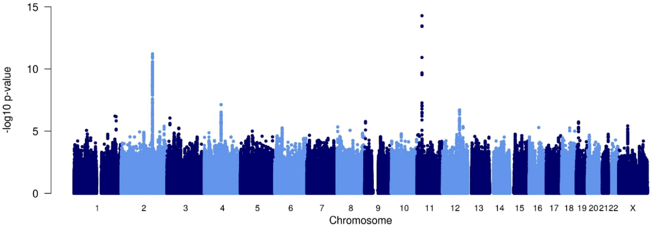
Association testing and GWAS



Line Skotte, Medical and Population Genetics Course, August 2018

Outline

What and why?

- **Goal: to identify (map) genetic variants that have an effect on a trait**
- Typically **disease related traits**, e.g. febrile seizures

What and why?

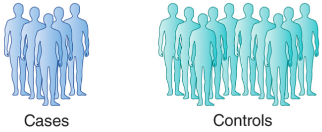
- **Goal: to identify (map) genetic variants that have an effect on a trait**
- Typically **disease related traits**, e.g. febrile seizures
- Motivation: reaching this goal can help
  - reveal the underlying genetic architecture
  - hopefully lead to better understanding of what **causes** the disease
  - in turn ideally lead to better treatment and/or prevention

What and why?

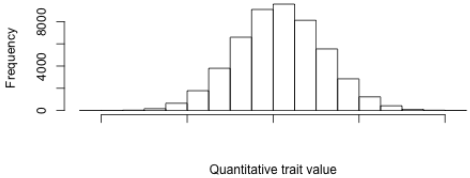
- **Goal: to identify (map) genetic variants that have an effect on a trait**
- Typically **disease related traits**, e.g. febrile seizures
- Motivation: reaching this goal can help
  - reveal the underlying genetic architecture
  - hopefully lead to better understanding of what **causes** the disease
  - in turn ideally lead to better treatment and/or prevention
- Note, **can also be used in e.g. evolutionary studies!**

Plan for today (to teach you how)

- **This afternoon:**
  - How to test if a genetic variant potentially affects a trait (single SNP tests)
  - How to search the genome for variants that affect a given trait (GWAS)
  - We will assume we have genotyping data (e.g. from SNP chip)
  - We will assume there is no population structure
  - We will look at disease status traits:



  - And quantitative traits:



Outline

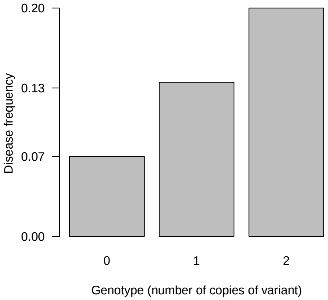
Overall idea in association testing

How do we test if a genetic variant potentially has an effect on a disease?

Overall idea in association testing

How do we test if a genetic variant potentially has an effect on a disease?

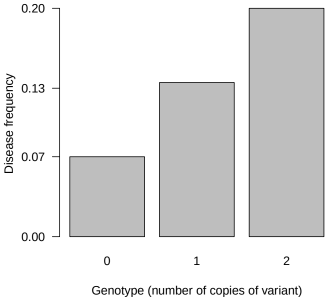
- Idea: test for **association** between the variant and disease status (case/control)



Overall idea in association testing

How do we test if a genetic variant potentially has an effect on a disease?

- Idea: test for **association** between the variant and disease status (case/control)

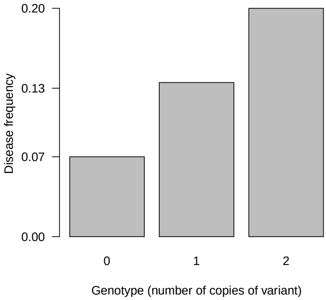


- Rationale: this is what we expect if the variant affects the trait

Overall idea in association testing

How do we test if a genetic variant potentially has an effect on a disease?

- Idea: test for **association** between the variant and disease status (case/control)



- Rationale: this is what we expect if the variant affects the trait
- Approach: test null hypothesis, H_0, of no association (independence)

χ^2 test for independence

- A test which you can apply to counts tables for two categorical variables
  E.g. disease status and genotypes:

	AA	Aa	aa	Total
Case	441	418	141	1000
Control	749	611	140	1500
Total	1190	1029	281	2500

χ^2 test for independence

- A test which you can apply to counts tables for two categorical variables
  E.g. disease status and genotypes:

	AA	Aa	aa	Total
Case	441	418	141	1000
Control	749	611	140	1500
Total	1190	1029	281	2500

- The null hypothesis, H_0, of the test, is **no association** (independence)

χ^2 test for independence

- A test which you can apply to counts tables for two categorical variables
  E.g. disease status and genotypes:

	AA	Aa	aa	Total
Case	441	418	141	1000
Control	749	611	140	1500
Total	1190	1029	281	2500

- The null hypothesis, H_0, of the test, is **no association** (independence)
- Has the test statistic $X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$
  (measures how far your observed data is from what you expect if H_0 is true)

χ^2 test for independence

- A test which you can apply to counts tables for two categorical variables
  E.g. disease status and genotypes:

	AA	Aa	aa	Total
Case	441	418	141	1000
Control	749	611	140	1500
Total	1190	1029	281	2500

- The null hypothesis, H_0, of the test, is **no association** (independence)
- Has the test statistic $X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$
  (measures how far your observed data is from what you expect if H_0 is true)
- If H_0 is true then $X^2 \sim \chi^2$
  (This means we can use χ^2 to translate X^2 to p-value
  i.e. the probability of seeing $\geq X^2$ if H_0 is true)

χ^2 test for independence

- A test which you can apply to counts tables for two categorical variables
  E.g. disease status and genotypes:

	AA	Aa	aa	Total
Case	441	418	141	1000
Control	749	611	140	1500
Total	1190	1029	281	2500

- The null hypothesis, H_0, of the test, is **no association** (independence)
- Has the test statistic $X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$
  (measures how far your observed data is from what you expect if H_0 is true)
- If H_0 is true then $X^2 \sim \chi^2$
  (This means we can use χ^2 to translate X^2 to p-value
  i.e. the probability of seeing $\geq X^2$ if H_0 is true)
- We use this to decide whether we reject the null hypothesis
  (we reject when p is small and see it as evidence for association)

χ^2 tests - test with genotype counts

- Can be applied to genotype count tables
- So assume we have **observed** this data:

	AA	Aa	aa	Total
Case	O_1=441	O_3=418	O_5=141	1000
Control	O_2=749	O_4=611	O_6=140	1500
Total	1190	1029	281	2500

χ^2 tests - test with genotype counts

- Can be applied to genotype count tables
- So assume we have **observed** this data:

	AA	Aa	aa	Total
Case	O_1=441	O_3=418	O_5=141	1000
Control	O_2=749	O_4=611	O_6=140	1500
Total	1190	1029	281	2500

- **Expected under H_0:** if there is no association between the SNP and disease we would expect proportions of cases within the genotype categories to be the same (here 1000/2500=0.4, i.e. 40%).

χ^2 tests - test with genotype counts

- Can be applied to genotype count tables
- So assume we have **observed** this data:

	AA	Aa	aa	Total
Case	O_1=441	O_3=418	O_5=141	1000
Control	O_2=749	O_4=611	O_6=140	1500
Total	1190	1029	281	2500

- **Expected under H_0:** if there is no association between the SNP and disease we would expect proportions of cases within the genotype categories to be the same (here 1000/2500=0.4, i.e. 40%).
- So e.g. we would expect 40% of those with genotype AA to be cases and the rest to be controls. Thus E_1=0.4×1190=476 and E_2=1190-476=714

χ^2 tests - test with genotype counts

- Can be applied to genotype count tables
- So assume we have **observed** this data:

	AA	Aa	aa	Total
Case	O_1=441	O_3=418	O_5=141	1000
Control	O_2=749	O_4=611	O_6=140	1500
Total	1190	1029	281	2500

- **Expected under H_0:** if there is no association between the SNP and disease we would expect proportions of cases within the genotype categories to be the same (here 1000/2500=0.4, i.e. 40%).
- So e.g. we would expect 40% of those with genotype AA to be cases and the rest to be controls. Thus E_1=0.4×1190=476 and E_2=1190-476=714
- **Small exercise:** what would E_3 and E_4 be?

χ^2 tests - test with genotype counts

- So we have:

Observed	AA	Aa	aa	Total
Case	O_1=441	O_3=418	O_5=141	1000
Control	O_2=749	O_4=611	O_6=140	1500
Total	1190	1029	281	2500

Expected	AA	Aa	aa	Total
Case	E_1=476	E_3=411.6	E_5=112.4	1000
Control	E_2=714	E_4=617.4	E_6=168.6	1500
Total	1190	1029	281	2500

χ^2 tests - test with genotype counts

- So we have:

Observed	AA	Aa	aa	Total
Case	O_1=441	O_3=418	O_5=141	1000
Control	O_2=749	O_4=611	O_6=140	1500
Total	1190	1029	281	2500

Expected	AA	Aa	aa	Total
Case	E_1=476	E_3=411.6	E_5=112.4	1000
Control	E_2=714	E_4=617.4	E_6=168.6	1500
Total	1190	1029	281	2500

- $X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(441-476)^2}{476} + \frac{(749-714)^2}{714} + \ldots + \frac{(140-168.6)^2}{168.6} = 16.5838$

χ^2 tests - test with genotype counts

- So we have:

Observed	AA	Aa	aa	Total
Case	O_1=441	O_3=418	O_5=141	1000
Control	O_2=749	O_4=611	O_6=140	1500
Total	1190	1029	281	2500

Expected	AA	Aa	aa	Total
Case	E_1=476	E_3=411.6	E_5=112.4	1000
Control	E_2=714	E_4=617.4	E_6=168.6	1500
Total	1190	1029	281	2500

- $X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(441-476)^2}{476} + \frac{(749-714)^2}{714} + \ldots + \frac{(140-168.6)^2}{168.6} = 16.5838$
- Using the χ^2-distribution with 2 df we get a p-value for X^2 (p$\simeq$0.00025)

χ^2 tests - test with genotype counts

- So we have:

Observed	AA	Aa	aa	Total
Case	O_1=441	O_3=418	O_5=141	1000
Control	O_2=749	O_4=611	O_6=140	1500
Total	1190	1029	281	2500

Expected	AA	Aa	aa	Total
Case	E_1=476	E_3=411.6	E_5=112.4	1000
Control	E_2=714	E_4=617.4	E_6=168.6	1500
Total	1190	1029	281	2500

- $X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(441-476)^2}{476} + \frac{(749-714)^2}{714} + \ldots + \frac{(140-168.6)^2}{168.6} = 16.5838$
- Using the χ^2-distribution with 2 df we get a p-value for X^2 (p$\simeq$0.00025)
- Tells us that the probability of getting a X^2 value 16.5838 or higher **if** there is no association is low (p$\simeq$0.00025$<$0.05)

χ^2 tests - test with genotype counts

- So we have:

Observed	AA	Aa	aa	Total
Case	O_1=441	O_3=418	O_5=141	1000
Control	O_2=749	O_4=611	O_6=140	1500
Total	1190	1029	281	2500

Expected	AA	Aa	aa	Total
Case	E_1=476	E_3=411.6	E_5=112.4	1000
Control	E_2=714	E_4=617.4	E_6=168.6	1500
Total	1190	1029	281	2500

- $X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(441-476)^2}{476} + \frac{(749-714)^2}{714} + \ldots + \frac{(140-168.6)^2}{168.6} = 16.5838$
- Using the χ^2-distribution with 2 df we get a p-value for X^2 ($p \simeq 0.00025$)
- Tells us that the probability of getting a X^2 value 16.5838 or higher **if** there is no association is low ($p \simeq 0.00025 < 0.05$)
- We therefore reject the null hypothesis of no association and conclude that the variant is associated with the disease status

χ^2 tests - specific inheritance models

- In a similar way we can test for association **assuming specific inheritance models** by rewriting the table accordingly and doing a χ^2 test

χ^2 tests - specific inheritance models

- In a similar way we can test for association **assuming specific inheritance models** by rewriting the table accordingly and doing a χ^2 test
- E.g. assuming a **recessive model** we can rewrite the genotype counts table:

Our genotype counts	AA	Aa	aa	Total
Case	441	418	141	1000
Control	749	611	140	1500
Total	1190	1029	281	2500

to

Counts of homozygous carriers vs others	AA or Aa	aa	Total
Case	441+418=859	141	1000
Control	749+611=1360	140	1500
Total	2219	281	2500

χ^2 tests - specific inheritance models

- In a similar way we can test for association **assuming specific inheritance models** by rewriting the table accordingly and doing a χ^2 test
- E.g. assuming a **recessive model** we can rewrite the genotype counts table:

Our genotype counts	AA	Aa	aa	Total
Case	441	418	141	1000
Control	749	611	140	1500
Total	1190	1029	281	2500

to

Counts of homozygous carriers vs others	AA or Aa	aa	Total
Case	441+418=859	141	1000
Control	749+611=1360	140	1500
Total	2219	281	2500

- Then the same as before: we use a χ^2 test for association w. df=1

χ^2 tests - specific inheritance models

- In a similar way we can test for association **assuming specific inheritance models** by rewriting the table accordingly and doing a χ^2 test
- E.g. assuming a **recessive model** we can rewrite the genotype counts table:

Our genotype counts	AA	Aa	aa	Total
Case	441	418	141	1000
Control	749	611	140	1500
Total	1190	1029	281	2500

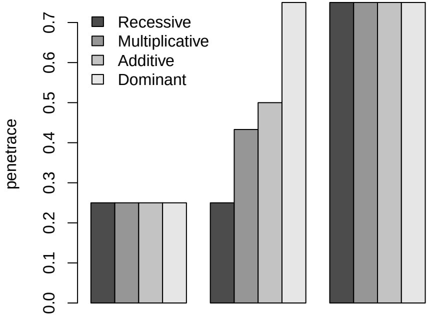
to

Counts of homozygous carriers vs others	AA or Aa	aa	Total
Case	441+418=859	141	1000
Control	749+611=1360	140	1500
Total	2219	281	2500

- Then the same as before: we use a χ^2 test for association w. df=1
- How would you test assuming a dominant model?

Other inheritance models

- Commonly considered genetic inheritance models:



- Testing under an additive genetic inheritance models is more tricky can be done using e.g. an Armitage trend test
- Testing under a multiplicative model can be done using logistic regression

Logistic regression

- Based on the following general model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1^i + \ldots + \beta_n x_n^i$$

where the βs are *regression coefficients* (effect sizes).

- The x^i's are determined by the genotype of individual i and the inheritance model

Logistic regression

- Based on the following general model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1^i + \ldots + \beta_n x_n^i$$

where the βs are *regression coefficients* (effect sizes).

- The x^is are determined by the genotype of individual i and the inheritance model
- E.g. for a simple multiplicative inheritance model we have

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1^i$$

where x_1^i is the number number of copies of the variant so 0, 1 or 2

Logistic regression

- Based on the following general model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1^i + \ldots + \beta_n x_n^i$$

where the βs are *regression coefficients* (effect sizes).

- The x^is are determined by the genotype of individual i and the inheritance model
- E.g. for a simple multiplicative inheritance model we have

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1^i$$

where x_1^i is the number number of copies of the variant so 0, 1 or 2

- Test if β_1 is zero (no association between the variant and the trait)

Why is logistic regression a good framework to use?

Logistic regression is very convenient due to its flexibility:

- Most inheritance models can be tested (by recoding x^i):

Genotypes	multiplicative	dominant	recessive	genotypes	
AA	0	0	0	0	0
Aa	1	1	0	1	0
aa	2	1	1	1	1

Why is logistic regression a good framework to use?

Logistic regression is very convenient due to its flexibility:

- Most inheritance models can be tested (by recoding x^i):

Genotypes	multiplicative	dominant	recessive	genotypes	
AA	0	0	0	0	0
Aa	1	1	0	1	0
aa	2	1	1	1	1

- Can incorporate other factors in the model
  - discrete factors such as gender
  - continuous factors such as age

  Can be used to correct for possible confounding factors
  Can be used for metaanalysis by incl a factor for the different studies

Exercise

Let's try to perform some of these tests in R:

Solve exercise 1A, 1B, 1C and 1D (+ 1E if you have time)

Causality?

Causality?

- No, not necessarily!

Causality?

- No, not necessarily!
- We expect to see some loci highly correlated w. causal variant, e.g:

Causal	Other locus
A	G
A	G
A	G
A	G
A	G
C	T
C	T
C	T

Causality?

- No, not necessarily!
- We expect to see some loci highly correlated w. causal variant, e.g:

Causal	Other locus
A	G
A	G
A	G
A	G
A	G
C	T
C	T
C	T

- This means that we see association in loci that are in high LD with the causal SNP
  So you have to be careful what you conclude from an association signal!

Other important limitations

One also has to be aware of the underlying assumptions:

- In all the tests there is an assumption that the individuals are independent (unrelated) and from a homogenous (unstructured) population

Other important limitations

One also has to be aware of the underlying assumptions:

- In all the tests there is an assumption that the individuals are independent (unrelated) and from a homogenous (unstructured) population
- If these assumptions are violated you risk getting false positives!

Other important limitations

One also has to be aware of the underlying assumptions:

- In all the tests there is an assumption that the individuals are independent (unrelated) and from a homogenous (unstructured) population
- If these assumptions are violated you risk getting false positives!
- Hence Quality Control (QC) and appropriate modelling is crucial!

Effect sizes for case-control data - relative risk

Relative risk - definition

$$RR = \frac{P(Case|Exposed)}{P(Case|Not\ exposed)}$$

where exposed depends on model, e.g. exposed=aa under recessive model

I.e. how many times higher the *risk* of disease is for exposed

Relative risk - example with recessive model

	Cases	Controls	Total
Exposed (g=aa)	100	100	200
Not exposed (g=AA or Aa)	400	3600	4000

- $P(Case|Exposed) = \frac{100}{200} = \frac{1}{2}$
- $P(Case|Not\ exposed) = \frac{400}{4000} = \frac{1}{10}$
- $RR = \frac{1/2}{1/10} = 5$

Effect sizes for case-control data - odds ratio

Odds ratio - definition

$$OR = \frac{ODD_{Exposed}}{ODD_{Not\ \ Exposed}} = \frac{\frac{P(Case|Exposed)}{P(Control|Exposed)}}{\frac{P(Case|Not\ \ exposed)}{P(Control|Not\ \ exposed)}}$$

where exposed depends on model, e.g. exposed=aa under recessive model

I.e. how many times higher the *odds* of disease is for exposed

Odds ratio - example with recessive model

	Cases	Controls	Total
Exposed (g=aa)	100	100	200
Not exposed (g=AA or Aa)	400	3600	4000

- $\frac{P(Case|Exposed)}{P(Control|Exposed)} = \frac{100/200}{100/200} = \frac{100}{100} = 1$
- $\frac{P(Case|Not\ \ exposed)}{P(Control|Not\ \ exposed)} = \frac{400/4000}{3600/4000} = \frac{400}{3600} = 1/9$
- $OR = \frac{1}{1/9} = 9$ (very high for an association study!)

Effect size estimates from logistic regression

- In logistic regression the ORs are estimated directly:
  In the model we estimate the effect size β_1

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1^i \ldots$$

Effect size estimates from logistic regression

- In logistic regression the ORs are estimated directly:
  In the model we estimate the effect size β_1

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1^i \ldots$$

- Example: recessive model

$$\frac{\mathrm{ODD}_{aa}}{\mathrm{ODD}_{aA/AA}} = \frac{\frac{p_{aa}}{1-p_{aa}}}{\frac{p_{aA/AA}}{1-p_{aA/AA}}} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

Effect size estimates from logistic regression

- In logistic regression the ORs are estimated directly:
  In the model we estimate the effect size β_1

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1^i \ldots$$

- Example: recessive model

$$\frac{\mathrm{ODD}_{aa}}{\mathrm{ODD}_{aA/AA}} = \frac{\frac{p_{aa}}{1-p_{aa}}}{\frac{p_{aA/AA}}{1-p_{aA/AA}}} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

- So we can get OR by taking the exp() of β_1

Effect size estimates from logistic regression

- In logistic regression the ORs are estimated directly:
  In the model we estimate the effect size β_1

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1^i \dots$$

- Example: recessive model

$$\frac{\mathrm{ODD}_{aa}}{\mathrm{ODD}_{aA/AA}} = \frac{\frac{p_{aa}}{1-p_{aa}}}{\frac{p_{aA/AA}}{1-p_{aA/AA}}} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

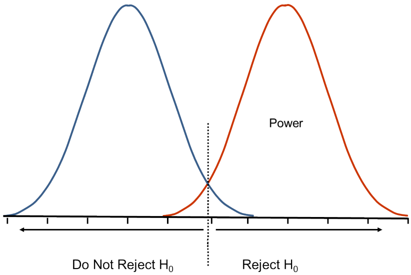
- So we can get OR by taking the exp() of β_1
- If time allows do exercise 1F

Design

- Will your study answer your research question? **Key: power**

Design

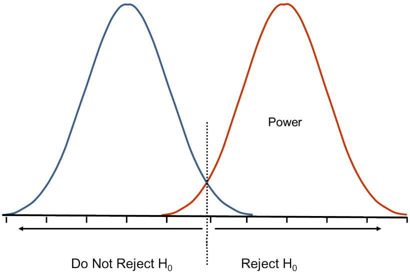
- Will your study answer your research question? **Key: power**
- Power is the probability that a true association is found when testing



Crucial for whether the study is worth performing!

Design

- Will your study answer your research question? **Key: power**
- Power is the probability that a true association is found when testing



Crucial for whether the study is worth performing!

- Before you start your study: calculate power for your study and assess it
  Rule of thumb: power should be at least 0.8

Power and power calculations

- Power depends on
  - the inheritance mode, e.g. recessive effect
  - the effect size, e.g. OR of 1.3 (the bigger the higher power)
  - the frequency of allele, e.g. 0.04 (the bigger the higher power)
  - **the rejection criterion**, e.g. $p < 0.05$ (the bigger the higher power)
  - **the number of samples** (the bigger the higher power)
  - **the test you use**

Power and power calculations

- Power depends on
  - the inheritance mode, e.g. recessive effect
  - the effect size, e.g. OR of 1.3 (the bigger the higher power)
  - the frequency of allele, e.g. 0.04 (the bigger the higher power)
  - **the rejection criterion**, e.g. $p < 0.05$ (the bigger the higher power)
  - **the number of samples** (the bigger the higher power)
  - **the test you use**
- Can often be calculated using "power-calculators"

Power and power calculations

- Power depends on
  - the inheritance mode, e.g. recessive effect
  - the effect size, e.g. OR of 1.3 (the bigger the higher power)
  - the frequency of allele, e.g. 0.04 (the bigger the higher power)
  - **the rejection criterion**, e.g. $p < 0.05$ (the bigger the higher power)
  - **the number of samples** (the bigger the higher power)
  - **the test you use**
- Can often be calculated using "power-calculators"
- So before you start:
  Do power calculations to make sure you will have enough samples!

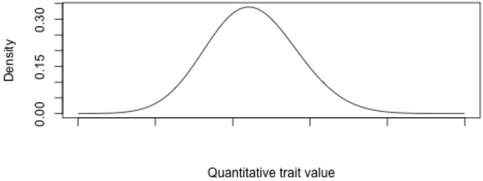
Power and power calculations

- Power depends on
  - the inheritance mode, e.g. recessive effect
  - the effect size, e.g. OR of 1.3 (the bigger the higher power)
  - the frequency of allele, e.g. 0.04 (the bigger the higher power)
  - **the rejection criterion**, e.g. $p < 0.05$ (the bigger the higher power)
  - **the number of samples** (the bigger the higher power)
  - **the test you use**
- Can often be calculated using "power-calculators"
- So before you start:
  Do power calculations to make sure you will have enough samples!
- To detect association we might not choose the model that is most correct, but instead choose the model that has the most power

Outline

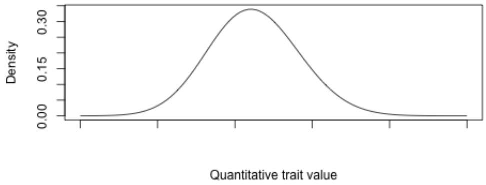
Quantitative trait

- Distribution of the trait in the population

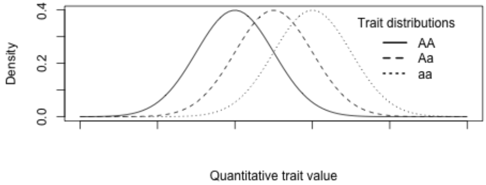


Quantitative trait

- Distribution of the trait in the population



- If a variant influence the trait value, we expect:



Linear regression

- Based on the following general model

$$E(y_i) = \beta_0 + \beta_1 x_1^i + \ldots + \beta_n x_n^i$$

  where the βs are *regression coefficients* (effect sizes).
- The x^is are determined by the genotype of individual i and the inheritance model

Linear regression

- Based on the following general model

$$E(y_i) = \beta_0 + \beta_1 x_1^i + \ldots + \beta_n x_n^i$$

  where the βs are *regression coefficients* (effect sizes).
- The x^is are determined by the genotype of individual i and the inheritance model
- E.g. for a simple additive inheritance model we have

$$E(y_i) = \beta_0 + \beta_1 x_1^i$$

  where x_1^i is the number number of copies of the variant so 0, 1 or 2

Linear regression

- Based on the following general model

$$E(y_i) = \beta_0 + \beta_1 x_1^i + \ldots + \beta_n x_n^i$$

where the βs are *regression coefficients* (effect sizes).
- The x^is are determined by the genotype of individual i and the inheritance model
- E.g. for a simple additive inheritance model we have

$$E(y_i) = \beta_0 + \beta_1 x_1^i$$

where x_1^i is the number number of copies of the variant so 0, 1 or 2
- Test if β_1 is zero (no association between the variant and the trait)

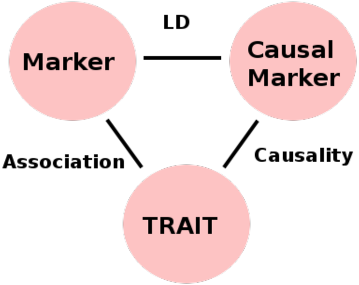
Outline

Types of association studies

- Candidate causative genetic variant
  - 1 SNP or deletion, duplication. Evidence from other study
- Candidate causative gene
  - 5-50 SNPs. Evidence from other study or function
- Candidate causative region
  - 100s of SNPs Evidence from other study
- Genome-wide (GWAS)
  - $>$500,000 SNPs. No prior evidence required

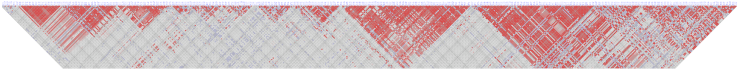
Why GWAS?

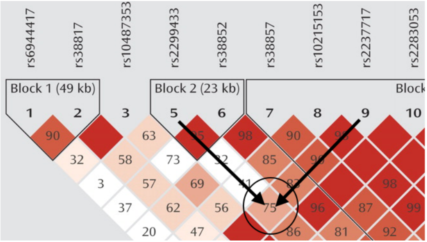
- If we look at 500.000 SNPs we are likely not to have the causal SNP!
- But, remember SNPs in high LD with a causal SNP will also be associated:



Why GWAS?

- SNPs are in high LD in blocks along the human genome





Why GWAS?

- By testing a few SNPs in each block most common SNPs are indirectly tested

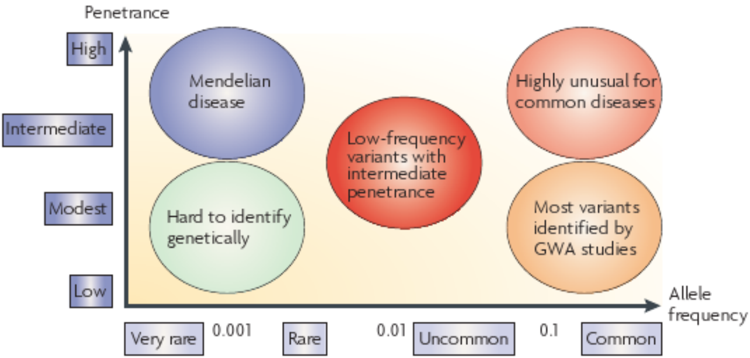
Why GWAS?

- By testing a few SNPs in each block most common SNPs are indirectly tested
- We can test most common SNPs (indirectly) by using $\geq 500,000$ SNPs

Why GWAS?

- By testing a few SNPs in each block most common SNPs are indirectly tested
- We can test most common SNPs (indirectly) by using $\geq 500,000$ SNPs
- Pro: Cheap! (only need to genotype $\geq 500,000$ SNPs)
  Con: We are far from sure the identified SNPs (if any) are causal!

When GWAS?



Strategies for locating disease loci

How GWAS (step-by-step overview)

1. Collect samples and traits of interest (based on power calculations!)
2. Genotype samples at a number of SNP loci ($\geq$500,000)
3. Lots and lots of quality control (QC)!
4. Statistically test each SNP for association
5. Assess the results:
    - make sure things went OK
    - identify associated SNPs
6. Identify causal variant (if possible)
7. Replicate associations in a different dataset
8. Investigate what the underlying biological mechanism is
9. Ideal longterm goal/hope: better prevention or treatment

GWAS step-by-step

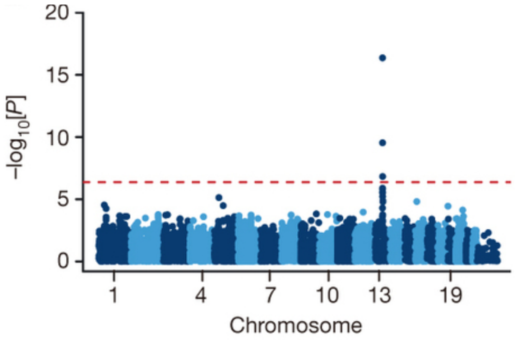
1. Collect samples and traits of interest (based on power calculations!)
2. Genotype samples at a number ($\geq$500,000) of SNP loci
3. **Lots and lots of quality control (QC)!**
4. **Statistically test each SNP for association**
5. **Assess the results:**
   - **make sure things went OK**
   - **identify associated SNPs**
6. Identify causal variant (if possible)
7. Replicate associations in a different dataset
8. Investigate what the underlying biological mechanism is
9. Ideal longterm goal/hope: better prevention or treatment

Statistically test each SNP for association

- Use one of the tests you just learned how to perform
- There are programs like PLINK2 that will help you do this
- Can be done using one 1-line command
- Also offers functions for doing QC (we'll see that later)

Identify associated SNPs

Manhattan plot

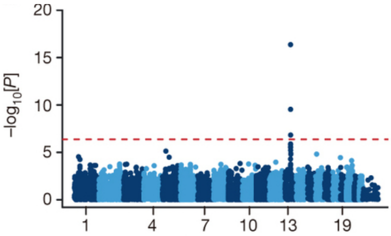


What p-value threshold to use

- Usually for a single test we use a p-value threshold of $\alpha = 0.05$

What p-value threshold to use

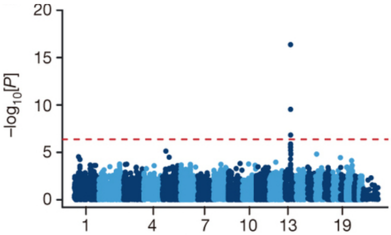
- Usually for a single test we use a p-value threshold of $\alpha = 0.05$
- If you perform many tests w. this α some will be falsely rejected
  With threshold 0.05 thousands of false positives!! (-log(0.05)=1.3)



So we have to **correct for multiple testing**

What p-value threshold to use

- Usually for a single test we use a p-value threshold of $\alpha = 0.05$
- If you perform many tests w. this α some will be falsely rejected
  With threshold 0.05 thousands of false positives!! (-log(0.05)=1.3)

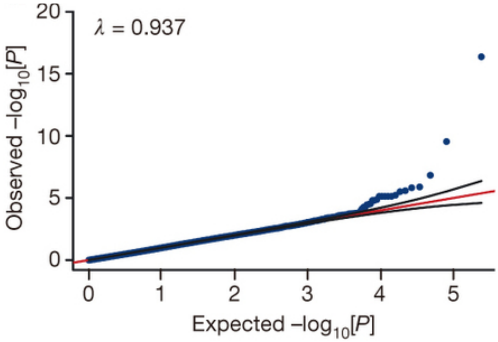
  So we have to **correct for multiple testing**
- Often **Bonferroni correction** is used; α is divided by the number of tests:
  - E.g. 100000 SNPs and $\alpha = 0.05$
  - Bonferroni corrected $\alpha = 0.05/100000 = 0.0000005 = 5 \times 10^{-7}$
  - Which on the Manhattan plot is $-log_{10}(5 \times 10^{-7}) = 6.3$

Exercise

Solve exercise 2A, i.e. perform your first GWAS analysis :)

Make sure things went OK!

QQ-plots and genomic control inflation factor λ



If so most of the dots will be on the x=y line and $\lambda \simeq 1$

Exercise

Solve exercise 2B, i.e. check if your results look OK...

Lots and lots of QC

This shows why we usually do QC first ...! :)

Let's therefore return to that step
(we wont go through all QCs, but some important ones)

Sample mislabling?

- One thing that can go wrong is the samples can be misslabled
- If so, genotypes won't match phenotypes
- This is difficult to catch
- But a simple check is to see of gender is correct
- If not the disease status is likely not to be either...
- We can check this using PLINK2

Sample mislabling?

- One thing that can go wrong is the samples can be misslabled
- If so, genotypes won't match phenotypes
- This is difficult to catch
- But a simple check is to see of gender is correct
- If not the disease status is likely not to be either...
- We can check this using PLINK2
- **Exercise:** try checking it for your data (exercise 2C)

Closely related individuals or duplicates?

- All association tests mentioned assume that the participants are **independent** samples from a population
- This would not be the case if some participants
  - are closely related
  - represented more than once
- One way to check if this is the case is to use PLINK2 (again)

Closely related individuals or duplicates?

- All association tests mentioned assume that the participants are **independent** samples from a population
- This would not be the case if some participants
  - are closely related
  - represented more than once
- One way to check if this is the case is to use PLINK2 (again)
- **Exercise:** try checking it for your data (exercise 2D)

Batch biases/non-random genotyping error?

- Sometimes the data handling/generation process can lead to non-random genotyping errors
- E.g. if all cases were genotyped first and then all controls, then changes in genotyping procedure along the way may lead to non-random differences in genotypes between cases and controls
- This may lead the false positive association test results

Batch biases/non-random genotyping error?

- Sometimes the data handling/generation process can lead to non-random genotyping errors
- E.g. if all cases were genotyped first and then all controls, then changes in genotyping procedure along the way may lead to non-random differences in genotypes between cases and controls
- This may lead the false positive association test results
- **Exercise:** try checking it for your data (exercise 2E+F if there is time)

Additional important checks?

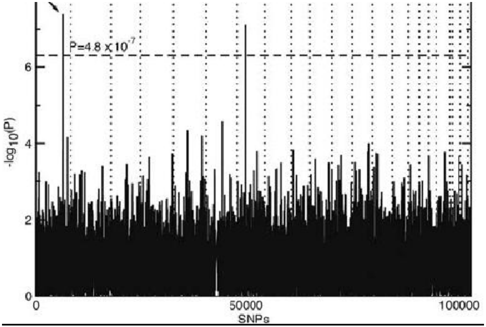
- Other additional signs of something being wrong include:
  - high missingness in specific loci/individuals
  - loci (strongly) out of Hardy-Weinberg Equilibrium (why?)
- Furthermore, low frequency variants tend to be difficult to genotype
- Removing such loci/individuals can help a lot

Additional important checks?

- Other additional signs of something being wrong include:
  - high missingness in specific loci/individuals
  - loci (strongly) out of Hardy-Weinberg Equilibrium (why?)
- Furthermore, low frequency variants tend to be difficult to genotype
- Removing such loci/individuals can help a lot
- **Exercise:** try rerunning your analyses with these QC filters (exercise 2G)

First study went extremely well!

- Study of age-related Macular Degeneration (Klein et al. 2005)
- 96 cases and 50 controls, 100K SNPs



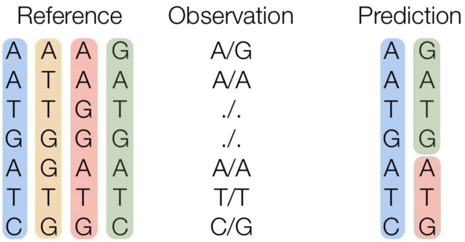
- SNP in *CFH* w large effect (OR=7.4)+led to new biological insight

Turned out to be unusual...

- MANY studies and many associations
- But in the beginning few were replicated
  (underpowered, population structure, insufficient corr. for multiple tests)
- So later studies have many more samples and are much stricter
- And most found small effect sizes and limited biological insight

NGS enters the stage

- Reference panels
  - 1000 genomes project
  - Haplotype reference consortium
- Imputation:



- Results in posterior genotype probabilities.

Dealing with uncertain genotypes in associations

- The easy solution: DOSAGE

$$E[g] = \sum_{g=0}^{2} g \; p(G = g|X)$$

- The complicated solution: Full likelihood model

$$p(y|X) = \prod_i \sum_g p(y_i|G_i = g)p(G_i = g|X_i)$$

- Same goes for association studies based on directly on sequencing data.