

Genome-Wide Association analysis significance, power and coverage

Yurii Aulchenko

Outline

- Significance of GWA study
- Power of GWA study
- Coverage of GWA study
- Concluding remarks

Bonferroni for GW significance?

- Bonferroni correction
 - GW type 1 error rate of 0.05 corresponds to nominal $P = 0.05/(\text{\# SNPs})$
- Problems:
 - Bonferroni assume that tests are independent
 - **SNPs are not** (because of LD)
 - Therefore Bonferroni is conservative correction (meaning you loose power and can miss association when it is truly there)
 - 550K SNPs were typed, and imputations were done to 2.5M SNPs using HapMap panel. How many tests are done? 0.5M or 2.5M? ... or neither?

Empirical GW significance?

- Empirical estimation of GW (experiment-wise) significance gives exact answer, taking the LD structure and phenotype distribution into account
- Works very well for a single one-stage study
- Problems:
 - May be technically demanding (no problem for few dozens of traits, but is a problem for 100s)
 - More complex design: e.g. two-stage, or multiple independent studies
 - Knowledge accumulation (meta-analysis)

Multiple testing burden: fixed threshold

- Pe'er et al, Genetic Epi, 2008, 32: 381-385
- If we measured all common SNPs in the genome, what number of “independent” SNPs could mimic the null distribution of the test statistics?
- ~1M tests → GW 5% ~ nominal $P = 0.05/1M = 5 \cdot 10^{-8}$
- To keep in mind:
 - Above is true for CEU (2M for Yoruba)
 - Estimated using 1/600th of the genome (ENCODE)

So is my *p*-value significant or not?!

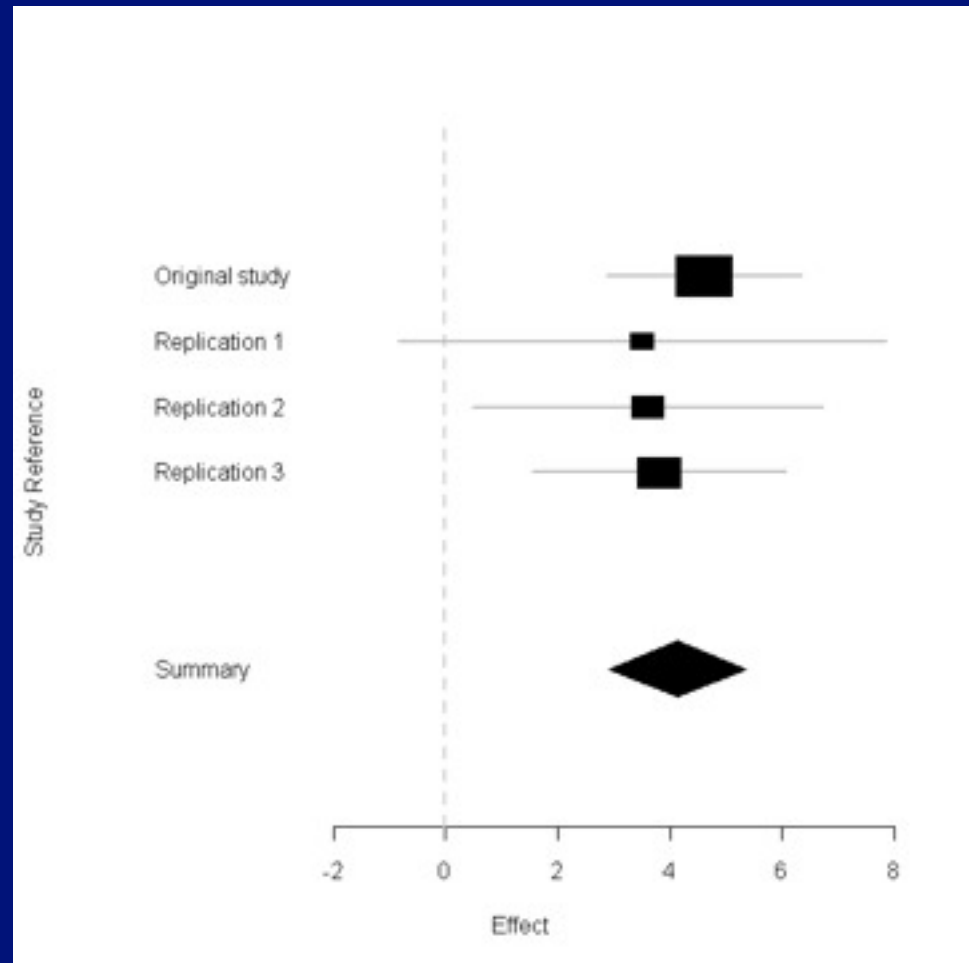
- You (your referees) may be convinced (or not) by a *p*-value which pass
 - Permutation procedure
 - Bonferroni correction
 - $P < 5 \times 10^{-8}$
 - ...
- Ultimate answer: **replication**
- This is a way to
 - Achieve “overwhelming significance”
 - Exclude possibility that the finding is “study-specific”

Example

- A genetic study estimates effect of the SNP rs724016*C allele on height as +4.6 mm (s.e. = 0.88)
 - Nominal p -value = 2×10^{-7}
 - Permutation-based p -value = **0.045**
 - Bonferroni p -value = **0.06**
 - Fixed threshold: **$2 \times 10^{-7} > 5 \times 10^{-8}$**
- Is that a true finding or not?
- Replicate!

Replication in three populations

Study	Effect	S.E.	<i>P</i> -value
Original	4.6	0.88	2×10^{-7}
Rep 1	3.5	2.21	0.11
Rep 2	3.6	1.59	0.02
Rep 3	2.8	1.15	0.001
Total	4.14	0.62	2×10^{-11}



Yurii Aulchenko

Outline

Significance of GWA study

Power of GWA study

Coverage of GWA study

Concluding remarks

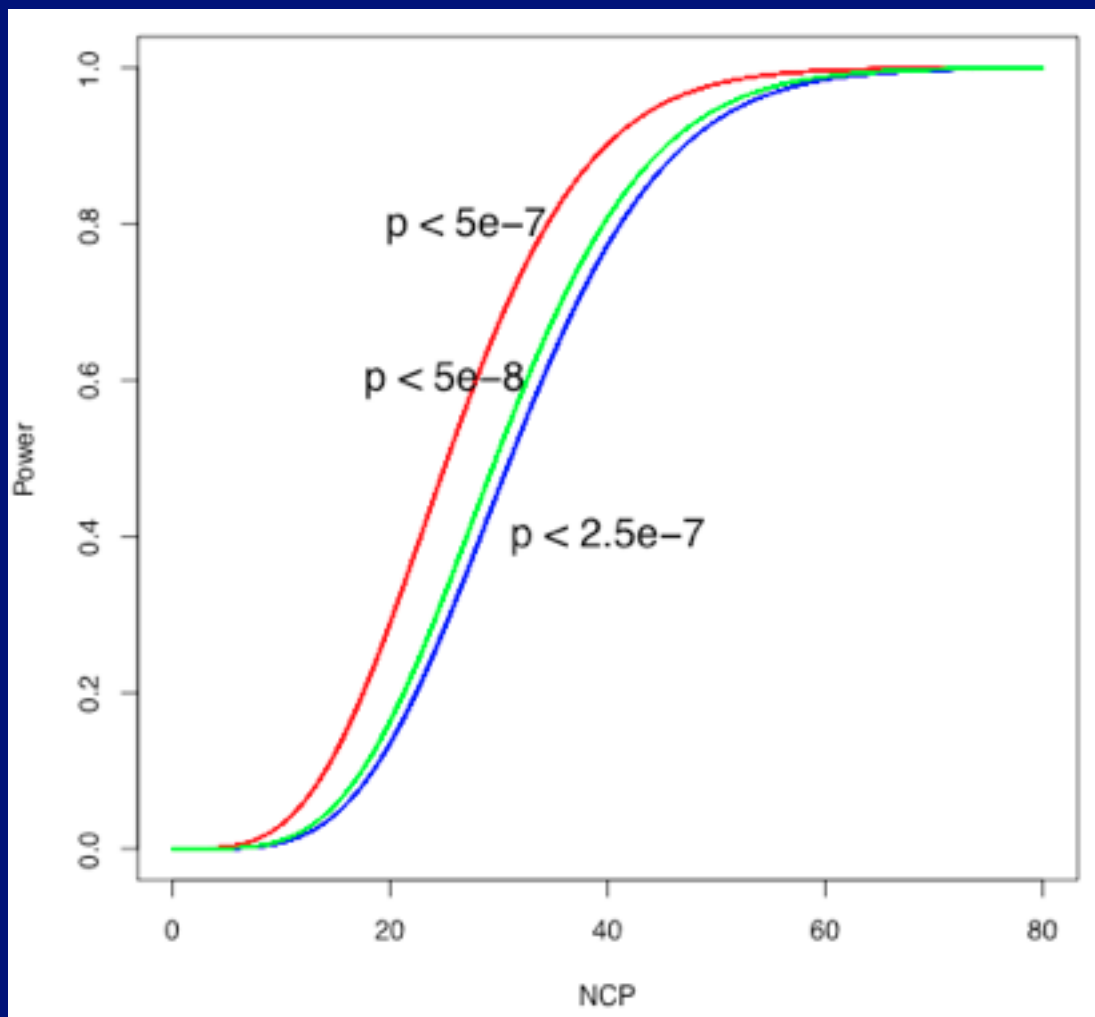
Estimating power

Is study large enough to achieve statistical significance?

- Proportion of trait variance (V_{SNP}) explained by the SNP (this is the coefficient of determination, R^2 !)
- The non-centrality parameter (NCP)
 - Measures (under alternative) how much the $(\chi^2 - \sigma^2/\alpha)$ test statistic is expected to deviate from its expectation under the null
 - $NCP = (\text{no. samples}) \times V_{\text{SNP}}$
- Power to achieve critical threshold X is $\Pr(T^2_{NCP} > X)$
Can be computed in R using `pchisq(X,df=1,ncp=NCP,low=FALSE)`

Exact (not known!) model of the gene action is to be assumed -- need to pick up some reasonable model

Power as function of NCP



Yurii Aulchenko

Power of GWA study

	Sample size	V_{SNP}	NCP	Power to achieve $p < 5 \times 10^{-8}$
“Biggest common loci”:	1,000	3%	30	51%
		1%	10	1%
		0.5%	5	<1%
• HDL: <i>CETP</i> ~ 2.5%	5,000	0.1%	1	<1%
		3%	150	100%
• Total chol.: <i>APOE</i> ~ 0.5%		1%	50	95%
	10,000	0.5%	25	33%
		0.1%	5	<1%
• Height: <i>HMGA2</i> ~ 0.3%		3%	300	100%
		1%	100	100%
		0.5%	50	95%
		0.1%	10	1%

A note on adjustment for the covariates

- Consider *HMGA2* which explains 0.15% of height variation
- Expected power in a study of 14000 people is 20%
- Sex and age together explain ~50% of height variation
- Therefore in the adjusted data the QTL explains 0.3%
- The power to detect it GW is thus 84%

Outline

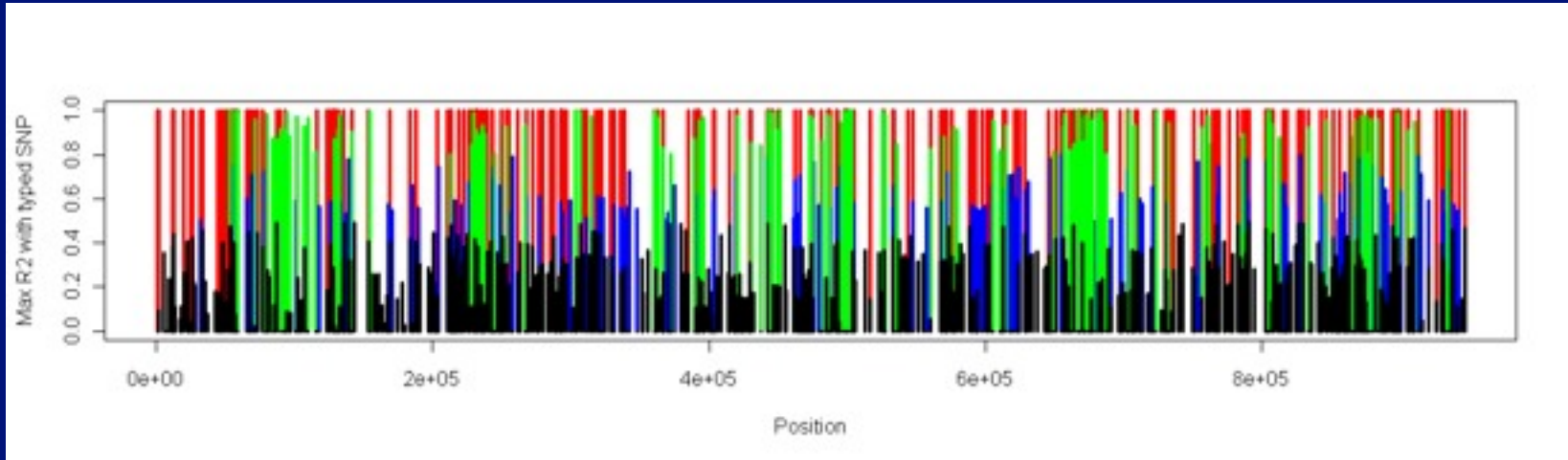
Significance of GWA study

Power of GWA study

Coverage of GWA study

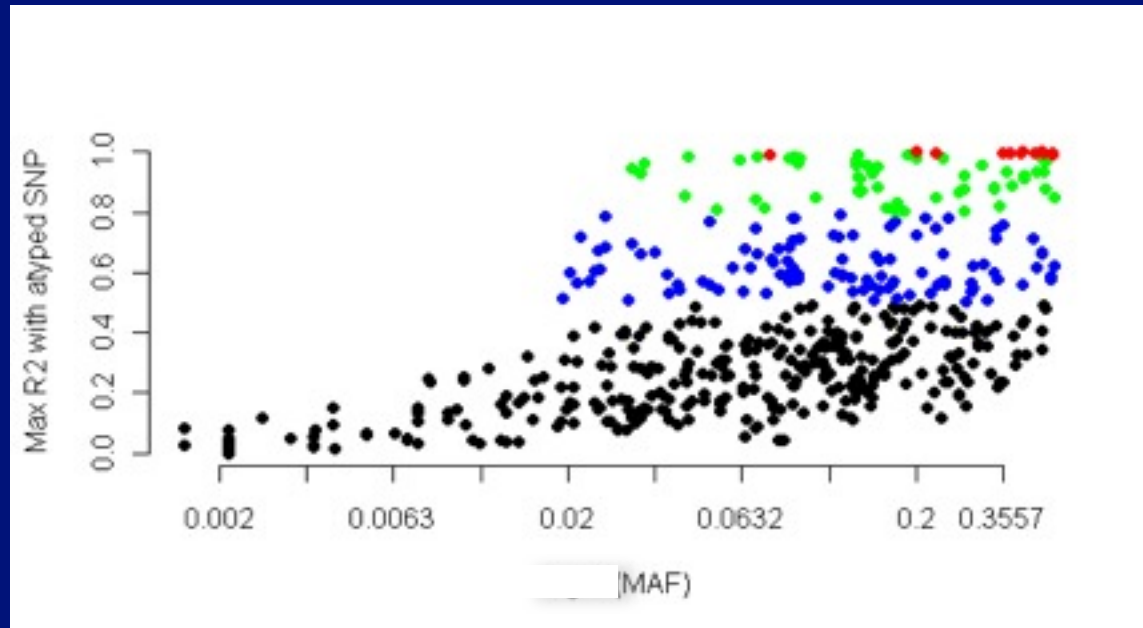
Concluding remarks

How many SNPs we capture?



- Red: typed SNPs
- Green: SNPs with $R^2 \geq 0.8$ with a typed SNP (well-captured)
- Blue: SNPs with $0.8 > R^2 \geq 0.5$ with a typed SNP (captured)
- Black: SNPs with $R^2 < 0.5$

Max R^2 with a typed SNP depends on MAF



- Selected SNPs are likely to be common (if it is very rare, it is not likely to be known!)
- High R^2 between two SNPs is possible only if their frequencies are similar

Genomic coverage by standard panels

- What proportion of common SNPs ($MAF \geq 0.05$) are in the genotyped set or are in high LD ($r^2 > 0.8$) with at least one genotyped SNP?

SNP panel	Type	HapMap population		
		CEU	JPT+CHB	YRI
Affymetrix 111K	Random	31	31	15
Affymetrix 500K	Random	65	66	41
Affymetrix 1M	Combined	80		
Illumina 300K	Tag	75	63	28
Illumina 550K	Tag	87		
Illumina 1M	Tag	91		

Barret & Cardon, NatGenet, 2006
Anderson et al., AJHG, 2008

Outline

Significance of GWA study

Power of GWA study

Coverage of GWA study

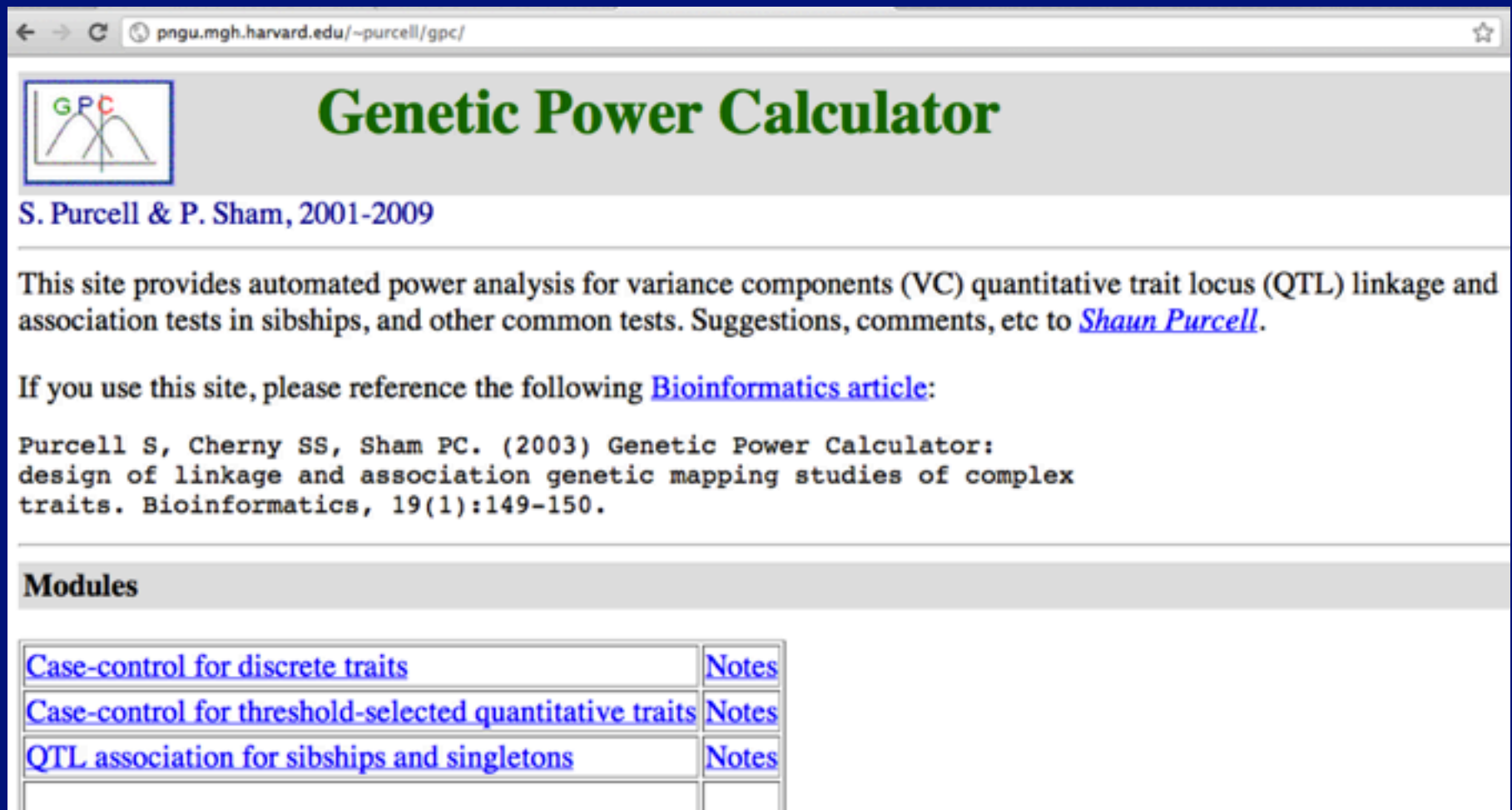
Concluding remarks

Coverage pitfalls

- With 1,000K-2,000K SNP panels we may expect good coverage of common variants for any human population
- Some diseases/traits may be expected to be explained in large part by common variants
- For other disease multiple rare variants may play large role
- Coverage is poor for rare variants

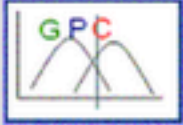
Power for binary traits

Google: “genetic power calculator”



The screenshot shows a web browser window with the address bar displaying `pngu.mgh.harvard.edu/~purcell/gpc/`. The page features a logo on the left with the letters 'GPC' and a graph of two overlapping normal distribution curves. The main heading is 'Genetic Power Calculator' in a large green font. Below this, the authors 'S. Purcell & P. Sham, 2001-2009' are listed. A paragraph describes the site's purpose: 'This site provides automated power analysis for variance components (VC) quantitative trait locus (QTL) linkage and association tests in sibships, and other common tests. Suggestions, comments, etc to [Shaun Purcell](#).' It then asks users to reference a 'Bioinformatics article' if they use the site. The article citation is: 'Purcell S, Cherny SS, Sham PC. (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19(1):149-150.' A 'Modules' section follows, containing a table with three rows of links to specific calculator modules and their corresponding notes.

← → ↻ ⓘ pngu.mgh.harvard.edu/~purcell/gpc/ ☆



Genetic Power Calculator

S. Purcell & P. Sham, 2001-2009

This site provides automated power analysis for variance components (VC) quantitative trait locus (QTL) linkage and association tests in sibships, and other common tests. Suggestions, comments, etc to [Shaun Purcell](#).

If you use this site, please reference the following [Bioinformatics article](#):

Purcell S, Cherny SS, Sham PC. (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19(1):149-150.

Modules

Case-control for discrete traits	Notes
Case-control for threshold-selected quantitative traits	Notes
QTL association for sibships and singletons	Notes