

# Genome-wide association studies (GWAS) - Part 1

Heather J. Cordell

Population Health Sciences Institute  
Faculty of Medical Sciences  
Newcastle University, UK

[heather.cordell@ncl.ac.uk](mailto:heather.cordell@ncl.ac.uk)



# Genome-wide association studies (GWAS)

- Popular (and highly successful) approach over past  $\sim$  15 years
- Enabled by advances in high-throughput (microarray-based) genotyping technologies
- Idea is to measure the genotype at a set of single nucleotide polymorphisms (SNPs) across the genome, in a large set of **unrelated** individuals
  - Cases and controls
  - Or population cohort measured for relevant quantitative phenotypes (height, weight, blood pressure etc)
  - Or **related** individuals (family data) – but need to analyse differently

# Genome-wide association studies (GWAS)

## Two individuals

Person 1    ACCTGTGTGCCCAATGGCGTCCCATACTATCGG  
             ACCTGTGCGCCCAATGGCGTCCCATACTATCGG

Person 2    ACCTGTGCGCCCAGTGGCGTCCCATACTATCGG  
             ACCTGTGCGCCCAGTGGCGTCCCATAAGTATCGG

- Test each SNP for association/correlation with disease or quantitative phenotype

# Association testing: case/control studies

- Collect sample of affected individuals (cases) and unaffected individuals (controls)
  - Or else a sample of random “population” controls
    - Most of whom will not have the disease of interest
- Examine the association (correlation) between alleles present at a genetic locus and presence/absence of disease
  - By comparing the distribution of genotypes in affected individuals with that seen in controls

# Case/control studies

- Each person can have one of 3 possible genotypes at a diallelic genetic locus

Genotype	Cases	Controls
2 2	500 (= a)	200 (= b)
1 2	1100 (= c)	820 (= d)
1 1	400 (= e)	980 (= f)
Total	2000	2000

# Case/control studies

- Each person can have one of 3 possible genotypes at a diallelic genetic locus

Genotype	Cases	Controls
2 2	500 (= a)	200 (= b)
1 2	1100 (= c)	820 (= d)
1 1	400 (= e)	980 (= f)
Total	2000	2000

- Test for association (correlation) between genotype and presence/absence of disease using standard  $\chi^2$  test for independence on 2 df

# Case/control studies

- Each person can have one of 3 possible genotypes at a diallelic genetic locus

Genotype	Cases	Controls
2 2	500 (= a)	200 (= b)
1 2	1100 (= c)	820 (= d)
1 1	400 (= e)	980 (= f)
Total	2000	2000

- Test for association (correlation) between genotype and presence/absence of disease using standard  $\chi^2$  test for independence on 2 df
  - Defined as  $\sum_{i=1,6} \frac{(O_i - E_i)^2}{E_i}$  where  $O_i$  and  $E_i$  are observed and expected counts (calculated from the row and column totals) respectively
  - Generates a *p value* indicating how significant the association/correlation appears to be

# Case/control studies

- Each person can have one of 3 possible genotypes at a diallelic genetic locus

Genotype	Cases	Controls
2 2	500 (= a)	200 (= b)
1 2	1100 (= c)	820 (= d)
1 1	400 (= e)	980 (= f)
Total	2000	2000

- Test for association (correlation) between genotype and presence/absence of disease using standard  $\chi^2$  test for independence on 2 df
  - Defined as  $\sum_{i=1,6} \frac{(O_i - E_i)^2}{E_i}$  where  $O_i$  and  $E_i$  are observed and expected counts (calculated from the row and column totals) respectively
  - Generates a *p value* indicating how significant the association/correlation appears to be

- Two odds ratios can be estimated

- $OR(2|2 : 1|1) = \frac{af}{be}$
- $OR(1|2 : 1|1) = \frac{cf}{de}$

# Odds ratios

- Odds of disease are defined as  $P(\text{diseased})/P(\text{not diseased})$ 
  - Odds ratio OR(2|2 : 1|1) represents the factor by which your **odds** of disease must be multiplied, if you have genotype 2|2 as opposed to 1|1
    - i.e. the 'effect' of genotype 2|2

# Odds ratios

- Odds of disease are defined as  $P(\text{diseased})/P(\text{not diseased})$ 
  - Odds ratio OR(2|2 : 1|1) represents the factor by which your **odds** of disease must be multiplied, if you have genotype 2|2 as opposed to 1|1
    - i.e. the 'effect' of genotype 2|2
- Similarly, we can define the OR for 1|2 vs 1|1
  - As the factor by which your odds of disease must be multiplied, if you have genotype 1|2 as opposed to 1|1
    - i.e. the 'effect' of genotype 1|2

# Odds ratios

- Odds of disease are defined as  $P(\text{diseased})/P(\text{not diseased})$ 
  - Odds ratio OR(2|2 : 1|1) represents the factor by which your **odds** of disease must be multiplied, if you have genotype 2|2 as opposed to 1|1
    - i.e. the 'effect' of genotype 2|2
- Similarly, we can define the OR for 1|2 vs 1|1
  - As the factor by which your odds of disease must be multiplied, if you have genotype 1|2 as opposed to 1|1
    - i.e. the 'effect' of genotype 1|2
- ORs are closely related (often  $\approx$ ) genotype relative risks
  - The factor by which your **probability** of disease must be multiplied, if you have genotype 1|2 as opposed to 1|1 (say)
- If your genotype has no effect on your probability (and therefore on your odds) of disease, then the ORs=1.
  - So the association test can be thought of as a test of the null hypothesis that the ORs=1

## Genotype relative risks

- If a disease is reasonably rare, the odds ratio approximates the genotype relative risk (GRR, RR)

Genotype	Penetrance	GRR	Odds	OR
1/1	0.01	1.0	$0.01/0.99 = 0.0101$	1.00
1/2	0.02	2.0	$0.02/0.98 = 0.0204$	2.02
2/2	0.05	5.0	$0.05/0.95 = 0.0526$	5.21

- If your genotype has no effect on your probability (and therefore your RR) of disease, then both the ORs and the GRRs=1.

# Dominant/recessive effects

Dominant:

Genotype	Cases	Controls	Total
2 2 and 1 2	500+1100	200+820	700+1920
1 1	400	980	1380
Total	2000	2000	4000

Recessive:

Genotype	Cases	Controls	Total
2 2	500	200	700
1 2 and 1 1	1100+400	820+980	1920+1380
Total	2000	2000	4000

- Can also rearrange table to examine effects of alleles (1 df tests):

# Counting alleles

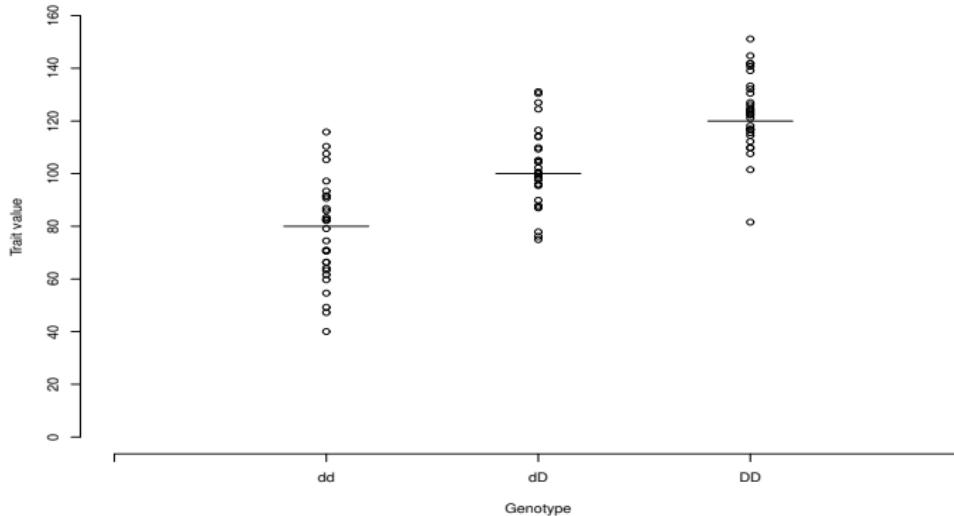
Allele	Counts in	
	Cases	Controls
2	2100 (=a)	1220 (=b)
1	1900 (=c)	2780 (=d)
Total	4000	4000

$$\text{Allelic OR} = ad/bc$$

- $\chi^2$  test statistic on 1 df =  $\sum_i (O_i - E_i)^2/E_i$  where  $O_i$  and  $E_i$  are the observed and expected values in cell  $i$ .
  - Assumes HWE under null and multiplicative allelic effects under alternative: considers chromosomes as independent units
  - **Better approach:** use counts in previous genotype table to perform a Cochran-Armitage trend test
  - **Even better approach:** use linear or logistic regression

# Testing for association: quantitative traits

- Linear regression provides a natural test for quantitative traits
  - Testing the null hypothesis that the slope = 0



# Logistic regression

- Used in case/control studies
  - Outcome is affected or unaffected
  - Model probability (and thus odds) of disease  $p$  as function of variable  $x$  coding for genotype:

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x \equiv c + mx$$

- Use observed genotypes in cases and controls to estimate the values of regression coefficients  $\beta_0$  and  $\beta_1$ 
  - And to test whether  $\beta_1 = 0$

# Logistic regression

- Standard method used in standard epidemiological studies e.g. of risk factors such as smoking in lung cancer
- Main advantage is you can include **more than one predictor** in the regression equation e.g.

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where  $x_1, x_2, x_3$  code for

- genotypes at 3 loci
- measured environmental covariates (e.g. age, sex, smoking etc),
- genetic principal component scores (to adjust for population substructure),
- interactions between loci etc. etc.

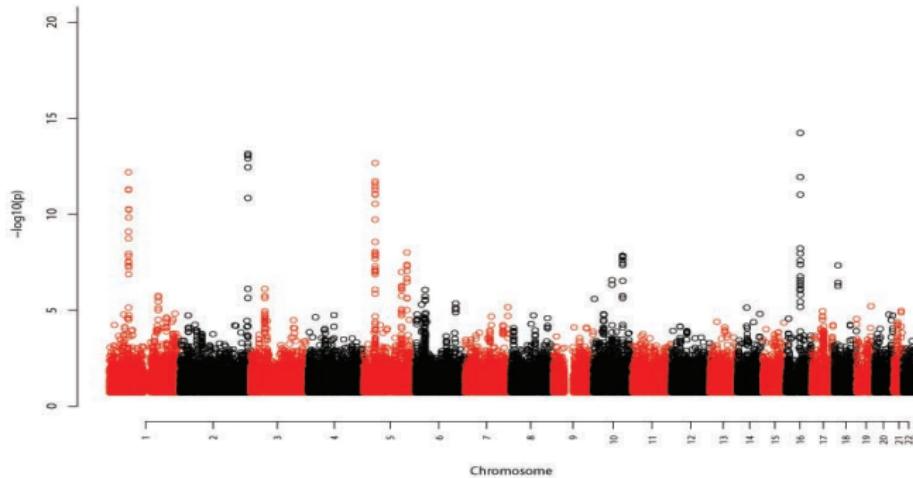
# Testing for association

- All methods produce a **test statistic** and a **p value** at each SNP, indicating how significant the association/correlation observed appears to be
  - i.e. how likely it was to have occurred by chance
  - The threshold to declare ‘genome-wide significance’ is usually around  $p = 5 \times 10^{-8}$
  - To account for multiple testing of many SNPs across the genome

# Testing for association

- All methods produce a **test statistic** and a **p value** at each SNP, indicating how significant the association/correlation observed appears to be
  - i.e. how likely it was to have occurred by chance
  - The threshold to declare 'genome-wide significance' is usually around  $p = 5 \times 10^{-8}$ 
    - To account for multiple testing of many SNPs across the genome
- Alternative (Bayesian) methods produce a **Bayes Factor**
  - Indicates how likely the data is under the alternative hypothesis (of **association** between genotype and phenotype)
    - Compared to under the null hypothesis (of **no association** between genotype and phenotype)
  - Requires you to make some prior assumptions regarding the likely strength of associations (i.e. the value of the  $\beta$ 's)
  - Choosing a sensible threshold (e.g.  $\log_{10} BF > 4$ ) requires you to make some prior assumptions regarding what proportion of SNPs in the genome are likely to be associated with the phenotype

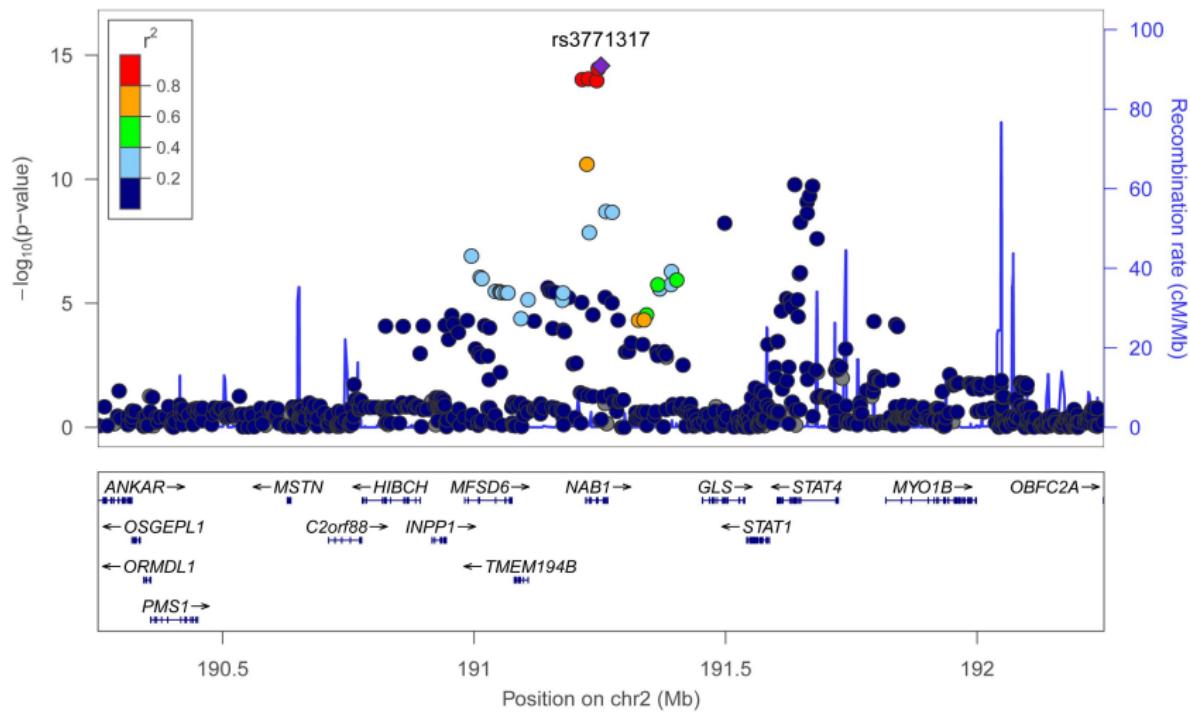
# Manhattan Plots



- At any location showing 'significant' association, we expect to see several SNPs in the same region showing association/correlation with phenotype
  - Due to the correlation or **linkage disequilibrium** (LD) between neighbouring SNPs

# Close-up of hit region

Plotted SNPs



## Historical Perspective: Complement Factor H in AMD

- First (?) GWAS was by Klein et al. (2005) Science 308:385-389
- Typed 116,204 SNPs in 96 cases (with age-related macular degeneration, AMD) and 50 controls
  - Very small sample size – they were very lucky to find anything!
  - Luck was due to the fact the polymorphism has a very large effect (recessive OR=7.4)
- Klein et al. followed up on two SNPs passing threshold ( $p < 4.8 \times 10^{-7}$ )
  - Plus a third SNP that just failed to pass significance threshold, but lay in same region as first SNP

# Complement Factor H in AMD

- Of the 3 SNPs followed up:
  - One appeared to be due to genotyping errors: significance disappeared on filling in some missing genotypes
  - First and third SNP lie in intron of Complement Factor H (*CFH*) gene
    - Lies in region previously implicated by family-based linkage studies
- Resequencing of the region identified a polymorphism of plausible functional effect
- Immunofluorescence experiments in the eyes of AMD patients supported the involvement of *CFH* in disease pathogenesis.

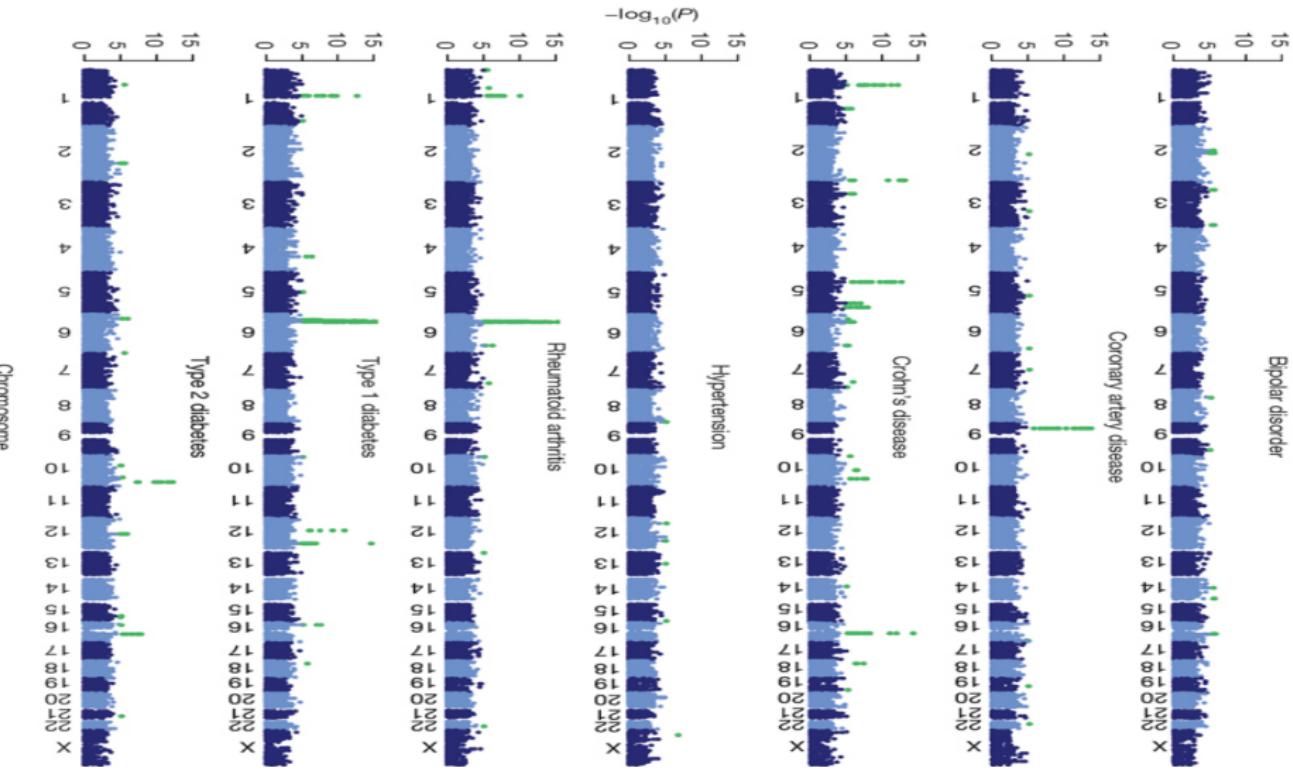
# GWAS

- GWAS really got going in around 2007
  - Visscher et al. (2012) AJHG 90:7-24 "Five Years of GWAS Discovery"
  - Visscher et al. (2017) AJHG 101:5-22 "10 Years of GWAS Discovery: Biology, Function and Translation"
  - Abdellaoui et al. (2023) AJHG 110:179-194 "15 Years of GWAS Discovery: Realizing the promise"
- 2007/2008 saw a slew of high-profile GWAS publications
  - Breast cancer (Easton et al. 2007)
  - Rheumatoid Arthritis (Plenge et al. 2007)
  - Type 1 and Type 2 diabetes (Todd et al. 2007; Zeggini et al. 2008)
- Arguably the most influential was the Wellcome Trust Case Control Consortium (WTCCC) study of 7 different diseases
  - <http://www.wtccc.org.uk/>

# WTCCC

- Nature 447: 661-678 (2007)
- Considered 2000 cases for each of the following diseases:
  - Bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, type 2 diabetes
- Compared each disease cohort to common control panel
  - 3000 population-based controls
  - From 1958 birth cohort and National Blood Service
- Highly successful
  - WTCCC found 24 separate association signals
  - Including highly convincing signals in 5 out of the 7 diseases studied
  - All were replicated in subsequent independent follow-up studies

# Manhattan plots for 7 diseases



## Lessons from WTCCC (and others)

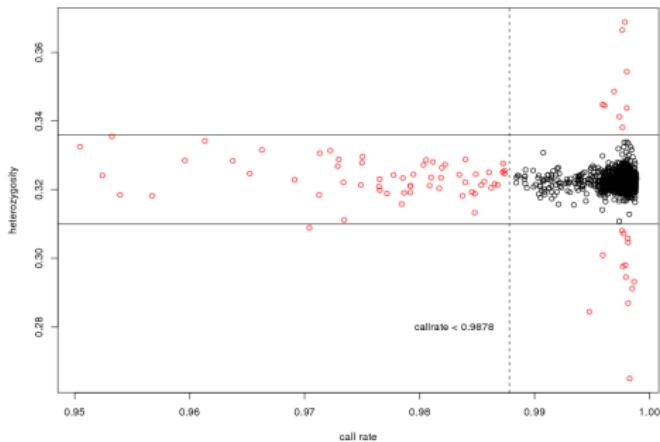
- Typically used rather standard statistical/epidemiological methods ( $\chi^2$  tests, *t* tests, logistic regression etc.)
- Success largely due to:
  - An appreciation of the importance of **large sample size** ( $> 2000$  cases, similar or greater number of controls)
  - Stringent **quality control** procedures for discarding low-quality SNPs and/or samples
  - Stringent **significance thresholds** ( $p=5\times 10^{-8}$ ) to account for multiple testing and/or low prior prob of true effect
  - Importance of **replication** in an independent data set

# Short break

# Quality Control

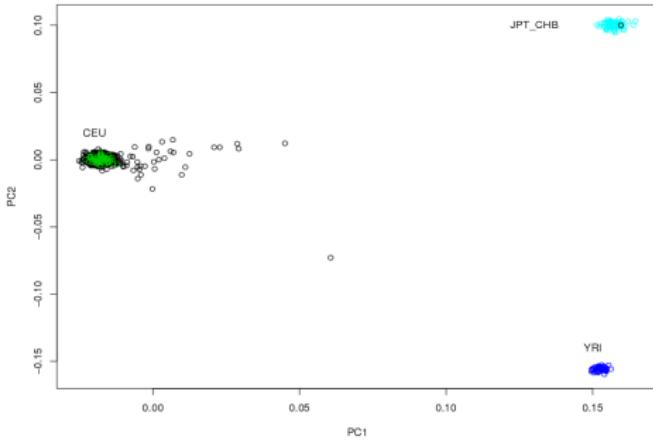
- Stringent QC checks are required for GWAS data
- Discard samples (people) deemed unreliable
  - Low genotype call rates, excess heterozygosity etc.
  - X chromosomal markers useful for checking gender
    - Males should 'appear' homozygous at all X markers
  - Genome-wide SNP data useful for checking relationships and ethnicity
- Discard data from SNPs deemed unreliable
  - On basis of genotype call rates, Mendelian misinheritances, Hardy-Weinberg disequilibrium
  - Exclude SNPs with low minor allele frequency (MAF)

# QC: call rates and heterozygosity



- 61 sample exclusions (low call-rate); 23 exclusions (heterozygosity)
- SNP exclusions also made based on call-rates, MAF and Hardy-Weinburg equilibrium (HWE)

# QC: ethnicity tests



- Multidimensional scaling (with 210 HapMap individuals) identifies 33 samples with non-Caucasian ancestry
- MDS or similar multivariate methods can also be used to model more subtle population differences between samples...

# Multivariate Analysis

- Several related multivariate analysis techniques have been proposed for detecting **population structure** in genome-wide association studies
  - Principal components analysis (PCA)
  - Principal coordinates analysis (PCoA)
  - Multidimensional scaling (MDS)

# Multivariate Analysis

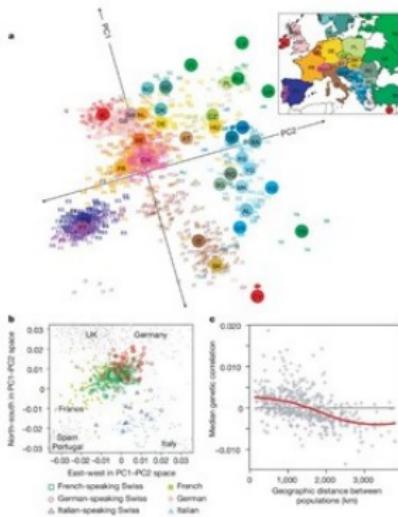
- Several related multivariate analysis techniques have been proposed for detecting **population structure** in genome-wide association studies
  - Principal components analysis (PCA)
  - Principal coordinates analysis (PCoA)
  - Multidimensional scaling (MDS)
- If population differences can be detected (and adjusted for) in association analysis, this offers a way to deal with the problem of **population stratification**
  - Population sampled actually consists of several ‘sub-populations’ that do not really intermix
  - Can lead to spurious false positives (type 1 errors) in case/control studies

# Multivariate Analysis

- Several related multivariate analysis techniques have been proposed for detecting **population structure** in genome-wide association studies
  - Principal components analysis (PCA)
  - Principal coordinates analysis (PCoA)
  - Multidimensional scaling (MDS)
- If population differences can be detected (and adjusted for) in association analysis, this offers a way to deal with the problem of **population stratification**
  - Population sampled actually consists of several ‘sub-populations’ that do not really intermix
  - Can lead to spurious false positives (type 1 errors) in case/control studies
- These techniques can also be used in quality control (QC) procedures, to check for (and discard) gross population outliers

# Principal components analysis (PCA)

Genes mirror geography within Europe



J Novembre *et al.* (2008) *Nature* **456**(7218):98-101, doi:10.1038/nature07331

# Principal Components Analysis

- Price et al. (2006) Nature Genetics 38:904-909; Patterson et al. (2006) PLoS Genetics 2(12):e190
  - Based on popn genetics ideas from Cavalli-Sforza (1978)
- Idea is to form a large matrix  $M$  of SNP counts (0,1,2) corresponding to the genotype at a  $L$  loci (=rows) for  $n$  individuals (=columns)

$$M = \begin{pmatrix} g_{11} & g_{12} & \cdot & g_{1n} \\ g_{21} & g_{22} & \cdot & g_{2n} \\ g_{31} & g_{32} & \cdot & g_{3n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ g_{L1} & g_{L2} & \cdot & g_{Ln} \end{pmatrix}$$

# Principal Components Analysis

- Subtract row means and normalise by function of row allele frequency  $\sqrt{f_l(1 - f_l)}$  to give matrix X

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ x_{31} & x_{32} & \dots & x_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{L1} & x_{L2} & \dots & x_{Ln} \end{pmatrix}$$

- This matrix will be used as starting point for PCA
  - In principle we could start with a different matrix – in particular not all PCA approaches would normalise by  $\sqrt{f_l(1 - f_l)}$

# Multivariate Analysis

- Estimate covariance matrix  $\Psi = X^T X$  between all pairs of individuals, with entries  $\psi_{ij}$  defined as the covariance (summing over SNPs) between column  $i$  and  $j$  of  $X$ 
  - Represents average genome-wide identity by descent (IBD) (estimated from identity by state, IBS)

# Multivariate Analysis

- Estimate covariance matrix  $\Psi = X^T X$  between all pairs of individuals, with entries  $\psi_{ij}$  defined as the covariance (summing over SNPs) between column  $i$  and  $j$  of  $X$ 
  - Represents average genome-wide identity by descent (IBD) (estimated from identity by state, IBS)
  - Compute the eigenvectors  $\vec{v}_j$  and eigenvalues  $\lambda_j$  of matrix  $\Psi$ 
    - Co-ordinate  $j$  of the  $k$ th eigenvector represents the ancestry of individual  $j$  along 'axis'  $k$

# Multivariate Analysis

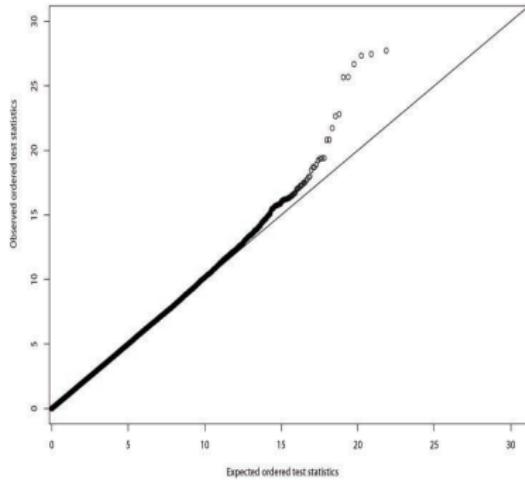
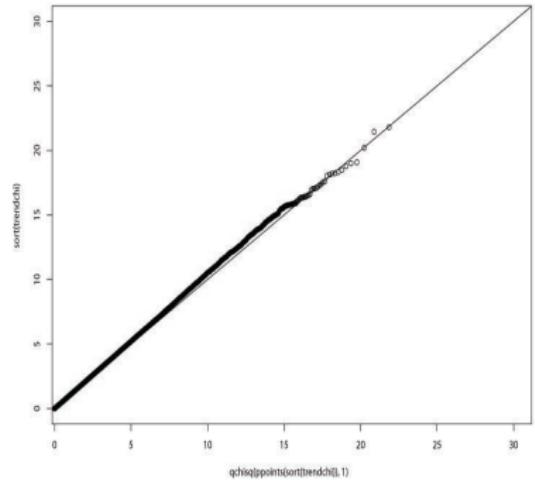
- Estimate covariance matrix  $\Psi = X^T X$  between all pairs of individuals, with entries  $\psi_{ij}$  defined as the covariance (summing over SNPs) between column  $i$  and  $j$  of  $X$ 
  - Represents average genome-wide identity by descent (IBD) (estimated from identity by state, IBS)
  - Compute the eigenvectors  $\vec{v}_j$  and eigenvalues  $\lambda_j$  of matrix  $\Psi$ 
    - Co-ordinate  $j$  of the  $k$ th eigenvector represents the ancestry of individual  $j$  along 'axis'  $k$
- For technical details, see McVean (2009) PLoS Genetics 5;10:e1000686

# Multivariate Analysis

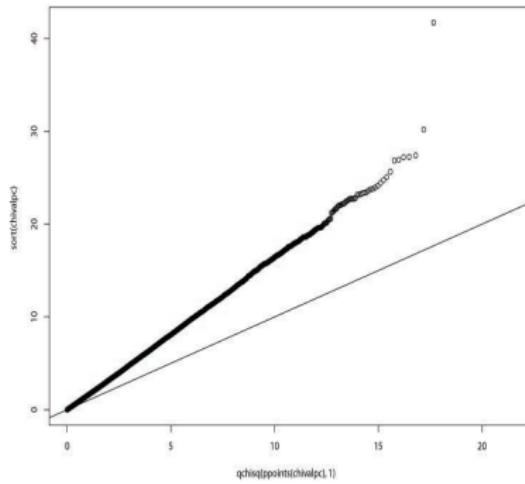
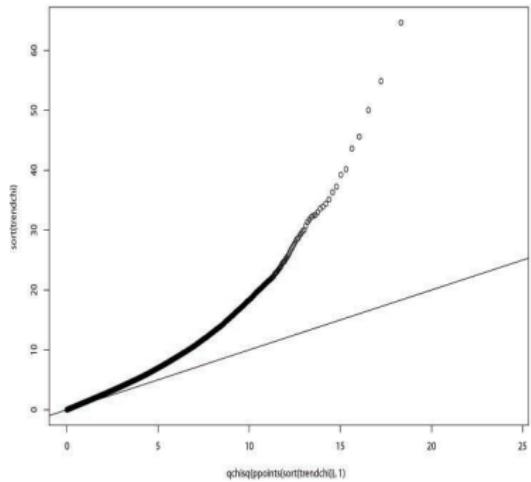
- Estimate covariance matrix  $\Psi = X^T X$  between all pairs of individuals, with entries  $\psi_{ij}$  defined as the covariance (summing over SNPs) between column  $i$  and  $j$  of  $X$ 
  - Represents average genome-wide identity by descent (IBD) (estimated from identity by state, IBS)
  - Compute the eigenvectors  $\vec{v}_j$  and eigenvalues  $\lambda_j$  of matrix  $\Psi$ 
    - Co-ordinate  $j$  of the  $k$ th eigenvector represents the ancestry of individual  $j$  along 'axis'  $k$
- For technical details, see McVean (2009) PLoS Genetics 5;10:e1000686
- Many genetics packages e.g. (PLINK) will allow you to calculate the top 10 (or more) PCs
  - Different geographic populations can often be well separated by just the first two or three PCs
    - Useful for outlier detection
  - For more subtle differences, you may need to calculate more PCs
    - And include them as covariates in the regression equation
    - Post-GWAS QC can determine whether you have included 'enough'

# Post GWAS QC: Q-Q Plots (good)

- Plot ordered test statistics (y axis) against their expected values under the null hypothesis (x axis)



# Q-Q Plots (bad)



# Population stratification

- A QQ plot showing constant inflation (straight line with slope  $> 1$ ) can indicate population stratification/population substructure
- Simple solution: Genomic Control (Devlin and Roeder 1999)
  - Use your observed test statistics to estimate the slope (=inflation factor  $\lambda$ )
  - Divide each test statistic by  $\lambda$  to get an adjusted (deflated) test statistic
- More complicated solution: use PCA/MDS or similar
- Even more complicated solution: use linear mixed models

# Relatedness

- With genome-wide data, can also infer relationships based on average identity by descent (IBD)  $\Psi = X^T X$  or identity by state (IBS)
  - Using 'thinned' subset of markers with high minor allele frequency (MAF) and in approximate linkage equilibrium
  - Simple relationships (PO, FS, MZ/duplicates) can be identified with only a few hundred markers
  - More complicated relationships require 10,000-50,000 SNPs
- Various software packages, including PLINK, KING and TRUFFLE

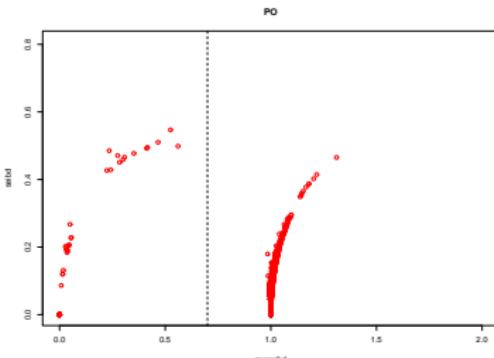
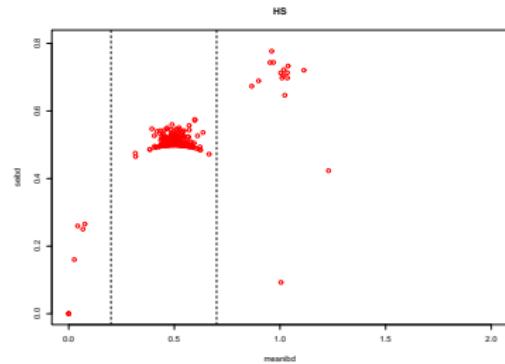
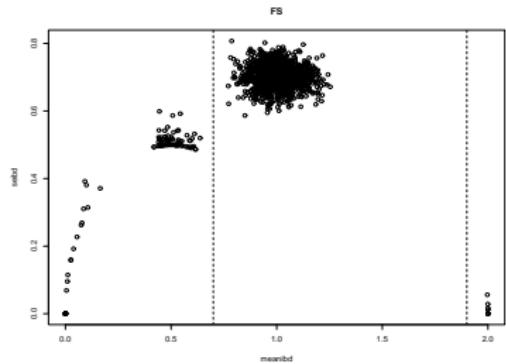
# Expected IBD sharing

- Assuming no inbreeding, the IBD state probabilities are:

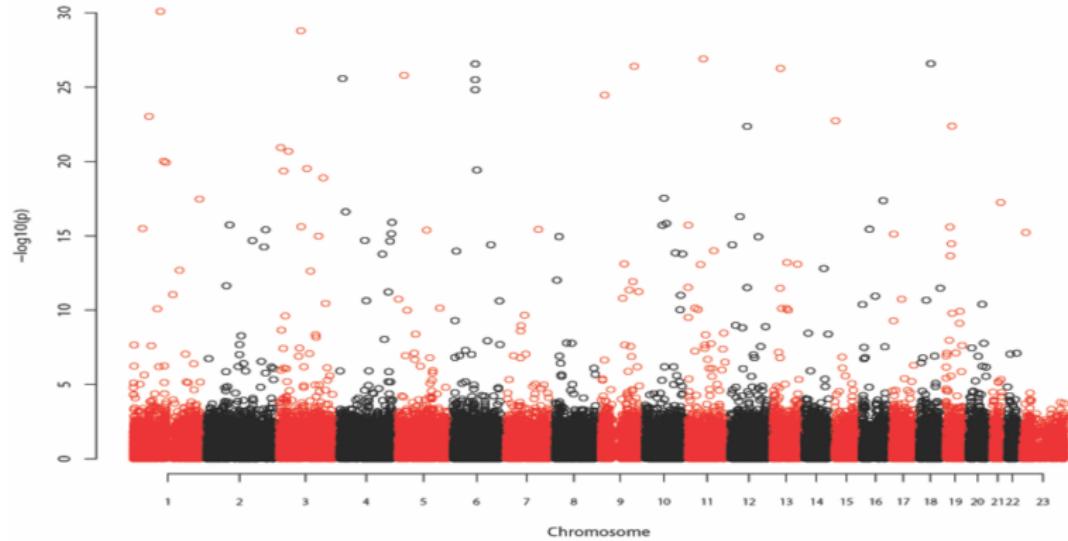
Relationship	Number of alleles shared IBD		
	2	1	0
MZ twins	1	0	0
Parent–Offspring	0	1	0
Full siblings	1/4	1/2	1/4
<i>Half siblings</i>	0	1/2	1/2
<i>Grandchild–grandparent</i>	0	1/2	1/2
<i>Uncle/aunt–nephew/niece</i>	0	1/2	1/2
First cousins	0	1/4	3/4
Second cousins	0	1/16	15/16
Double 1st cousins	1/16	6/16	9/16

- A useful visualisation tool is to plot  $SE(\text{IBD})$  vs  $\text{mean}(\text{IBD})$  (as estimated across the genome)
  - Or kinship coefficient  $\{\frac{1}{2}P(\text{IBD}=2) + \frac{1}{4}P(\text{IBD}=1)\}$  against  $P(\text{IBD}=0)$

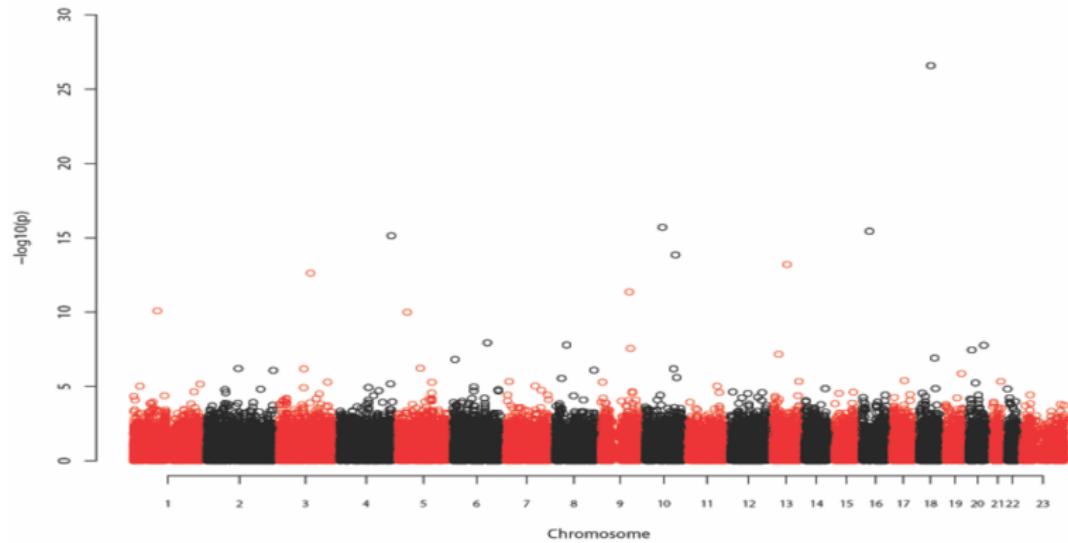
# Full/half sibs and parent-offspring



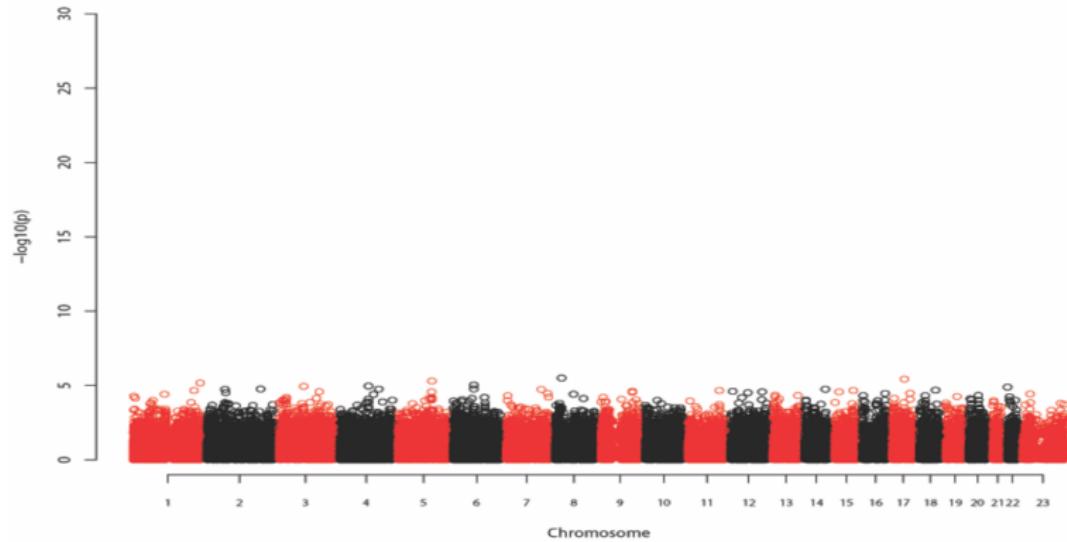
# CHD GWAS results (low QC)



# CHD GWAS results (better QC)



# CHD GWAS results (final QC)



# Genome-wide meta-analysis

- Puts together data (or results) from a number of different studies
  - Could analyse as one big study
  - But preferable to analyse using **meta-analytic** techniques
    - At each SNP construct an overall test based on the results (log ORs and standard errors) from the individual studies

# Genome-wide meta-analysis

- Puts together data (or results) from a number of different studies
  - Could analyse as one big study
  - But preferable to analyse using **meta-analytic** techniques
    - At each SNP construct an overall test based on the results (log ORs and standard errors) from the individual studies
- Meta-analysis is often made easier by using *imputation*
  - Inferring (probabilistically) the genotypes at SNPs which have not actually been genotyped
    - On the basis of their known correlations with nearby SNPs that have been genotyped
    - Using a reference panel of people (e.g. 1000 Genomes) who have been genotyped at all SNPs

# Genome-wide meta-analysis

- Puts together data (or results) from a number of different studies
  - Could analyse as one big study
  - But preferable to analyse using **meta-analytic** techniques
    - At each SNP construct an overall test based on the results (log ORs and standard errors) from the individual studies
- Meta-analysis is often made easier by using *imputation*
  - Inferring (probabilistically) the genotypes at SNPs which have not actually been genotyped
    - On the basis of their known correlations with nearby SNPs that have been genotyped
    - Using a reference panel of people (e.g. 1000 Genomes) who have been genotyped at all SNPs
- Enables meta-analysis of studies that used different genotyping platforms
  - By imputing to generate data at a **common set** of SNPs
    - Ideally while accounting for the imputation uncertainty in the downstream statistical analysis
    - In practice often don't bother - use post-imputation QC to remove poorly-imputed SNPs