Session 2: Steps involved in Genome-wide association studies

Teresa Ferreira and Nilufer Rahmioglu

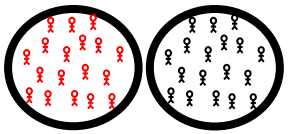
Wellcome Centre for Human Genetics/Big Data Institute

Learning objective

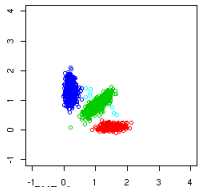
- Have a working knowledge of the different steps involved in the conduct of genome-wide association studies, including:
  - Study design
  - Quality control
  - Basic analyses
  - Replication/meta-analysis

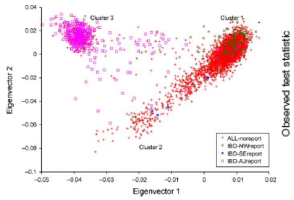
Genome-wide Association Study (GWAS) Recipe

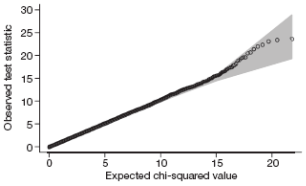
Genotype 100,000s **common** SNPs in 1000s of cases+controls



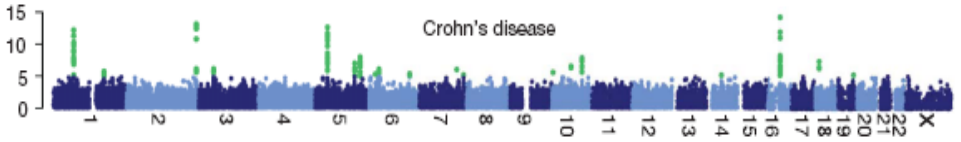
Quality-control analyses: e.g. genotype calling, population biases



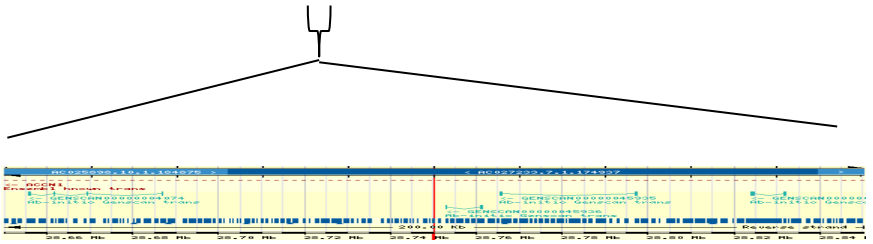




At each SNP test for allele frequency difference between cases& controls (chi^2, logistic regression)

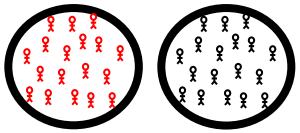


Identify significant associations, nominal p-value (5x 10^{-8})



Assess genomic info: genes, SNP density, regulatory regions, etc

Genotype selected SNPs in different case+control samples of same population: replication/meta-analysis



GWAS Study Design

Important steps

- Define the case phenotype in adequate detail
  - At least sufficient for replication studies.....
  - In practice.... paid variable attention to in GWAS

    'Lumpers' vs. 'Splitters'!

'Lumpers':
Very large case sets result in ↑ power that outweighs ↓ power due to (a degree of) misclassification and genetic heterogeneity of disease

'Splitters':
Lack of accurate phenotypic definitions result in need for much greater sample sizes, but also hide differential heterogeneity of 'subtypes'

GWAS Study Design

Important steps

- Define the case phenotype in adequate detail
  - ➢ At least sufficient for replication studies.....
  - ➢ In practice.... paid variable attention to in GWAS

    'Lumpers' vs. 'Splitters'!

'Lumpers':
Very large case sets result in ↑ power that outweighs ↓ power due to (a degree of) misclassification and genetic heterogeneity of disease

'Splitters':
Lack of accurate phenotypic definitions result in need for much greater sample sizes, but also hide differential heterogeneity of 'subtypes'

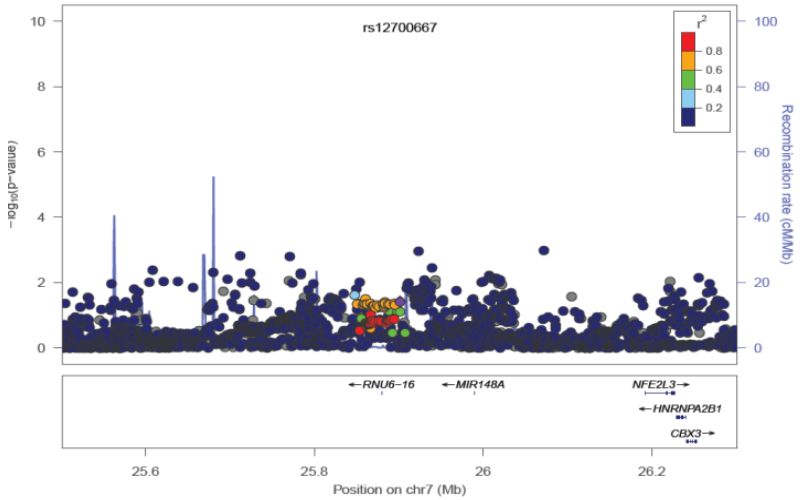
- Impact of definition accuracy likely disease/trait-specific
- Well-phenotyped datasets including sub-types disease/correlated traits are useful to dissect differential genetic origins (e.g. CAD and cholesterol levels, endometriosis and surgical stage of the disease)

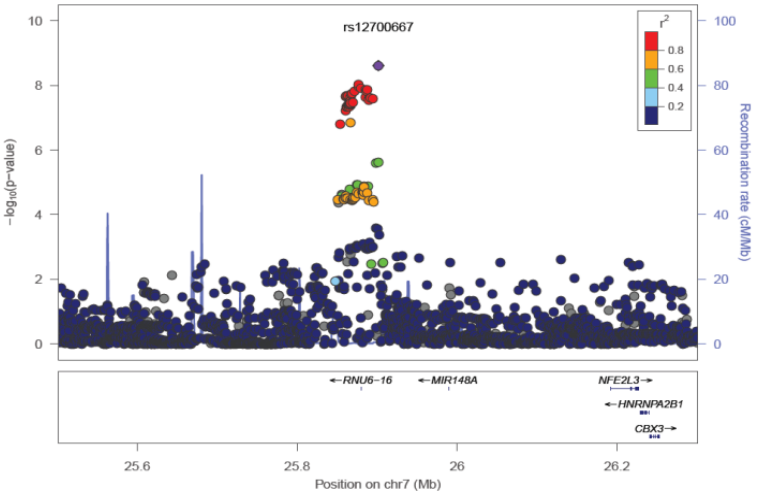
GWAS in Endometriosis (Sub-types)

Stage I/II endometriosis (1.6K cases)

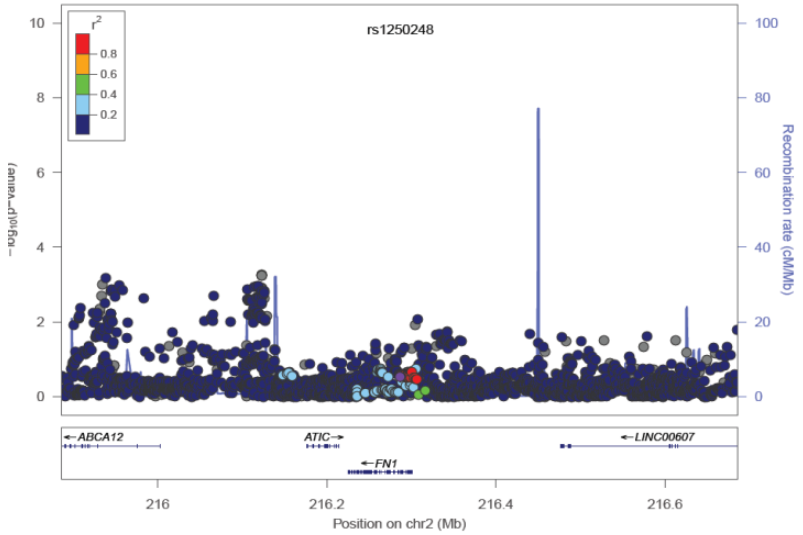
Stage III/IV endometriosis (1.3K cases)

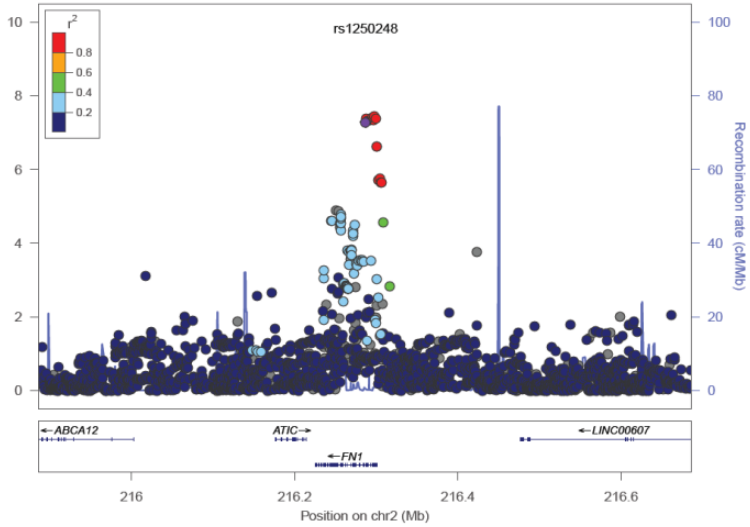
7p15.2 (rs12700667)





FN1, intronic (rs1250248)





GWAS Study Design

Important steps

- Define the case phenotype in adequate detail
- [Check the heritability of the disease in question]
  - Most diseases are ‘heritable’

GWAS Study Design

Important steps

- Define the case phenotype in adequate detail
- [Check the heritability of the disease in question]
- Consider whether a population-based study is the appropriate design for the research question
  - Are you interested in a disease sub-type that looks to be highly familial?

*

GWAS Study Design

Important steps

- Define the case phenotype in adequate detail
- [Check the heritability of the disease in question]
- Consider whether a population-based study is the appropriate design for the research question
- Select appropriate controls
  - Same ethnic population from which cases arose
  - 'Common controls' principle: publicly available genomic datasets on large numbers of individuals of a certain ancestry (e.g. WTCCC sets of blood donors and 1958BC members; many such datasets now available).
  - Female controls for female-specific conditions? Not necessary for analysis of autosomes (chromosomes 1-22)!

GWAS Study Design

Important steps

- Define the case phenotype in adequate detail
- [Check the heritability of the disease in question]
- Consider whether a population-based study is the appropriate design for the research question
- Select appropriate controls
- Calculate required sample size
  - Previous GWAS: allow for allelic ORs in the 1.1-1.5 range
  - Typically you need at least 2,000 cases (and 1:1 to 1:3 control ratio)

*

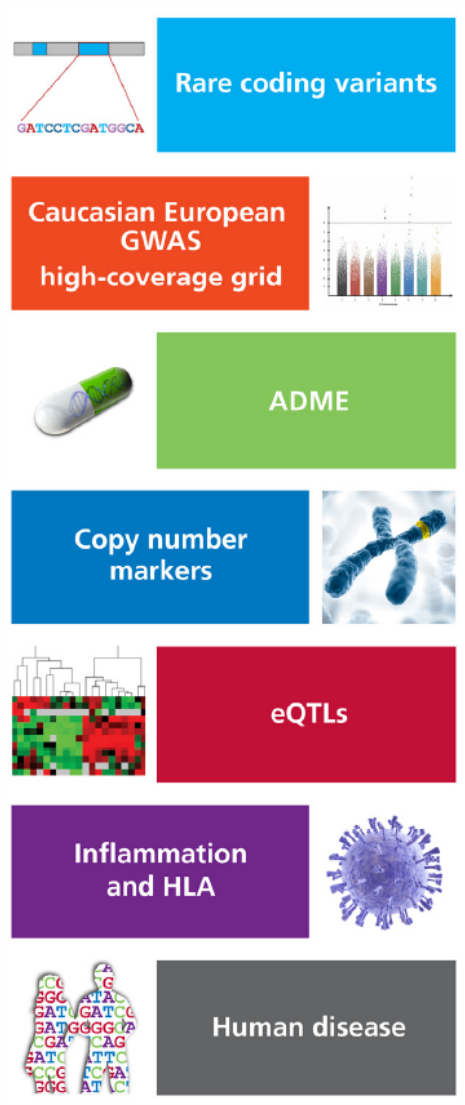
GWAS Marker (Chip) Selection

Considerations

- Many different genotyping arrays/chips on the market
- Evolved from increasingly large tagSNP panels (LD based), at ever decreasing cost
- To include supplements specific to certain analyses, as knowledge of the genomic aspects of these analyses increases
  - Low frequency and exome SNPs (following 1000G)
  - Disease-specific chips (e.g. Illumina Psych-chip/Metabo-chip)
  - Combination panels for Biobanking (e.g. Affymetrix UKBiobank Axiom 800K)

UK Biobank Axiom array

820,967 SNPs and indel markers, covering areas:



GWAS Marker (Chip) Selection

Considerations

- Many different genotyping arrays/chips on the market
- Evolved from increasingly large tagSNP panels (LD based), at ever decreasing cost
- To include supplements specific to certain analyses, as knowledge of the genomic aspects of these analyses increases
  - Low frequency and exome SNPs (following 1000G)
  - Disease-specific chips (e.g. Illumina Psych-chip/Metabo-chip)
  - Combination panels for Biobanking (e.g. Affymetrix UKBiobank Axiom 800K)
- Make sure cases and controls are genotyped on the same platform, and together (randomly distributed over chips, to avoid batch effects)

GWAS Quality Control (QC)

Important steps

- The most important part of GWAS analysis.

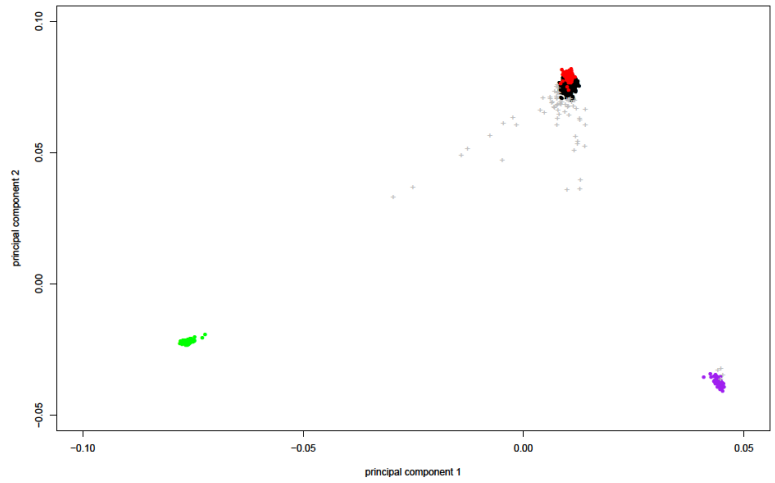
GWAS Quality Control (QC)

Important steps

- The most important part of GWAS analysis.
- Per-Individual QC:
  - Discordant sex information (comparing homozygosity rates for X chromosome)
  - Outlying missing genotypes (typically >5%) or heterozygosity rates
  - Duplicated or related individuals
  - Divergent ancestry:

    Using multivariate methods (e.g. principal component analysis). PCs each absorb decreasing amount of variance. Implemented in software to produce graphical representation of population stratification.

Remove population outliers



GWAS Quality Control (QC)

Important steps

- THE most important part of GWAS analysis.....
- Per-Individual QC:
  - Discordant sex information (comparing homozygosity rates for X chromosome)
  - Outlying missing genotypes (typically >5%) or heterozygosity rates
  - Duplicated or related individuals
  - Divergent ancestry
- Per-SNP QC:
  - Excessive missing genotype rates (call rates < 95%)
  - Deviation from HWE (significance thresholds vary, most commonly 5×10^{-7} combined with post-analysis checks of genotype cluster plots)
  - Different missing genotype rates between cases and controls
  - MAF < 1%

GWAS Analysis

Important steps

- Select genotypic disease models to test. *Multiplicative model most powerful under most disease models!*

Genetic Data Analysis – Models

'per-allele' assumes an additive genetic model, i.e. an effect for each allele copy

looks for an incremental effect across the genotype groups

Allele Counts		
	C	T
Case	2a+b	b+2c
Control	2d+e	e+2f

Cochran Armitage Test for Trend			
	CC ➔	CT ➔	TT
Case	a	b	c
Control	d	e	❖

Full Data – Genotype Counts			
	CC	CT	TT
Case	a	b	c
Control	d	e	f

'per genotype' looks for any difference across the genotype groups without making any assumptions about the direction of the effect

Dominant Model (T risk)		
	CC	CT or TT
Case	a	b+c
Control	d	e+f

Recessive Model (T risk)		
	CC or CT	TT
Case	a+b	c
Control	d+e	f

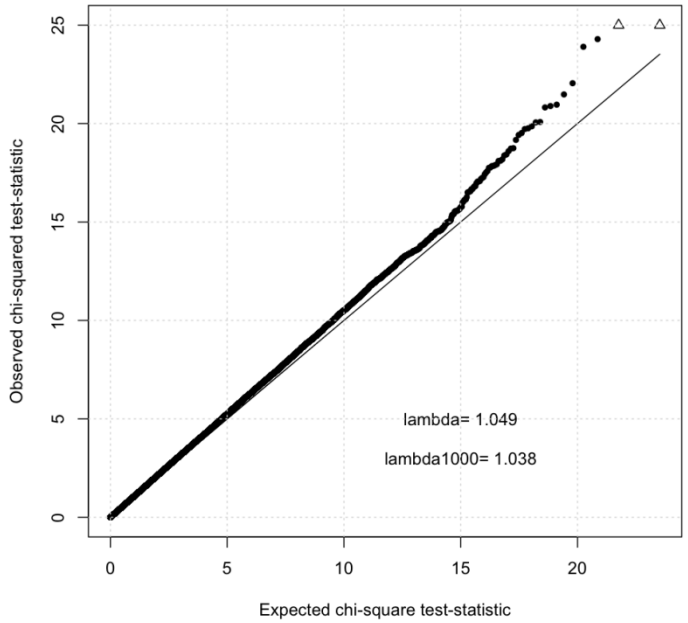
Statistical Analysis

- Logistic regression (SNPTEST/PLINK): Binary Traits
- Linear regression (SNPTEST/PLINK): Continuous Traits
- Firth-test (EPACTS): Datasets with case/control imbalance.
- Linear Mixed Model (BOLT-LMM): Datasets with related individuals

GWAS Analysis

Important steps

- Select genotypic disease models to test (e.g. multiplicative, recessive, dominant). *Multiplicative model most powerful under most disease models!*
- Post-analysis QC: QQ plots and lambda inflation score



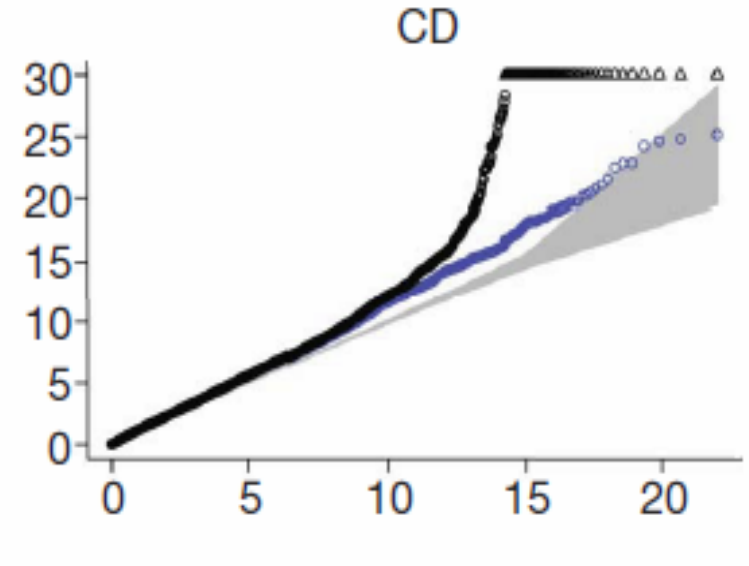
Lambda >1 :
population structure

GWAS Analysis

Important steps

- Select genotypic disease models to test (e.g. multiplicative, recessive, dominant). *Multiplicative model most powerful under most disease models!*
- Need to adjust for confounders*?*

  Principal components related to ancestry



Q-Q plot of GWAS of Crohn's Disease before (**black**) and after (**blue**) adjusting for principal components relating to ancestry (population stratification)

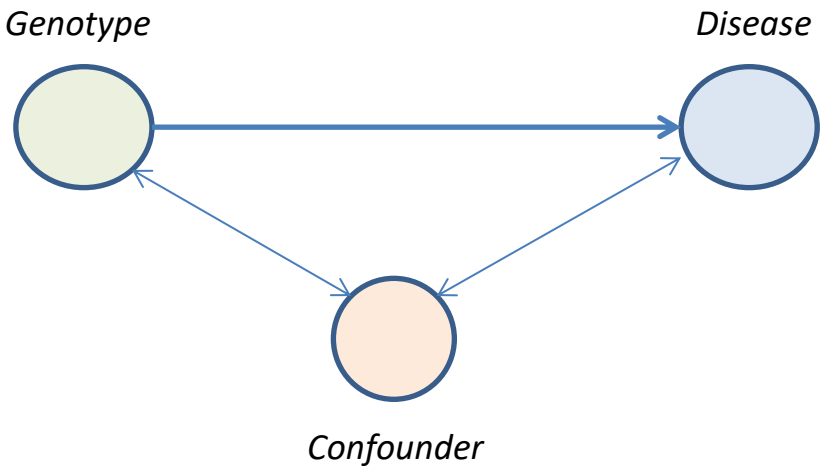
WTCCC, Nature 2007

GWAS Analysis

Important steps

- Select genotypic disease models to test (e.g. multiplicative, recessive, dominant). *Multiplicative model most powerful under most disease models!*
- Need to adjust for confounders*? In the absence of indicators for ancestry-based population stratification: rarely....*

  Other types of confounding difficult to argue



GWAS Analysis

Important steps

- Select genotypic disease models to test (e.g. multiplicative, recessive, dominant). *Multiplicative model most powerful under most disease models - apart from recessive!*
- Need to adjust for covariates*? In the absence of indicators for ancestry-based population stratification: rarely....*
- Adjustment for non-confounding covariates to 'absorb phenotypic noise'?

  See: Pirinen et al., Nature Genetics 2012; 44: 848-850

  – In case-control setting (logistic regression models):
    - Will reduce power when disease prevalence is < 2% (many diseases!)
    - Will only gain power when disease prevalence is > ~20%

GWAS Analysis

Important steps

- Select genotypic disease models to test (e.g. multiplicative, recessive, dominant). *Multiplicative model most powerful under most disease models - apart from recessive!*
- Need to adjust for covariates*? In the absence of indicators for ancestry-based population stratification: rarely....*
- Adjustment for non-confounding covariates to 'absorb phenotypic noise'?

  See: Pirinen et al., Nature Genetics 2012; 44: 848-850

  – In case-control setting (logistic regression models):
    - Will reduce power when disease prevalence is < 2% (many diseases!)
    - Will only gain power when disease prevalence is > ~20%

  – Can be of value:
    - In linear regression models of quantitative traits
    - In individual studies prior to meta-analyses, to avoid effect-size heterogeneity
    - Where interaction effects between genetic variants and covariate exist (e.g. gender effects)

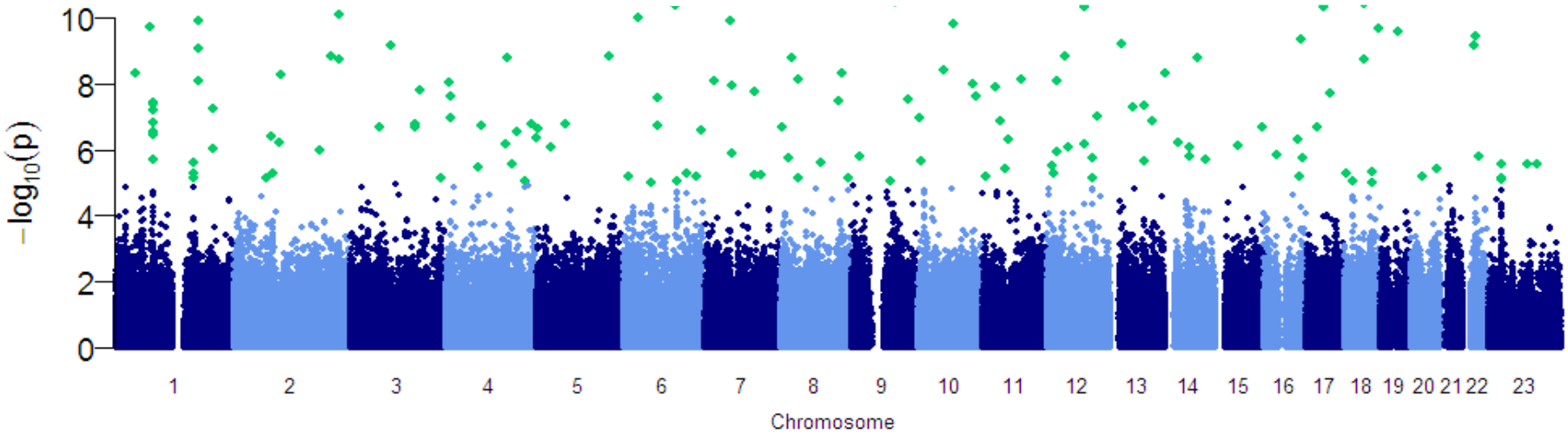
    - **Conclusion: think very carefully about adjustments...!**

GWAS Analysis

Important steps

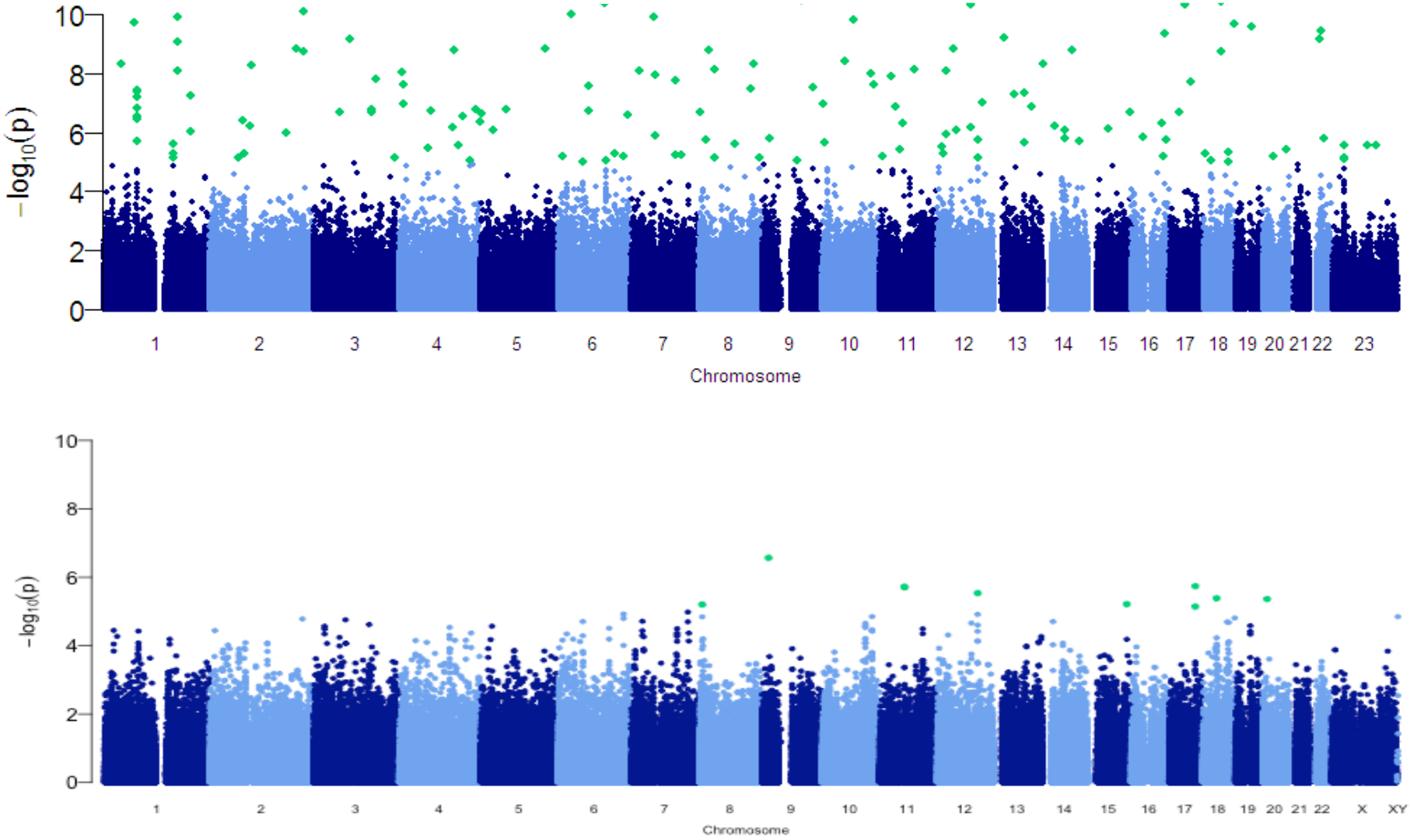
- Select genotypic disease models to test (e.g. multiplicative, recessive, dominant). *Multiplicative model most powerful under most disease models!*
- Need to adjust for covariates*? In the absence of indicators for ancestry-based population stratification: rarely....*
- Visualisation and interpretation of results

Manhattan Plots



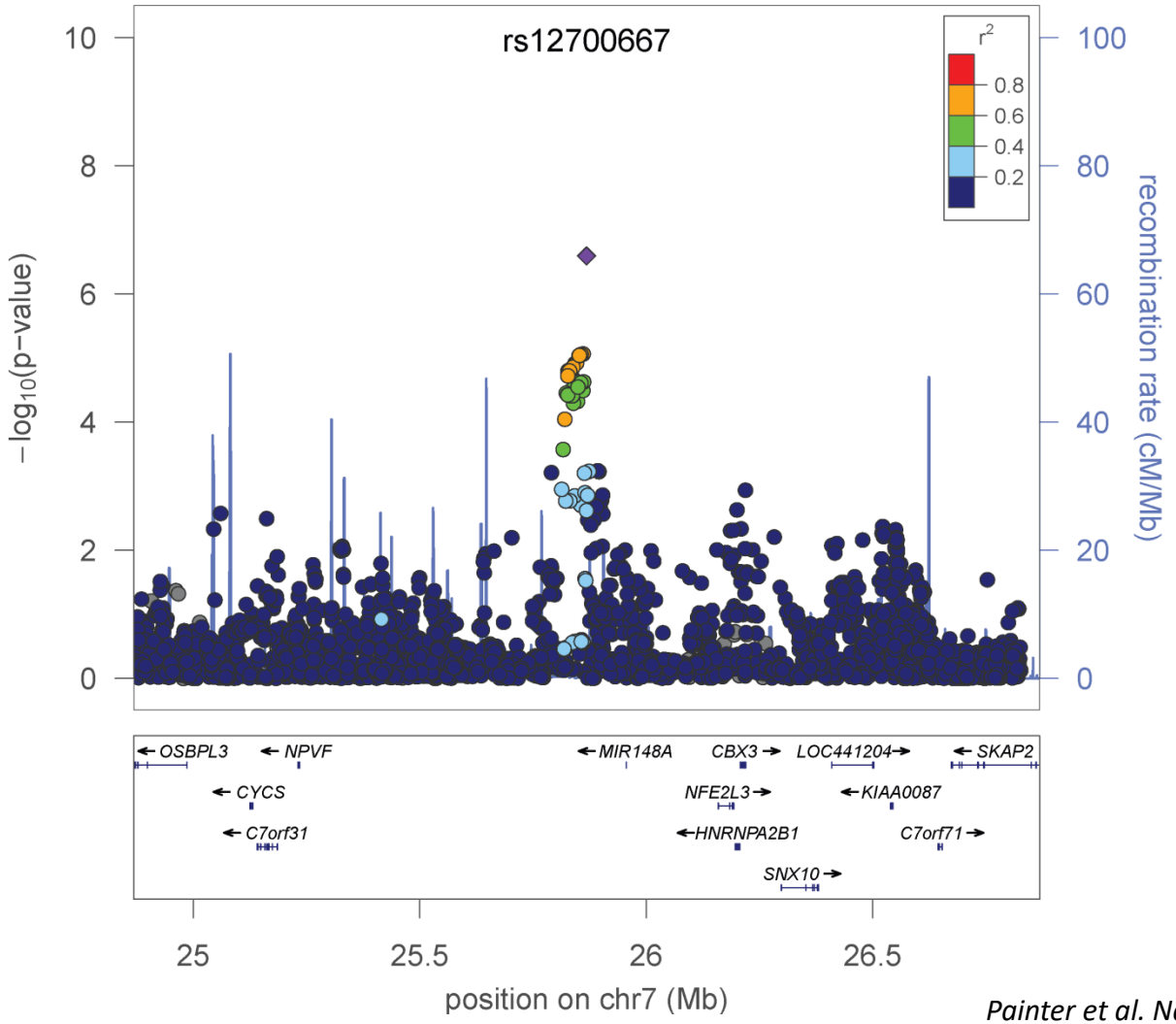
Manhattan Plots

Before and after QC….



*

Regional association plots (e.g. Locuszoom)



Painter et al. Nat Genet 2011; 43: 51-4

GWAS Analysis

Important steps

- Select genotypic disease models to test (e.g. multiplicative, recessive, dominant). *Multiplicative model most powerful under most disease models!*
- Need to adjust for covariates*? In the absence of indicators for ancestry-based population stratification: rarely.*...
- Visualisation and interpretation of results
- What is ‘significant’? ‘Multiple testing’?

GWAS Analysis – reducing false positives

- Adjustment of ‘genome-wide significance’ threshold for # tests conducted?

*

GWAS Analysis – reducing false positives

- Adjustment of 'genome-wide significance' threshold for # tests conducted?
- **Much more complex....** WTCCC paper: factor determining the threshold is not the number of tests performed, but the a-priori probability that there is likely to be a true association at any specified location in the genome ('Bayesian' statistical theory)
- Different significance thresholds proposed, converging on the one most commonly used now: 5×10^{-8} (independent of # SNPs tested)

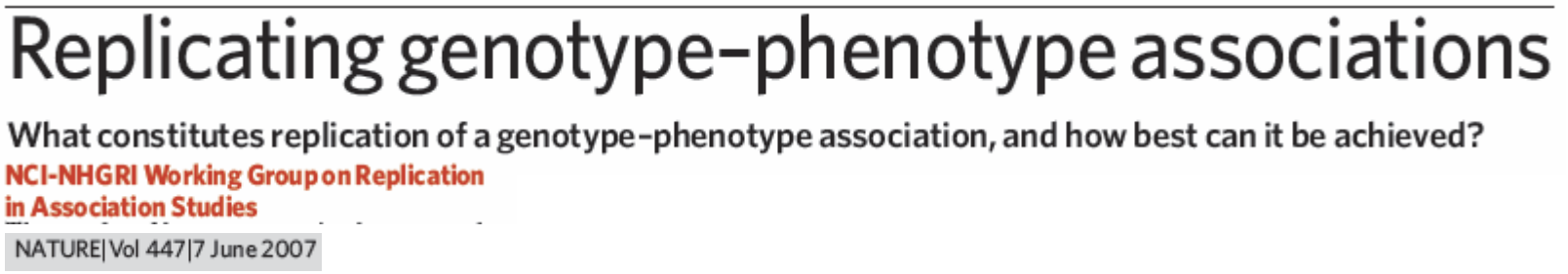
GWAS Analysis

Important steps

- Select genotypic disease models to test (e.g. multiplicative, recessive, dominant). *Multiplicative model most powerful under most disease models!*
- Need to adjust for confounders*? In the absence of indicators for ancestry-based population stratification: rarely.*...
- Visualisation and interpretation of results
- What is 'significant'? 'Multiple testing'?
- Replication replication replication (and meta-analysis....)

*

Replication



Replicating genotype–phenotype associations

What constitutes replication of a genotype–phenotype association, and how best can it be achieved?

NCI-NHGRI Working Group on Replication in Association Studies

NATURE|Vol 447|7 June 2007

- Same study population as original finding
- Same/similar case definition
- Same marker (with supporting evidence from others in high LD)
- Good study design practices, including sufficiently large sample size

Genome-wide Association Meta-analysis

- Aims:
  - Assess the strength of evidence of association across multiple GWAS and replication datasets
  - Leverage maximum power of detection
  - Check for heterogeneity of results

Genome-wide Association Meta-analysis

- Aims:
  - Assess the strength of evidence of association across multiple GWAS and replication datasets
  - Leverage maximum power of detection
  - Check for heterogeneity of results
- Methodology and QC (*Winkler et al., Nat Prot 2014*).
  - Typically, now many different studies/datasets involved
  - Central analysis plan shared
  - Often, QC + imputation + analyses performed by the individual centres (to comply with data sharing policies)
  - Association statistics for each SNP calculated and shared:

    (effect size, SE/CIs, allele freqs, sample size, p-values)
  - Specific analysis tools to detect errors from aggregated data

Genome-wide Association Meta-analysis

- Aims:
  - Assess the strength of evidence of association across multiple GWAS and replication datasets
  - Leverage maximum power of detection
  - Check for heterogeneity of results
- Methodology and QC (*Winkler et al., Nat Prot 2014*).
  - Typically, now many different studies/datasets involved
  - Central analysis plan shared
  - Often, QC + imputation + analyses performed by the individual centres (to comply with data sharing policies)
  - Association statistics for each SNP calculated and shared:

    (effect size, SE/Cis, allele freqs, sample size, p-values)
  - Specific analysis tools to detect errors from aggregated data

Important steps (beyond GWAS-specific QC):

1. File-level QC (cleaning and checks)
2. Meta-level QC: comparison study-specific results (Identification of analytical issues by SE-N and P-Z plots; allele frequency or strand problems; population stratification through lambda-N plots)
3. Meta-analysis QC (identifying analytical issues)

Fixed-effects Meta-analysis

- Let β_i denote the allelic effect (aligned to a fixed baseline allele) of the *i*th study, with variance denoted v_i.
- Estimate of the allelic effect over all *N* studies is then given by

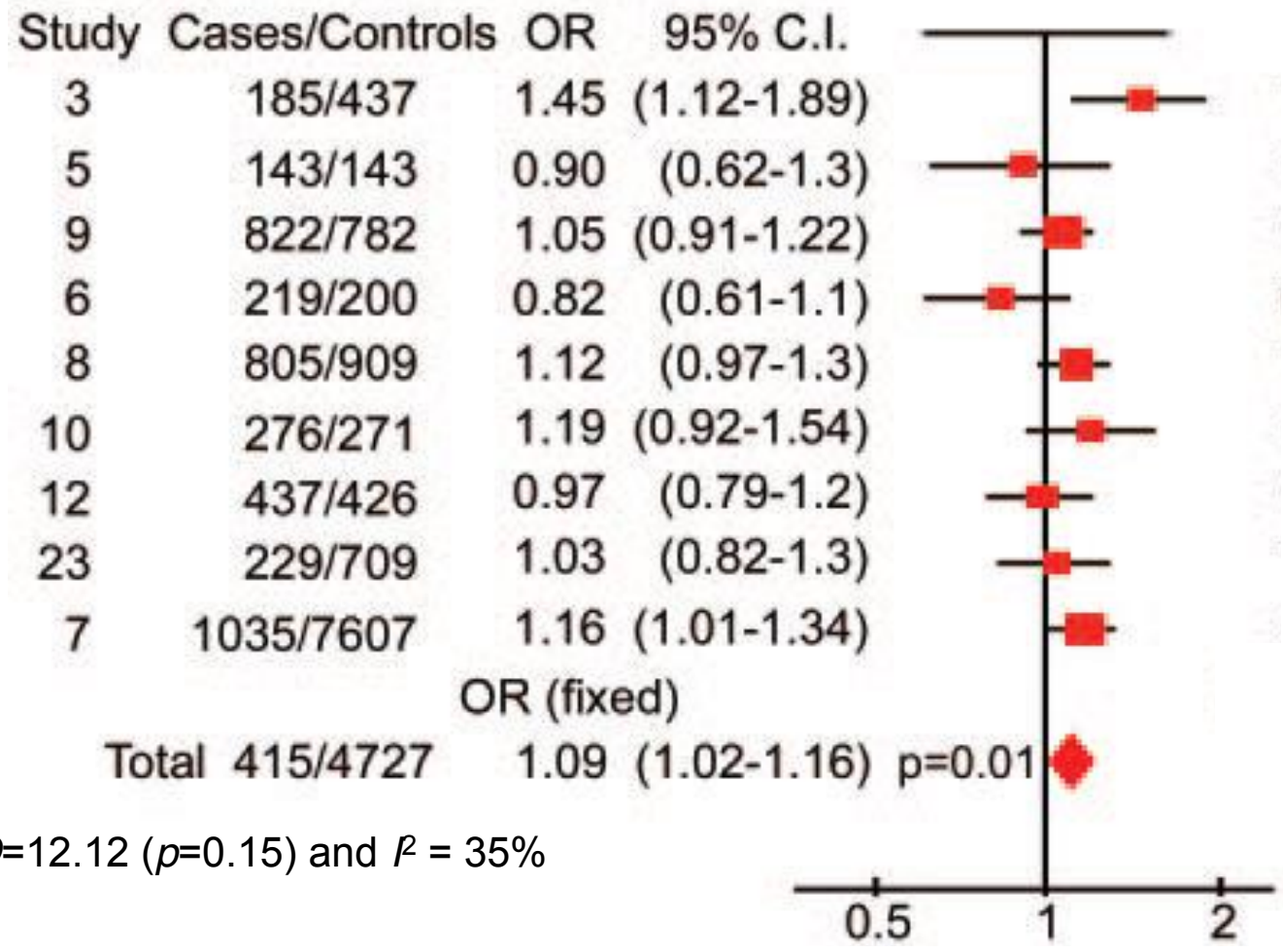
$$B = \frac{\sum_i w_i \beta_i}{\sum_i w_i}$$

  where $w_i = 1/v_i$, with variance given by $V = \left[\sum_i w_i\right]^{-1}$.
- Test for association over all studies given by $X^2 = B^2/V$, having an approximate chi-squared distribution with one degree of freedom.

Assessing heterogeneity

- Fixed effects meta-analysis assumes the same odds ratio (allelic effect) over all studies.
- We can test for heterogeneity between effects using the statistic $Q = \sum_i w_i (B - \beta_i)^2$, which has an approximate chi-squared distribution with *N*-1 degrees of freedom.
- An alternative statistic, I^2=[*Q*-(*N*-1)]/*Q*, quantifies the extent of heterogeneity from a collection of allelic effect sizes.
- Important to investigate source of potential heterogeneity.

Example: sporadic amyotrophic lateral sclerosis



Q=12.12 (p=0.15) and I^2 = 35%

Random-effects Meta-analysis

- Random effect meta-analysis often utilised when a SNP demonstrates significant evidence of heterogeneity in allelic effects between studies.
    - Random-effects meta-analysis: assume distribution of true allelic effects instead of a single underlying true effect size.
    - Random-effects variance component given by

$$\tau^2 = \max\left(0, \frac{Q-(N-1)}{\sum_i w_i - \left(\sum_i w_i^2 / \sum_i w_i\right)}\right)$$

    - Weight assigned to *i*th study then given by

$$w_i^* = (\tau^2 + v_i)^{-1}.$$

A comment on random effects

- Important to investigate the source of heterogeneity between studies: variability may be due to phenotype definition, population background, interaction with exposure to environmental risk factor.

Strand alignment

- Study 1: OR of 1.1 for allele A relative to allele G (aligned to + strand).
- Study 2: OR of 1.3 for allele C relative to allele T (aligned to – strand).
- Effect in study 2 is in opposite direction to study 1 since A is not complementary to C.
- It is straightforward to overcome this issue for non-AT or non-GC SNPs: otherwise rely on correct strand information or matching of allele frequencies (possibly with reference to HapMap or 1000 Genomes data).

Software

- Fixed- and random-effects meta-analysis can be performed for individual SNPs in standard statistical software packages such as R.
- Specialised software for genome-wide association meta-analysis that can handle large numbers of SNPs and studies, and can incorporate checks for strand alignment:
  - METAL: http://genome.sph.umich.edu/wiki/METAL_Program
  - GWAMA: http://www.geenivaramu.ee/en/tools/gwama
  - METASOFT: http://genetics.cs.ucla.edu/meta/

Summary

- Define the case phenotype in adequate detail.
- Select appropriate control group.
- The larger the sample size, the more power to identify variants of smaller effects. Typically at least 2000 cases needed with 1:1 or 1:3 control ratio.
- Quality control is the most important part of GWAS analysis: Sample QC and variant QC .
- Choose appropriate statistical test for association dependent on your phenotype (binary vs. linear, case/control imbalance, related individuals?)
- Post-GWAS: QQ plots and lambda to check for population stratification. Consider adjustment for any additional covariates?
- Replication and meta-analysis for strengthening of evidence for findings.