	69	57	133
Controls	20	56	33	109
Total	27	125	90	242

	1	2	Total
Cases	83	183	266
Controls	96	122	218
Total	179	305	484

$$r_1 = 2 \cdot r_{11} + r_{12} = 2 \cdot 7 + 69 = 83$$
$$r_2 = 2 \cdot r_{22} + r_{12} = 2 \cdot 57 + 69 = 183$$
$$\dots$$

- The test is a χ^2 test for independence in a 2 × 2 table of alleles.

$$X^2 = \sum_{i,j} \frac{(observed - expected)^2}{expected}$$

$$e_{11} = 266 \cdot 179/484 = 98.38$$
$$e_{12} = 218 \cdot 179/484 = 80.62$$
$$\dots$$

	1	2	Total
Cases	98.38	167.62	266
Controls	80.62	137.38	218
Total	179	305	484

$$X^2 = \frac{(83 - 98.38)^2}{98.38} + \dots + \frac{(122 - 137.38)^2}{137.38} = 8.4671$$

$$p - value = P(\chi^2 \leq 8.4761) = 0.0036$$

Expected allele count:
e_{ij} = total row i × total colum j / total of table

We REJECT the null hypothesis that $H_0 : P_{cases}(A) = P_{control}(A)$

Genetic association studies

Allele based tests - Pearson's χ^2 test

Example in R

- A polymorphism in the Dopamine receptor is supposed to be involved in Schizophrenia. In a case-control study, the following data were obtained:

	11	12	22	Total
Cases	7	69	57	133
Controls	20	56	33	109
Total	27	125	90	242

	1	2	Total
Cases	83	183	266
Controls	96	122	218
Total	179	305	484

- The test is a χ^2 test for independence in a 2 × 2 table of alleles.

```
> X <- matrix(c(7,69,57,20,56,33),byrow=TRUE,ncol=3)
> colnames(X) <- c("11","12","22")
> rownames(X) <- c("Cases","Controls")

> Y <- cbind(2*X[,1]+X[,2],2*X[,3]+X[,2])
> colnames(Y) <- c("1","2")

> chisq.test(Y,correct=FALSE)
Pearson's Chi-squared test
data: Y
X-squared = 8.4671, df = 1, p-value = 0.003616
```

Genetic association studies

Allele based tests - Fisher exact test

- Recall: $H_0 : P_{cases}(A) = P_{control}(A)$
- Calculation of Fisher's exact test involves direct calculation of the probability P from the number of samples observed n and its counts.
- Often used for tables with low counts with a 2x2 matrix, or when $e_{ij} < 5$.
- Like the χ^2 test, events must be independent

Example in R (same data)

```
>Y
          1   2
cases    83 183
controls 96 122

> fisher.test(Y)
Fisher's Exact Test for Count Data
data: Y

p-value = 0.00448

alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.3903016 0.8512261
sample estimates: odds ratio 0.5770451
```

Genetic association studies

Allele based tests - OR for the effect size

- Odds of an event p:

$$Odds(p) = \frac{p}{1-p} \qquad \text{for example:} \quad Odds(disease) = \frac{P(disease)}{P(nodisease)}$$

- The odds ratio (OR) compares the odds of an event in two groups and provides a relative **measure of effect** in case-control studies. Quantifies the strength of an association between two events. Since $0 \leq p \leq 1$, odds can range from 0 to infinity.

- From the data, we can estimate the probability $\hat{p}$ of an event from the frequency counts. Starting from a 2 x 2 contingency table:

	Event (Yes)	Event (No)	Total
Group 1	a	b	(a+b)
Group 2	c	d	(c+d)

$$\hat{p}_1 = \frac{a}{a+b} \text{ and } \hat{p}_2 = \frac{c}{c+d}$$

Odds of group 1 $\hat{O}_1 = \frac{\hat{p}_1}{1-\hat{p}_1} = \frac{a/(a+b)}{b/(a+b)} = \frac{a}{b}$

$$OR = \frac{O_1}{O_2} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{b \cdot c}$$

Odds of group 2 $\hat{O}_2 = \frac{\hat{p}_2}{1-\hat{p}_2} = \frac{c/(c+d)}{d/(c+d)} = \frac{c}{d}$

- OR = 1 no association, odds are the same in both groups
- OR > 1 outcome is more likely in group 1
- OR < 1 outcome is more likely in group 2

Genetic association studies

Allele based tests - OR for the effect size

- Odds of an event p:

$$Odds(p) = \frac{p}{1-p} \qquad \text{for example:} \quad Odds(disease) = \frac{P(disease)}{P(nodisease)}$$

- The odds ratio (OR) compares the odds of an event in two groups and provides a relative **measure of effect** in case-control studies. Quantifies the strength of an association between two events. Since $0 \leq p \leq 1$, odds can range from 0 to infinity.

- From the data, we can estimate the probability $\hat{p}$ of an event from the frequency counts. Starting from a 2 x 2 contingency table:

	Event (Yes)	Event (No)	Total
Group 1	a	b	(a+b)
Group 2	c	d	(c+d)

$$\hat{p}_1 = \frac{a}{a+b} \text{ and } \hat{p}_2 = \frac{c}{c+d}$$

$$OR = \frac{O_1}{O_2} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{b \cdot c}$$

Odds of group 1 $\hat{O}_1 = \frac{\hat{p}_1}{1-\hat{p}_1} = \frac{a/(a+b)}{b/(a+b)} = \frac{a}{b}$

Odds of group 2 $\hat{O}_2 = \frac{\hat{p}_2}{1-\hat{p}_2} = \frac{c/(c+d)}{d/(c+d)} = \frac{c}{d}$

- The distribution of the log odds ratio $ln(OR)$ is approximately normal $ln(OR) \sim N(ln(OR), V(ln(OR)))$

- Variance of $ln(OR)$ is known, which allows for the calculation of confidence intervals for the OR

$$V(log(OR)) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \text{ and } CI_{ln(OR)} = ln(OR) \pm z_{\alpha/2} \cdot SD(ln(OR))$$

Genetic association studies

Allele based tests - OR for the effect size

- Odds of an event p:

$$Odds(p) = \frac{p}{1-p} \qquad \text{for example:} \quad Odds(disease) = \frac{P(disease)}{P(nodisease)}$$

- The odds ratio (OR) compares the odds of an event in two groups and provides a relative **measure of effect** in case-control studies. Quantifies the strength of an association between two events. Since $0 \leq p \leq 1$, odds can range from 0 to infinity.
- The OR compares the odds of the disease for the two alleles:

	A	B
Cases	n_{11}	n_{12}
Controls	n_{21}	n_{22}

Odds of disease with allele A

$$OR = \frac{odds_{cases}}{odds_{controls}} = \frac{\frac{n_{11}}{n_{21}}}{\frac{n_{12}}{n_{22}}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}$$

Odds of disease with allele B

- OR = 1 no association between A and B
- OR > 1 outcome is more likely in group with allele A
- OR < 1 outcome is more likely in group with allele B

Genetic association studies

Allele based tests - OR for the effect size

Example

- A polymorphism in the Dopamine receptor is supposed to be involved in Schizophrenia. In a case-control study, the following data were obtained:

	11	12	22	Total
Cases	7	69	57	133
Controls	20	56	33	109
Total	27	125	90	242

	1	2	Total
Cases	83	183	266
Controls	96	122	218
Total	179	305	484

- Odds ratio

$$OR = \frac{n_{11}/n_{21}}{n_{12}/n_{22}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}} = \frac{83 \cdot 122}{96 \cdot 183} = 0.576$$

Odds that a person with allele 1 will suffer the disease to the odds that a person will still get a disease with allele 2.

- Calculation of 95% confidence intervals for the OR:

Sample $SD(ln(OR)) = \sqrt{V(ln(OR))} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = 0.19$

$$CI = e^{ln(OR) \pm z_{\alpha/2} SD(ln(OR))} = e^{ln(0.5764) \pm 1.96 \cdot 0.19} = (0.397; 0.837)$$

There is a probability of 0.95 that we've captured the true OR in the interval (0.397; 0.837)

Content

Genetic Association Studies

1. Introduction
2. Allele based tests
3. Genotype based tests
4. Quantitative traits and multiple polymorphisms
5. Computer exercise

Genetic association studies

The data

The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

r_0, r_1 and r_2 refer to the number of observed individuals with genotype aa, aA and AA respectively, in the case group

s_0, s_1 and s_2 refer to the number of observed individuals with genotype aa, aA and AA respectively, in the control group

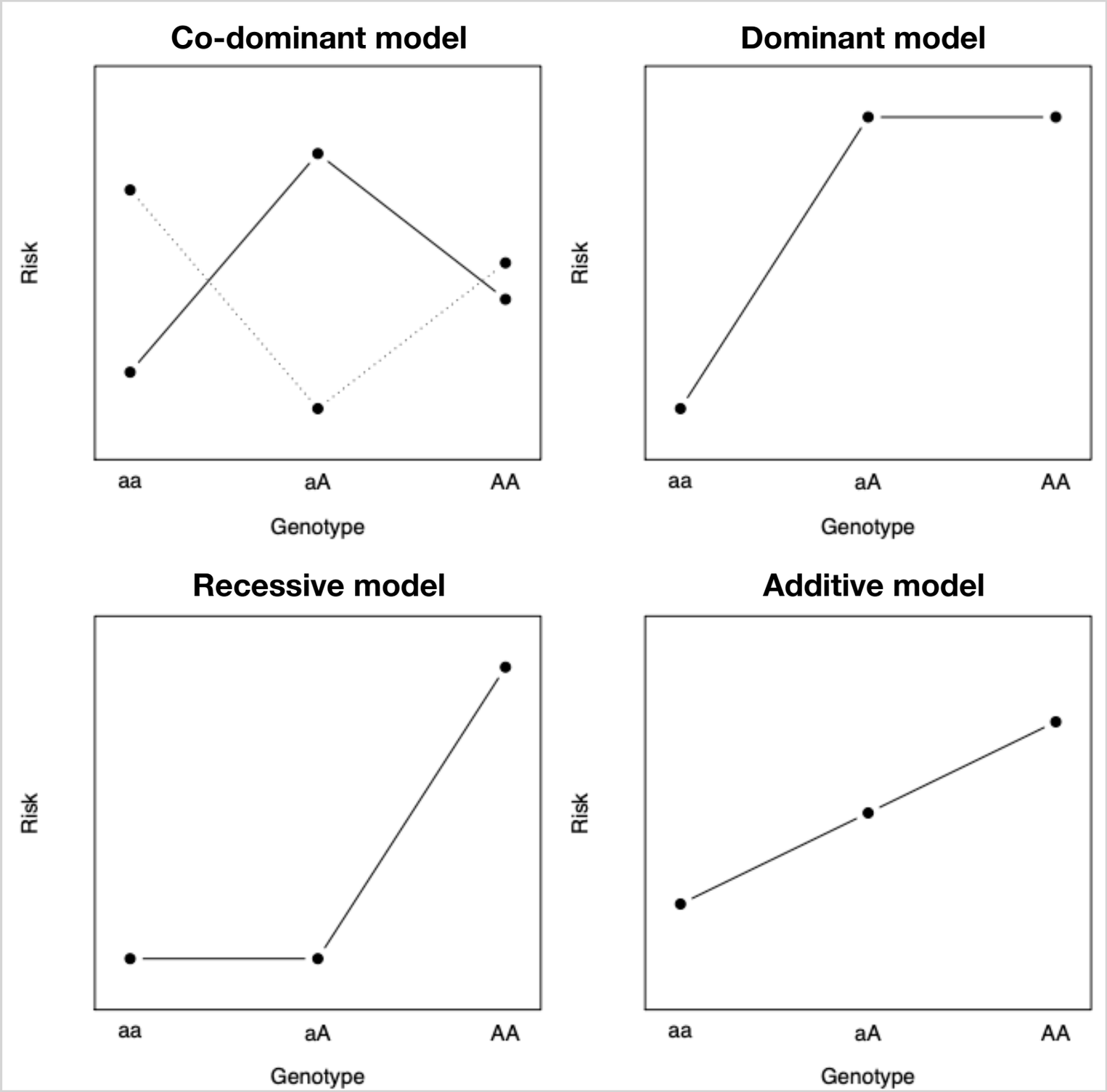
n_0, n_1 and n_2 refer to the total number of observed individuals with genotype aa, aA and AA respectively

We can test for association using different genetic models:

- Codominant model
- Dominant model
- Recessive model
- Additive model

Genetic association studies

Genotype based tests



Genetic association studies

Genotype based tests - Codominant test

- We test the null hypothesis of no effect of the marker on the trait. Formally:
  - $H_0 : P(Y = 1 \mid AA) = P(Y = 1 \mid aA) = P(Y = 1 \mid aa)$
  - H_1 : at least one pair different
- Test statistic:

$$X^2 = \sum_{i,j} \frac{(observed - expected)^2}{expected}$$

- Under H_0, the test statistic X^2 follows a χ^2_2 distribution (df=2)
- The test makes no assumptions about the relationship between genotype and trait.
- Under H_1, each genotype can have a different disease rate.
- The test can reject the null if the data support heterozygote advantage (overdominance).

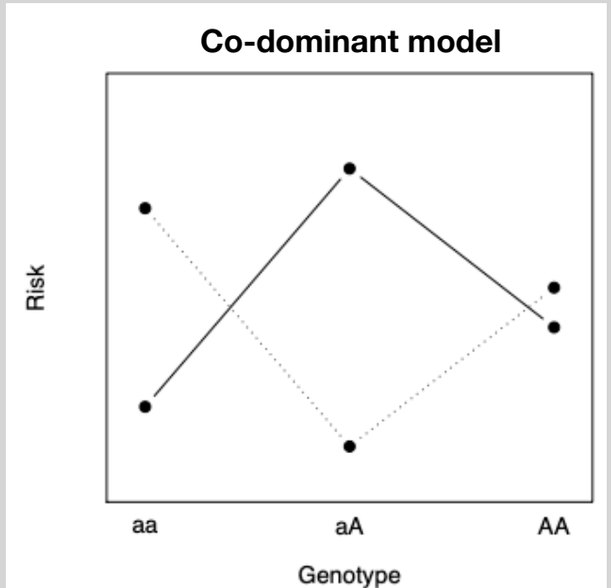
The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The original data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n



Genetic association studies

Genotype based tests - Codominant

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

	GG	GA	AA	Total
Cases	66	43	4	113
Controls	99	15	0	114
Total	165	58	4	227

Example

- TNF genotype (G/A polymorphism) is on a study on acne patients and controls
- Estimate $\hat{p}(Y=1) = \frac{r}{n} = \frac{113}{227} = 0.497$
- The expected counts for cases:
  - $exp(Y=1\,|\,aa) = \hat{p}(Y=1) \cdot n_0 = \frac{r \cdot n_0}{n} = 113 \cdot 165/227 = 82.136$
  - $exp(Y=1\,|\,aA) = \hat{p}(Y=1) \cdot n_1 = \frac{r \cdot n_1}{n} = 113 \cdot 58/227 = 28.872$
  - $exp(Y=1\,|\,AA) = \hat{p}(Y=1) \cdot n_2 = \frac{r \cdot n_2}{n} = 113 \cdot 4/227 = 1.991$
- The expected counts for controls:
  - $exp(Y=0\,|\,aa) = (1-\hat{p}(Y=1)) \cdot n_0 = \frac{s \cdot n_0}{n} = 114 \cdot 165/227 = 82.863$
  - $exp(Y=0\,|\,aA) = (1-\hat{p}(Y=1)) \cdot n_1 = \frac{s \cdot n_1}{n} = 114 \cdot 58/227 = 29.127$
  - $exp(Y=0\,|\,AA) = (1-\hat{p}(Y=1)) \cdot n_2 = \frac{s \cdot n_2}{n} = 114 \cdot 4/227 = 2.008$
- The chi-square statistics:
- $X^2 = \sum_{i,j} \frac{(observed - expected)^2}{expected} = \frac{(66-82.136)^2}{82.136} + \ldots + \frac{(0-2.008)^2}{2.008} = 24.113$
- P-value = $P(\chi^2_2 \geq 24.113) = 5.806e-06$

Genetic association studies

Genotype based tests - Codominant test

Example - R code

- TNF genotype (G/A polymorphism) is on a study on acne patients and controls

	GG	GA	AA	Total
Cases	66	43	4	113
Controls	99	15	0	114
Total	165	58	4	227

```
> X <- matrix(c(66,43,4,99,15,0),byrow=TRUE,ncol=3)
> colnames(X) <- c("GG","GA","AA")
> rownames(X) <- c("Acne","Control")
> X
        GG GA AA
Acne    66 43 4
Control 99 15 0
> results <- chisq.test(X)
Warning message:
In chisq.test(X) : Chi-squared approximation may be incorrect

> print(results)

Pearson's Chi-squared test
X-squared = 24.1133, df = 2, p-value = 5.806e-06

> results$expected
              GG       GA       AA
Acne    82.13656 28.87225 1.991189
Control 82.86344 29.12775 2.008811
```

```
> fisher.test(X)
Fisher's Exact Test for Count Data
data: X
p-value = 1.97e-06
alternative hypothesis: two.sided
```

We reject the null hypothesis that the probability of disease with all the genotypes is the same

Genetic association studies

Genotype based tests - Dominant test

- Columns in the original data table are combined to reflect the dominant model.

	aa	aA or AA	Total
Cases	r_0	$r_1 + r_2$	r
Controls	s_0	$s_1 + s_2$	s
Total	n_0	$n_1 + n_2$	n

- Hypothesis:
  - $H_0 : P(Y = 1 \mid aa) = P(Y = 1 \mid (aA + AA))$, there is NO association between A and the outcome
  - H_1 : there is an association between A and the outcome
- Test statistic:

$$X^2 = \sum_{genotypes} \frac{(observed - expected)^2}{expected}$$

- Under H_0, the test statistic X^2 follows a χ^2_1 distribution (df=1)

The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

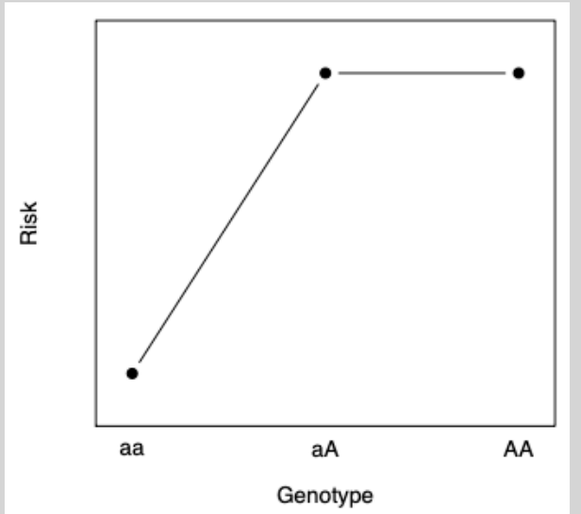
- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The original data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

Dominant model



Genetic association studies

Genotype based tests - Dominant test

Example - R code

- TNF genotype (G/A polymorphism) is on a study on acne patients and controls

	GG	GA	AA	Total
Cases	66	43	4	113
Controls	99	15	0	114
Total	165	58	4	227

```
> Y <- cbind(X[,1],X[,2]+X[,3])
> colnames(Y) <- c(“GG","GA or AA")
> rownames(Y) <- c("Acne","Control")
> Y
        GG GA or AA
Acne    66     47
Control 99     15
> results <- chisq.test(Y)
> print(results)

Pearson’s Chi-squared test with Yates’ continuity correction
data: Y
X-squared = 21.7021, df = 1, p-value = 3.184e-06


> results <- chisq.test(Y,correct=FALSE)
> print(results)
Pearson’s Chi-squared test
data: Y
X-squared = 23.1122, df = 1, p-value = 1.528e-06
```


We reject the null hypothesis that the probability of disease does not depend on A

Genetic association studies

Genotype based tests - Recessive test

- Columns in the original data table are combined to reflect the dominant model.

	aa or aA	AA	Total
Cases	$r_0 + r_1$	r_2	r
Controls	$s_0 + s_1$	s_2	s
Total	$n_0 + n_1$	n_2	n

- Hypothesis:
  - $H_0 : P(Y = 1 \mid AA) = P(Y = 1 \mid (aA + aa))$, so that the probability of disease does not depend on being homozygote AA
  - H_1 : there is an association between the presence of AA and the outcome
- Test statistic:

$$X^2 = \sum_{genotypes} \frac{(observed - expected)^2}{expected}$$

- Under H_0, the test statistic X^2 follows a χ^2_1 distribution (df=1)

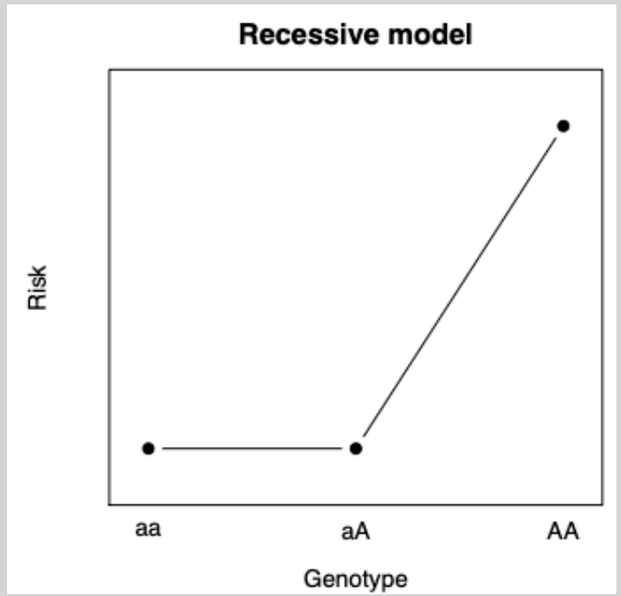
The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The original data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n



Genetic association studies

Genotype based tests - Recessive test

Example - R code

- TNF genotype (G/A polymorphism) is on a study on acne patients and controls

	GG	GA	AA	Total
Cases	66	43	4	113
Controls	99	15	0	114
Total	165	58	4	227

```
> Y <- cbind(X[,1]+X[,2],X[,3])
> colnames(Y) <- c("GG","GA or AA")
> rownames(Y) <- c("Acne","Control")
> Y
        GG or GA    AA
Acne         109     4
Control      114     0
> results <- chisq.test(Y)
Warning message:
In chisq.test(Y) : Chi-squared approximation may be incorrect

> print(results)

        Pearson's Chi-squared test with Yates' continuity correction
data: Y
X-squared = 2.3174, df = 1, p-value = 0.1279
```

```
> fisher.test(Y)
        Fisher's Exact Test for Count Data
data: X
p-value = 0.05977
alternative hypothesis: true odds
ratio is not equal to 1
95 percent confidence interval:
 0.000000 1.485382
sample estimates:
odds ratio
         0
```

We FAILED to reject the null hypothesis and thus, the probability of disease does not depend on being homozygote AA

Genetic association stud

Genotype based tests - summary

TNF genotype (G/A polymorphism) is on a study on acne patients and controls

	GG	GA	AA	Total
Cases	66	43	4	113
Controls	99	15	0	114
Total	165	58	4	227

Significant association between G/A polymorphism and acne risk under dominant model

Allele model test (G vs A):

$H_0 : P_{cases}(A) = P_{control}(A)$
$H_1 : P_{cases}(A) \neq P_{control}(A)$

p-value = 1.072e-06 (Fisher Exact Test)
odds ratio (G/A) = 0.2423696

We REJECT the null hypothesis that the probability of A is the same in both cases and controls groups
Disease is 0.2 more frequent with allele A

Codominant model test (GG vs GA vs AA):

$H_0 : P(Y = 1 \mid AA) = P(Y = 1 \mid GA) = P(Y = 1 \mid GG)$
H_1 : at least one pair different

p-value = 1.97e-06 (Fisher Exact Test)

We REJECT the null hypothesis that the probability of disease with all the genotypes is the same.
There has to be a genotype with higher prevalence.

Recessive model test (GG + GA vs AA):

$H_0 : P(Y = 1 \mid AA) = P(Y = 1 \mid (GG + GA))$ so that probability of disease does not depend on being homozygote AA
H_1 : disease depends on being homozygote AA

p-value = 0.05977 (Fisher Exact Test)

We FAILED to reject the null hypothesis that the probability of disease does not depend on being homozygote AA.
Probability of disease is the same in AA vs GA/GG.

Dominant model test (GG vs GA + AA):

$H_0 : P(Y = 1 \mid GG) = P(Y = 1 \mid (GA + AA))$ so that the probability of disease does not depend on being homozygote GG
H_1 : there is an association between the presence of AA and the outcome

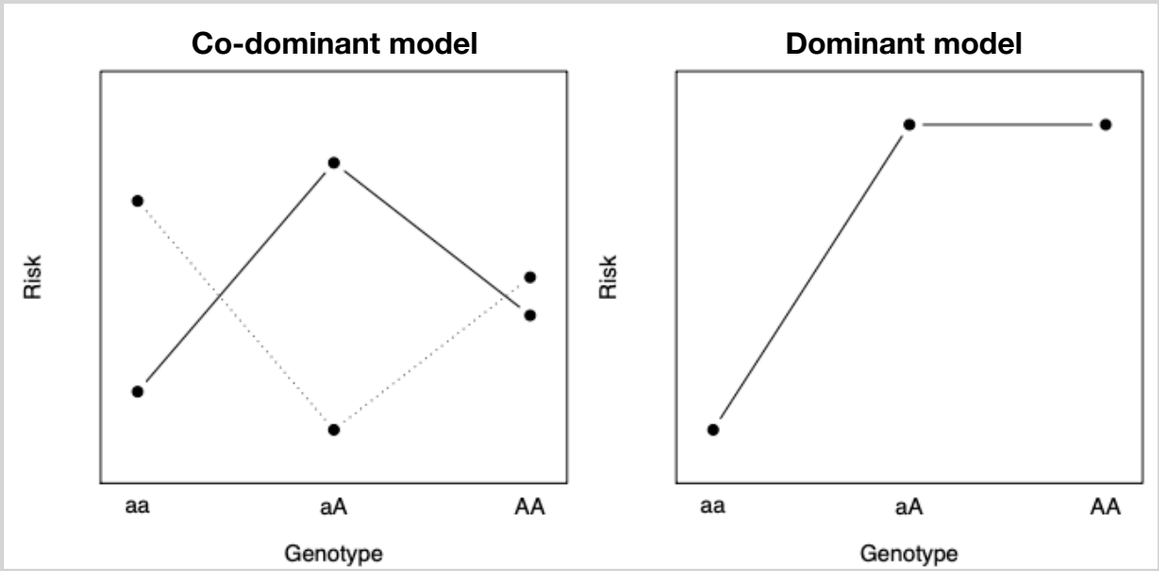
p-value = 1.528e-06 (χ^2 without cc)

We REJECT the null hypothesis that the probability of disease does not depend on presence of A.
Probability of disease is NOT the same in GG vs GA/AA

Genetic association studies

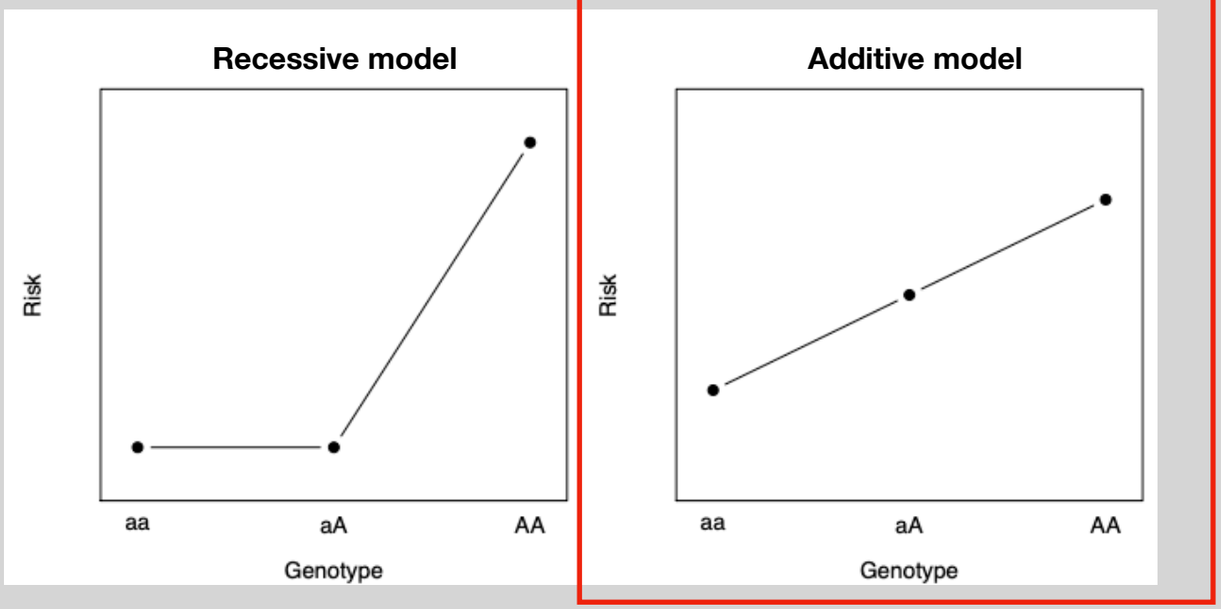
Genotype based tests

Some codominant inheritance traits
- Alleles A and B in blood type
- HLA protein (cell surface antigen)
- …



Common autosomal dominant inheritance disorders
- Huntington's disease
- Marfan syndrome (cardiovascular)
- Acondroplasia
- …about 200 disorders
-

Common autosomal recessive inheritance disorders
- Spinal muscular atrophy
- Cystic fibrosis
- ….about 450 disorders
-



Genetic association studies

Genotype based tests - Additive test

- Basic idea: disease risk increases as a function of the number of alleles (0,1 or 2).
- There are two tests for the additive genetic model
  - The alleles test
  - Cochran-Armitage trend test

The alleles test:

- Given that $P_{cases}(A)$ denotes the frequency of A alleles among cases and $P_{control}(A)$ denotes the frequency of A alleles among cases in the population, the hypothesis:

$$H_0 : P_{cases}(A) = P_{control}(A)$$

$$H_0 : P_{cases}(A) \neq P_{control}(A)$$

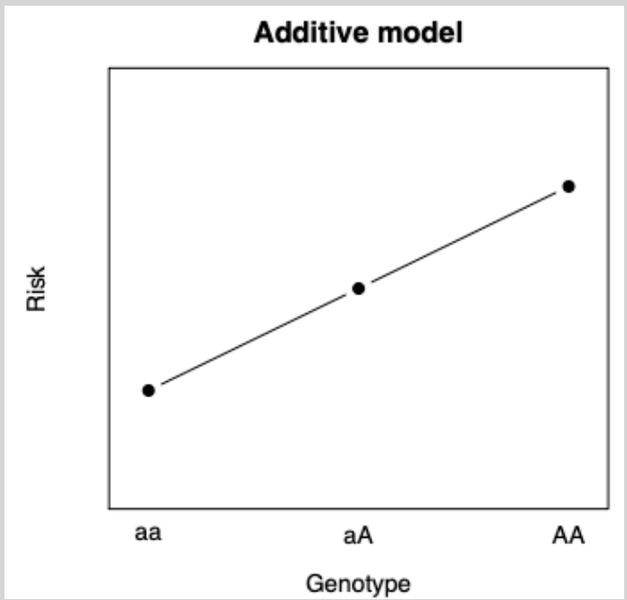
The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The original data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n



Genetic association studies

Genotype based tests - Additive test

- Basic idea: disease risk increases as a function of the number of alleles (0,1 or 2).
- There are two tests for the additive genetic model
  - The alleles test
  - **Cochran-Armitage trend test**

- The trend test is based on the linear regression model:

$$p(Y = 1 \mid X) = \beta_0 + \beta_1 X + \epsilon$$

- Where X is the number of alleles A (0, 1 or 2) and Y is disease status, so that $Y_i = 1$ describes the individual that has the trait.
- Several options to resolve the additive test based on regression:

  1) Logistic regression

  2) Armitage trend test

  3) Assign weights by hand

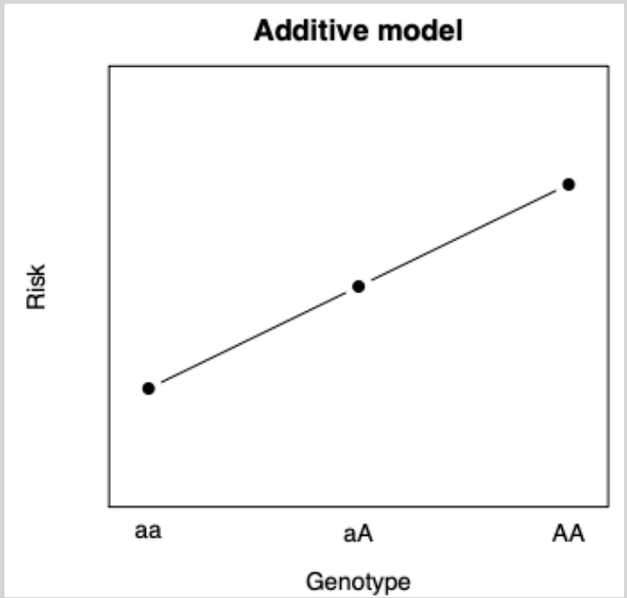
The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The original data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n



Genetic association studies

Genotype based tests - Logistic regression

- **Option 1:** Consider a general linear regression model:

$$p(Y = 1 \mid X) = \beta_0 + \beta_1 X + \epsilon$$

- Y is the disease status ($Y_i = 1$ individual i has the trait, $Y_i = 0$, individual i does not have the trait)
- X is the number of A alleles (0=BB, 1=AB, 2=AA).
- Hypothesis:
  - $H_0 : \beta_1 = 0$
  - $H_1 : \beta_1 \neq 0$

LOGISTIC REGRESSION

- Test statistic for a linear regression model:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$ where $\hat{\beta}_1$= estimated slope of the regression line and

$SE(\hat{\beta}_1)$= standard error of the slope estimate

- Under H_0, the test statistic t follows a T-distribution (df=n-2) from where we can obtain the p-value.

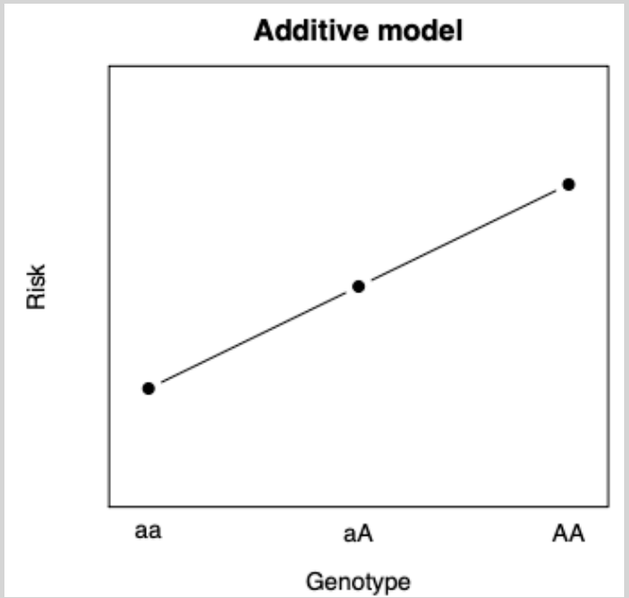
The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The original data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n



Genetic association studies

Genotype based tests - Logistic regression

- **Logistic regression:** (subset of GLM) linear regression adjusted for binary outcomes by using the logistic or logit function as a link function:

$$logit(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i = \mathbf{x_i^T}\beta + \epsilon_\mathbf{i}$$

- Logit (or logistic) function:

$$logit(p) = \frac{p}{1-p}$$

- Inverse of the logit function:

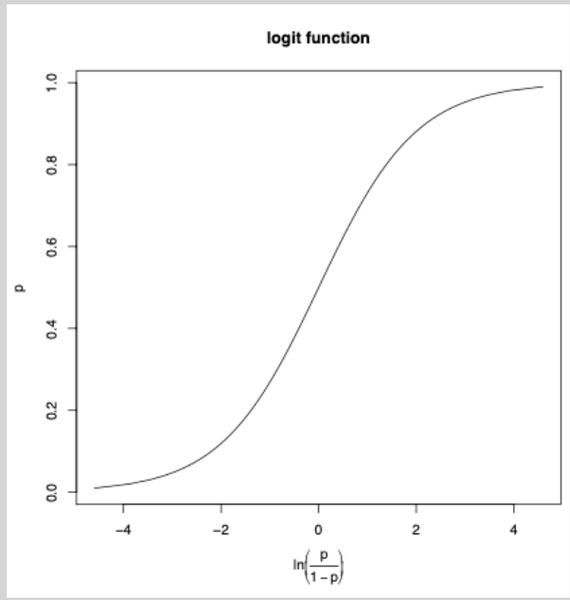
$$logit^{-1}(p) = \frac{e^p}{e^p + 1}$$

- And our linear model

$$Y = g(x) = \beta_0 + \beta_1 X + \epsilon$$

becomes:

$$g(x) = ln(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 X + \epsilon$$



Genetic association studies

Genotype based tests - OR in Logistic regression

- Take one genotype as the reference genotype (e.g. AA)
- Recall: odds of an event p:

$$Odds(p) = \frac{p}{1-p} \qquad \text{for example:} \quad Odds(disease) = \frac{P(disease)}{P(nodisease)}$$

- The odds ratio (OR) compares the odds of an event in two groups and provides a relative measure of effect in case-control studies. Quantifies the strength of an association between two events. Odds can range from 0 to infinity.
- The OR compares the odds of the disease for the two alleles:

$$OR_{BB} = \frac{\text{Odss disease for a BB person}}{\text{Odds disease AA person}}$$

$$OR_{AB} = \frac{\text{Odss disease for a AB person}}{\text{Odds disease AA person}}$$

- The OR can be estimated from the logistic regression:
- Given the model: $g(x) = ln(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_{AB} I_{AB} + \beta_{BB} I_{BB}$:
- $\hat{\beta}_{BB}$ is already the log odds ratio per unit increase, so $OR_{BB} = e^{\hat{\beta}_{BB}}$
- $\hat{\beta}_{AB}$ is already the log odds ratio per unit increase, so $OR_{AB} = e^{\hat{\beta}_{AB}}$

Genetic association studies

Genotype based tests - Logistic regression

Example - R code

```
> Cases    <- c(MM=149,Mm=269,mm=91)
> Controls <- c(MM=153,Mm=325,mm=180)

> cas <- rep(c("MM","Mm","mm"),Cases)
> con <- rep(c("MM","Mm","mm"),Controls)

> ncas <- length(cas)
> ncon <- length(con)

> y <- c(rep(1,ncas),rep(0,ncon))
> x <- factor(c(cas,con))

> out.lm <- glm(y~x, family = binomial(link = "logit"))
> summary(out.lm)

Call:
glm(formula = y ~ x, family = binomial(link = "logit"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6821     0.1286  -5.303 1.14e-07 ***
xMm           0.4930     0.1528   3.227 0.001251 **
xMM           0.6556     0.1726   3.798 0.000146 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1598.7  on 1166  degrees of freedom
Residual deviance: 1582.7  on 1164  degrees of freedom
AIC: 1588.7

Number of Fisher Scoring iterations: 4
```

```
> or <- exp(coefficients(out.lm))
> or
(Intercept)          xMm          xMM
  0.5055556   1.6371936   1.9263090
```

Genetic association studies

Genotype based tests - Additive test

- Basic idea: disease risk increases as a function of the number of alleles (0,1 or 2).
- There are two tests for the additive genetic model
  - The alleles test
  - **Cochran-Armitage trend test**

- The trend test is based on the linear regression model:

$$p(Y = 1 \mid X) = \beta_0 + \beta_1 X + \epsilon$$

- Where X is the number of alleles A (0, 1 or 2) and Y is disease status, so that $Y_i = 1$ describes the individual that has the trait.
- Several options to resolve the additive test based on regression:

  1) Logistic regression

  2) Armitage trend test

  3) Assign weights by hand

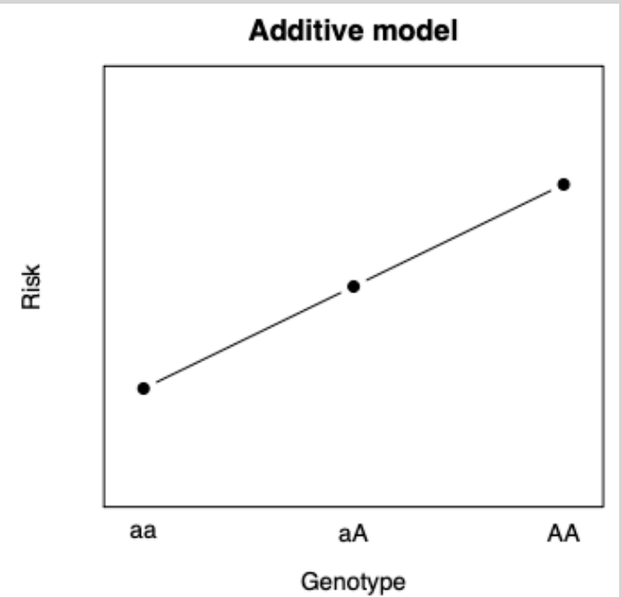
The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The original data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n



Genetic association studies

Genotype based tests - Armitage trend test

- **Option 2:** Consider a general linear regression model:

$$p(Y = 1 \mid X) = \beta_0 + \beta_1 X + \epsilon$$

- Y is the disease status ($Y_i = 1$ individual i has the trait, $Y_i = 0$, individual i does not have the trait)
- X is the number of A alleles (0=BB, 1=AB, 2=AA).
- Hypothesis:
  - $H_0 : \beta_1 = 0$
  - $H_1 : \beta_1 \neq 0$
- Test statistic for the Armitage trend test:

$$A = \frac{\hat{\beta}_1^2}{V(\hat{\beta}_1)}$$ where $\hat{\beta}_1$= estimated slope of the regression line and

$V(\hat{\beta}_1)$= variance of the slope estimate

- Under H_0, the test statistic A follows a χ_1^2 distribution (df=1) from where we can obtain the p-value.
- **Additionally:** $A = n \cdot r_{xy}^2$

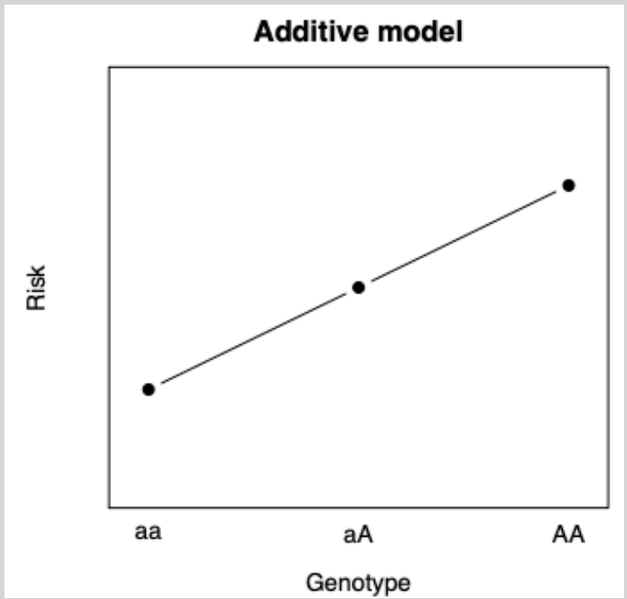
The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The original data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n



Genetic association studies

Genotype based tests - Armitage trend test

- Specialized test that assumes that the groups (alleles) are ordered, so that we're testing whether there is a linear increase/decrease across ordered categories (i.e. treatment dosage or genotype). Assumes there is an explanatory model. No covariates, large sample sizes.

- Assumptions:
  - Categories are ordered
  - Outcome is binary
  - Assesses if proportions increase or decrease linearly across the ordered categories.

- Attention: does not assume HWE hold in H_0

- Alternative computation of the Armitage trend test statistic (A):

$$A = \frac{\sum_i w_i(O_i - E_i)}{\sqrt{(\sum_i w_i V_i)}}$$

- Where
  - w_i: score assigned to the i-th group (0,1,2)
  - O_i: observed number of cases in the i-th group
  - E_i: expected number of cases in the i-th group
  - V_i: variance of the binary outcome in the i-th group

- Under the null hypothesis, A approximately follows a standard normal distribution $N(0,1)$, from where we obtain the p-values.

Genetic association studies

Genotype based tests - Armitage tre

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

	GG	GA	AA	Total
Cases	66	43	4	113
Controls	99	15	0	114
Total	165	58	4	227

Example - Option 2: Armitage trend test

- TNF genotype (G/A polymorphism) is on a study on acne patients and controls
- Test statistic: $A = n \cdot r_{xy}^2 = 227 \cdot (0.3257)^2 = 24.02$
- P-value $p - value = P(\chi^2 \geq 24.02) = 9.49e - 07$

```
Cases    <- c(66,43,4)
Controls <- c(99,15,0)

X <- rbind(Cases,Controls)
rownames(X) <- c("Cases","Controls")
colnames(X) <- c("GG","GA","AA")
n <- sum(X)

cas <- rep(c(0,1,2),Cases)
con <- rep(c(0,1,2),Controls)

y <- c(rep(1,sum(Cases)), rep(0,sum(Controls)))

x <- c(cas,con)


r <- cor(x,y)
A <- n*(r^2)

pvalue <- pchisq(A,df=1,lower.tail=FALSE)
```

SHORTCUT: The coefficient of determination is equivalent to the correlation coefficient between x, y when a single intercept is included and the sample standard deviations (SD) are the same.

$$\hat{\beta} = cor(x, y) \cdot \frac{SD(y)}{SD(x)}$$

Genetic association studies

Genotype based tests - Additive test

- Basic idea: disease risk increases as a function of the number of alleles (0,1 or 2).
- There are two tests for the additive genetic model
  - The alleles test
  - **Cochran-Armitage trend test**

- The trend test is based on the linear regression model:

$$p(Y = 1 \mid X) = \beta_0 + \beta_1 X + \epsilon$$

- Where X is the number of alleles A (0, 1 or 2) and Y is disease status, so that $Y_i = 1$ describes the individual that has the trait.
- Several options to resolve the additive test based on regression:

  1) Logistic regression

  2) Armitage trend test

  3) Assign weights by hand

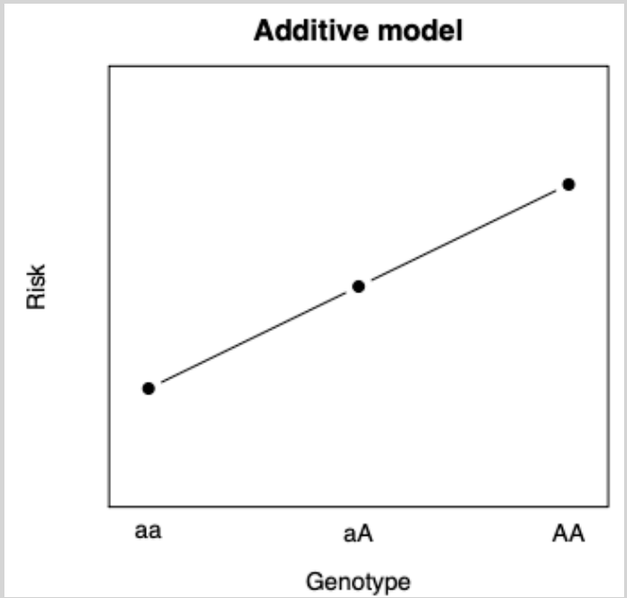
The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The original data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n



Genetic association studies

Genotype based tests - Weights by hand

- RECALL alternative Armitage test statistic: $A = \frac{\sum_i w_i(O_i - E_i)}{\sqrt{(\sum_i w_i V_i)}}$
- **Option 3:** Assign weights by hand
  - Test statistic $T = \sum_{i=1}^{k} t_i(r_i \cdot s - s_i \cdot r)$
  - Variance $Var(T) = \frac{r \cdot s}{n}(\sum_{i=1}^{k} t_i r_i(n - n_i) - 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} t_i t_j n_i n_j)$
  - If we're testing for linear trend on a bi-allelic marker $k = 3$ and $t = (0,1,2)$

- Under H_0, the test statistic T follows a χ_1^2 distribution (df=1), which corresponds to a standard normal distribution (zero mean and variance of 1) so that we can use the Z- score:
  - The Z- score $Z = \frac{T}{\sqrt{(var(T))}} \sim N(0,1)$

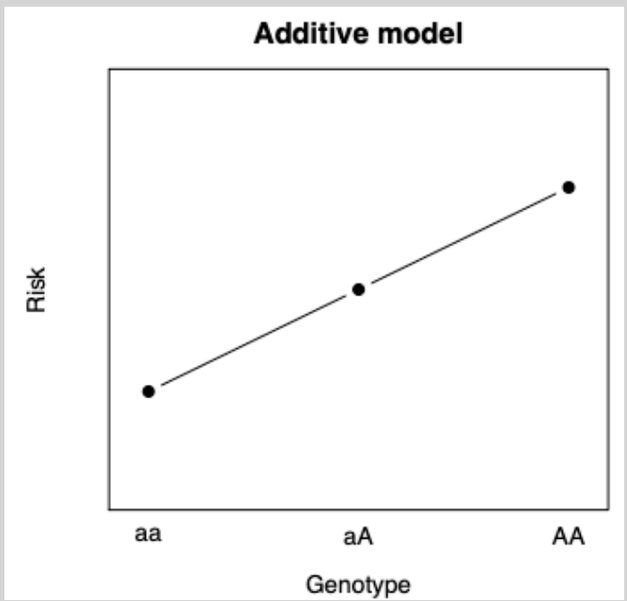
The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous) so that:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The original data table:

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n



Genetic association studies

Genotype based tests - Weights by h

	aa	aA	AA	Total
Cases	r_1	r_2	r_3	r
Controls	s_1	s_2	s_3	s
Total	n_1	n_2	n_3	n

	GG	GA	AA	Total
Cases	66	43	4	113
Controls	99	15	0	114
Total	165	58	4	227

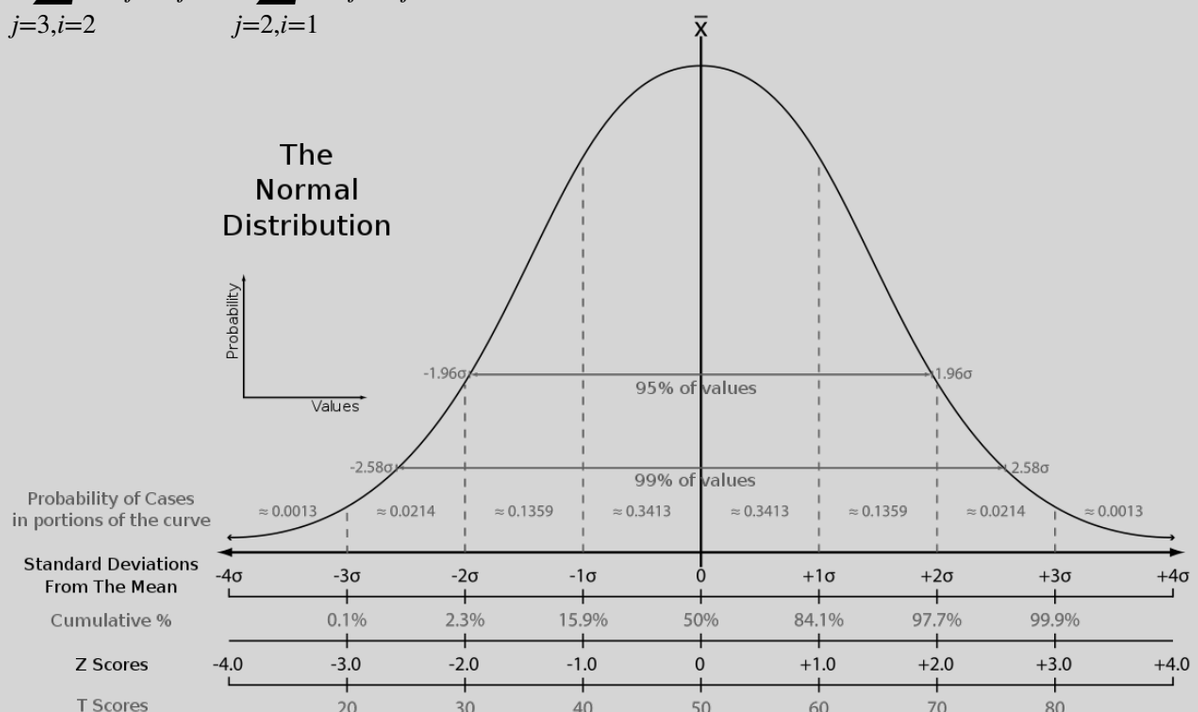
Example - Option 3

- TNF genotype (G/A polymorphism) is on a study on acne patients and controls
- As we're testing for linear trend on a bi-allelic marker $k = 3$ then $t = (0,1,2)$
- Test statistic and its variance:
  - $T = \sum_{i=1}^{k} t_i(r_i \cdot s - s_i \cdot r) = 0 \cdot (66 \cdot 114 - 99 \cdot 113) + 1 \cdot (43 \cdot 114 - 15 \cdot 113) + 2 \cdot (4 \cdot 114 - 0 \cdot 113) = 4119$
  - $Var(T) = \frac{r \cdot s}{n}(\sum_{i=1}^{k} t_i r_i(n - n_i) - 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} t_i t_j n_i n_j) = \ldots = 706081.3$
    - $\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} t_i t_j n_i n_j = \sum_{j=i+1,i=1}^{k} t_i t_j n_i n_j + \sum_{j=i+1,i=2}^{k} t_i t_j n_i n_j = \sum_{j=2,i=1}^{k} t_i t_j n_i n_j + \sum_{j=3,i=2}^{k} t_i t_j n_i n_j = \sum_{j=2,i=1}^{k} t_i t_j n_i n_j + t_2 t_3 n_2 n_3 = t_1 t_2 n_1 n_2 + t_1 t_3 n_1 n_3 + t_2 t_3 n_2 n_3 = t_2 t_3 n_2 n_3$
- The z-score $Z \sim N(0,1)$

$$Z = \frac{T}{\sqrt{(var(T))}} = \frac{4119}{\sqrt{(var(T))}} = 4.9019$$

- P-value $p - value = P(Z \geq 4.9019) = 9.49e - 07$

We reject the null hypothesis that the probability of disease is the same with different number of A



Step by step process:
https://www.youtube.com/watch?v=3UPqMfuboH4

Genetic association studies

Summary - population-based studies

Tests of association at the level of **alleles**

- We are sampling alleles
- Alleles assumed to be independent
- Rely on the Hardy-Weinberg equilibrium assumption
- Statistics on the alleles by trait cross table
    - Chi-square test
    - Fisher exact test
    - Odds ratio

Tests of association at the level of the **genotypes**

- We are sampling individuals
- Hardy-Weinberg equilibrium assumption is not needed
- Co-dominant, dominant and recessive Chi-square tests
- Cochran-Armitage trend test
- Logistic regression

Content

Genetic Association Studies

1. Introduction
2. Allele based tests
3. Genotype based tests
4. Quantitative traits and multiple polymorphisms
5. Computer exercise

Until now….

The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

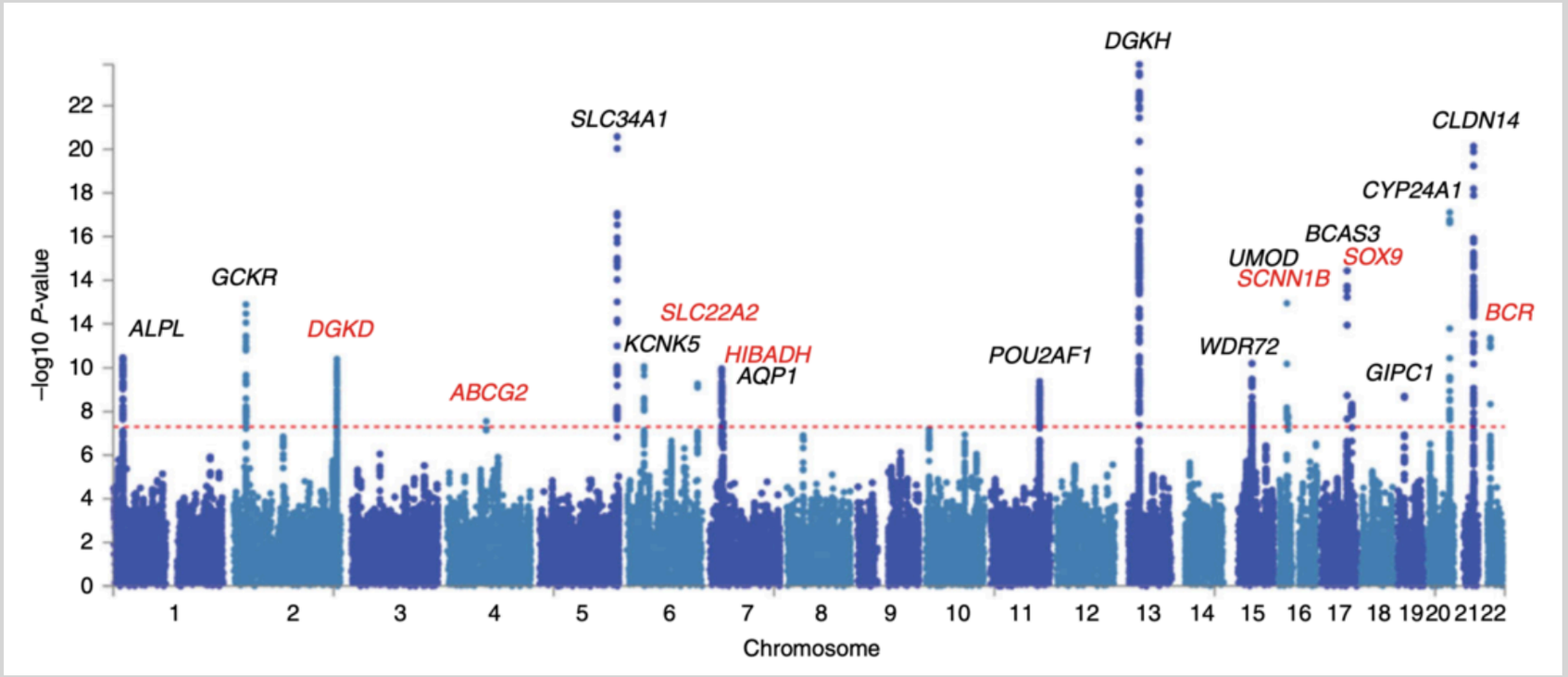
The phenotype is binary:

- $Y_i = 1$ individual i has the trait
- $Y_i = 0$, individual i does not have the trait.

Genetic association studies

RECALL - population-based studies

- We will focus on population-based association studies, executed on unrelated subjects. Two types:
  - **Allele-based tests and genotype-based tests**: are hypothesis driven, where a candidate locus is being tested.
  - **Genome-wide association studies (GWAS)**: involve the analysis of multiple polymorphisms conducted without prior hypothesis.



Typically, a single test statistic (for case–control studies, a chi-squared ($\chi 2$) comparison of absolute genotype counts) is calculated for each variant passing quality control.

Increasing number of studies are being extended from case-control studies to population-based cohorts.

Manhattan plot depicting several strongly associated risk loci. Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level. The peaks indicate genetic variants that are found more often in individuals with kidney stones.

When we are looking for regions of the genome or SNP that is causal for a gene, we often find that a whole bunch of SNPs are associated with the disease. Its not that they all cause disease, it is just that a whole bunch are correlated with the causal SNP (passenger mutations). Thus it is our job to identify the causal needle in the haystack.

Genetic association studies

Quantitative traits and multiple SNPs

The trait or disease of interest can be quantitative (e.g. height, lipid levels, blood pressure, etc…).

The trait or disease of interest can involve multiple SNPs.

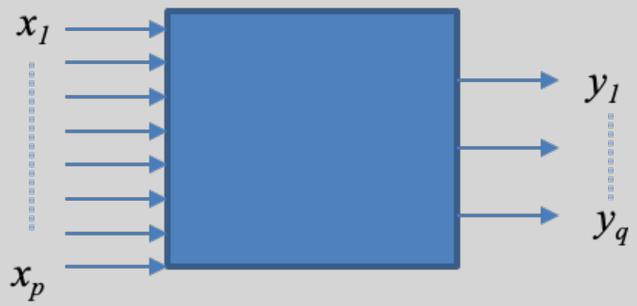
How to deal with quantitative outcomes and multiple SNPs?

- Multiple linear regression (small amounts of SNPs)
- Mixed effects models (to account for correlation between individuals)
- Test all variants: GWAS
- Regression with haplotypes
- …

Genetic association studies

Regression models

- **Multivariate regression analysis:** estimate relationship between dependent variables $y_1, \ldots, y_q$ (outcome, output, response or label) and one or more independent variables $x_1, \ldots, x_p$ (predictors, features, covariates or explanatory variables)…where q can be $q = 1$, depending on the problem.



- **Example:**
  - Objective: prediction of the country median household income given a set of socio-economic independent variables (or predictors)
  - Variables: Geographic border to metro county, educational attainment measures, interstate highway density, population density, labor force participation rate, …
  - Outcome: Country median household income

- **Multiple linear regression:** a response $y,$ is modeled as a linear combination of the predictors $x_1, \ldots, x_p$ plus a random fluctuation ϵ (residuals) so that:

$$y = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i = \mathbf{x_i^T}\beta + \epsilon_\mathbf{i}$$

NOTES:
- The standard linear model (LM) assumes a **fixed** relationship between the predictor(s) and the outcome variable
- β_i are assumed to be constant
- $\epsilon_i \sim N(0,\sigma)$: so that residuals are normally distributed with constant variance
- Observations are assumed to be independent of each other

Genetic association studies

Regression models

- **Multivariate linear regression:** each response $y_1, \ldots, y_q$, is modeled as a linear combination of the input $x_1, \ldots, x_p$ plus a random fluctuation ϵ so that:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i = \mathbf{x_i^T}\beta + \epsilon_\mathbf{i}$$

Or in matrix notation, considering n individuals:

$$\begin{bmatrix} y_{11} & y_{1k} & y_{1q} \\ \vdots & \vdots & \vdots \\ y_{n1} & y_{nk} & y_{nq} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{1k} & \beta_{1q} \\ \beta_{21} & \beta_{2k} & \beta_{2q} \\ \vdots & \vdots & \vdots \\ \beta_{p1} & \beta_{pk} & \beta_{pq} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \varepsilon_{1k} & \varepsilon_{1q} \\ \vdots & \vdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{nk} & \varepsilon_{nq} \end{bmatrix}$$

$\vec{y}_k \quad \vec{x}_1 \quad \vec{x}_p \quad \vec{\varepsilon}_k$

$$\vec{y}_k = X b_k + e_k \quad k = 1, \ldots, q$$

$$\underset{(n,q)}{Y} = \underset{(n,p)}{X}\underset{(p,q)}{\mathrm{B}} + \underset{(n,q)}{\mathrm{E}}$$

Y matrix of q response variables (centered)

X matrix of p predictors variables (centered)

B is the matrix of $p*q$ b_{jk} parameters

E random fluctuation matrix

Genetic association studies

Regression models

Linear mixed models: extension of linear models that allow for both fixed and random effects.

- **Fixed Effects:** Represent the population-level effects (e.g., overall mean, slope).
- **Random Effects:** Represent the group-specific deviations from the fixed effects. These allow intercepts (or slopes) to vary across groups

- Used to analyze hierarchical data, longitudinal or correlated.
  - Mixed-effects models can account for nested data (hierarchical structure within predictors accounted through random effects)
  - Observations within each group are assumed to be correlated

- Each response $y_1, \ldots, y_q$ is modeled as a linear combination of the fixed effects $x_1, \ldots, x_p$, plus plus a linear combination of random effects $z_1, \ldots, z_p$ and a random fluctuation ϵ so that:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + u_0 + u_1 z_{i1} + \ldots + u_p z_{ip} + \epsilon_i = \mathbf{x_i^T}\beta + \mathbf{z_i^T}\mathbf{u} + \epsilon_\mathbf{i}$$

  - β_i represent the fixed effects and assumed to be constant (population-level parameters)
  - u_i represent the random effects for group u_i (group-specific intercept).
  - $u_i \sim N(0,\sigma_u)$: so that the coefficients are normally distributed with constant variance
  - $\epsilon_i \sim N(0,\sigma)$: so that residuals are normally distributed with constant variance
  - Observations are assumed to be independent of each other

Genetic association studies

Regression models

- Multiple linear regression helps understanding genetic mechanisms that contribute to a trait, providing an effect size for each SNP.
- Only works for small amount of SNPs, as we need more participants (data points) than SNPs (parameters).
- Marginal SNP effects can be estimated one SNP at a time but….multiple regression allows for multiple estimations.
- Additional covariate terms may be included (mixed effects models)

Genetic association studies

Multiple polymorphisms - Example study

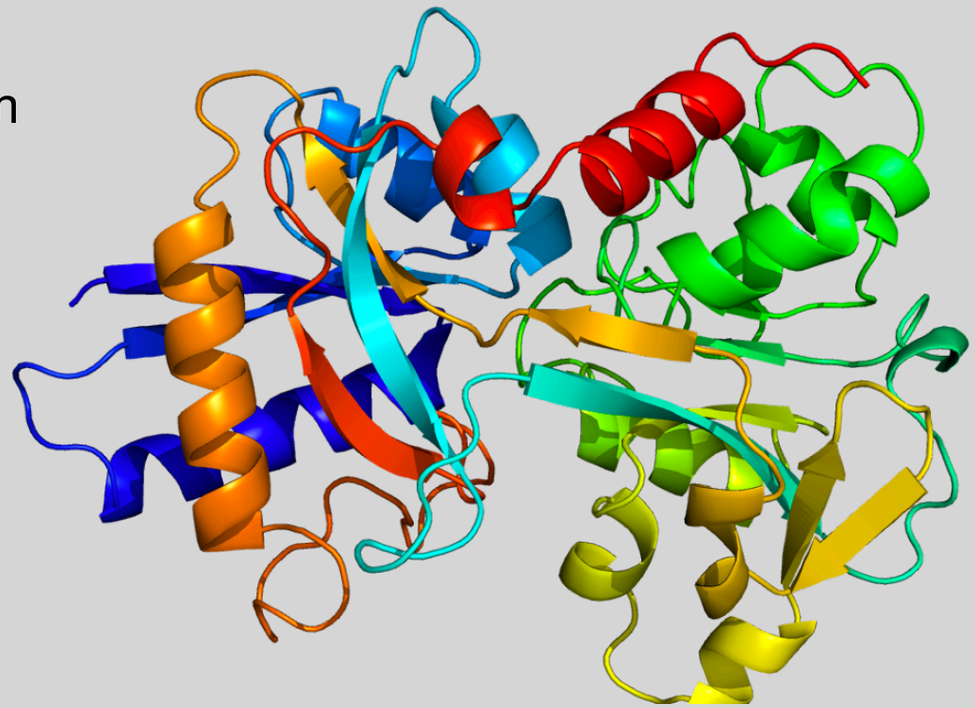
GWAS for transferrin

- 2,362 individuals for which (adjusted) transferrin serum levels are available.
- 281,313 SNPs from all chromosomes.

Data pre-processing

- Filters: missing rate < 0.01;
- MAF > 0.05;
- HWE p-value > 0.001.
- We use an **additive model** for each SNP and fit this model with PLINK.

Fit a linear model for **each** SNP and estimate whether the slope of the linear model is significant. Obtain p-value for the test $H_1 : \beta_1 \neq 0$ and apply Bonferroni correction



Transferrins are a family of proteins that mediate iron transport.

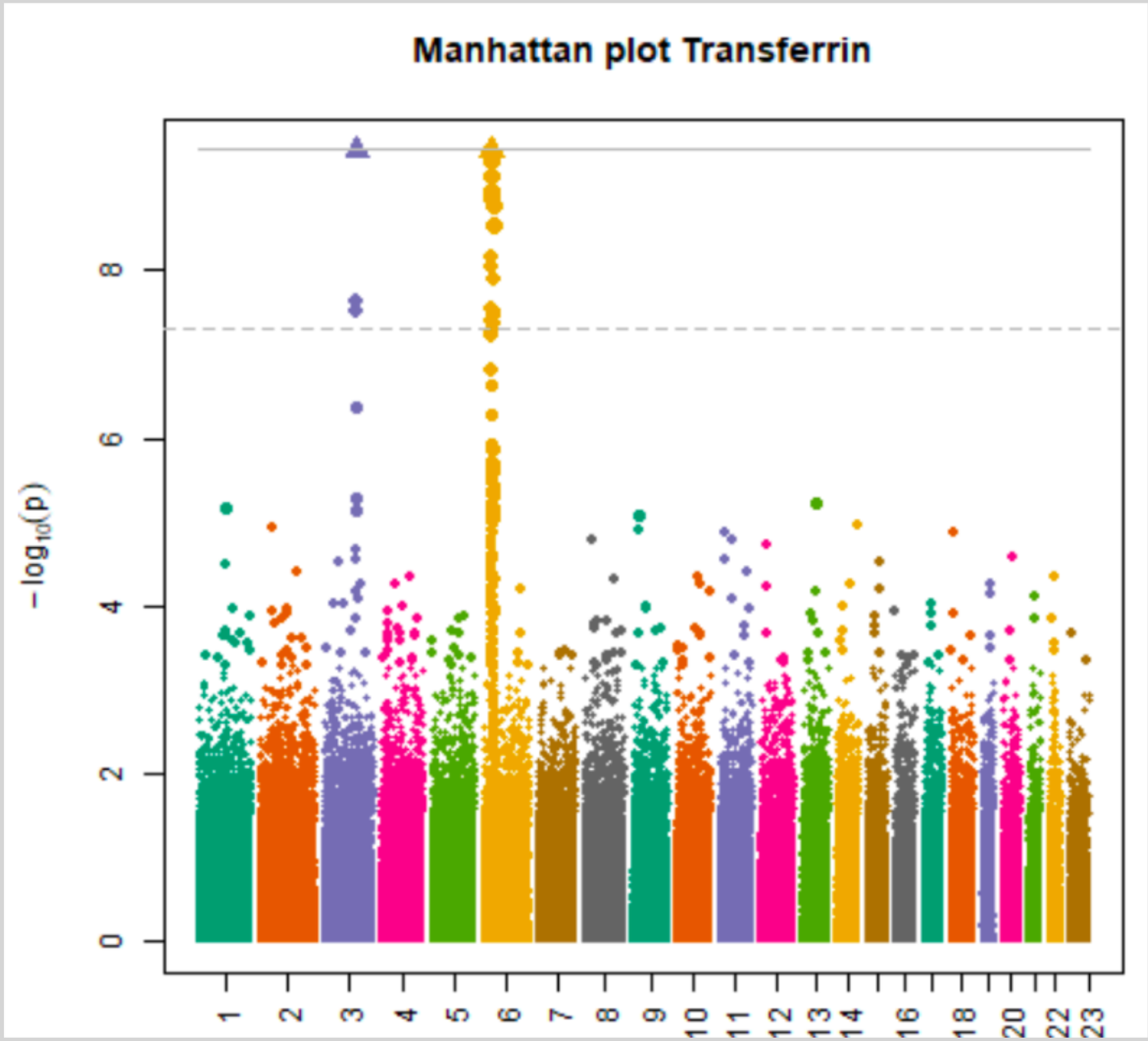
Gene coding for transferrin in humans is located in the chromosome band 3q21.

- The first point of reference is a number (or letter) which denotes the chromosome (e.g. 3q21 refers to chromosome 3)
- The second point of reference is a letter (p or q) to denote which arm the locus is positioned on (e.g. 3q21 is on the q arm)
- The third point of reference is a number corresponding to the G band location (e.g. 3q21 is at the longitudinal position 21)

Genetic association studies

Multiple polymorphisms - Example study

GWAS for transferrin



Genetic association studies

Multiple polymorphisms - Example study

GWAS for transferrin - top 25

	SNP	CHR	BP	B	SE	R2	T	P	-log10(P)
1	rs3811647	3	134966719	0.3832	0.02889	0.06936	13.260	8.965e-39	38.047450
2	rs6794945	3	135001153	0.3652	0.02940	0.06136	12.420	2.324e-34	33.633764
3	rs1800562	6	26201120	-0.5884	0.04988	0.05572	-11.800	2.968e-31	30.527536
4	rs13214703	6	28049366	-0.4378	0.04886	0.03292	-8.961	6.390e-19	18.194499
5	rs1358024	3	134966878	0.3290	0.03745	0.03168	8.785	2.941e-18	17.531505
6	rs2274089	6	25596562	-0.3791	0.04551	0.02856	-8.330	1.352e-16	15.869023
7	rs4525863	3	134918826	0.2399	0.03017	0.02609	7.951	2.845e-15	14.545918
8	rs1867503	3	134893338	0.2039	0.02864	0.02103	7.120	1.428e-12	11.845272
9	rs1867504	3	134893351	0.2039	0.02864	0.02103	7.120	1.428e-12	11.845272
10	rs9853615	3	135002671	-0.2083	0.02929	0.02098	-7.111	1.523e-12	11.817300
11	rs12216125	6	26105437	-0.1974	0.02891	0.01936	-6.826	1.107e-11	10.955852
12	rs9379818	6	26131185	-0.1931	0.02838	0.01925	-6.805	1.276e-11	10.894149
13	rs13194984	6	26608542	-0.2719	0.04060	0.01865	-6.698	2.638e-11	10.578725
14	rs932316	6	25749179	-0.2371	0.03557	0.01849	-6.664	3.309e-11	10.480303
15	rs17270561	6	25928418	-0.2183	0.03292	0.01829	-6.631	4.108e-11	10.386370
16	rs2013063	6	26102077	-0.1798	0.02823	0.01690	-6.369	2.285e-10	9.641114
17	rs1543680	6	26211156	-0.2036	0.03259	0.01627	-6.247	4.944e-10	9.305922
18	rs10484432	6	26116855	-0.2013	0.03256	0.01595	-6.183	7.390e-10	9.131356
19	rs2009610	6	26075047	-0.1959	0.03205	0.01558	-6.111	1.158e-09	8.936291
20	rs707889	6	26203910	-0.1969	0.03238	0.01543	-6.082	1.383e-09	8.859178
21	rs1029328	6	28555894	-0.2509	0.04150	0.01526	-6.047	1.709e-09	8.767258
22	rs11757000	6	28592848	-0.2307	0.03868	0.01486	-5.966	2.806e-09	8.551912
23	rs169219	6	26065371	0.1669	0.02870	0.01413	5.816	6.861e-09	8.163613
24	rs7748771	6	25463078	-0.2678	0.04645	0.01389	-5.765	9.249e-09	8.033905
25	rs3130253	6	29741991	-0.2769	0.04845	0.01365	-5.715	1.238e-08	7.907279

(top 25) Significant polymorphisms in the TF gene on CHR 3

Statistical concerns:

- Effect of filters applied?
- Multiple testing problem?
- X-chromosome adequately dealt with?
- Family structure accounted for?
- Adjustment for covariates?
- Power?

Further information: GWAS in PLINK
https://www.youtube.com/watch?v=7QMSZx3io-Q

Content

Genetic Association Studies

1. Introduction
2. Allele based tests
3. Genotype based tests
4. Quantitative traits and multiple polymorphisms
5. Computer exercise

Genetic association studies

References

- https://www.mv.helsinki.fi/home/mjxpirin/GWAS_course/

- Uffelmann, E., Huang, Q.Q., Munung, N.S., De Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T. and Posthuma, D., 2021. Genome-wide association studies. Nature Reviews Methods Primers, 1(1), p.59.

- Zondervan, K.T. and Cardon, L.R., 2007. Designing candidate gene and genome-wide case–control association studies. Nature protocols, 2(10), pp.2492-2501.

- McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P. and Hirschhorn, J.N., 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews genetics, 9(5), pp.356-369.

- Laird, N.M. & Lange, C. (2011) The fundamentals of modern statistical genetics. Springer.