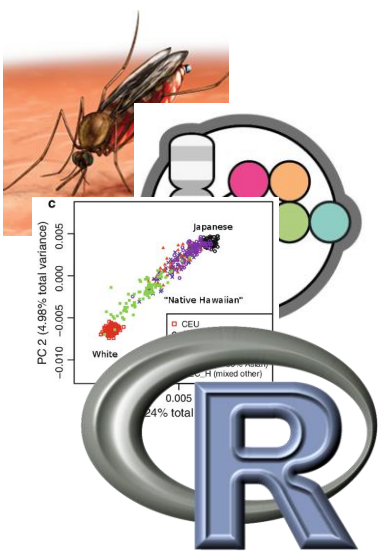
Selection: phenotypes and genotypes

Andrea Manica

UNIVERSITY OF CAMBRIDGE

Outline

- Some classic examples
- Selection scans vs GWAS
- Challenges
- Practical

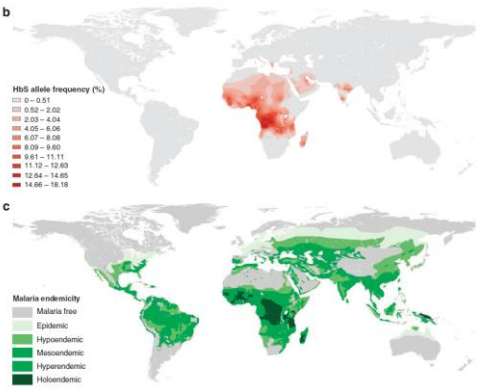


Some “classics”

Sickle cell anaemia – **heterozygote advantage** in the presence of malaria

But is this the driver of this mutation? Beware of **“just so stories”**

Recent test suggest significant association in terms of geographic spread in Africa, but not Asia or Americas



Piel et al Nat Comm (2010)

Some “classics”

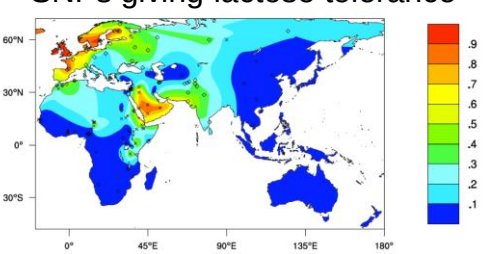
Lactose tolerance – ability to digest milk in adulthood

Multiple mutations giving the same phenotype

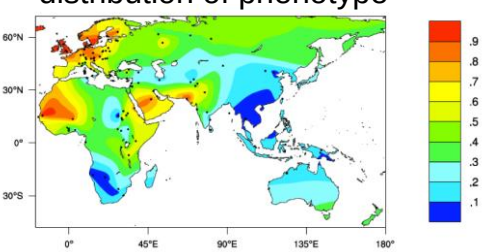
European version suggested to arrive with Neolithic farmers

But high frequency only following the Bronze age arrival of Steppe ancestry

SNPs giving lactose tolerance



distribution of phenotype

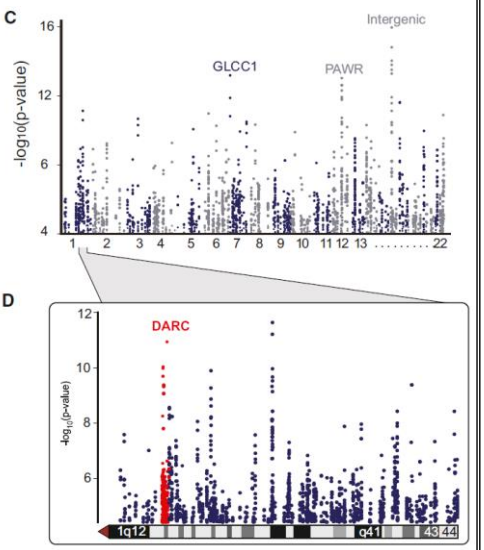


Selection in AMHs populations

Many scans for regions under selection

Look for genomic regions with **unusual characteristics** that are likely signatures of selection

Many selection statistics, with different properties. Some focus on individual loci, others on haplotypes



Grossman et al Cell (2013)

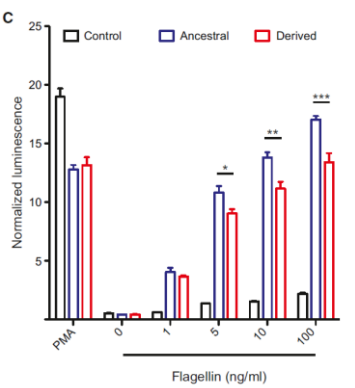
Problem with selection scans

Different scans give **very different lists** of loci under selection:

- Different statistics detecting different types of selection events?
- Different population panels?
- Lots of false positives?

Ideally use **functional studies** to confirm candidates (e.g. Toll-Like Receptor TLR5 response to flagellin in transgenic mice)

But **expensive** and **time** consuming



Grossman et al Cell (2013)

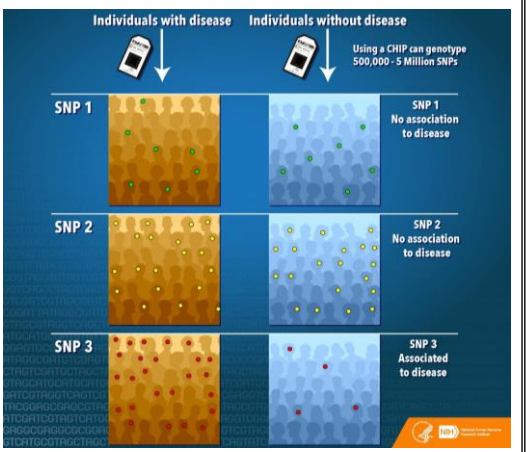
Genome Wide Association Studies

For **known phenotypes**, we can look for Single Nucleotide Polymorphisms (SNPs) associated with them

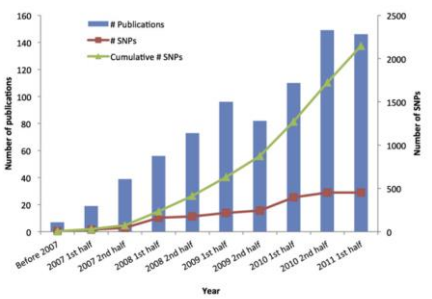
GWAS assume that common variants have important effects

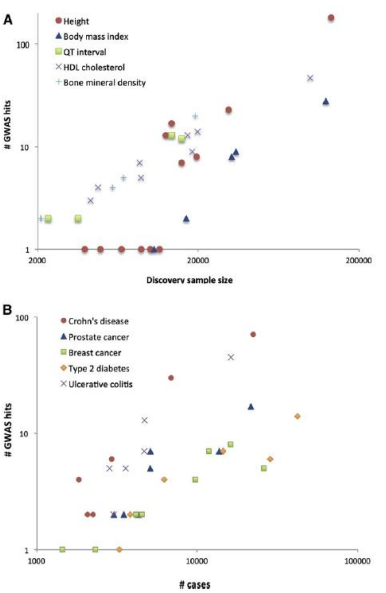
Big debate on whether they delivered useful information

[...] the bulk of heritability in these conditions cannot be ascribed to loci that have emerged from GWAS [...]
Sir Alec Jeffreys



Genome Wide Association Studies





New SNPs being added constantly

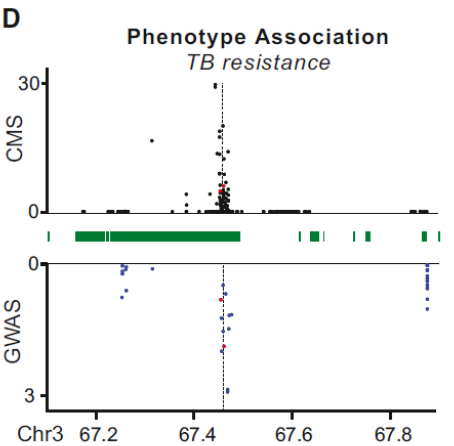
Sample size matters a lot

Most SNPs have VERY small effects

Visscher et al AJHG (2012)

Selection in AMHs populations

Compare results to GWAS: large hits from selection scans often associated with a SNP from GWAS (but smaller hits not as easily classified)



Grossman et al Cell (2013)

GWAS – the basics

Simple summary for a binary phenotype

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

H_0: Genotypes and phenotypes are independent

H_1: Genotypes frequencies differ between cases and controls

Simple **contingency table**, χ^2 test

It assumes codominance

GWAS – the basics

Additive model – simple allele test

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

Simple 2x2 table, χ^2 test

	a	A
Cases	$2*r_0+r_1$	$2*r_2+r_1$
Control	$2*s_0+s_1$	$2*s_2+s_1$

Fancier approaches, e.g. Cochran-Armitage test

Continuous phenotypes

Continuous phenotypes – e.g. BMI, cholesterol, etc.

We can model a Guassian response with a simple linear model (regression) framework:

$Y = \alpha + \beta X + \varepsilon$

where X defines the genotype (# of A alleles in an individual), and ε is the error term

$\beta \neq 0$ means that the number of A alleles is a predictor of the phenotype

The advantage of the regression framework is that we could also include **additional covariates** (smoking, # hours of exercise, etc.) that might affect the phenotype.

Regression and binary phenotypes

Binary responses can be modelled in a General Linear Model framework

Consider binary response:
Develop Coronary Heart Disease (CHD) vs Did NOT develop CHD

Y is the probability that an individual developed CHD

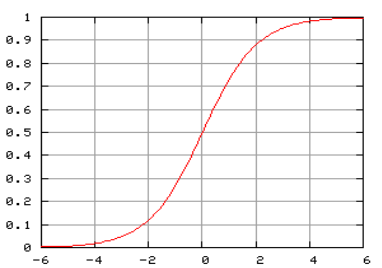
Phrase *Y* in terms of odds (*Y* to 1-*Y*):
Y=0.2 => 1 to 4

Regression and binary phenotypes

We can model the ln of the odds ratio (called a **logit**) as a linear response

$$\ln\left(\frac{Y}{1-Y}\right) = \alpha + \beta X + \varepsilon$$

$$Y = \frac{1}{1+e^{-(\alpha+\beta X+\varepsilon)}} = \frac{1}{1+e^{-z}}$$



Logistic equation

Regression and binary phenotypes

We could easily expand the model to include **additional covariates** besides the genotype information

$$Y = \frac{1}{1 + e^{-(\alpha+\beta_1 X+\beta_2 smoker+\beta_3 BMI+\varepsilon)}} = \frac{1}{1 + e^{-z}}$$

And by rephrasing *X*, we could model different types of **dominance**.

Challenges – Multiple testing

- A **type I error** occurs when we reject the H_0 of no association, when in fact the null hypothesis is true.
- It is important to correct for **multiple testing** to maintain the type I error rate for the experiment overall (i.e. all the SNPs tested in the association study).
- Several solutions:
  - **Bonferroni correction** (0.05/#SNPs)
  - **False Discovery Rate**
  - **Randomisation tests**
- **Replication** is necessary to confirm association.

Challenges - Stratification

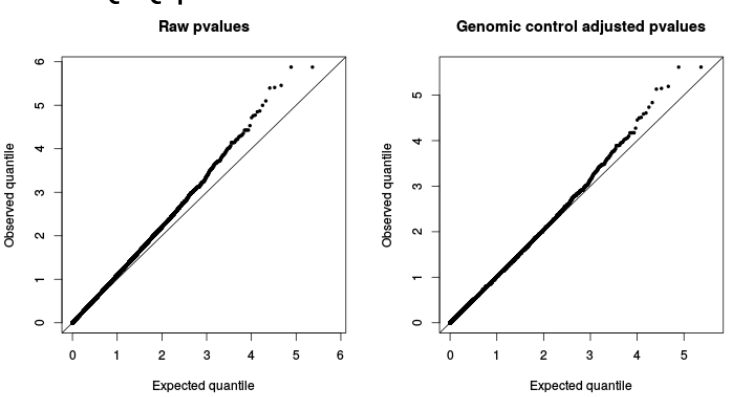
Stratification (e.g. from population structure) can generate spurious results.

For small levels of stratification, we can use the **genomic control** approach

Under the assumption that most loci are neutral and the ones under selection have relatively small effects, we expect a χ^2 distribution with 1df for our test statistics.

Challenges - Stratification

Look at a Q-Q plot



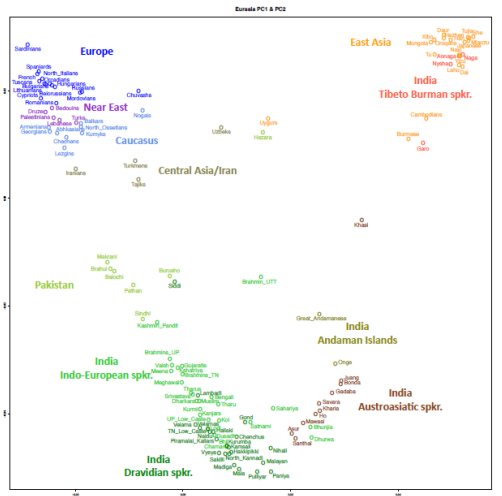
λ is the median of the test scores divided by the median of the expected χ^2_1

We can use **λ**, if ≠1, as a **correction** factor

Challenges - Stratification

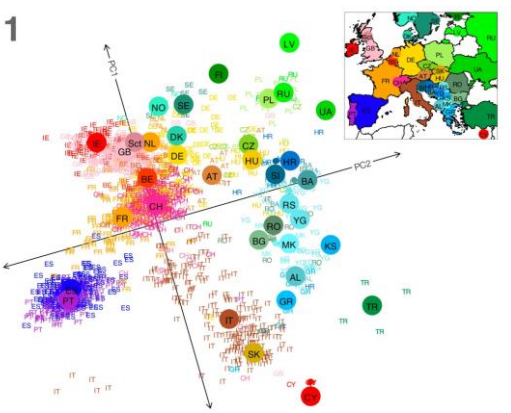
Genomic control not appropriate for strong stratification.

If **discrete clusters**, then we could simply compute stats within each stratum and pool stats.



Challenges - Stratification

If stratification not discrete, then we can use **PCA** to provide information on the relationships between samples.



Use PCs as covariates to account for stratification

$$Y = \frac{1}{1 + e^{-(\alpha+\beta_1 X+\beta_2 PC1+\beta_3 PC2+\varepsilon)}}$$

The importance of Quality Control

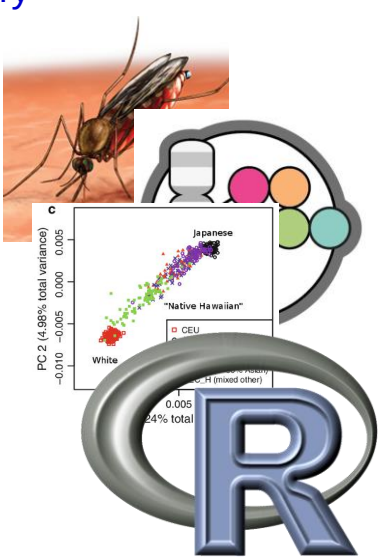
Data QC is **ESSENTIAL!** Significances are meaningless if assumptions are broken.

Olivier covered the key issues on the first day. You should always thoroughly clean your data before you do **ANY analysis** (GWAS is particularly affected by QC, but other approaches too!)

HWE is a key assumption, and something easy to test. Deviations from HWE are a good indication that your data might be problematic (but not always!).

Summary

- Some classic examples
- Selection scans vs GWAS
- Challenges
- Practical



Practical

- Use the R package GenABEL for GWAS
- Get an overview of your data
- Run a simple GWAS
- Run basic QC on data and see how it affects your results
- Correct for stratification in a few ways