Statistical Power and Type 1 errors

Pak Sham

2019 International Workshop on Statistical Genetic Methods for Human Complex Traits

March 6, 2019

Aims of genetic studies

- To identify genetic variants that influence traits
- To estimate the effect sizes of genetic variants on traits
- To use these genetic variants for trait prediction
- To characterize remaining sources of variation which lead to correlation between individuals
- To combine genetic and phenotypic (individual or family) information for prediction
- These aims feed into broader objectives such as new biological insights and improved human health

Empirical data

- Empirical data is needed to achieve these aims
  - Candidate gene association studies
  - Genome-wide association studies
  - Whole-genome sequencing studies
- How much data is needed?
- No single answer, depends on
  - Specific aim: detection, estimation or prediction
  - Complexity of trait (heterogeneity / polygenicity)
  - Study design: family unit, genotyping, phenotyping, sampling

Detecting an effect

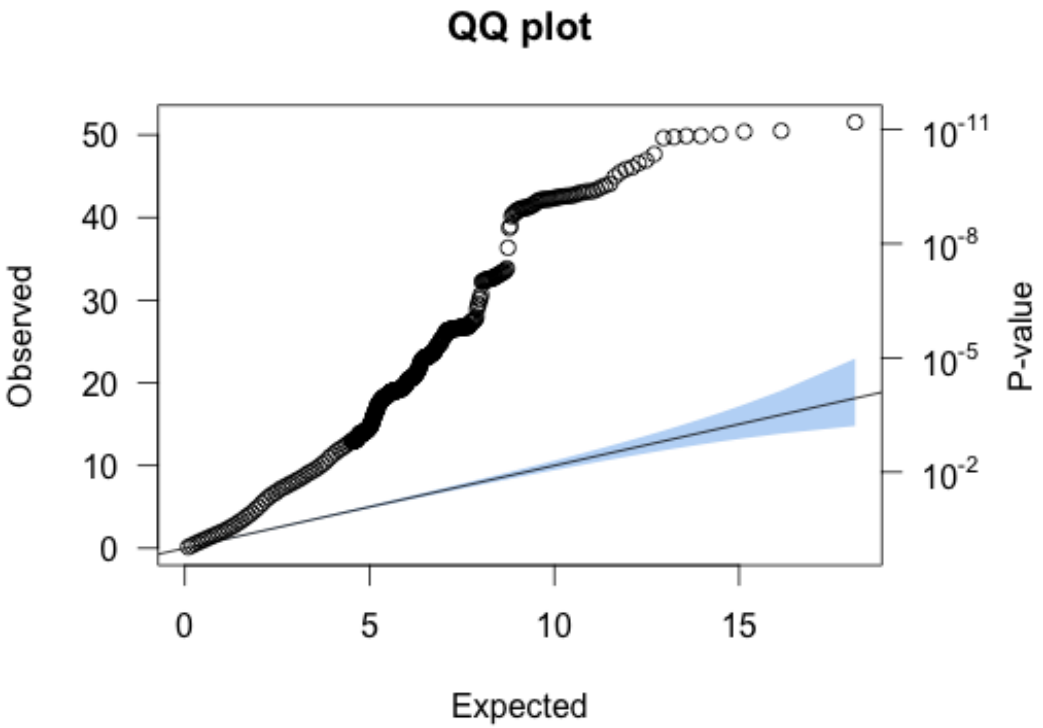
- Classical hypothesis testing
  - Decides whether to reject the null hypothesis
  - Decision based on value of test statistic in relation to its sampling distribution under the null
  - The p-value is the probability of test statistic more extreme than its observed value
  - Null hypothesis is rejected when the p-value is smaller than a desired cut-off (e.g. 0.05)
  - This cut-off p-value is the type 1 error rate of the test (probability of rejecting the null when it is true)

Fisher

Non-replicable findings

- Hypothesis testing was introduced to exert stringent control on type 1 errors (i.e. false positive findings).
- Despite this, non-replicable findings have been a major problem in many fields, including genetics
- Possible reasons:
  - Non-random errors (especially errors correlated with trait)
  - Uncontrolled confounding (e.g. population stratification)
  - Model misspecification (e.g. allele frequencies in linkage)
  - Ignoring dependencies in data (e.g. related individuals)
  - Testing many hypotheses
  - Selective reporting of positive results

Genome-wide studies



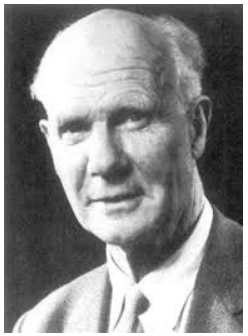
https://www.biostars.org/p/178536/

- Genome-wide studies allow check for inflated type 1 errors by QQ plots
- Multiple testing is explicit so that appropriate p-value threshold can be set
- p-value threshold of 5×10^{-8} was designed to control type 1 error rate to 1 per 20 genome scans in European populations

Statistical power

- Classical hypothesis testing requires only the null hypothesis to be clearly defined.
- A clearly defined alternative hypothesis was introduced later, to calculate the probability of a type 2 error (not rejecting the null hypothesis when the alternative hypothesis is true).
- Statistical power is the probability of rejecting the null under an assumed alternative hypothesis (1 – type 2 error probability)

Neyman



Pearson

Simple power calculation

- Genetic Power Calculator (GPC)

http://zzz.bwh.harvard.edu/gpc/

- Calculates statistical power for association analysis of discrete traits (case-control and case-parents) and continuous traits (singleton and sibships)
- Interactive input of sample size and assumed parameter values under alternative hypothesis (e.g. effect size, allele frequencies, linkage disequilibrium)

Shaun Purcell

Genetic Power Calculator

QTL Association for Sibships

Total QTL variance : (0 - 1)

Dominance : additive QTL effects : (0 - 1) No dominance (* see below)

QTL increaser allele frequency : (0 - 1)

Marker M1 allele frequency : (0 - 1)

Linkage disequilibrium (D-prime) : (0 - 1)

Sibling correlation : (0 - 1) (* see below)

Sample Size : (0 - 10000000) (*N=families, not individuals*)

Sibship Size : Pairs Both parents genotyped

User-defined type I error rate : 0.05 (0.00000001 - 0.5)

User-defined power: determine N : 0.80 (0 - 1)
(1 - type II error rate)

Process Reset

Genetic Power Calculator

QTL Association for Sibships

Total QTL variance	:	.01	(0 - 1)
Dominance : additive QTL effects	:	0	(0 - 1) ☑ No dominance (* see below)
QTL increaser allele frequency	:	.1	(0 - 1)
Marker M1 allele frequency	:	.2	(0 - 1)
Linkage disequilibrium (D-prime)	:	.9	(0 - 1)
Sibling correlation	:	.4	(0 - 1) (* see below)

Sample Size	:	5000	(0 - 10000000) (*N=families, not individuals*)
Sibship Size	:	Pairs ☐ Both parents genotyped	

User-defined type I error rate	:	0.00000005	(0.00000001 - 0.5)
User-defined power: determine N (1 - type II error rate)	:	0.80	(0 - 1)

Genetic Power Calculator

QTL Association : Sibships

Components of variance at QTL

Additive QTL variance	0.01
Dominance QTL variance	0
Shared residual variance	0.395
Nonshared residual variance	0.595

Components of variance at marker

Additive QTL variance	0.0036
Dominance QTL variance	0
Shared residual variance	0.3982
Nonshared residual variance	0.5982

Misc. statistics

Sibship Size	2
Sample Size	5000
QTL allele frequency (p)	0.1
Marker allele frequency (m1)	0.2
Linkage disequilibrium (D-prime)	0.9
Sibling correlation	0.4

Test Statisitics : Power Analysis

Between Sibships Association NCP = 19.32

Alpha	Power	Sample for 80% power
0.1	0.997	1600
0.05	0.9926	2031
0.01	0.9656	3022
0.001	0.8655	4418
5e-08	0.1456	1.025e+04

Within Sibship Association NCP = 15.02

Alpha	Power	Sample for 80% power
0.1	0.9872	2058
0.05	0.9723	2612
0.01	0.9032	3887
0.001	0.7208	5683
5e-08	0.05758	1.318e+04

Overall Association NCP = 34.35

Alpha	Power	Sample for 80% power
0.1	1	900
0.05	1	1143
0.01	0.9995	1700
0.001	0.9949	2486
5e-08	0.6588	5765

All tests are for additive effects only (1 df); two-tailed

Exercise

- Use GPC to calculate the statistical power of overall association test under the same assumptions, but on a sample of 10,000 unrelated singletons.
- Why does the power change, when the number of subjects is the same?
- Returning to 5000 sib pairs, investigate the power (or non-centrality parameter) for overall association test when the sib correlation is 0.005, 0.25, 0.50, 0.75
- How do you explain the impact of increasing sib correlation on power?

Non-centrality parameter is the difference in mean between the null and alternative distributions of the test statistic. It determines power for any p-value significance level, and is linearly related to sample size.

Genetic Power Calculator

QTL Association : Sibships

Components of variance at QTL

Additive QTL variance	0.01
Dominance QTL variance	0
Shared residual variance	0.395
Nonshared residual variance	0.595

Components of variance at marker

Additive QTL variance	0.0036
Dominance QTL variance	0
Shared residual variance	0.3982
Nonshared residual variance	0.5982

Misc. statistics

Sibship Size	1
Sample Size	10000
QTL allele frequency (p)	0.1
Marker allele frequency (m1)	0.2
Linkage disequilibrium (D-prime)	0.9
Sibling correlation	0.4

Test Statisitics : Power Analysis

Between Sibships Association NCP = 36.06

Alpha	Power	Sample for 80% power
0.1	1	1714
0.05	1	2176
0.01	0.9997	3238
0.001	0.9967	4734
5e-08	0.7102	1.098e+04

Within Sibship Association NCP = 0

Alpha	Power	Sample for 80% power
0.1	0.1	inf
0.05	0.05	inf
0.01	0.01	inf
0.001	0.001	inf
5e-08	5e-08	inf

Overall Association NCP = 36.06

Alpha	Power	Sample for 80% power
0.1	1	1714
0.05	1	2176
0.01	0.9997	3238
0.001	0.9967	4734
5e-08	0.7102	1.098e+04

All tests are for additive effects only (1 df); two-tailed

Results

Sibling correlation	Non-centrality parameter
0.005	35.99
0.25	33.05
0.50	36.06
0.75	51.58

- Having family data does not necessarily decrease statistical power
- The sibship association test is partitioned into between-sibships and within-sibships components (Fulker et al, 1999, 64, 259-267)
- High sib correlation decreases within-sibship variation, and this increases the power of the within-sibships component (Sham et al, 1999, AJHG, 66, 1616-1630)

How to increase power

- Increase sample size
- Improve accuracy of trait measurement
- Repeated measures (average out fluctuations)
- Reduce residual variation (e.g. age, sex)
- Joint analysis of multiple correlated phenotypes
- Select subjects at either extremes of trait values
- Increase SNP density (greater LD, improved imputation)
- Consider each p-value in relation to overall distribution of p-values - False Discovery Rate (FDR)
- Stratify SNPs into functional classes and perform separate FDR on each class

FDR - intuition

- Suppose that a study has performed 100 tests, and 20 of these are significant at $p<0.05$. How many of these 20 significant results would you guess constitute true discoveries?
- By chance, one would expect 5 out of 100 tests to be significant at $p<0.05$. Therefore one might guess that 15/20 of the significant results to be true discoveries (or in other words, 5/20 to be false discoveries).

Benjamini-Hochberg

Benjamini & Hochberg (1995) FDR Procedure:

1. Set FDR (e.g. to 0.05)
2. Rank the tests in ascending order of p-value, giving $p_1 \leq p_2 \leq \ldots \leq p_r \leq \ldots \leq p_m$
3. Then find the test with the highest rank, r, for which the p-value, p_r, is less than or equal to (r/m) × FDR
4. Declare the tests of rank 1, 2, …, r as significant
5. Define $q_m = p_m$, then calculate $q_r = \text{Min}\,(p_r m/r, q_{r+1})$ for r = m-1 to 1.

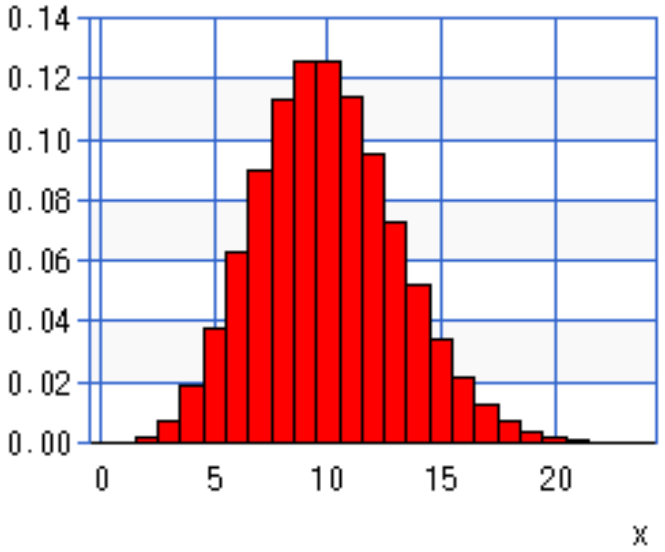
B & H FDR procedure

FDR=0.05

Rank	P-value	(Rank/m)×FDR	Reject H_0 ?	Q-value
1	.001	.005	1	.01
2	.010	.010	1	.05
3	.165	.015	0	0.51
4	.205	.020	0	0.51
5	.396	.025	0	0.75
6	.450	.030	0	0.75
7	.641	.035	0	0.916
8	.781	.040	0	0.953
9	.901	.045	0	0.953
10	.953	.050	0	0.953

Power under polygenicity

- Many SNPs contribute to complex traits
- A GWAS has multiple chances of detecting true associations
- Suppose a trait has 1,000 independent causal SNPs, and a study has only 1% power to detect each of these SNPs.
- The number of significant causal SNPs follows a binomial distribution with n=1,000 and p=0.01



- Study likely to detect 3 to 23 causal SNPs.
- These SNPs are no different from the other causal SNPs.
- Power of independent replication of each SNP is only 1%, with same sample size and p-value threshold

Estimation accuracy

- Another aim of GWAS is to estimate of effect size of a SNP
- Accuracy of estimate is usually measured by standard error
- With standardized dependent and independent variables, the standard error of the ordinary least squares (OLS) regression coefficient estimate is $1/\sqrt{n}$

$$\hat{\beta} = \beta + \epsilon \qquad \varepsilon \sim N(0, 1/n) \qquad z = \hat{\beta}/\sqrt{n} \qquad p = 2\Phi(-|z|)$$

- Accurate estimates of effect sizes are important for accurate prediction of the trait from SNP information

Prediction accuracy

- Assume all SNPs causal, with normally distributed effect sizes β
- Calculate polygenic risk score (PRS) by weighting each SNP by estimated effect size $\hat{\beta}$
- Correlation between PRS and "true" PRS equals r(β, $\hat{\beta}$)

$$Cov(\beta, \hat{\beta}) = Var(\beta) = \frac{h^2}{m}$$

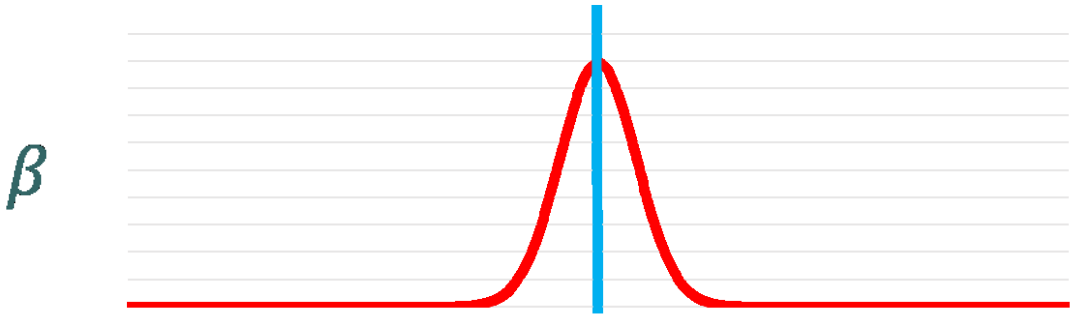
$$Var(\hat{\beta}) = Var(\beta) + \frac{1}{n} = \frac{h^2}{m} + \frac{1}{n}$$

$$r(\beta, \hat{\beta}) = \frac{Var(\beta)}{\sqrt{Var(\beta)(Var(\beta) + 1/n)}} = \frac{1}{\sqrt{1 + \frac{m}{nh^2}}}$$

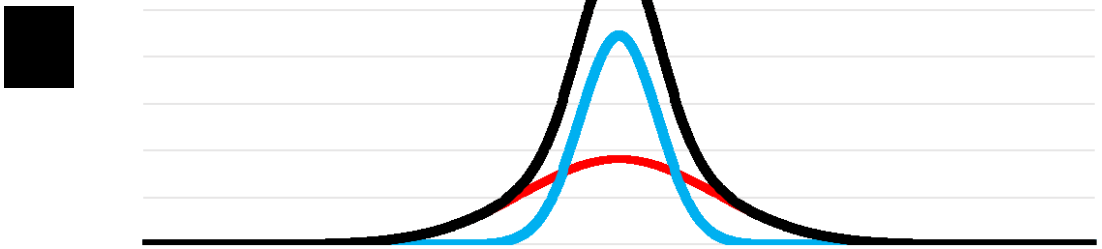
Dudbridge (2013) PLOS Genetics

Introducing null SNPs

- More realistically only a proportion of SNPs are causal and have a normal distribution of effect sizes



Distribution of true effect sizes: mixture of 0 (null SNPs), and normal (causal SNPs)



Distribution of estimates: mixture of normals with sampling, null SNPs having variance only, causal SNPs having both sampling variance plus effect size variance

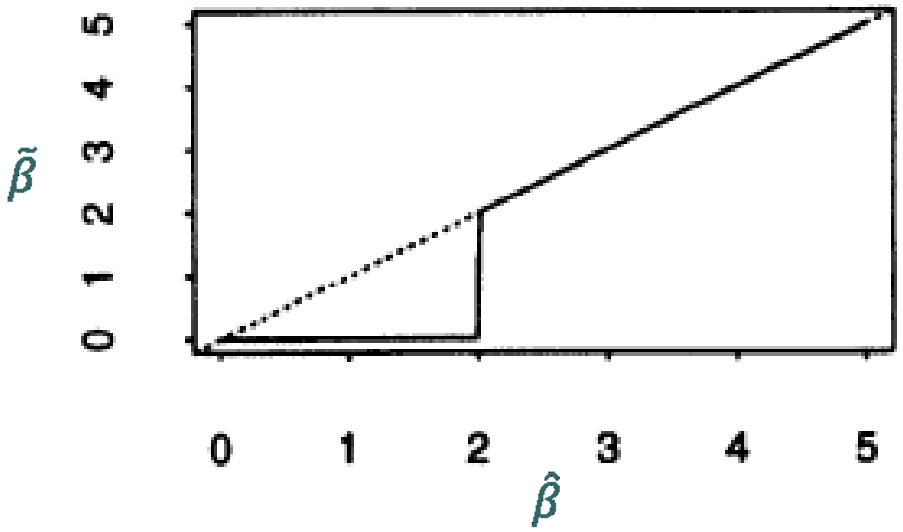
P-value thresholding

Set P_T

If $p < P_T$, then $\tilde{\beta} = \hat{\beta}$

Otherwise $\tilde{\beta} = 0$

The value of P_T may be fine-tuned to maximize r^2



Weight SNPs by $\tilde{\beta}$

Called “subset selection” by Tibshirani (1996)

Euesden, Lewis and O’Reilly (2014) PRSice: polygenic risk score software

Local true discovery rate (TDR)

If a proportion of SNPs, π_0 follow H_0 (i.e. have no effect), then it may be reasonable to shrink $\hat{\beta}$ by the posterior probability of H_1 given $\hat{\beta}$ (i.e. the local TDR)

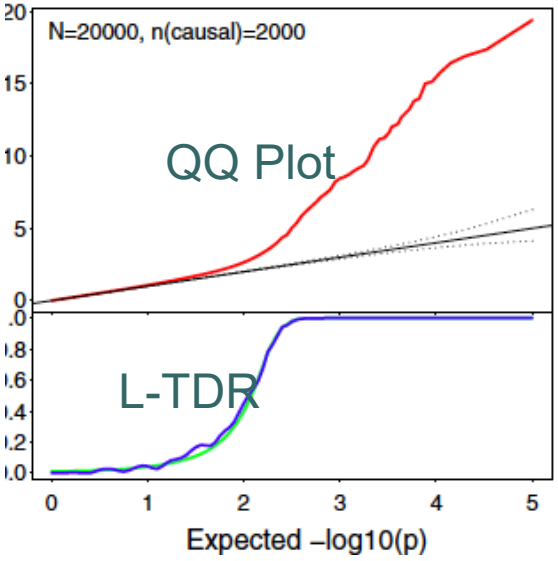
$$\tilde{\beta} = Prob(H_1|\hat{\beta})\hat{\beta}$$

If the effects the SNPs under H_1have a normal distribution, then

$$E(\beta|\hat{\beta}) = E(\beta|H_1,\hat{\beta})Prob(H_1|\hat{\beta})$$

$$= \frac{1}{\sqrt{1+\frac{(1-\pi_0)m}{nh^2}}}Prob(H_1|\hat{\beta})\hat{\beta}$$

Vilhjalmsson et al. (2015) called this Bpred. It is directly proportional to local TDR weighted $\hat{\beta}$ and therefore has the same prediction r^2

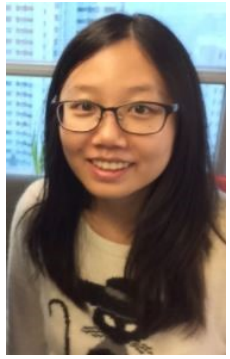


Mak et al. (2016) Local true discovery rate weighted polygenic scores using GWAS summary statistics

Vilhjalmsson et al. (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores

Comparison of methods

- In the presence of null SNPs, both p-value thresholding and local TDR weighting have better prediction accuracy than simple OLS weighting
- p-value thresholding and local TDR weighting have similar predictive accuracy
- Local TDR has slight advantage in not needing to optimize p-value threshold, which requires fine-tuning in a sample independent from the original GWAS



Tian Wu