Genome-Wide Association Studies (GWAS) – #3

Allan McRae

a.mcrae@uq.edu.au

Recap

The data is cleaned

- Individuals with poor quality genotyping are removed
- SNPs with poor genotyping quality are removed
- Genotypes are imputed

Time to do a GWAS!

GWAS

Association analysis is relatively straight forward...

At each SNP in the genome, a simple statistical test is performed to assess the association between the SNP and trait of interest.

Quantitative Traits

Test ***correlation*** between trait and SNPs

Typically uses a simple “additive” model

- each SNP is encoded 0, 1 or 2 representing the number of B alleles in the genotypes AA, AB and BB

- for imputed SNPs, calculate 2 x $P_{BB} + P_{AB}$

  P_{BB} and P_{AB} are the probabilities of genotypes BB and AB respectively

This is referred to as the additive model of association

- each copy of the B allele is adding to the association

Quantitative Traits

Additional covariates can be included in a linear regression model

e.g. age, sex, PCs, …

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \beta_3 D_i + \ldots$$

It is important that the assumptions of linear regression are met

→ normality of residuals

Phenotypes are often transformed with a rank-based inverse normal transformation

- could precorrect data for covariates with large effects first and then ensure the normality of the residuals from this model

Disease Traits

Test whether the proportion of B alleles at a SNP differs between cases and controls

- This is a multiplicative model of association

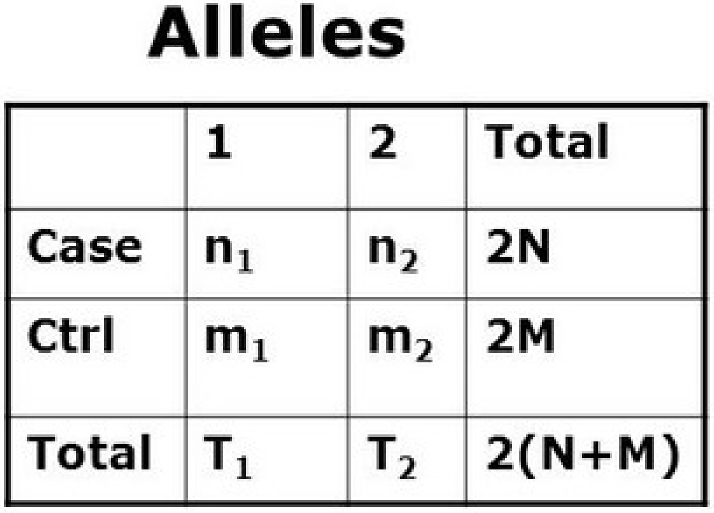
- The risk of developing the disease by a factor *r* for each B allele carried

i.e.

- baseline risk of *b* for genotype AA
- risk of *br* for genotype AB, and
- risk of br^2 for genotype BB

Disease Traits

Testing for association can be done using a simple chi-square contingency table test with a 2x2 matrix containing the counts of A and B alleles for cases and controls in each row



Alleles

	1	2	Total
Case	n_1	n_2	2N
Ctrl	m_1	m_2	2M
Total	T_1	T_2	2(N+M)

2x2 contingency table

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

1 degree of freedom

Disease Traits

Use a logistic regression model when covariates are to be included in the model

- logistic regression and contingency table are equivalent when no covariantes are used

$$P_i = \mathrm{E}(Y_i \mid X_i, C_i, D_i, \ldots)$$

$$\mathrm{logit}(P_i) = \beta_0 + \beta_1 X_i + \beta_2 C_i + \beta_3 D_i + \ldots$$

Including Genotype PC Covariates

Ancestral outliers were removed during the cleaning stage

Smaller scale differences in ancestry will still be present in the data and can be corrected for by including PCs from genotypes

PCs can also correct for possible biases induced by sample collection or non-genetic geographical effects on phenotype

How many PCs to include?

- 10 or 20 are common guidelines
- can test for association between each PC and your phenotype and include all PCs that explain significant variation in the phenotype

Significance

It is important to correct for the large number of tests performed in a GWAS study when assessing the significance of a result

Correcting for the number of SNPs tested using (e.g.) a Bonferroni correction is overly conservative due to the linkage disequilibrium between SNPs

A significance threshold of 5×10^{-8} corrects for the effective number of independent tests genome-wide

A less stringent threshold of 1×10^{-5} is widely used to indicate “suggestive” significance

Manhattan Plots

GWAS results are typically represented using a Manhattan plot

- genomic locations along the X-axis
- negative logarithm (base 10) of the p-value along the Y-axis
- each point is the result from a single SNP

The SNPs with the strongest associations will have the greatest negative logarithms, and will tower over the background of unassociated SNPs (like skyscrapers in the Manhattan skyline)

Manhattan Plots

Manhattan plots provides an additional check on the quality of the association test

- multiple SNPs should be under an association peak

- a single outlying SNP is usually the result of poor quality genotyping

Manhattan plots showing significant points occurring across the genome should be considered suspect:

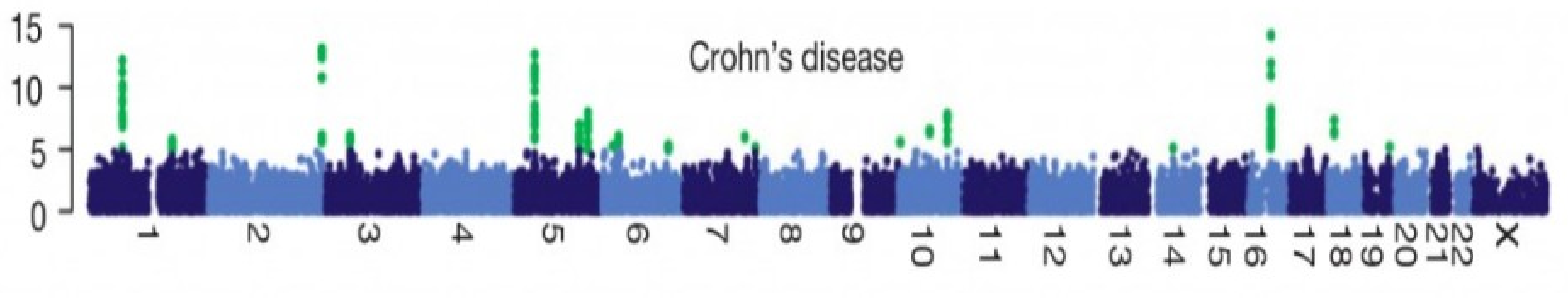
- genotyping batch effects (case/control)
- undetected relatedness
- sample duplications
- population stratification

Manhattan Plots

A ***good*** Manhattan plot

- Wellcome Trust Case Control Consortium,
  Crohn's disease, Nature 2007

- Shows signals supported by many neighboring SNPs

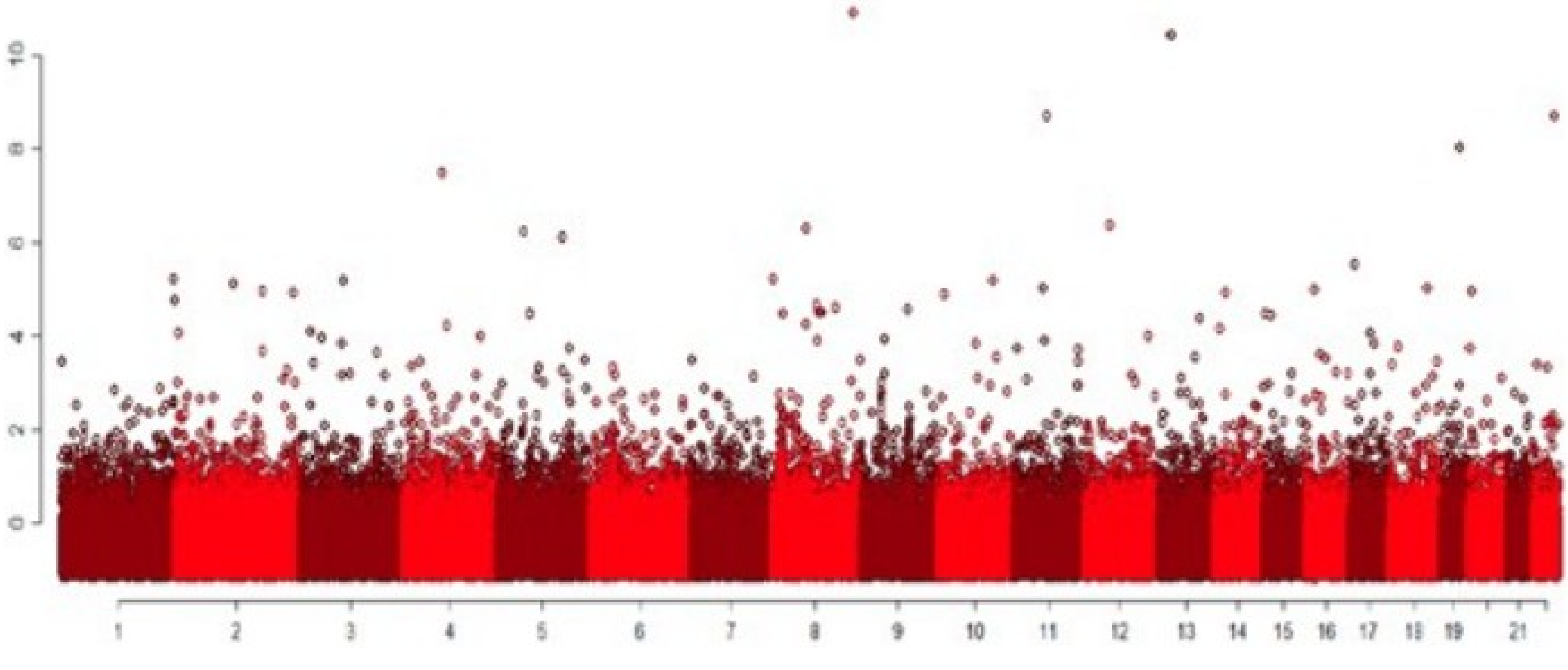


Manhattan Plots

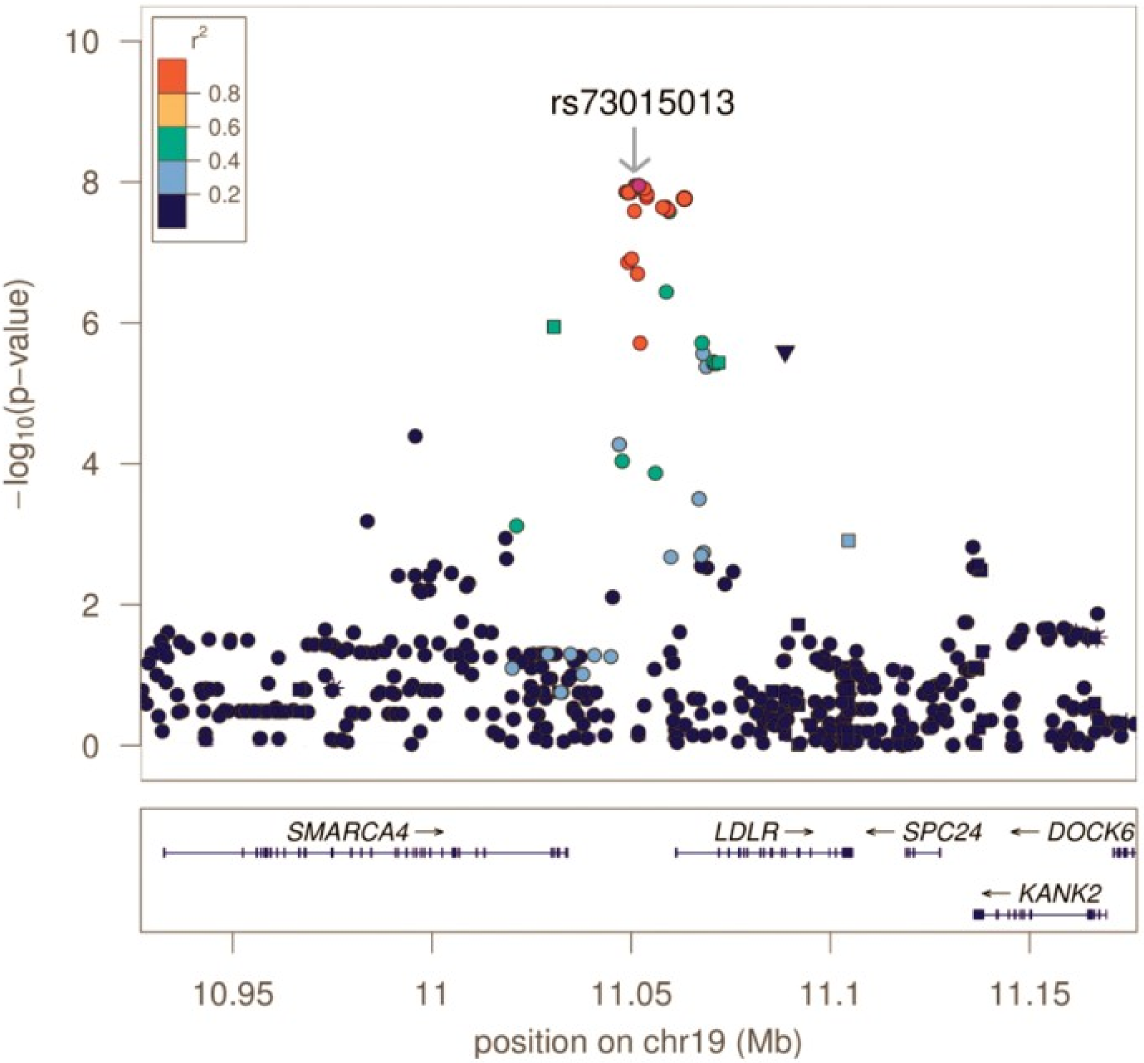
A ***bad*** Manhattan plot

-Sebastiani et al. “Genetic signatures of exceptional longevity in humans” Science July 2010

- Retracted July 2011 because of poor QC



Regional Association Plots



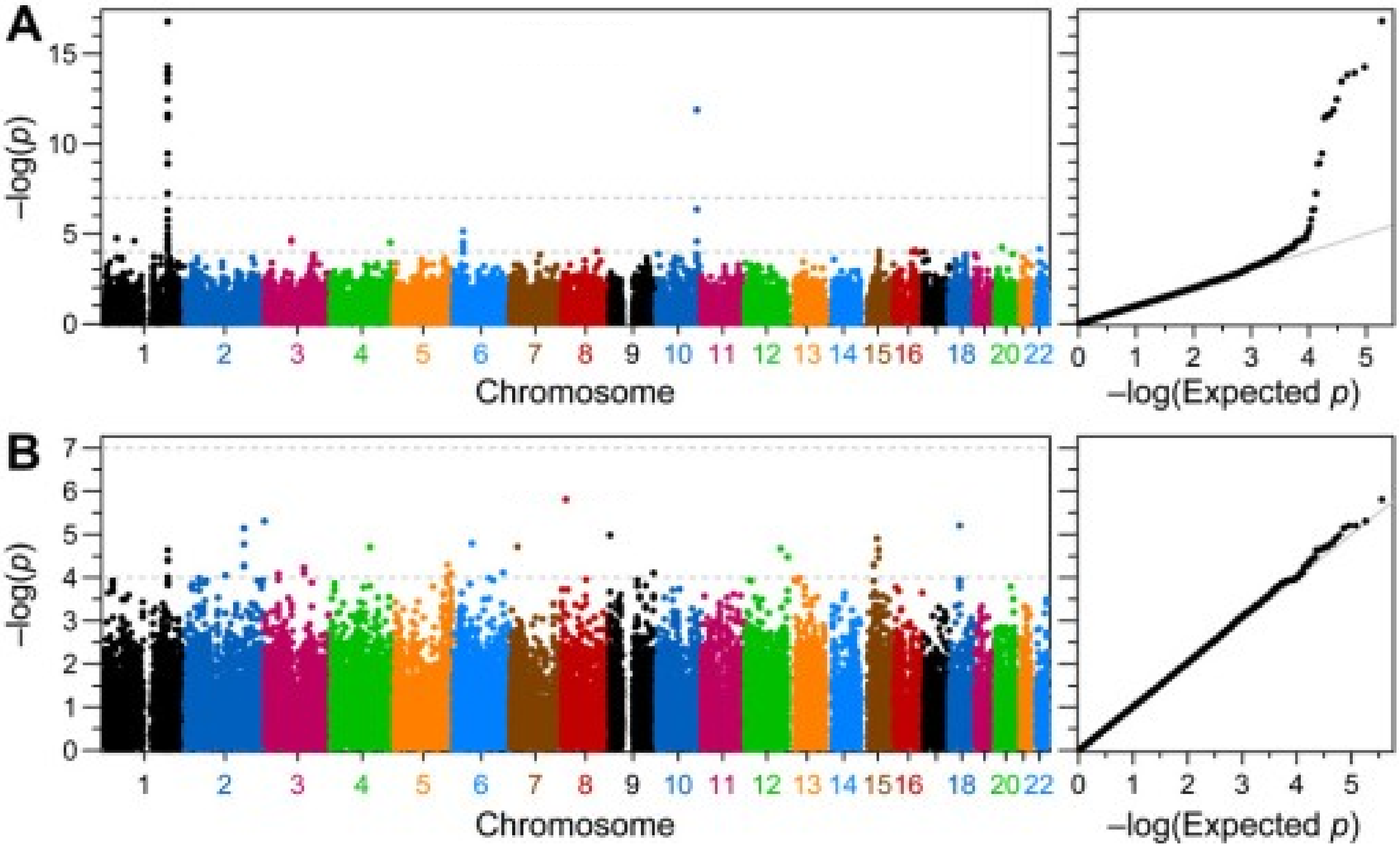
QQ Plot

A QQ plot is a common way to demonstrate the lack of confounding effects

The ordered observed negative logarithm of the p-values are plotted against the expected distribution under the null hypothesis of no association

Ideally, the points in the plot should align along the X = Y line, with deviation at the end for the significant associations

QQ Plot



A
B
-log(p)
0
5
10
15
1
2
3
4
5
6
7
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
20
22
–log(Expected p)
0
1
2
3
4
5

Genomic Inflation

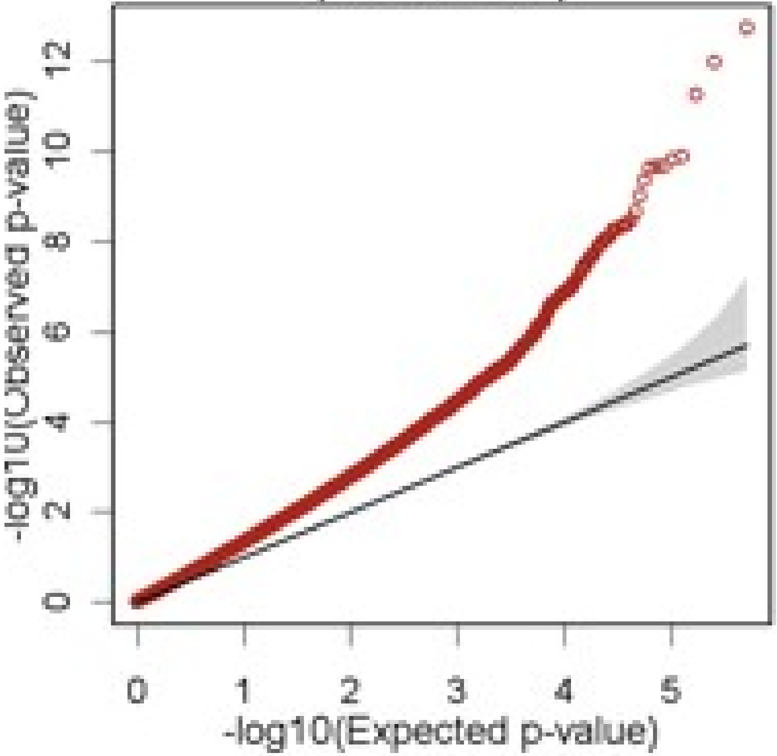
One way to quantify the lack of global inflation in the QQ plot is the genomic inflation factor (λ_{GC})

This is calculated by:
- determining the median p-value of GWAS test statistics
- calculating the quantile in a chi-squared distribution with one degree of freedom that would give this p-value
- divide this by the median of a chi-squared distribution with one degree of freedom (0.4549)

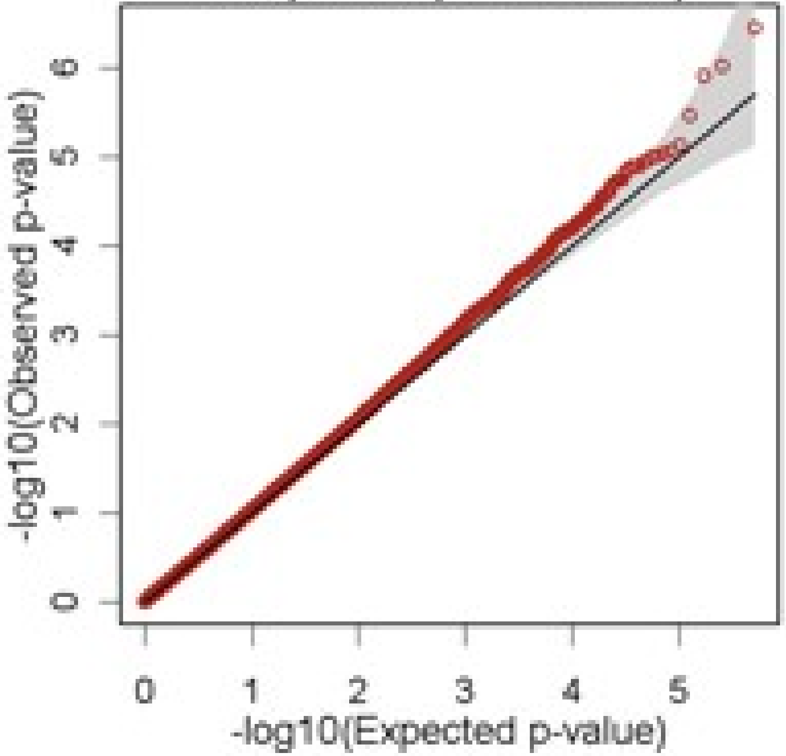
Deviations of this value away from 1.0 indicate genome-wide confounding in the data.

QQ Plot

Trend Test without Adjustment (lambda=1.447)



Logistic Regression Adjusted for Principal Components (lambda=1.037)

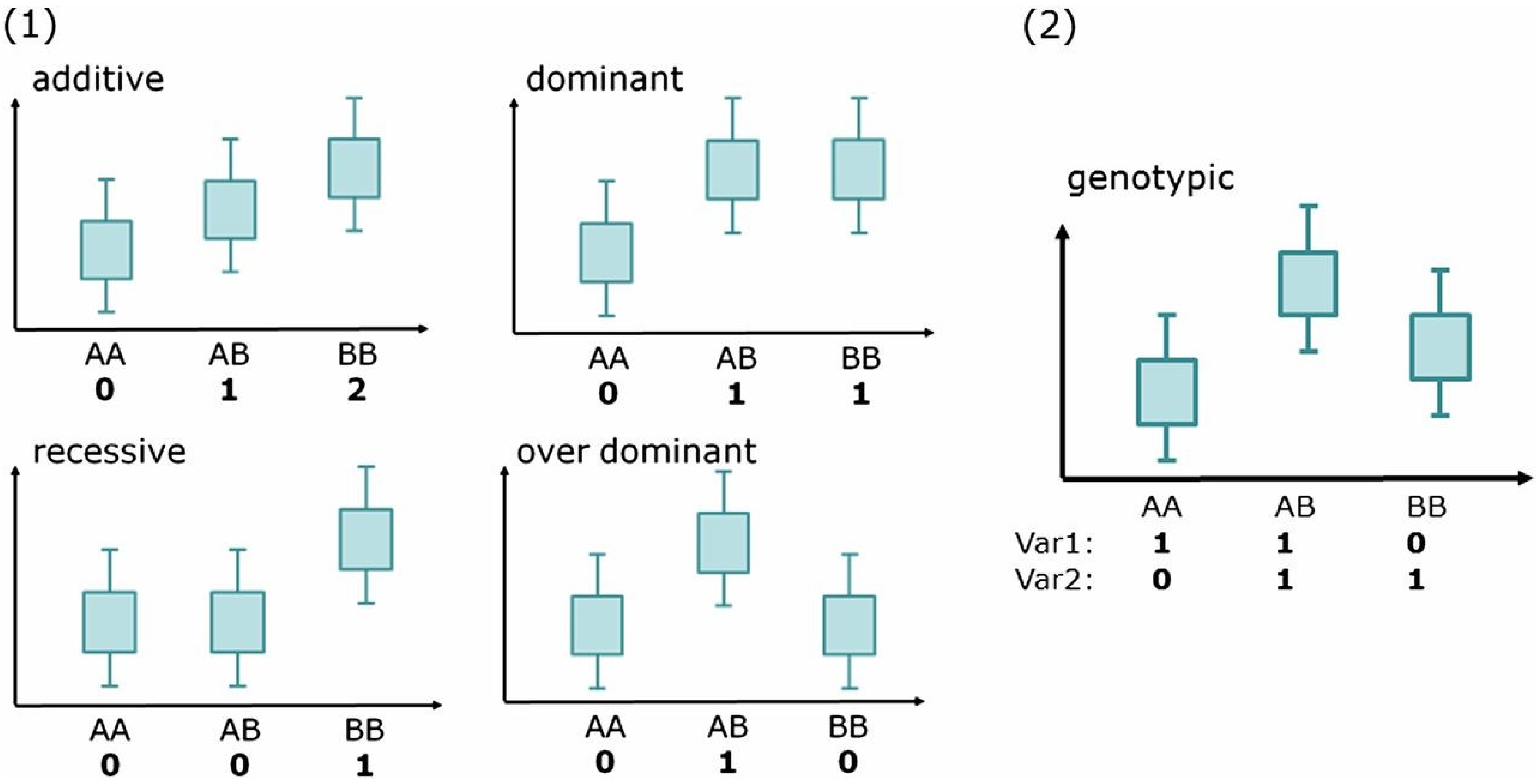


Non-Additive Genetic Models

The additive genetic model is most powerful under the range of plausible genetic inheritance models

Once significant GWAS peaks have been identified, you may want to explore more complex modes of inheritance

Non-Additive Genetic Models



How Many Association Signals?

The SNP showing the strongest statistical evidence for association in a genomic region (for example, a 2-Mb window centered on the locus) is reported to represent the association in this region.

This assumes that the detected association at the top SNP captures the maximum amount of variation in the region by its LD with an unknown causal variant and that other SNPs in the vicinity show association because they are correlated with the top SNP

However, there may be multiple causal variants at the locus

How Many Association Signals?

PLINK provides a LD-based result clumping procedure

SNPs are “clumped” into groups with high-linkage disequilibrium

Procedure:
- take most significant “unclumped” SNP (lowest p-value)
- look at all SNPs with R^2 > ***x*** and within ***y*** distance
- clump all significant SNPs in that group to one set

- Repeat until significant SNPs are all clumped

Somewhat arbitrary choice of thresholds…

How Many Association Signals?

Conditional or Joint analysis of significant SNP

- Perform GWAS
- Take most significant SNP
- Either:
    - add it to the covariate list, or
    - regress its effect out of the phenotype
- Repeat until no significant SNPs left

Can focus on just region of interest or whole genome

Can identify new signals in regions with no previous associtation signal

Power Considerations

Power of GWAS is determined by:

- Sample size

- Effect size
    - difference of means for quantitative traits
    - odds ratio of disease

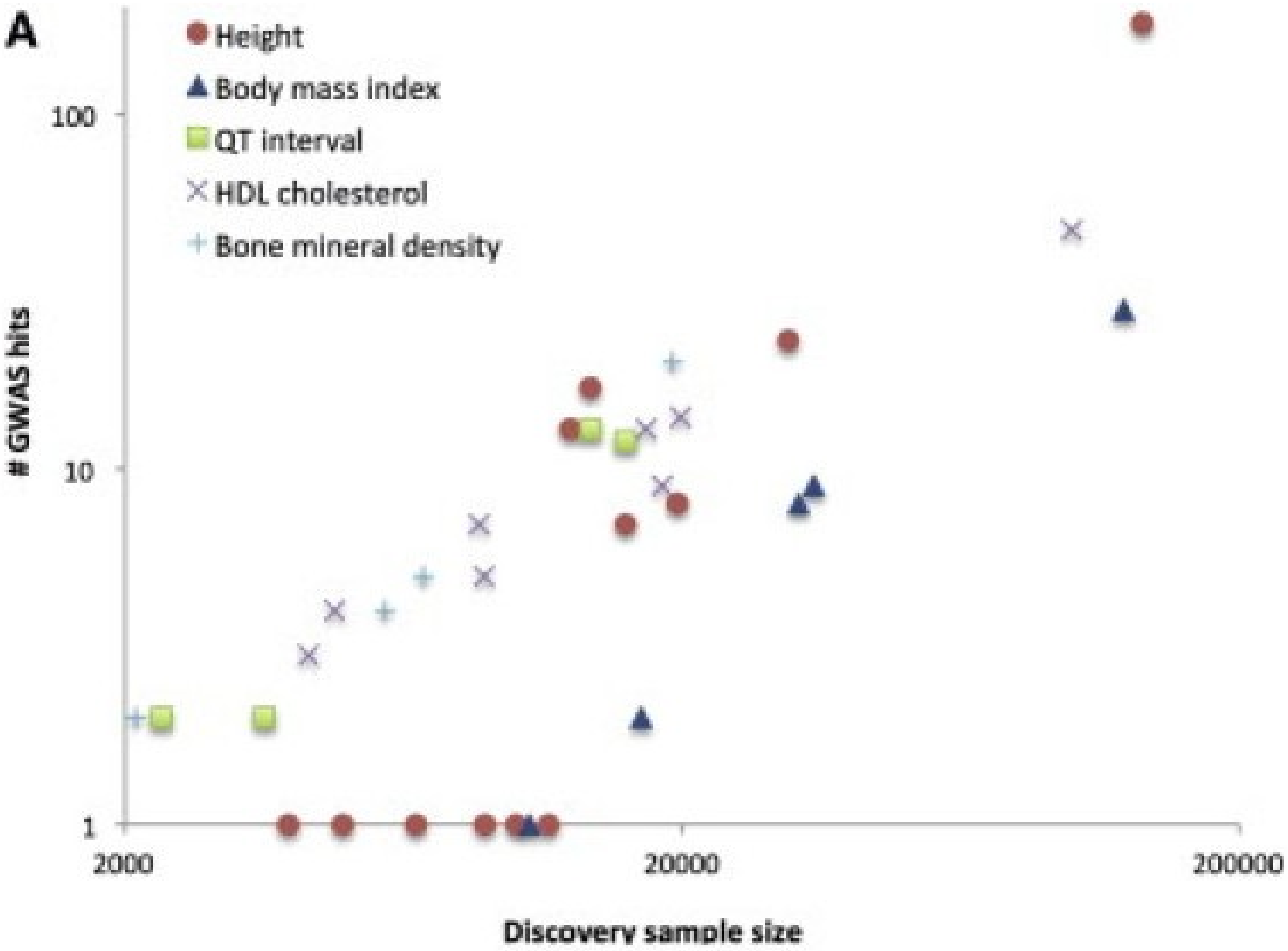
- Allele frequency

- Linkage disequilibrium / Imputation accuracy

- Disease prevelance (cases hidden in controls?)

- Case/Control ratio

Sample Size

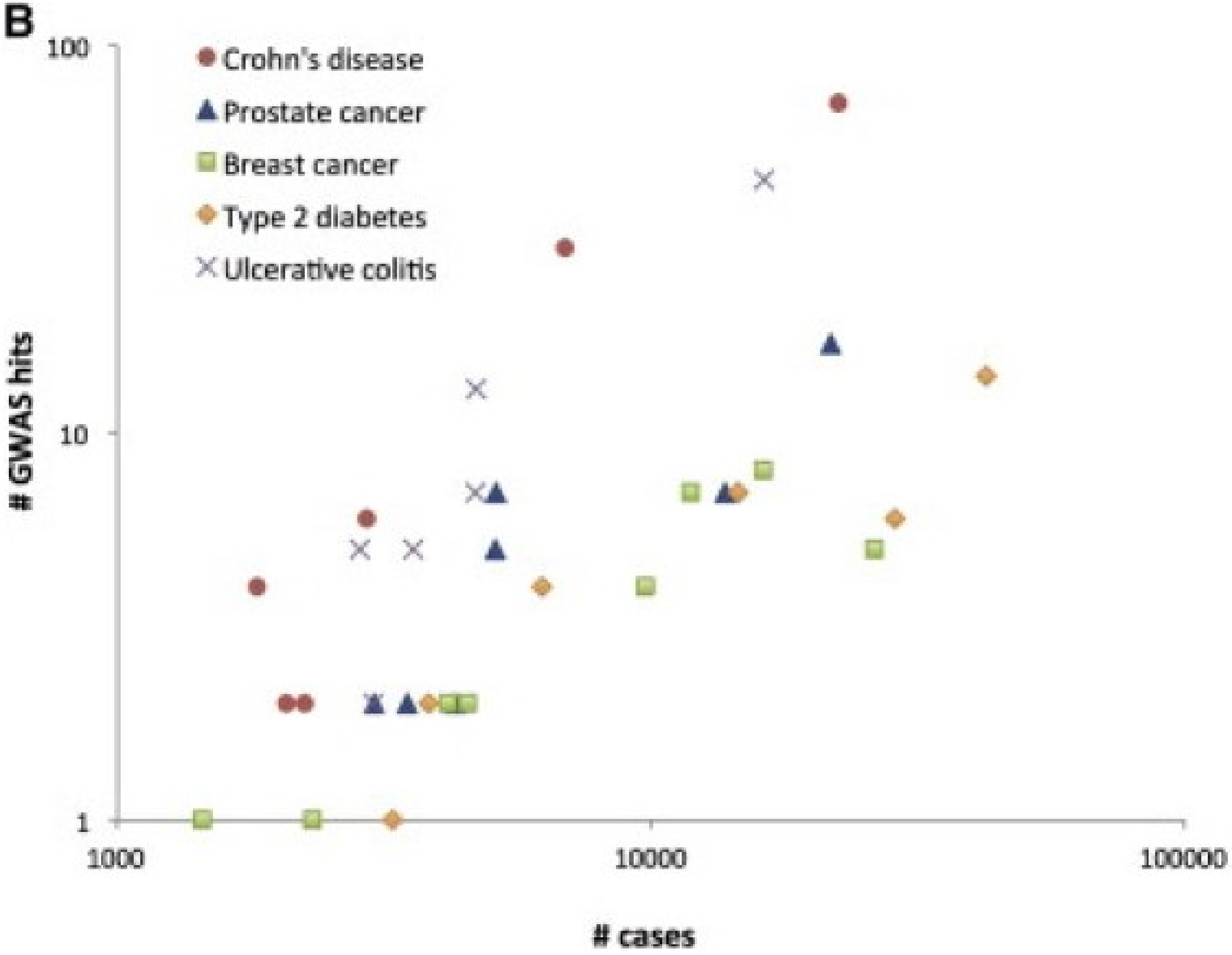


Visscher et al, 2012.

Five Years of GWAS Discovery.

American Journal of Human Genetics

Sample Size



B
Crohn's disease
Prostate cancer
Breast cancer
Type 2 diabetes
Ulcerative colitis
100
10
1
# GWAS hits
1000
10000
100000
# cases

Replication

The gold standard for validation of any genetic study is replication in an additional independent sample

Replication helps ensure that a genotype-phenotype association represents a credible association and is not a chance finding or an artifact due to uncontrolled biases

Usually only testing a few variants for replication
→ smaller multiple testing burden

Replication

Replication studies should have sufficient sample size to detect the effect of the susceptibility allele

Initial GWAS suffer from winner's curse, where the detected effect is likely stronger in the GWAS sample than in the general population

This means that replication samples should ideally be larger to account for the over-estimation of effect size

Lack of power and a negative replication result means we can not call the variant a false positive

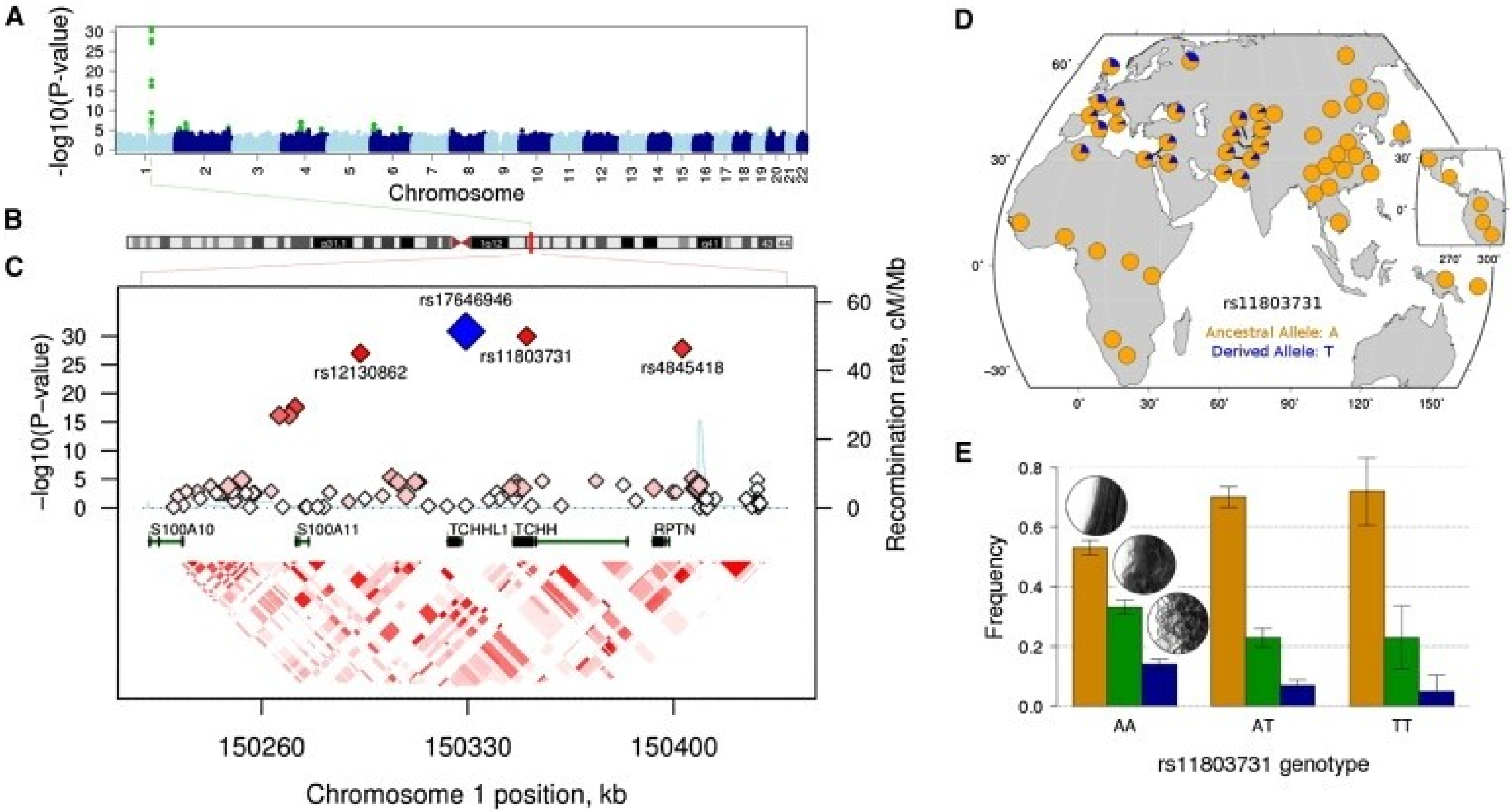
Replication

Replication studies should be conducted in an independent dataset drawn from the same population as the GWAS, in an attempt to confirm the effect in the GWAS target population.

Once an effect is confirmed in the target population, other populations may be sampled to determine if the SNP has an ethnic-specific effect.

Replication of a significant result in an additional population is sometimes referred to as generalization.

Example – Hair Curliness


A
-log10(P-value)
Chromosome
B
p31.1
q41
C
rs17646946
rs11803731
rs12130862
rs4845418
-log10(P-value)
Recombination rate, cM/Mb
S100A10
S100A11
TCHHL1
TCHH
RPTN
Chromosome 1 position, kb
D
60°
30°
0°
-30°
0°
30°
60°
90°
120°
150°
30°
0°
270°
300°
rs11803731
Ancestral Allele: A
Derived Allele: T
E
Frequency
AA
AT
TT
rs11803731 genotype