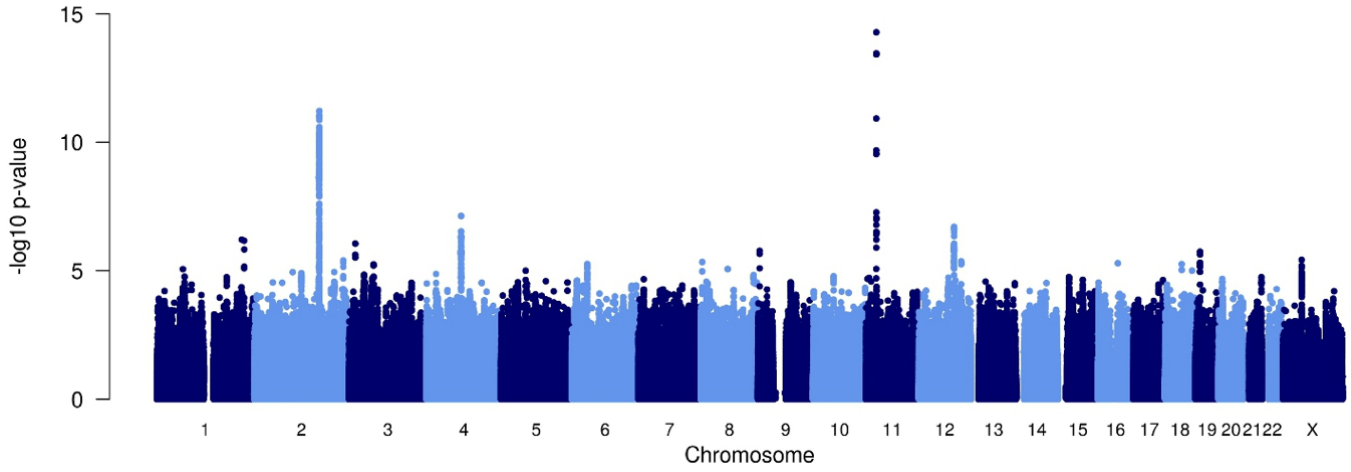
Association testing and GWAS



Line Skotte, Medical and Population Genetics Course, August 2019

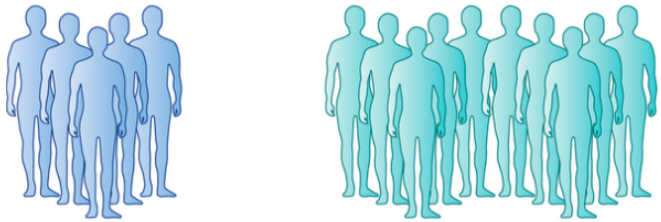
Outline

What and why?

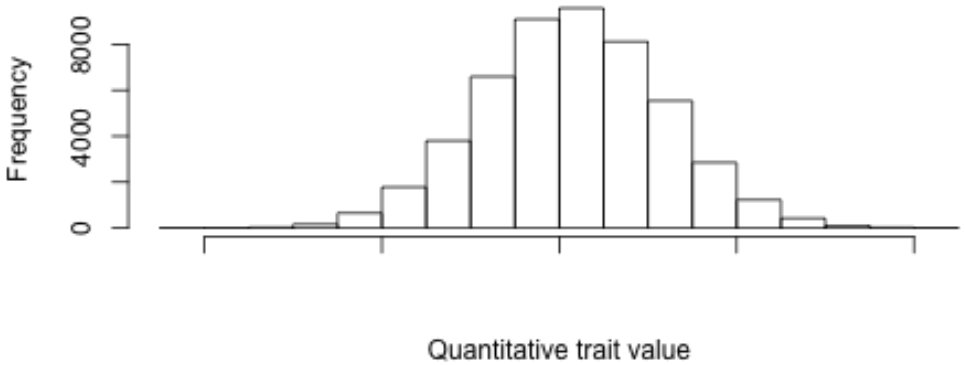
- **Goal: to identify (map) genetic variants that have an effect on a trait**
- Typically **disease related traits**, e.g. febrile seizures
- Motivation: reaching this goal can help
  - reveal the underlying genetic architecture
  - hopefully lead to better understanding of what **causes** the disease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

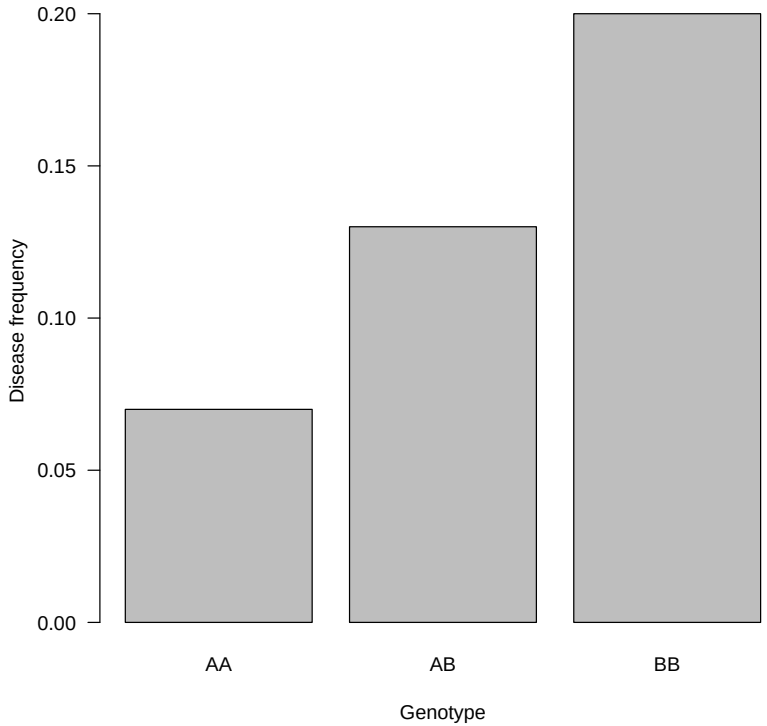


Outline

Overall idea in association testing

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

- 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)

Probability of disease given genotype

What is the probability of disease for the different genotypes?

- In the previous figure, $P(D|AA) = 0.07$, $P(D|AB) = 0.13$ and $P(D|BB) = 0.20$.
- More formally we model the probability of disease for an individual

$$p = P(D|g)$$

- We can use a logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{AB}x_{AB} + \beta_{BB}x_{BB}$$

  where the βs are *regression coefficients* (effect sizes).
- The xs are determined by the genotype of the individual considered:

Genotypes	x_{AB}	x_{BB}
AA	0	0
AB	1	0
BB	0	1

Exercise

What is the probability of disease for the different genotypes?

- An individual has genotype AA. Can you express $p = p(D|AA)$ in terms of the effect sizes?
- How do you find the probability of disease when the individual has genotype AB?

Solution

What is the probability of disease for the different genotypes?

- We can rephrase the logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{AB}x_{AB} + \beta_{BB}x_{BB}$$

into

$$p = \frac{\exp\left(\beta_0 + \beta_{AB}x_{AB} + \beta_{BB}x_{BB}\right)}{1 + \exp\left(\beta_0 + \beta_{AB}x_{AB} + \beta_{BB}x_{BB}\right)}$$

- For individual with genotype AA, we see that $x_{AB} = x_{BB} = 0$ and thus

$$p(D|AA) = \frac{\exp\left(\beta_0\right)}{1 + \exp\left(\beta_0\right)}$$

- For individual with genotype AB, we see that $x_{AB} = 1$ and thus

$$p(D|AB) = \frac{\exp\left(\beta_0 + \beta_{AB}\right)}{1 + \exp\left(\beta_0 + \beta_{AB}\right)}$$

Testing for association between disease and genotype

How do we test for association between disease and genotype?

- The logistic regression model allows the probability of disease to depend on genotype:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{AB}x_{AB} + \beta_{BB}x_{BB}.$$

- If there is **no association** then the probability of disease is the same regardless of genotype, i.e.: $P(D|AA) = P(D|AB) = P(D|BB)$.
- The corresponding logistic regression **null model** is

$$\log\left(\frac{p}{1-p}\right) = \beta_0$$

  and this null model assumes $\beta_{AB} = \beta_{BB} = 0$.
- The **likelihood ratio test** compares the likelihood of the data under these two models and here it has two degrees of freedom.

Exercise

How do we test for association between disease and genotype?

- Load a test dataset into R with one line per person and three columns. First colums is the disease status of each individual and the other two columns contain counts of allele B for two different variants, SNP1 and SNP2.

```
# Read data
dat<-read.table(file = "case_control_snp1_snp2.txt", header = TRUE,
     as.is = TRUE)

# Check out the data
str(dat)
table(dat$status,dat$snp1)
table(dat$status,dat$snp2)
```

- Which variant looks most associates with disease status?

Exercise

How do we test for association between disease and genotype?

- Do the association test for SNP1:

```
# Fit null model
null <- glm(status ~ 1, family = binomial(link = "logit"), data =
    dat)
summary(null)

# Fit full genotype model for snp1
full_snp1 <- glm(status ~ as.factor(snp1), family = binomial(link =
    "logit"), data = dat)
summary(full_snp1)

# Do the LRT test
anova(null, full_snp1, test = "LRT")
```

- Is the variant associated to disease status?

Exercise

How do we test for association between disease and genotype?

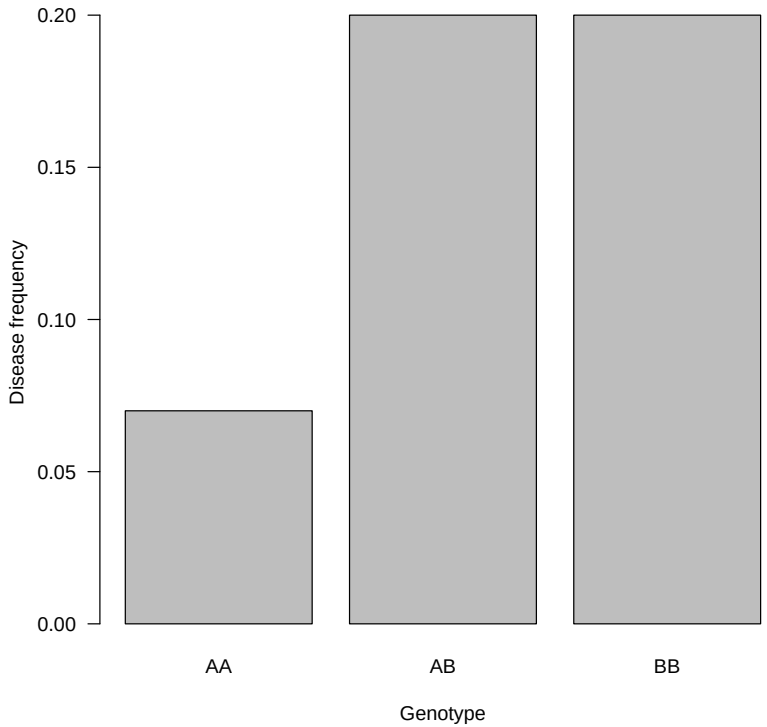
- Do the association test for SNP2:

```
# Fit null model
null <- glm(status ~ 1, family = binomial(link = "logit"), data =
    dat)
summary(null)

# Fit full genotype model for snp2
full_snp2 <- glm(status ~ as.factor(snp2), family = binomial(link =
    "logit"), data = dat)
summary(full_snp2)

# Do the LRT test
anova(null, full_snp2, test = "LRT")
```

- Is the variant associated to disease status?

Assuming recessive, dominant or "additive" effects

- The full genotype model just described allows different disease probabilities for each genotype.
- Genetic effects can be **dominant**:



Assuming recessive, dominant or "additive" effects

- Genetic effects can be **dominant**, meaning that we assume

$$P(D|AB) = P(D|BB).$$

- We can then use a simpler logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_D x_D$$

- The xs are determined by the genotype of the individual considered:

Genotypes	x_D
AA	0
AB	1
BB	1

- The corresponding **likelihood ratio test** has 1 degree of freedom.

Assuming recessive, dominant or "additive" effects

- Genetic effects can be **recessive**, meaning that we assume

$$P(D|AA) = P(D|AB).$$

- We can again use a simpler logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_R x_R$$

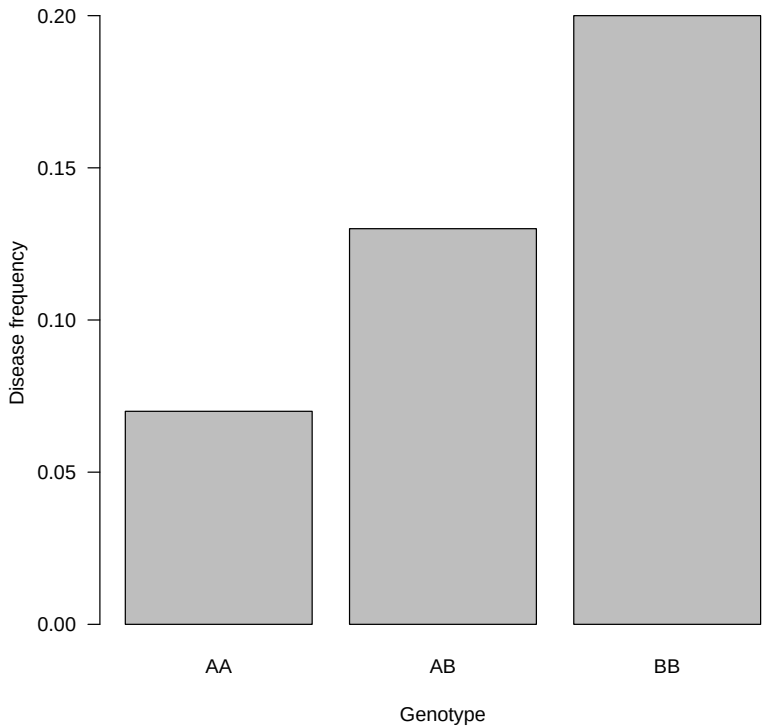
- The xs are determined by the genotype of the individual considered:

Genotypes	x_R
AA	0
AB	0
BB	1

- The corresponding **likelihood ratio test** has 1 degree of freedom.

Assuming recessive, dominant or "additive" effects

- In some cases genetic effects are expected to be near-**additive**.
- The the risk of disease for heterozygous carriers $P(D|AB)$ is midway between the two risks for homozygous carriers $P(D|AA)$ and $P(D|BB)$.



Assuming recessive, dominant or "additive" effects

Is there a 1 degree of freedom test for the near-additive effects?

- We can again use a simpler logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_A x_A$$

- The x_As are determined by the genotype of the individual considered:

Genotypes	x_A
AA	0
AB	1
BB	2

- The corresponding **likelihood ratio test** has 1 degree of freedom.

Exercise

How do we assume recessive, dominant or "additive" effects?

- Fit dominant and recessive models for SNP2 in R:

```
# Fit and test dominant model
dat$snp2_x_dom <- 1*(dat$snp2 > 0)
dom_snp2 <- glm(status ~ snp2_x_dom, family = binomial(link = "logit
    "), data = dat)
summary(dom_snp2)
anova(null, dom_snp2, test = "LRT")

# Fit and test recessive model
dat$snp2_x_rec <- 1*(dat$snp2 == 2)
rec_snp2 <- glm(status ~ snp2_x_rec, family = binomial(link = "logit
    "), data = dat)
summary(rec_snp2)
anova(null, rec_snp2, test = "LRT")
```

- Why does the association seem less significant when se assume that the effect is recessive?

Exercise

How do we assume recessive, dominant or "additive" effects?

- Fit the "additive" model for SNP2 in R:

```
# Fit and test "additive" model
add_snp2 <- glm(status ~ snp2, family = binomial(link = "logit"),
    data = dat)
summary(add_snp2)
anova(null, add_snp2, test = "LRT")
```

- Is the p-value lower than for the full genotype model?
- Compare the different nested models:

```
# Comparing nested models
anova(dom_snp2, full_snp2, test = "LRT")
anova(rec_snp2, full_snp2, test = "LRT")
anova(add_snp2, full_snp2, test = "LRT")
```

- What model describes "best" the genetic effect?

Model assumptions, degrees of freedom and power

What single-SNP test is the correct one to use?

- If a sub-model is correctly specified, the test with fewer degrees of freedom have better power than the full genotype test.
- It the model is strongly misspecified, the test may loose power.
- A slightly misspecified model with fewer degrees of freedom can have better power than a completely correct test with more degrees of freedom.

We will discuss statistical power again a bit later.

Logistic regression in general

- Based on the following general model

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

  βs are *regression coefficients* (effect sizes) and xs are covariates.

- They can also contain additional covariates and allow correcting the model for sex, population structure or batch effects.
- The **design matrix** is a matrix made of the xs as columns.
- Single-SNP design matrix calculation:

Genotypes	Additive	Dominant	Recessive	Full	
	x_A	x_D	x_R	x_{AB}	x_{BB}
AA	0	0	0	0	0
AB	1	1	0	1	0
BB	2	1	1	0	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).
- 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
- Possible to compare nested models using ANOVA

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.
- Example: In the recessive model we estimate the effect size β_R

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_R x_R$$

- The OR from the recessive model

$$\frac{\text{ODD}_{BB}}{\text{ODD}_{AA/AB}} = \frac{\frac{p_{BB}}{1-p_{BB}}}{\frac{p_{AA/AB}}{1-p_{AA/AB}}} = \frac{\exp(\beta_0+\beta_R)}{\exp(\beta_0)} = \exp(\beta_R)$$

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

Exercise

How do we extract effect size estimates from logistic regression models?

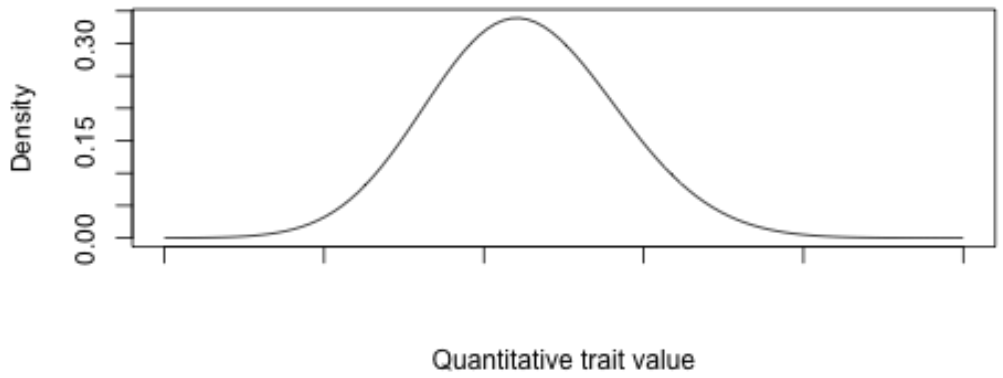
- Calculate ORs from the logistic regression in R:

```
# Inspect coefficients and find ORs in the full genotype model
full_snp2$coefficients
# OR for AB versus AA
exp(full_snp2$coefficients[2])
# OR for BB versus AA
exp(full_snp2$coefficients[3])

# OR from "additive" model
add_snp2$coefficients
exp(add_snp2$coefficients[2])
```

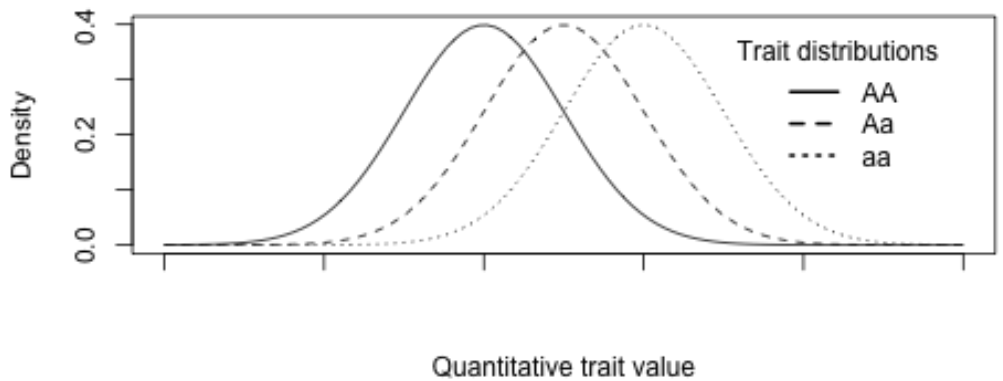
- How can we interpret the OR calculated in the "additive" model?

Quantitative trait

- Distribution of the trait in the population (assume normal distribution)



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



Linear regression

- Based on the following general model

$$\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

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

$$\mathrm{E}(y) = \beta_0 + \beta_A x_A$$

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

Outline

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

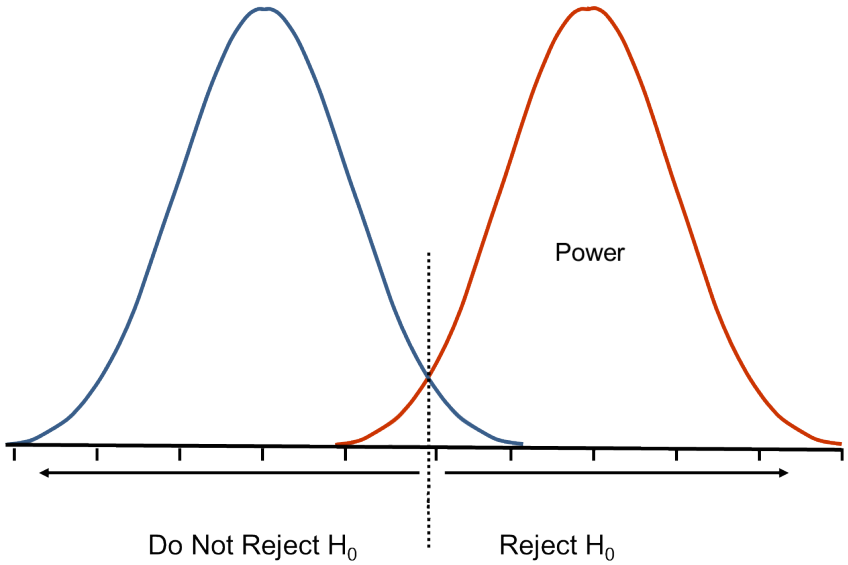
One also has to be aware of the underlying assumptions:

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

We will return to these issues a bit later.

Study design

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



Blue line: Distribution under H_0. Red line: Distribution under anticipated effects.

- Before you start your study: calculate power for your study and assess it Rule of thumb: power should be at least 0.8 (or the study is not really worth performing)!

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

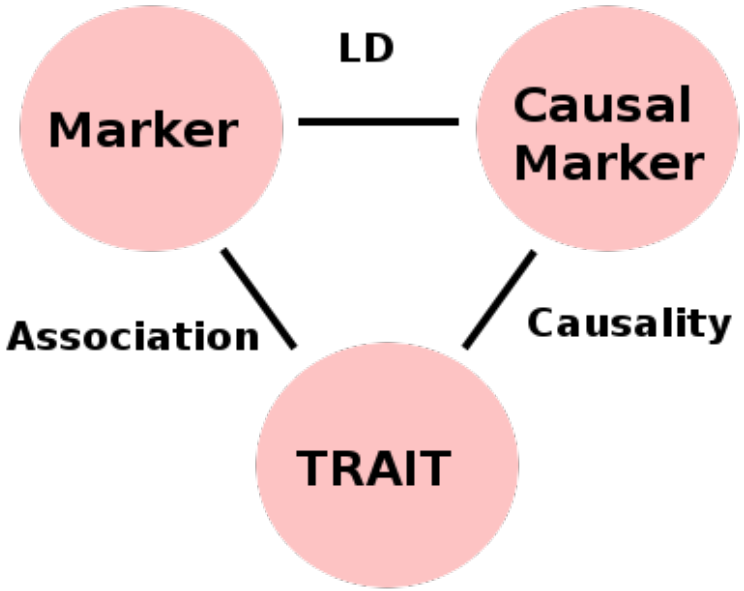
Outline

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)

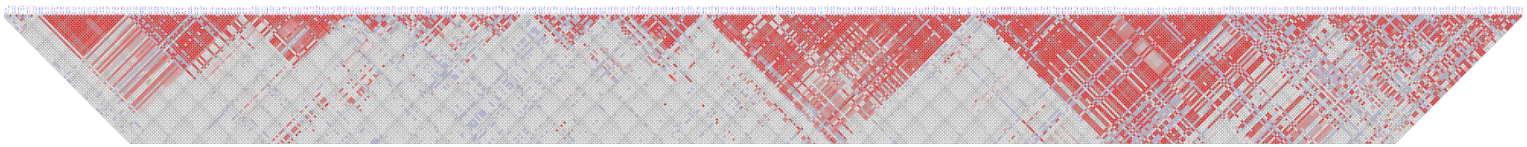
Why GWAS?

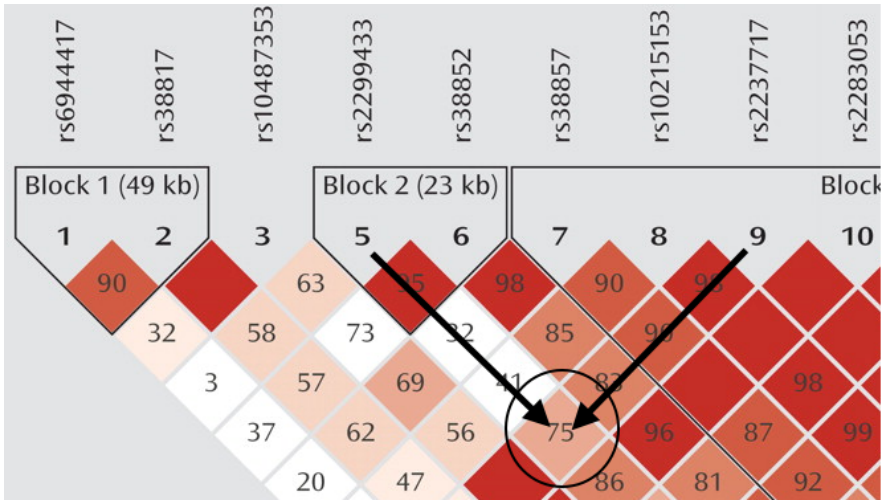
- If we look at 500.000 SNPs we are likely not to have the causal SNP!
- But, remember SNPs in high LD with a causal SNP will also be associated:



Why GWAS?

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

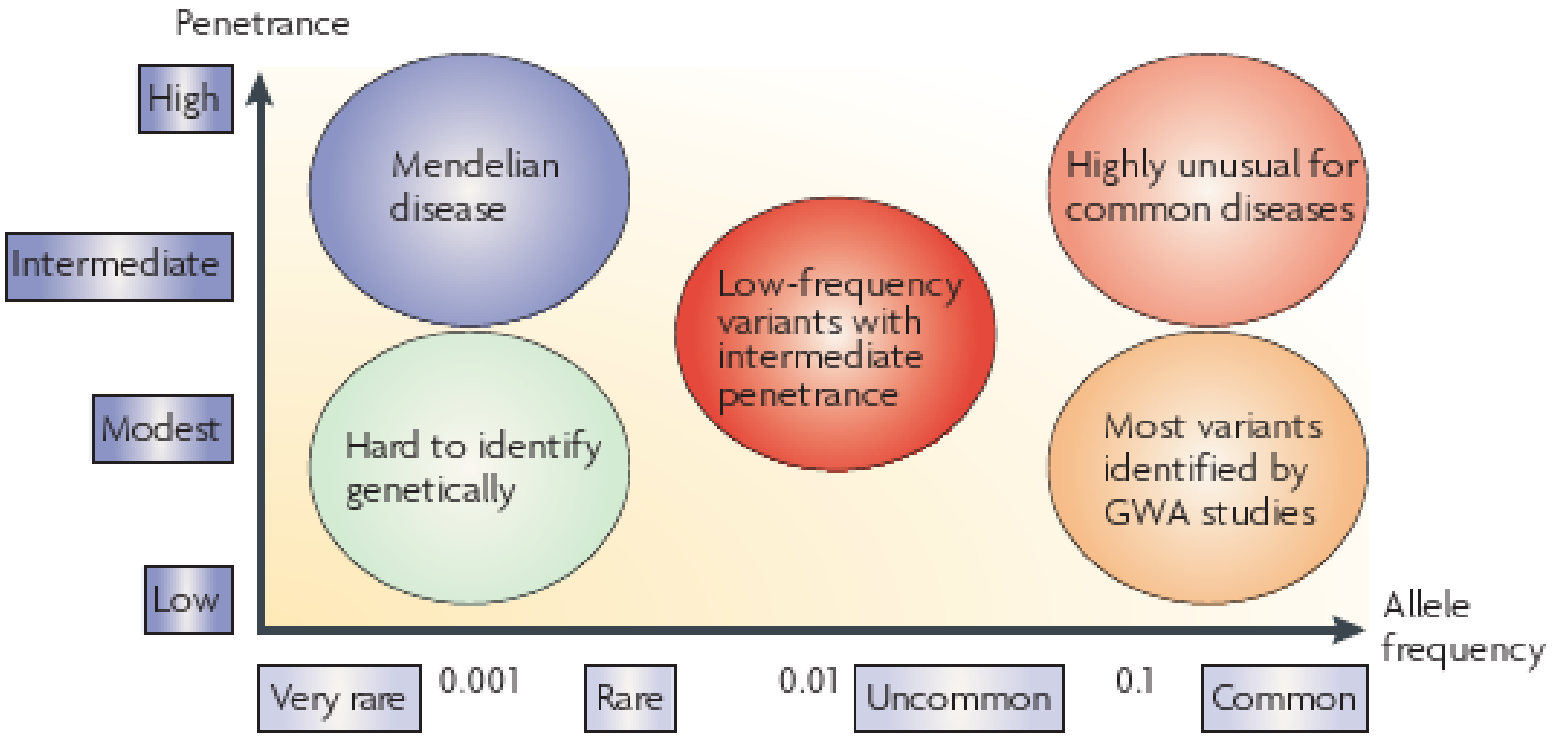




Why GWAS?

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

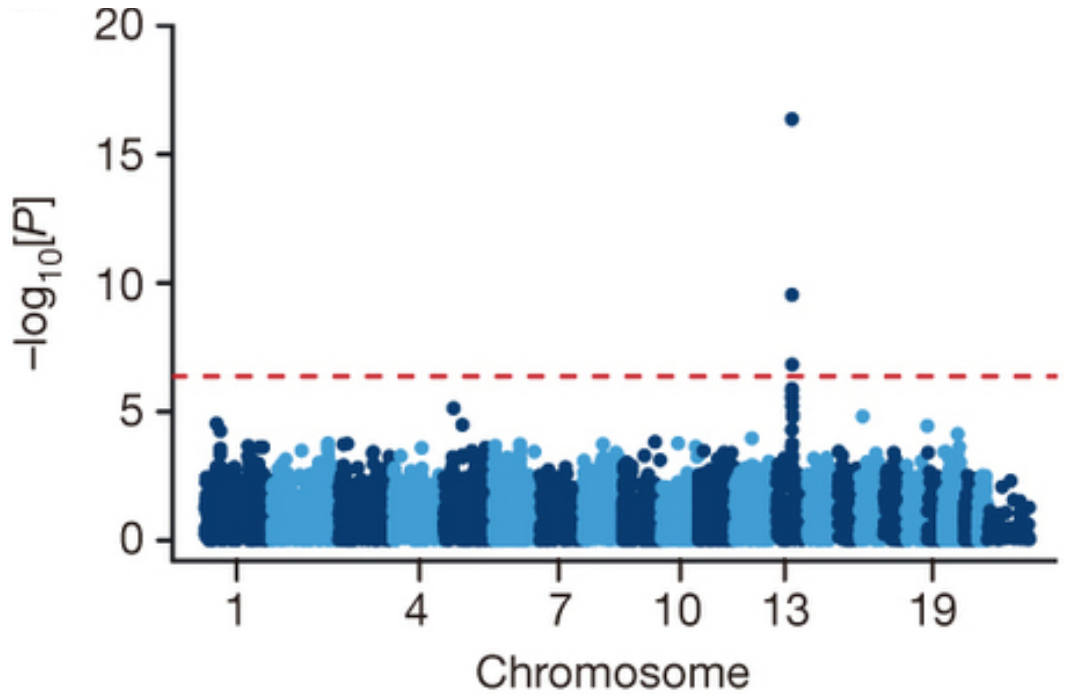
1. Collect samples and traits of interest (based on power calculations!)
2. Genotype samples at a number ($\geq$500,000) of SNP loci
3. **Lots and lots of quality control (QC)!**
4. **Statistically test each SNP for association**
5. **Assess the results:**
   - **make sure things went OK**
   - **identify 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)

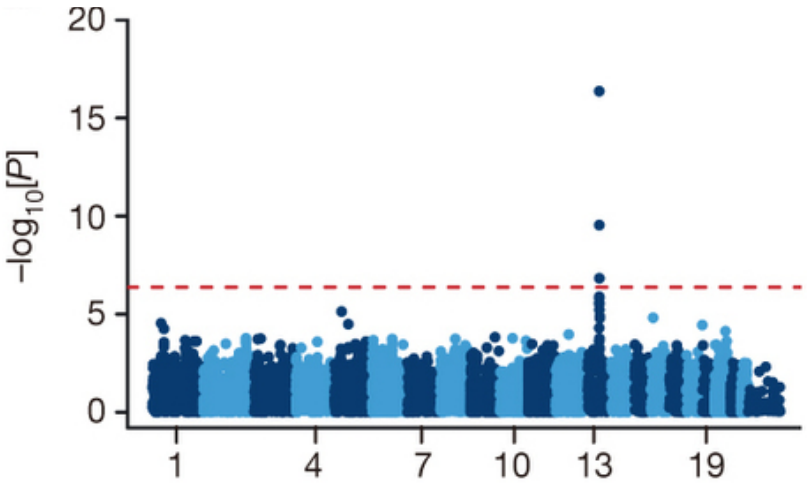
Identify associated SNPs

Manhattan plot



What p-value threshold to use

- Usually for a single test we use a p-value threshold of $\alpha = 0.05$
- If you perform many tests with 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**

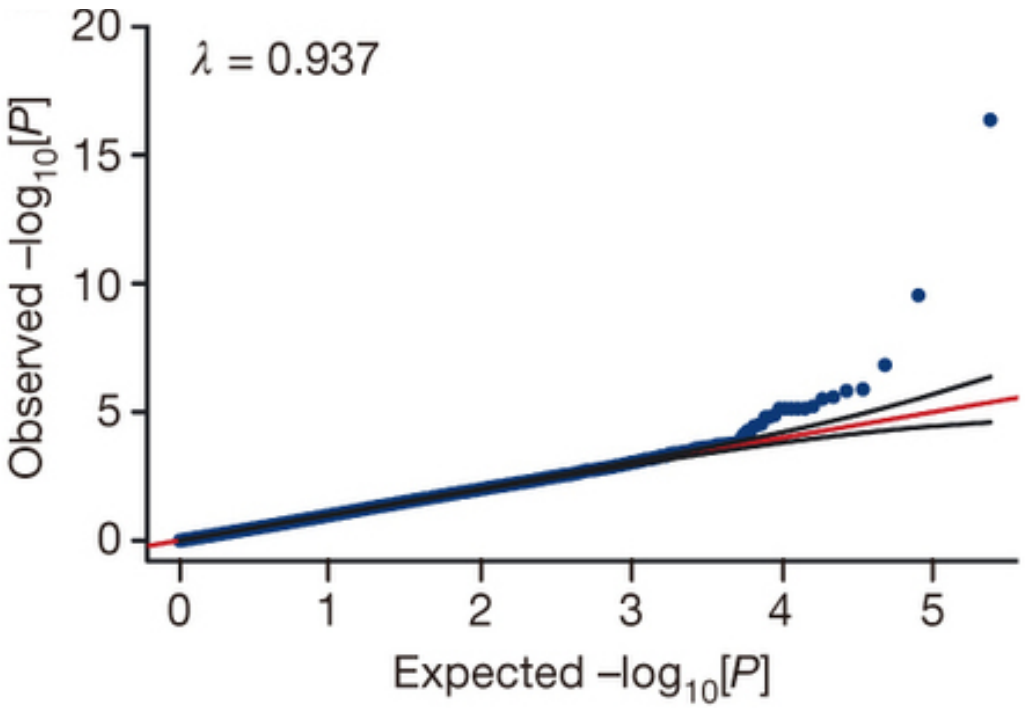
- Often **Bonferroni correction** is used; α is divided by the number of tests:
  - E.g. 100000 SNPs and $\alpha = 0.05$
  - Bonferroni corrected $\alpha = 0.05/100000 = 0.0000005 = 5 \times 10^{-7}$
  - Which on the Manhattan plot is $-log_{10}(5 \times 10^{-7}) = 6.3$

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
- **Exercise:** try checking it for your data (exercise 2E+F)

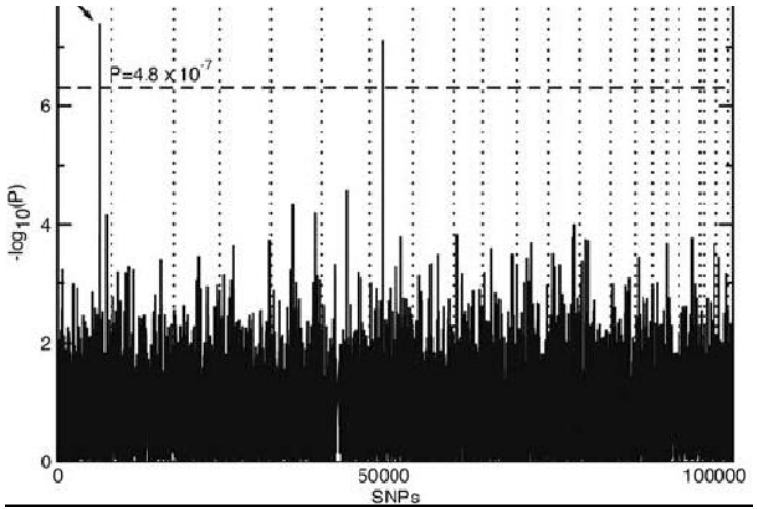
Additional important checks?

- 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)
- If you are done with the previous exercises, you can look into exercise 2F.

Outline

First study went extremely well!

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



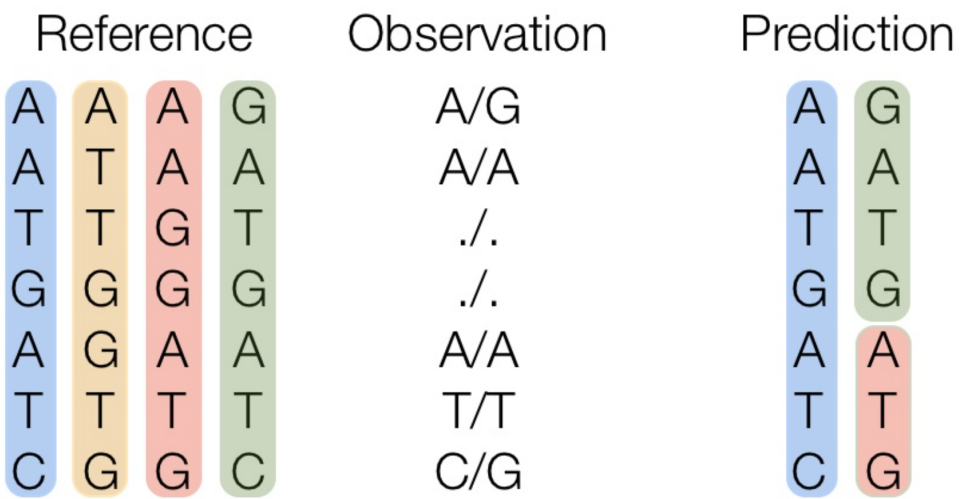
- SNP in *CFH* with large effect (OR=7.4) led to new biological insight

Turned out to be unusual...

- MANY studies and more than 50,000 associations with $p < 5x10^{-8}$
- 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
- Many studies find only small effect sizes and and some give only limited biological insight
- Some studies may be worth the effort and lead to:
  - Discovery of novel biological mechanisms
  - Clinical applications
  - Drug development and repurposing

NGS enters the stage

- NGS allowed generation of large **reference panels**
  - The 1000 Genomes Project $>$ 2500 genomes
  - Haplotype Reference Consortium $>$ 65000 haplotypes
- Imputation:



- Summarised in **posterior genotype probabilities**:

$$P(AA), \quad P(AB) \quad \text{and} \quad P(BB)$$

NGS enters the stage

How do we asses effects of rare or study-specific variants?

- NGS allows generation of **study specific reference panels**
- GWAS based directly on **NGS sequencing**
  - Example: Liu et al., 2018, Cell 175, 347−359
    $> 141K$ low-pass genomes $=>$ 16 novel associations
- Important to account appropriately lack of **full genotype information**
  - Increase **power**
  - Reduce **false positives**
- Summarised in **posterior genotype probabilities**:

$$P(AA), \quad P(AB) \quad \text{and} \quad P(BB)$$

Dealing with uncertain genotypes in associations

- The easy solution: Dosage aka **expected genotype**, which use

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

  in the single-SNP tests described above.

- The complicated solution: **Maximum likelihood**, based on

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

- Here $p(y_i|G_i = g)$ is the **probability of disease** in the case-control studies or the **normal distribution density** in the quantitative trait studies.
- and $p(G_i = g|X)$ is the **posterior genotype probability**.

Exercise

Solve exercise 3, How to run association tests on genotype probabilities.