GWAS Experimental Design: statistical tests

Outline

- Types of tests, quantitative & binary traits
- Power to detect loci
  - depends on LD, effect size, allele frequency, sample size
- Manhattan plots
- Other diagnostics
  - QQ plot, genomic inflation and FDR
- Replication

Quantitative traits – linear regression

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}\beta + \boldsymbol{\varepsilon}$$

$\mathbf{y}$ = vector of (corrected) phenotypes

$\mathbf{1}$ = vector of 1's

α = intercept

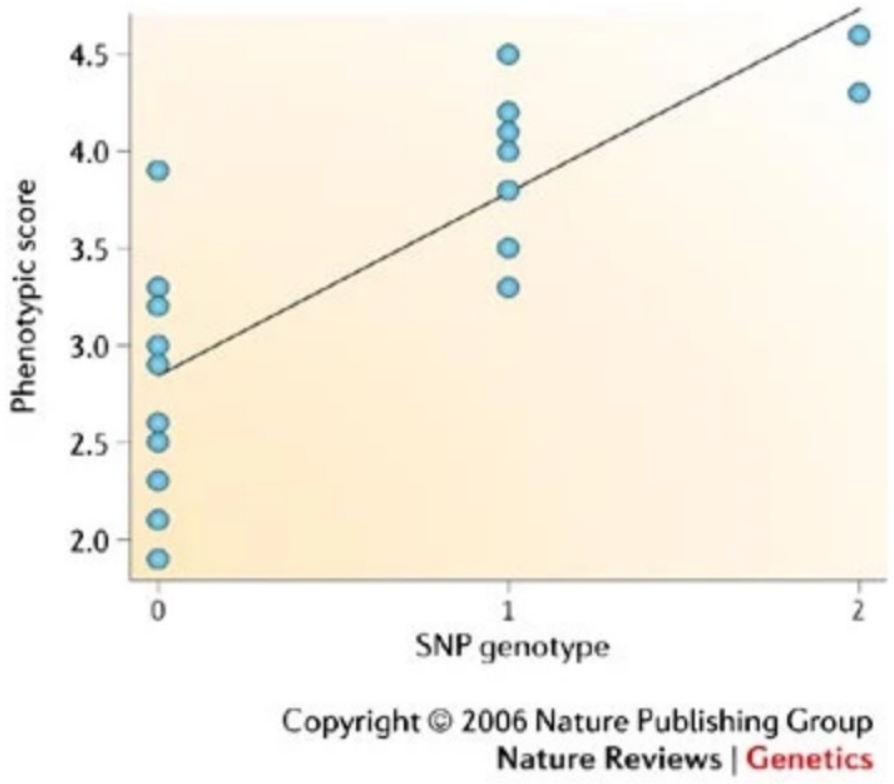
$\mathbf{x}$ = vector of SNP genotypes, encoded as 0, 1 or 2 copies of 'a' allele for AA, Aa or aa genotypes

β = SNP effect

$\boldsymbol{\varepsilon}$ = vector of errors

Null hypothesis, H_0: $\beta = 0$

Alternative hypothesis, H_1: $\beta \neq 0$



Balding (2006) *Nat Rev Genetics*

Binary traits

- Various options: chi-squared test, Armitage test, logistic regression etc.
- Make different assumptions about the mode-of-action of the allele -- this impacts power

e.g. chi-squared test; 2x2 contingency table

H_0: genotypes & case/control status are independent

H_1: genotypes & case/control status are dependent

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

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

- Use logistic model if need to correct for covariates

Power to detect loci

- Statistical **power** is the probability to correctly rejecting the null hypothesis when it is true
  - H_0 : there is no association between loci & trait
  - H_1 : this is a true association between the loci & trait

Power to detect loci

Power is a function of:

- LD between SNP and causal variant
- Proportion of phenotypic variance explained by causal variant
- Sample size
- Significance threshold (α)

Power – LD between SNP and causal variant

Usually, we don't expect the most significant GWAS variant in a region to be causal/functional

- i.e. tested SNP in LD with an ungenotyped 'causal variant'
- this reduces statistical power

- Sample size must increase by $1/r^2$ to detect an ungenotyped variant, compared to sample size required for testing causal variant itself
  - Hence increased SNP density (i.e. imputation, WGS) to maximise LD between causal variants & genotyped SNP

Power – LD between SNP and causal variant

Example:

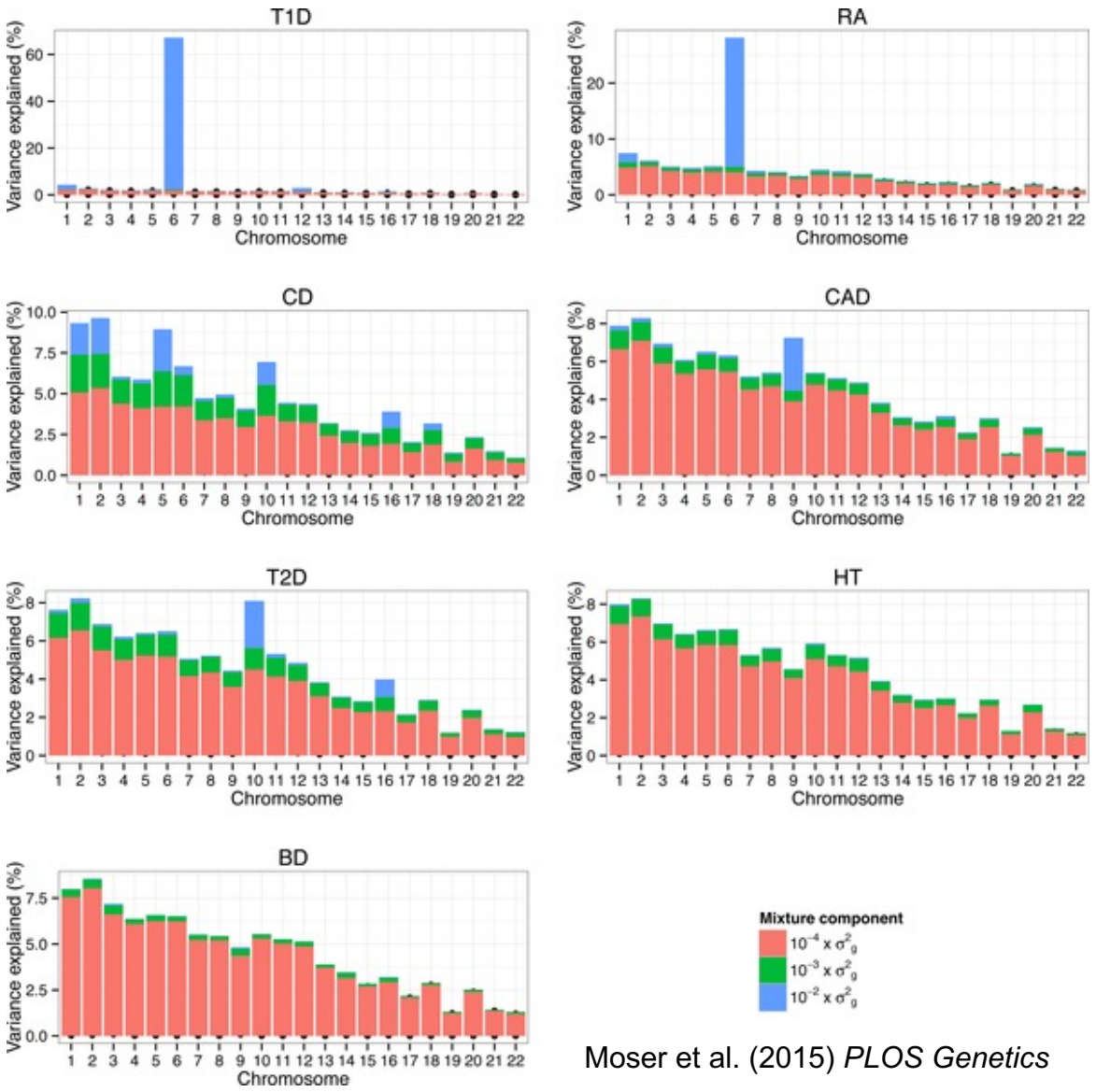
- The variance explained by a ‘causal variant’ is 1% of σ_P^2
- How much variance does a genotyped SNP explain when the LD between the causal variant and SNP is 0.2 or 0.8 ?
  - r^2 = 0.2 ; variance explained by SNP = 0.2 x 0.01 = 0.002 σ_P^2
  - r^2 = 0.8 ; variance explained by SNP = 0.8 x 0.01 = 0.008 σ_P^2

The r^2 between a SNP and a ‘causal variant’ is the proportion of the phenotypic variance which can be observed at the SNP

Power – effect size

How much of σ_P^2 is a marker expected to explain?

It is trait dependent



Moser et al. (2015) *PLOS Genetics*

Power – effect size

How much of σ_P^2 is a marker expected to explain?

It is trait dependent

For human height, the first detected (i.e. largest) effect explained 0.3% σ_P^2

nature genetics

LETTERS

A common variant of *HMGA2* is associated with adult and childhood height in the general population

Michael N Weedon[1,2,21], Guillaume Lettre[3,4,21], Rachel M Freathy[1,2,21], Cecilia M Lindgren[5,6,21], Benjamin F Voight[3,7], John R B Perry[1,2], Katherine S Elliott[5], Rachel Hackett[3], Candace Guiducci[3], Beverley Shields[2], Eleftheria Zeggini[5], Hana Lango[1,2], Valeriya Lyssenko[8,9], Nicholas J Timpson[5,10], Noel P Burtt[3], Nigel W Rayner[6], Richa Saxena[3,7,11], Kristin Ardlie[3], Jonathan H Tobias[12], Andrew R Ness[13], Susan M Ring[14], Colin N A Palmer[15], Andrew D Morris[16], Leena Peltonen[3,17,18], Veikko Salomaa[19], The Diabetes Genetics Initiative, The Wellcome Trust Case Control Consortium, George Davey Smith[10], Leif C Groop[8,9], Andrew T Hattersley[1,2], Mark I McCarthy[5,6,21], Joel N Hirschhorn[3,4,20,21] & Timothy M Frayling[1,2,21]

Human height is a classic, highly heritable quantitative trait. To begin to identify genetic variants influencing height, we examined genome-wide association data from 4,921 individuals. Common variants in the *HMGA2* oncogene, exemplified by rs1042725, were associated with height ($P = 4 \times 10^{-8}$). *HMGA2* is also a strong biological candidate for height, as rare, severe mutations in this gene alter body size in mice and humans, so we tested rs1042725 in additional samples. We confirmed the association in 19,064 adults from four further studies ($P = 3 \times 10^{-11}$, overall $P = 4 \times 10^{-16}$, including the genome-wide association data). We also observed the association in children ($P = 1 \times 10^{-6}$, $N = 6,827$) and a tall/short case-control study ($P = 4 \times 10^{-6}$, $N = 3,207$). We estimate that rs1042725 explains ~0.3% of population variation in height (~0.4 cm increased adult height per C allele). There are few examples of common genetic variants reproducibly associated with human quantitative traits; these results represent, to our knowledge, the first consistently replicated association with adult and childhood height.

Adult height is a classic polygenic trait. The genetics of height were central to the mendelian versus biometrician debate in the early part of the twentieth century that was resolved by Fisher, who proposed that height and other human phenotypes showed multifactorial inheritance[1]. Twin, family and adoption studies suggest that up to 90% of normal variation in human height within populations is due to genetic variation[2–6]. Severe mutations in several genes cause rare syndromes with extreme stature; however, these cannot explain normal population height variation[7]. Many regions of the genome have been linked with height based on numerous genome-wide linkage scans, with some overlap between studies[6], but thus far there have not been any examples of gene variants that are reproducibly associated with height variation in the general population.

The recent flood of data from many genome-wide association (GWA) studies offers new opportunities to identify genes influencing adult height. The identification of such genes will probably provide important insights into how best to dissect the genetics of polygenic quantitative traits. The identification of genes influencing growth may also have important medical implications. Height is associated with several common disorders, including a number of cancers[8,9].

Power – effect size

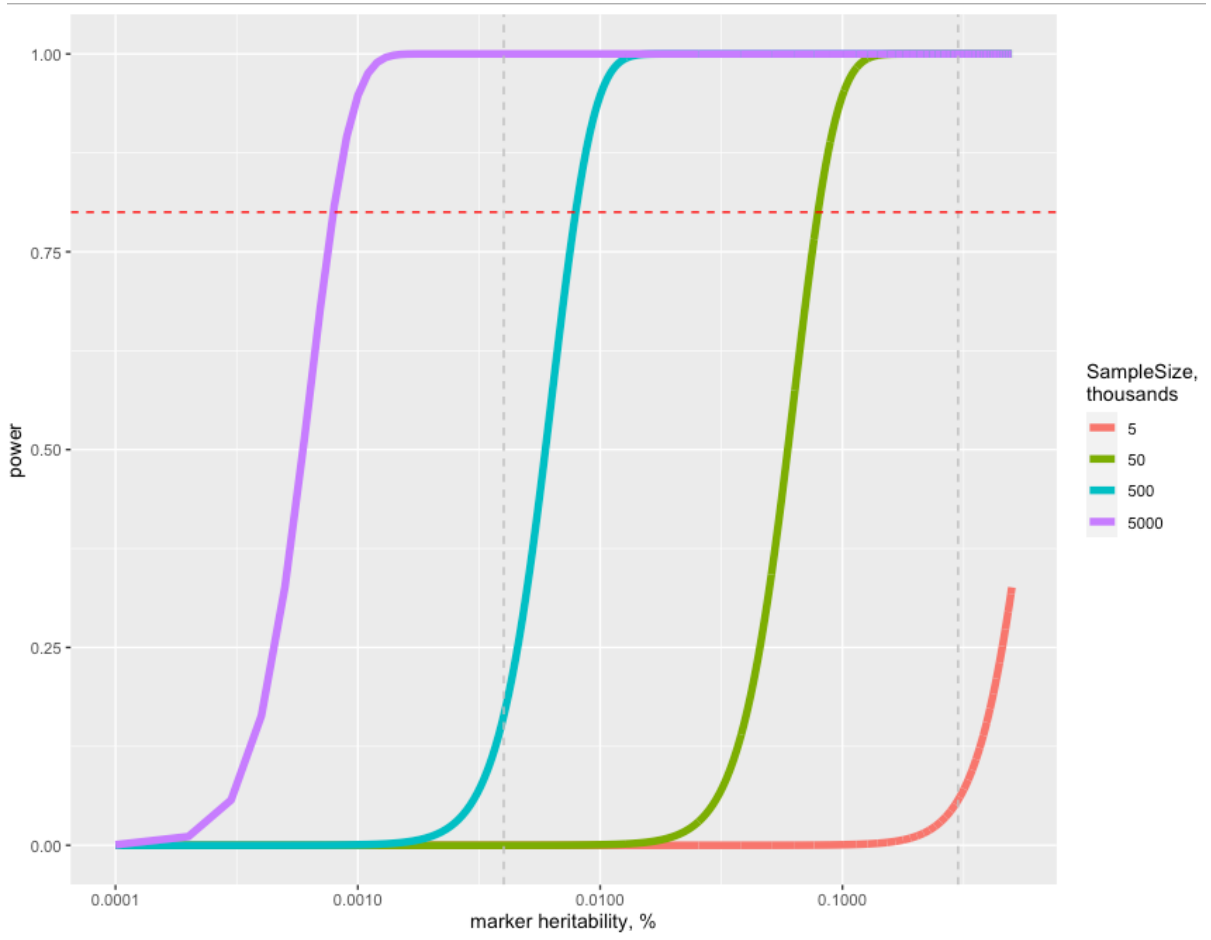
How much of σ_P^2 is a marker expected to explain?

It is trait dependent

For human height, the first detected (i.e. largest) effect explained 0.3% σ_P^2

Yengo et al. (2022) detected 12,111 SNP collectively explaining ~ 0.5 σ_P^2

i.e. 0.004 % σ_P^2 per SNP



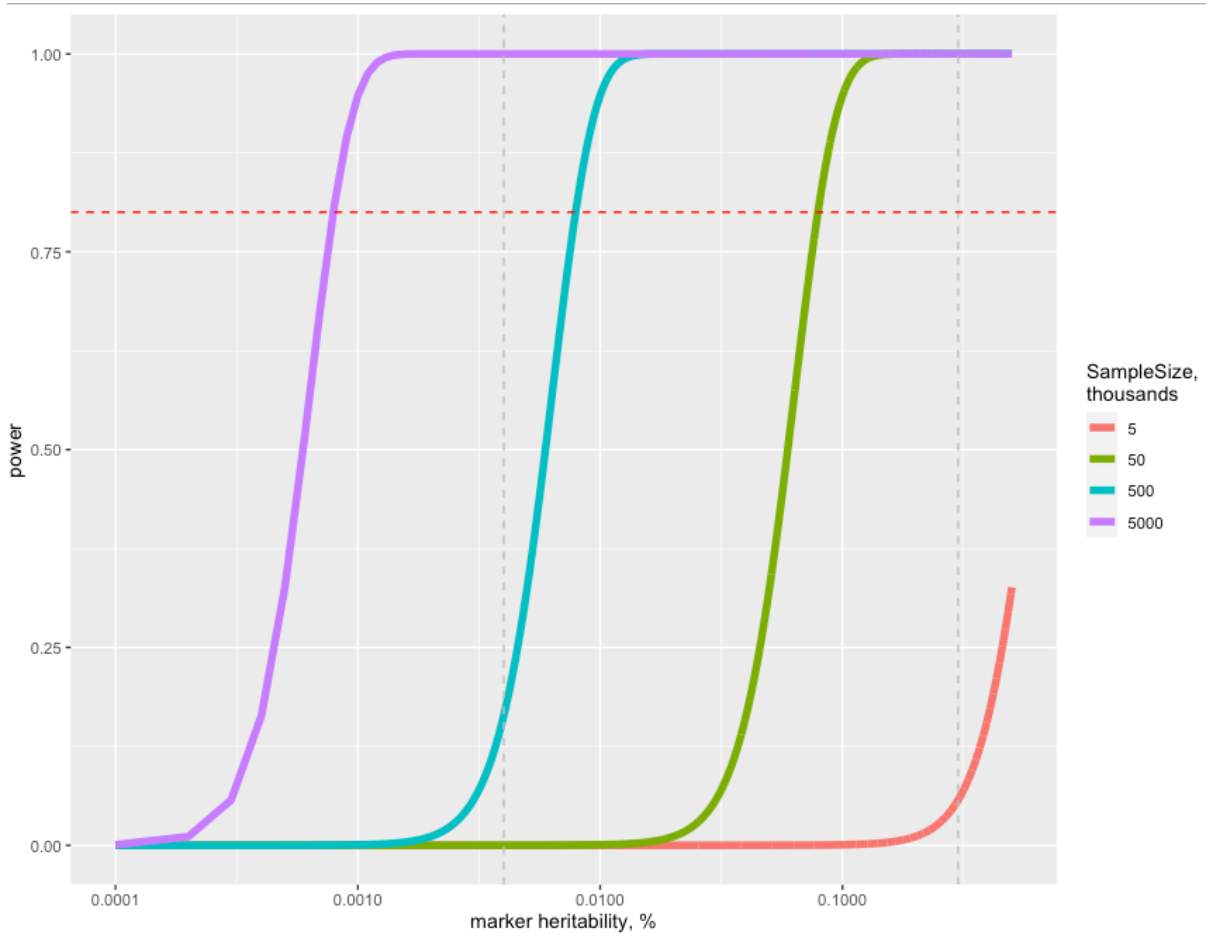
Power – sample size

How big do sample sizes need to be?

For human height,

5K individuals to detect loci 0.3% σ_P^2

5M to detect loci explaining ~ 0.004 % σ_P^2



Power - significance threshold

- GWAS performs millions of tests... many will be 'significant' ($P < 0.05$) by chance
- Easiest way to account for all these tests is to correct the significance threshold (α) for number of independent tests
  - correcting for the total number of tests is overly conservative due to the LD

- LD varies between populations, thus
  - EUR: 1 million independent tests ($0.05/1x10^6$) → sig. threshold p = $5x10^{-8}$
  - AFR: 2 million independent tests ($0.05/2x10^6$) → sig. threshold p = $2.5x10^{-8}$

Pe'er et al. (2008) *Genetic Epidemiology*

Power to detect loci

Power is a function of:

- LD between SNP and causal variant (dense SNPs to maximise LD)
- Proportion of phenotypic variance explained by SNP
  - Typically: < 0.005 σ_P^2 for quantitative traits, OR 1.1-1.2 binary traits
  - Can't change genetic architecture
- Sample size (bigger is more powerful)
- Significance threshold (α)

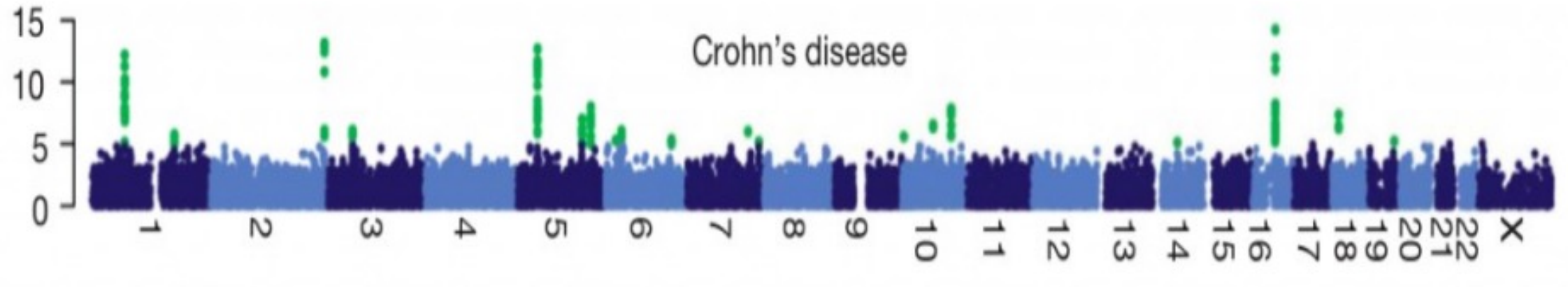
Manhattan Plots

- GWAS results are typically represented using a ‘Manhattan plot’
  - genomic locations/order 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 Manhattan →

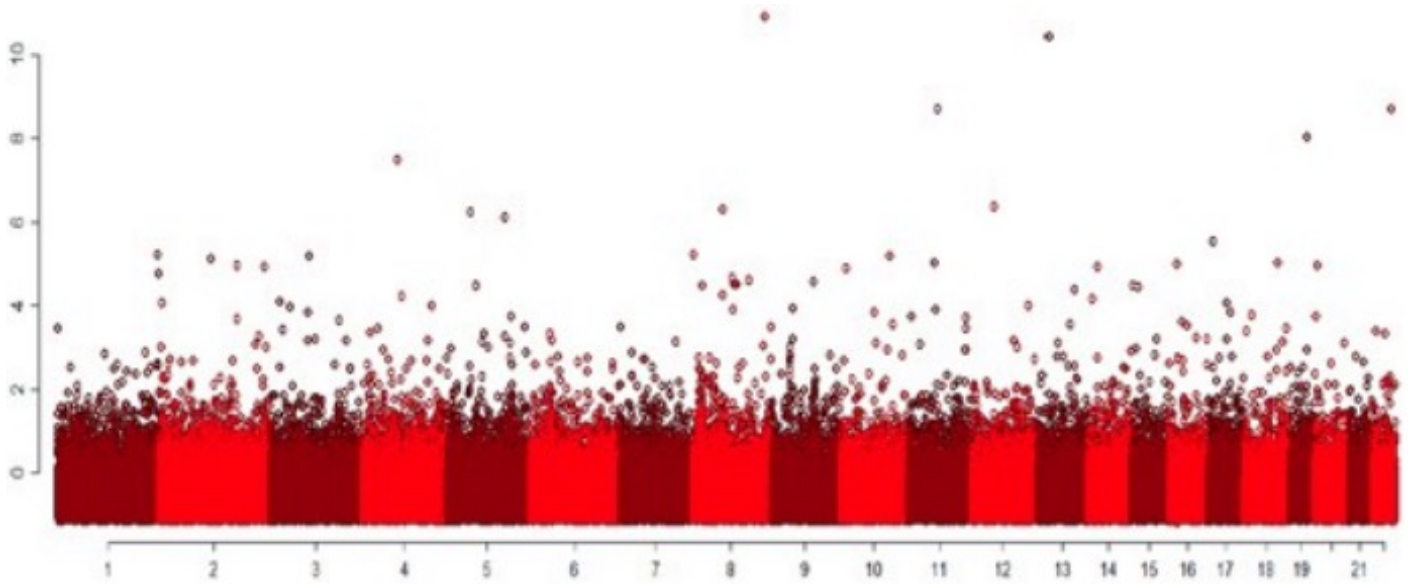
Manhattan Plots

- A ***good*** Manhattan plot
- Wellcome Trust Case Control Consortium, Crohn's disease, Nature 2007
- Shows signals supported by many neighboring SNPs

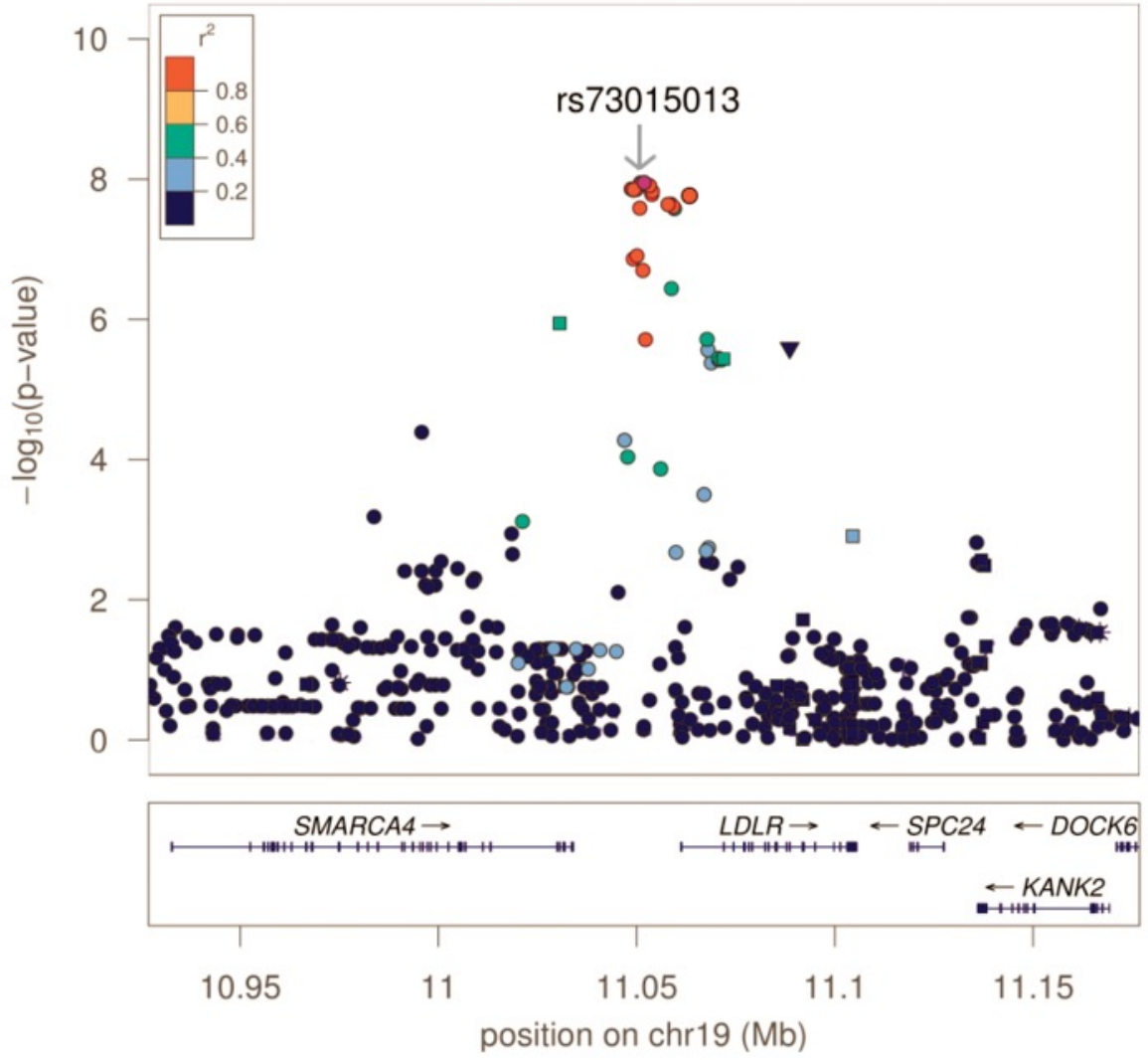


Manhattan Plots

- 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



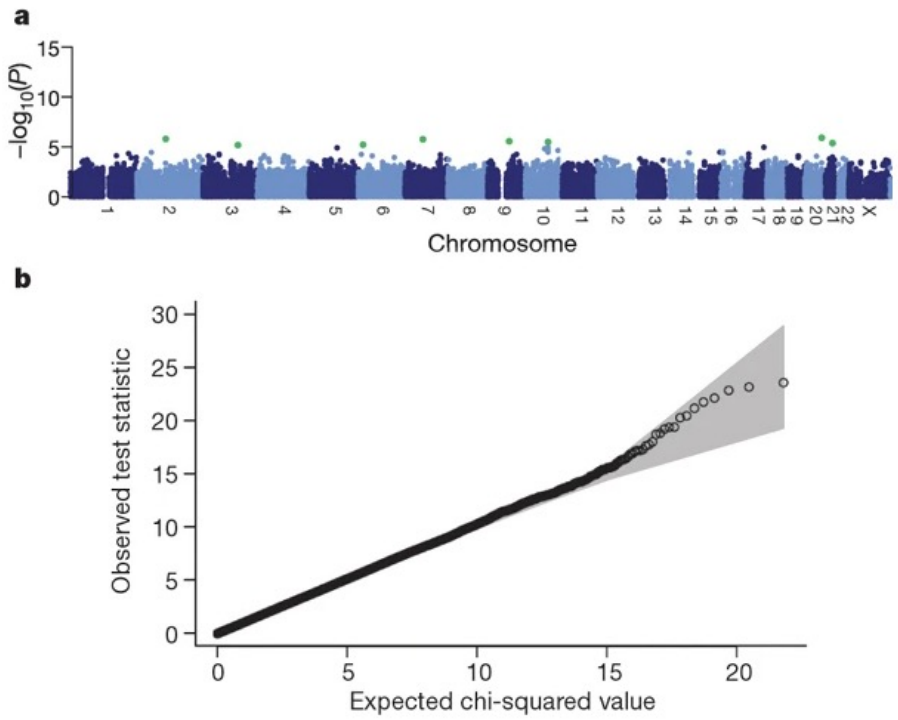
Interpreting GWAS signals & making biological insights is tricky, more on this tomorrow

Diagnostics (1) -- QQ Plot

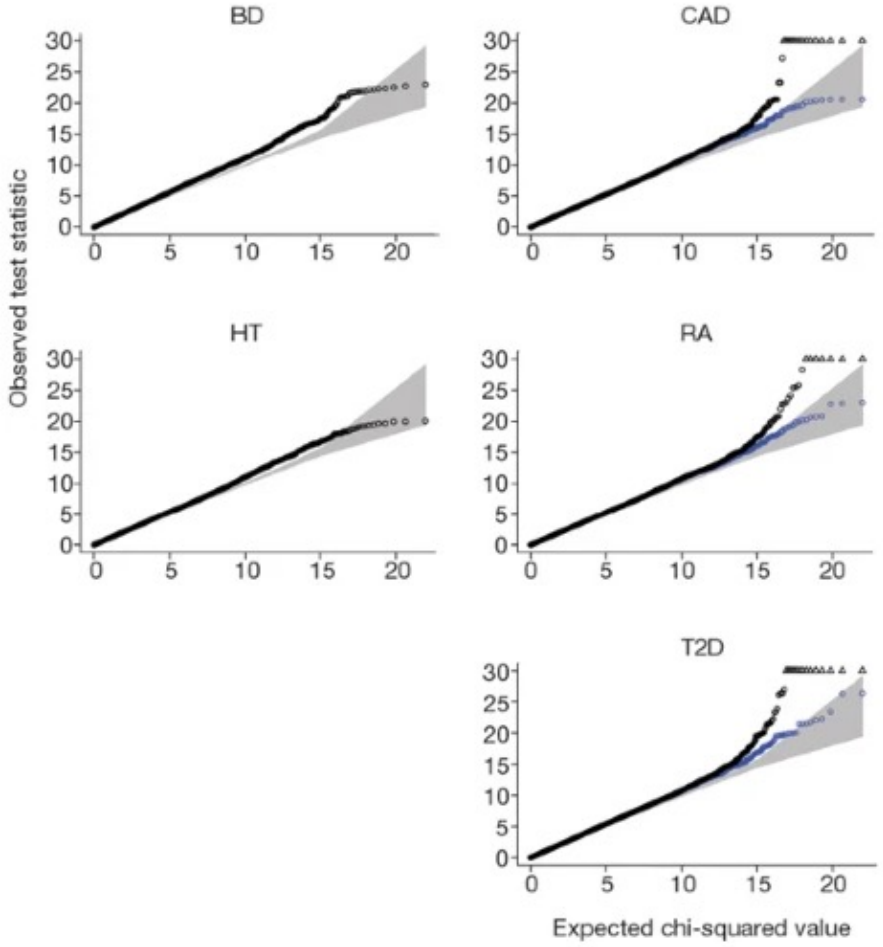
- A QQ plot is a common way to demonstrate the lack of confounding effects
- The ordered observed negative logarithm of the p-values are plotted against the expected distribution under the null hypothesis of no association
- Ideally, the points in the plot should align along the X = Y line, with deviation at the end for the significant associations

Diagnostics (1) -- QQ Plot

Figure 1: Genome-wide scan for allele frequency differences between controls.



a, *P* values from the trend test for differences between SNP allele frequencies in the two control groups, stratified by geographical region. SNPs have been excluded on the basis of failure in a test for Hardy–Weinberg equilibrium in either control group considered separately, a low call rate, or if minor allele frequency is less than 1%, but not on the basis of a difference between control groups. Green dots indicate SNPs with a *P* value $<1\times10^{-5}$. **b**, Quantile-quantile plots of these test statistics. In this and subsequent quantile-quantile plots, the shaded region is the 95% concentration band (see Methods).



WTCCC (2007) *Nature*

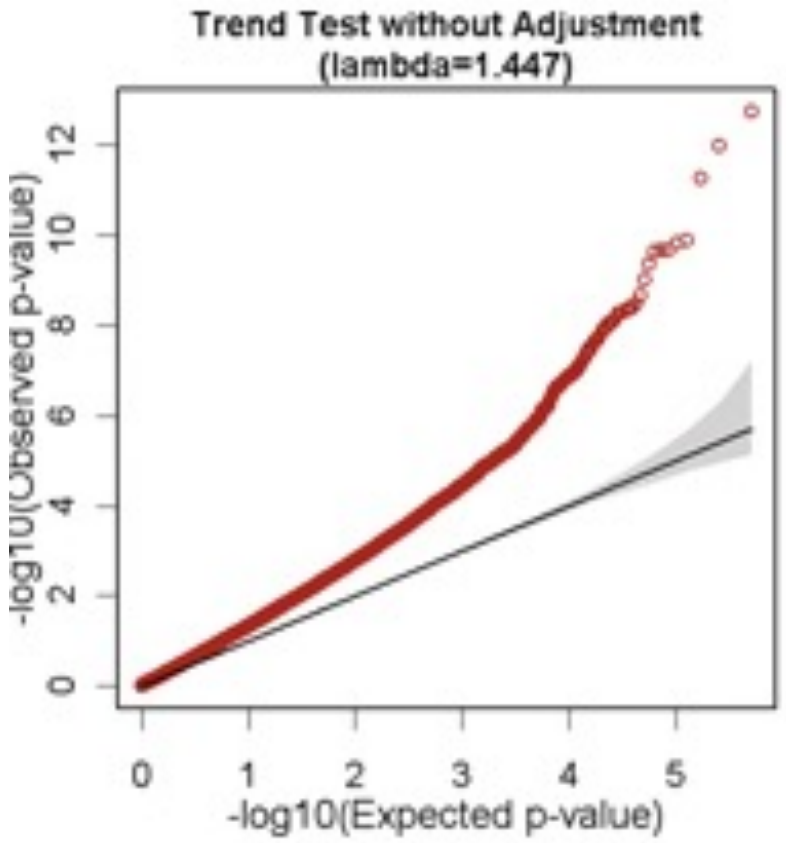
Diagnostics (2) -- Genomic Inflation

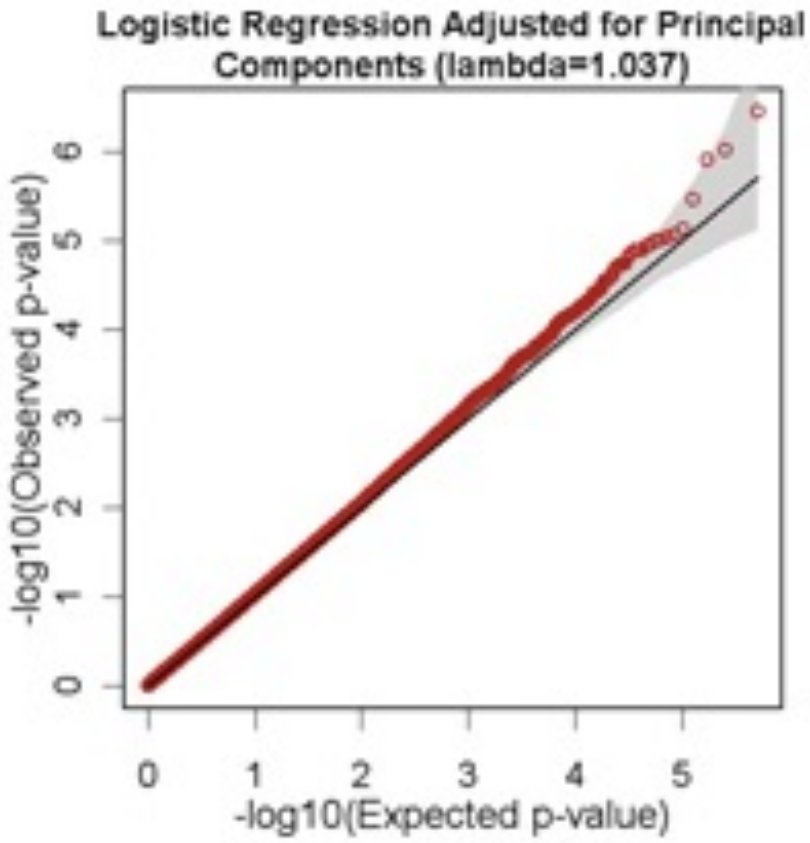
- One way to quantify the lack of global inflation in the QQ plot is the genomic inflation factor (λ_{GC})

- 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.

Diagnostics (2) -- Genomic Inflation





Diagnostics (3) -- FDR

Non-human species might use a **F**alse **D**iscovery **R**ate (FDR), thus at a given significant threshold (α) the FDR is

FDR = # expected 'significant' SNP / # observed 'significant' SNP

e.g. If we test 1M loci with α = 0.0001, we expect $1x10^6$ X 0.0001 = 100 sig. loci by chance

Say we observe 150 sig. loci at α = 0.0001

FDR = expected/observed = 100/150 = 0.67

Replication

- GWAS potentially have many false-positives
- Replication in an *independent* cohort is required
- Be mindful of sample size (is there enough power to replicate?)
- Replicate size and direction of effect
- *Question: What does 'Winner's curse' refer to in GWAS?*

Meta-analysis of CHARGE and Global BPgen of Top 10 Loci for Systolic and Diastolic Blood Pressure and Hypertension in CHARGE

SNP identifier	Chr	Position	Nearest Gene	Alleles (coded / other)	Freq. of coded allele	discovery			replication		
						Beta	SE	p-value	Beta	SE	p-value
Systolic blood pressure											
rs12046278	1	10,722,164	*CASZ1*	T/C	0.64	-0.84	0.18	1.84E-06	-0.29	0.15	5.71E-02
rs7571613	2	190,513,907	*PMS1*	A/G	0.82	-0.96	0.19	7.28E-07	-0.23	0.16	1.59E-01
rs448378	3	170,583,593	*MDS1*	A/G	0.52	-0.71	0.15	1.28E-06	-0.36	0.13	4.76E-03
rs2736376	8	11,155,175	*MTMR9*	C/G	0.13	-1.08	0.23	1.90E-06	-0.06	0.19	7.36E-01
rs1910252	8	49,569,915	*EFCAB1*	T/C	0.18	-0.93	0.19	1.70E-06	-0.07	0.17	6.80E-01
rs11014166	10	18,748,804	*CACNB2*	A/T	0.66	0.74	0.16	2.11E-06	0.33	0.13	1.31E-02
rs1004467	**10**	**104,584,497**	***CYP17A1***	**A/G**	**0.90**	**1.20**	**0.25**	**1.99E-06**	**0.94**	**0.21**	**1.08E-05**
rs381815	**11**	**16,858,844**	***PLEKHA7***	**T/C**	**0.26**	**0.84**	**0.17**	**5.76E-07**	**0.52**	**0.14**	**2.72E-04**
rs2681492	**12**	**88,537,220**	***ATP2B1***	**T/C**	**0.80**	**1.26**	**0.19**	**3.01E-11**	**0.50**	**0.17**	**4.07E-03**
rs3184504	**12**	**110,368,991**	***SH2B3***	**T/C**	**0.48**	**0.75**	**0.15**	**5.73E-07**	**0.45**	**0.13**	**6.36E-04**

Levy et al. (2009) *Nature Genetics*

Replication

- GWAS potentially have many false-positives
- Replication in an *independent* cohort is required
- Be mindful of sample size (is there enough power to replicate?)
- Replicate size and direction of effect
- ‘Winner’s curse’ -> effect size overestimated in discovery phase

Meta-analysis of CHARGE and Global BPgen of Top 10 Loci for Systolic and Diastolic Blood Pressure and Hypertension in CHARGE

SNP identifier	Chr	Position	Nearest Gene	Alleles (coded / other)	Freq. of coded allele	discovery			replication		
						Beta	SE	p-value	Beta	SE	p-value
Systolic blood pressure											
rs12046278	1	10,722,164	*CASZ1*	T/C	0.64	-0.84	0.18	1.84E-06	-0.29	0.15	5.71E-02
rs7571613	2	190,513,907	*PMS1*	A/G	0.82	-0.96	0.19	7.28E-07	-0.23	0.16	1.59E-01
rs448378	3	170,583,593	*MDS1*	A/G	0.52	-0.71	0.15	1.28E-06	-0.36	0.13	4.76E-03
rs2736376	8	11,155,175	*MTMR9*	C/G	0.13	-1.08	0.23	1.90E-06	-0.06	0.19	7.36E-01
rs1910252	8	49,569,915	*EFCAB1*	T/C	0.18	-0.93	0.19	1.70E-06	-0.07	0.17	6.80E-01
rs11014166	10	18,748,804	*CACNB2*	A/T	0.66	0.74	0.16	2.11E-06	0.33	0.13	1.31E-02
rs1004467	**10**	**104,584,497**	***CYP17A1***	**A/G**	**0.90**	**1.20**	**0.25**	**1.99E-06**	**0.94**	**0.21**	**1.08E-05**
rs381815	**11**	**16,858,844**	***PLEKHA7***	**T/C**	**0.26**	**0.84**	**0.17**	**5.76E-07**	**0.52**	**0.14**	**2.72E-04**
rs2681492	**12**	**88,537,220**	***ATP2B1***	**T/C**	**0.80**	**1.26**	**0.19**	**3.01E-11**	**0.50**	**0.17**	**4.07E-03**
rs3184504	**12**	**110,368,991**	***SH2B3***	**T/C**	**0.48**	**0.75**	**0.15**	**5.73E-07**	**0.45**	**0.13**	**6.36E-04**

Levy et al. (2009) *Nature Genetics*

Summary

- Different types of statistical tests, but all generate P-value per SNP
  - Linear model is the most common for quantitative traits
- Power considerations...
  - How many individuals? As many as you can
  - How many SNP? As many (good quality) SNP as you can
- Diagnostics (QQ-plots and genomic inflation) important but not perfect
- Replication is essential *why?*

Practical Session

Choose either Part 1, or Parts 2a & 2b

Part 1: power to detect loci
Part 2a: conduct a small GWAS in R
Part 2b: make a QQ-plot