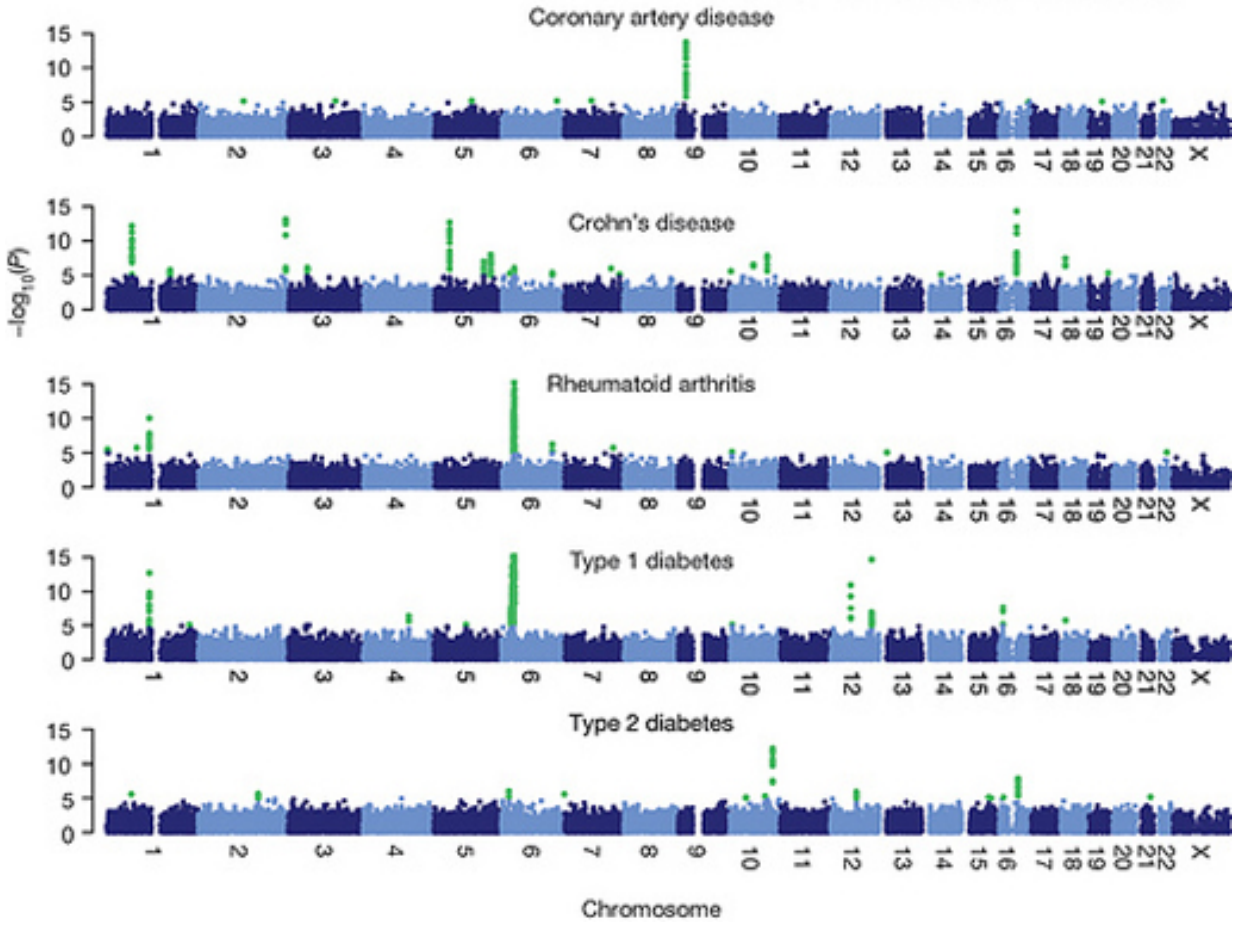
Genome-Wide Association Studies (GWAS) Workshop



Ida Moltke & Shyam Gopalakrishnan
Genomes and Biodiversity, Nov 2019

Learning objectives for today

By the end of today you should:

- be able to explain the basic idea behind a GWAS
- be able to perform the most basic steps of a GWAS
- be able to interpret the plots typically used in GWAS
  - Manhattan plots
  - QQ-plots
- be aware of some important potential pitfalls
  (poor quality data, batch effects, multiple testing, population structure)

Learning objectives for today

By the end of today you should:

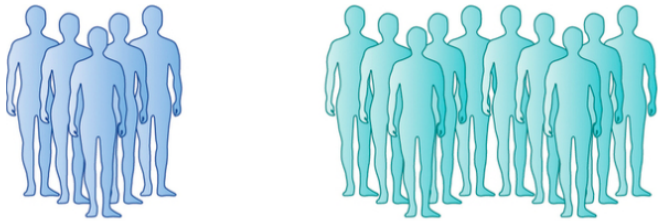
- be able to explain the basic idea behind a GWAS
- be able to perform the most basic steps of a GWAS
- be able to interpret the plots typically used in GWAS
  - Manhattan plots
  - QQ-plots
- be aware of some important potential pitfalls
  (poor quality data, batch effects, multiple testing, population structure)

The plan is to reach these goals via a lecture with exercises along the way

Outline

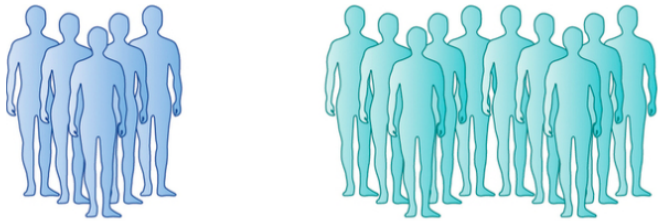
What is the goal?

- **To find (map) genetic variants that have an effect on a trait**, e.g. height
- Often **focused on disease traits** (also our main focus today)
  - i.e. studies where the phenotype of interest is whether the participants have a disease (cases) or not (controls)



What is the goal?

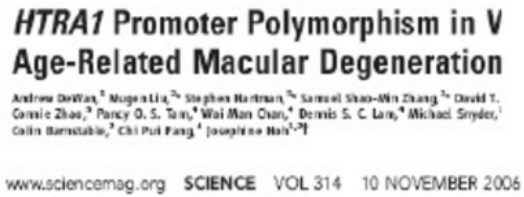
- **To find (map) genetic variants that have an effect on a trait**, e.g. height
- Often **focused on disease traits** (also our main focus today)
  - i.e. studies where the phenotype of interest is whether the participants have a disease (cases) or not (controls)



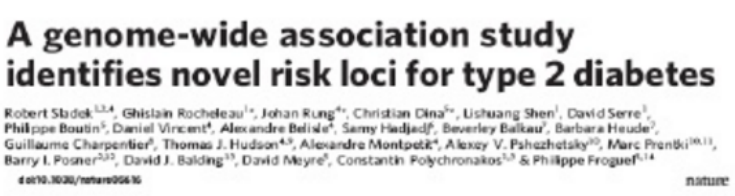
  - The motivation is that reaching this goal can hopefully
    - reveal the role genetics play in the disease
    - hopefully lead to better understanding the disease etiology
    - ideally lead to better treatment and/or prevention

Why learn about GWAS?

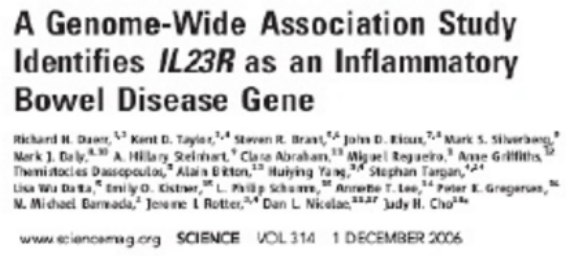
- Several approaches to such mapping, e.g. linkage mapping
- GWAS the most used approach to mapping since 2007:


HTRA1 Promoter Polymorphism in V Age-Related Macular Degeneration

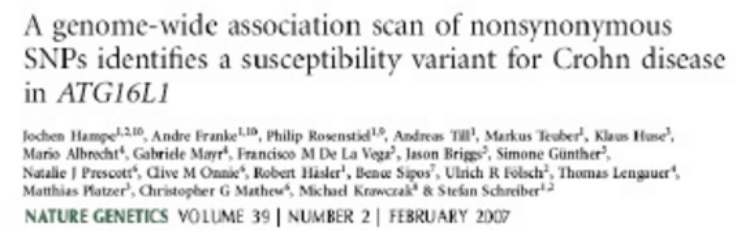
www.sciencemag.org SCIENCE VOL 314 10 NOVEMBER 2006


A genome-wide association study identifies novel risk loci for type 2 diabetes

nature


A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

www.sciencemag.org SCIENCE VOL 314 1 DECEMBER 2006


A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*

NATURE GENETICS VOLUME 39 | NUMBER 2 | FEBRUARY 2007

- So probably the single most important approach within disease mapping

Why did GWAS become so popular?

- Previous methods successful for **Mendelian diseases**, where
  - environmental factors are less important than genetic factors
  - causal genetic variants have **high penetrance*** (close to 1)
  - only one or a few causal variants are involved
  - causal variants are typically **rare**

*proportion of cases among individuals carrying the variant

Why did GWAS become so popular?

- Previous methods successful for **Mendelian diseases**, where
  - environmental factors are less important than genetic factors
  - causal genetic variants have **high penetrance*** (close to 1)
  - only one or a few causal variants are involved
  - causal variants are typically **rare**
- BUT previous methods not successful for **complex diseases**, where
  - environmental factors are very important
  - causal genetic variants have **low penetrance**
  - many causal genetic variants are involved
  - causal variants are potentially **common** (we do not know)

*proportion of cases among individuals carrying the variant

Why did GWAS become so popular?

- Previous methods successful for **Mendelian diseases**, where
  - environmental factors are less important than genetic factors
  - causal genetic variants have **high penetrance*** (close to 1)
  - only one or a few causal variants are involved
  - causal variants are typically **rare**
- BUT previous methods not successful for **complex diseases**, where
  - environmental factors are very important
  - causal genetic variants have **low penetrance**
  - many causal genetic variants are involved
  - causal variants are potentially **common** (we do not know)
- Most common diseases are complex! e.g. type 2 diabetes, CVD

*proportion of cases among individuals carrying the variant

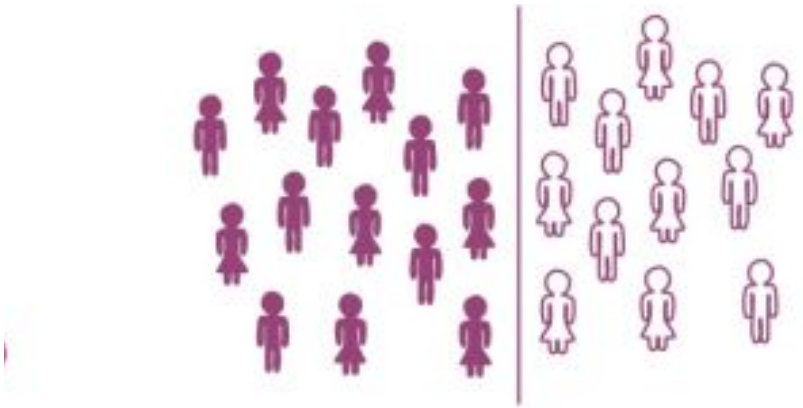
Why did GWAS become so popular?

- Previous methods successful for **Mendelian diseases**, where
  - environmental factors are less important than genetic factors
  - causal genetic variants have **high penetrance*** (close to 1)
  - only one or a few causal variants are involved
  - causal variants are typically **rare**
- BUT previous methods not successful for **complex diseases**, where
  - environmental factors are very important
  - causal genetic variants have **low penetrance**
  - many causal genetic variants are involved
  - causal variants are potentially **common** (we do not know)
- Most common diseases are complex! e.g. type 2 diabetes, CVD
- **GWAS works for common, low penetrance variants** (common disease)

*proportion of cases among individuals carrying the variant

The basic idea behind association studies

- We look at genetic data from unrelated, randomly sampled individuals

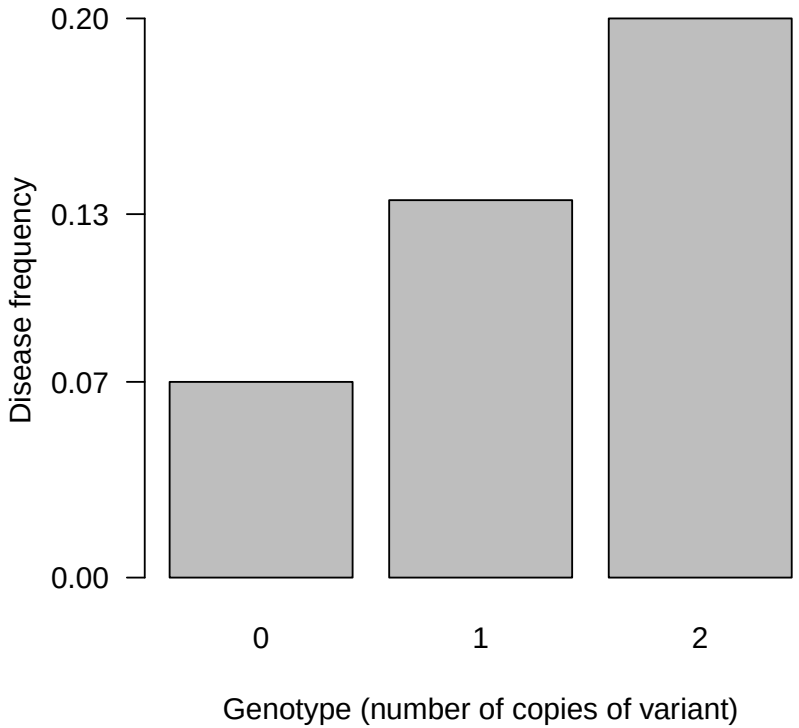


cases **controls**

- Typically 500,000-5,000,000 SNP variants along the genome

The basic idea behind association studies

- We look for SNP variants that are associated (correlated) with disease



- Test each SNP for association & identify those with low p-value

Why?

Why?

- Expect to see association in locus with a causal variant!

Why?

- Expect to see association in locus with a causal variant!
- Expect to see it in 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

- We expect to see it in loci that are in high LD with the causal SNP
- NB LD measure usually used is r^2 (ranges 0-1, where 1=fully correlated)

Outline

GWAS step-by-step

1. Collect samples and traits of interest
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 that passed QC 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

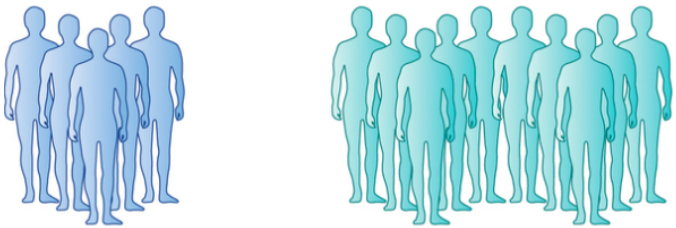
1. **Collect samples and phenotypic data**
2. **Genotype samples at a number ($\geq$500,000) of SNP loci**
3. **Lots and lots of quality control (QC)!**
4. **Statistically test each SNP for association**
5. **Assess the results:**
   - **make sure things went OK**
   - **identify associated SNPs**
6. 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

Step 1: Collect samples and phenotypic data

Two possible designs:

- Take a random sample from a population (population cohort):

- Select samples based on traits:



- Both require informed consent! (and ethical approval)

Outcome

The outcome of collecting data:

- Blood samples or similar (for later genotyping)
- Plus information about the traits of interest
  - **disease status**
    e.g. have diabetes (case) vs do not have diabetes (control)
  - **quantitative**
    e.g. cholesterol, blood pressure or blood sugar

Step 2: Genotype samples

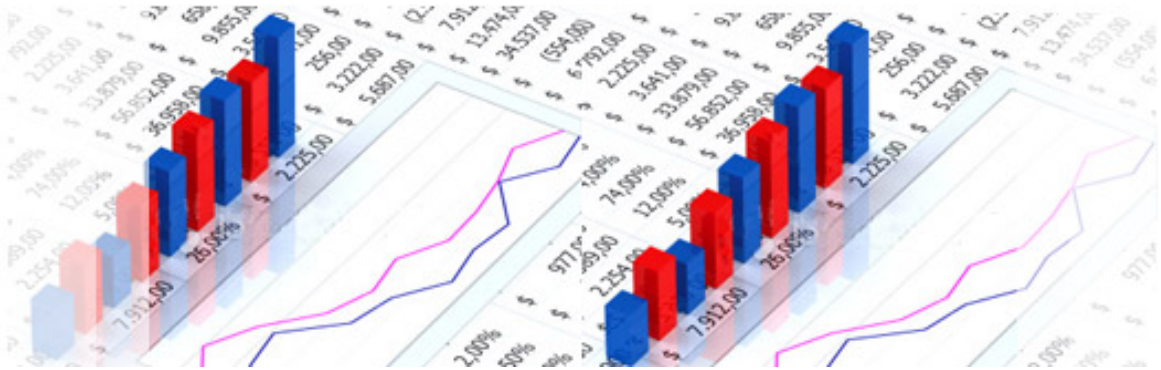
How does one choose genotyping platform?

- Want to (indirectly) test as many (common) SNPs as possible CHEAPLY
- Which and how many SNPs are needed depend on LD.
- Often data from standard SNP arrays (500K-5M, the smaller the cheaper!)

Step 3: Lots and lots of QC

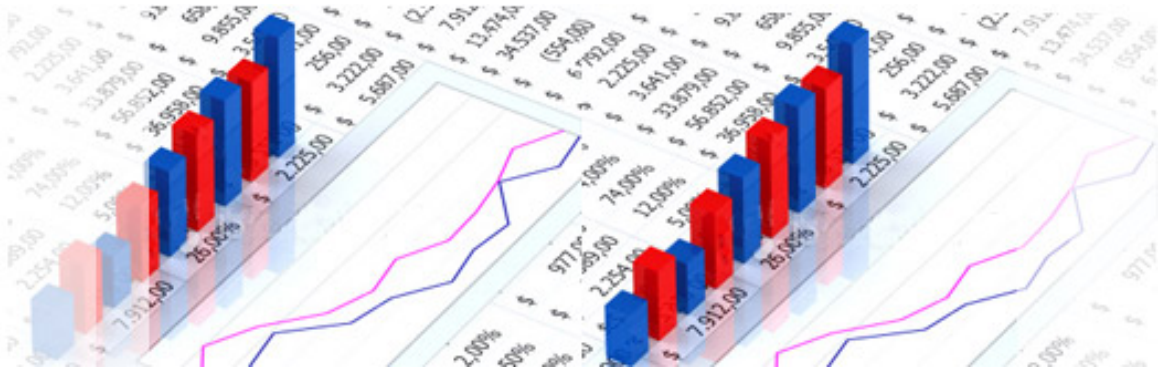
Skip for now :)

Overview of possible tests



- One can use many different tests
- Which to use depends on inheritance mode (additive, recessive, dominant)
- Mainly depends on type of data (case-control or quantitative)

Overview of possible tests



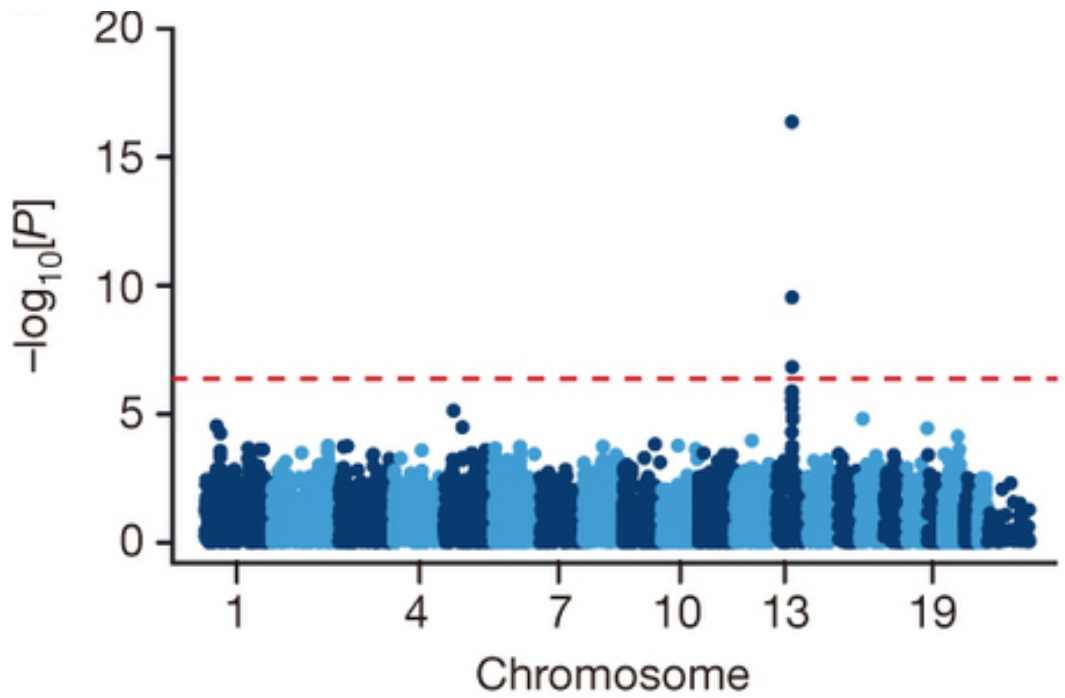
- One can use many different tests
- Which to use depends on inheritance mode (additive, recessive, dominant)
- Mainly depends on type of data (case-control or quantitative)
- **Additive model** for inheritance mode is almost always assumed
- For quantitative traits, e.g. height, **linear regression** often used
- For case-control traits, e.g. T2D, **logistic regression** often used

Test for association in case-control traits

Can you think of a simple statistic to find SNPs associated with disease status?

Identify associated SNPs

Manhattan plot

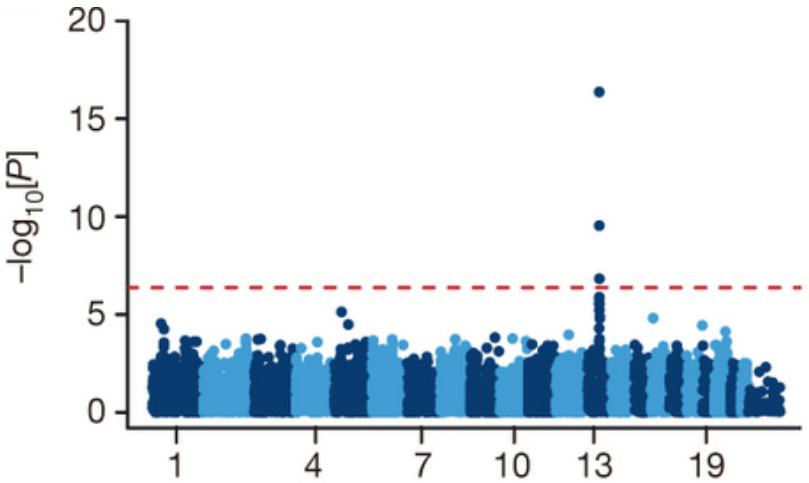


What p-value threshold to use

- Usually for a single test we use a p-value threshold of $\alpha = 0.05$
- If you perform many tests w. this α you will get a lot of false positives!
- 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 you will get a lot of false positives!
- So we have to **correct for multiple testing**
- Often **Bonferroni correction** is used; α is divided by the number of tests:
  - 100000 SNPs and $\alpha = 0.05$
  - Bonferroni corrected $\alpha = 0.05/100000 = 0.0000005 = 5 \times 10^{-7}$
  - which on the Manhattan plot is $-log_{10}(5 \times 10^{-7}) = 6.3$

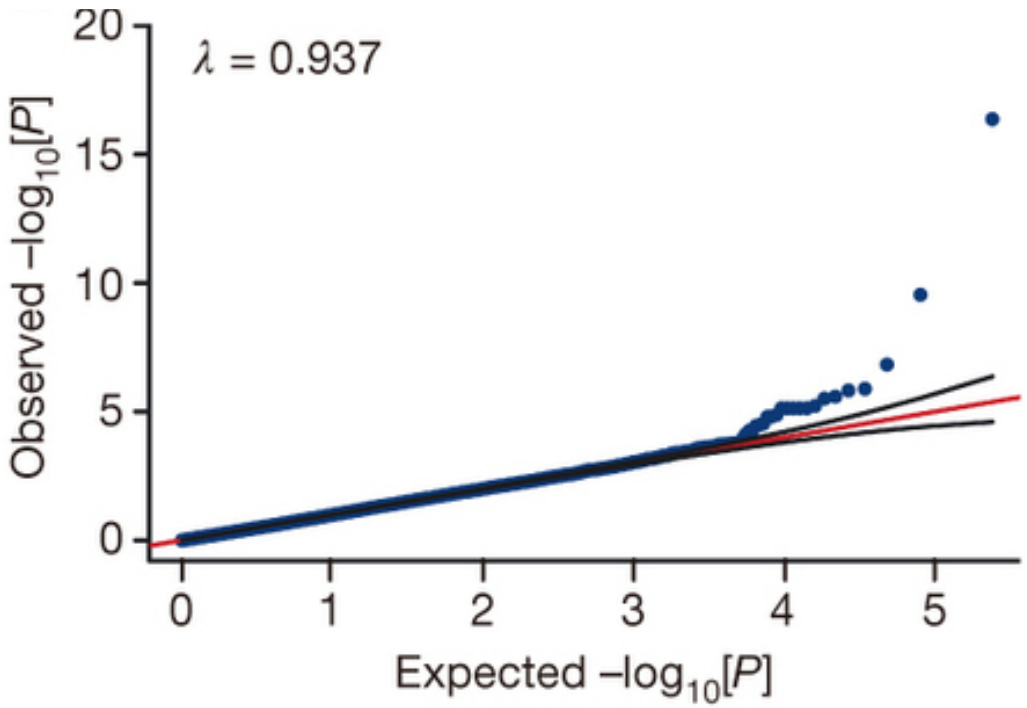


Exercise

Solve exercise A, i.e. run your first GWAS! :)

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 B, i.e. check if your results look OK...

Step 3 again: Lots and lots of QC

Now you have seen why one shouldn't skip QC...! :)

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

Sample mislabeling?

- One thing that can go wrong is that the samples can be mislabeled
- 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 PLINK
- **Exercise:** try checking it for your data (exercise C)

Closely related individuals or duplicates?

- Almost all association tests 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 PLINK (again)
- **Exercise:** try checking it for your data (exercise D)

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 (if there is time):** try checking it for your data (exercise E & F)

Additional important checks?

- Other additional signs of something being wrong include:
  - high missingness in specific loci/individuals
  - loci out of Hardy-Weinberg Equilibrium (in controls)
- 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 G)

Replication in a different dataset

- Some of the first GWAS results later turned out to be false positives
- Almost impossible to publish without replication now
- Basically consists of repeating the test of the SNP in a different dataset
- To make sure the result is not just a false positives
- This time correction for multiple testing not needed (unless you replicate more than one variant)

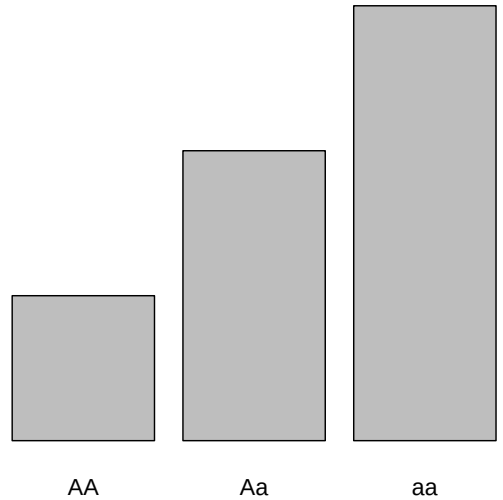
Outline

Confounding factors

- A few factors can sometimes confound your results
- One of these is gender
- We can correct for this by using a slightly more sophisticated test, namely logistic regression, which allows you to include "covariates" in your test and thereby take them into account
- If you want an example have a look at "Extra exercise if time allows"

Population structure

- All presented tests assume samples **are from one homogeneous population**
- If they are not this can lead to false positives
- E.g. if we look at height and mix Pygmies and Dutch in a GWAS.
- All loci where Pygmies mainly have *A* and Dutch *a* are associated:



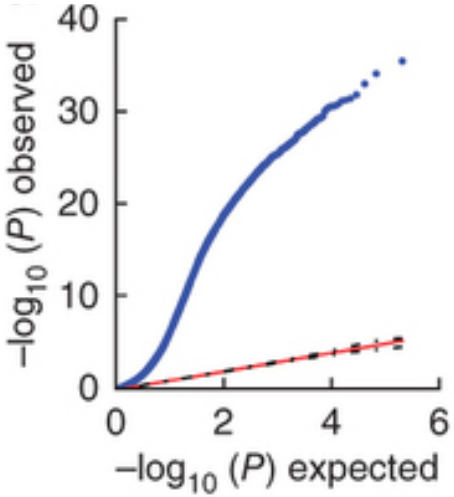
- Similar effect of **admixture and relatedness**

How can we deal with this?

Two ways

- Quality filtering: remove admixed individuals in QC
- Use other tests:
  - include first 5-10 PCs as covariates
  - a linear mixed model (extension of linear model)

QQ plot (linear model)



QQ plot (linear mixed model)