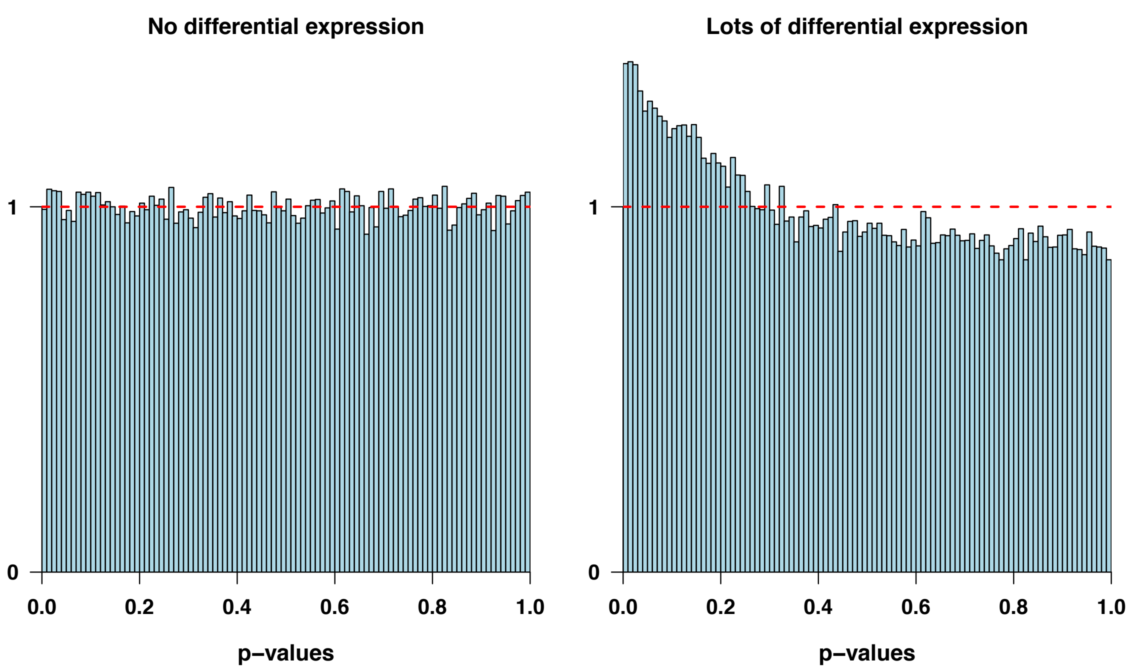
Multiple Hypothesis Testing

Type I error and false discovery control



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

Hypothetical example (no differential expression):

- Microarray with 10,000 genes.
- Calculate 10,000 p-values.
- Call genes "significant" if p-value < 0.05.
- Expected Number of False Positives:

  $10{,}000 \times 0.05 = 500$.

- Many procedures have been developed to control the Family Wise Error Rate (the probability of at least one type I error).
- Two general types of FWER corrections:
  - Single step: equivalent adjustments made to each p-value.
  - Sequential: adaptive adjustment made to each p-value.

Simple single step approach: Bonferroni.

- Very simple method for ensuring that the overall type I error rate of α is maintained when performing m hypothesis tests.
- Rejects any hypothesis with p-value $\leq \alpha$/m.
- The Bonferroni adjusted p-value is

$$p_j^{Bonf} = \min\{m \times p_j, 1\}$$

- For example, if we want to have an experiment wide type I error rate of 0.05 when we perform 10,000 hypothesis tests, we needed a p-value of 0.05 / 10,000 = 5×10^{-6} to declare significance.

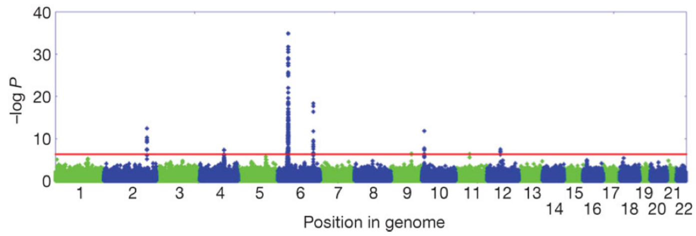
Simple sequential method: Holm-Bonferroni.

- Order the unadjusted p-values such that $p_1 \leq p_2 \leq \cdots \leq p_m$.
- Holm-Bonferroni uniformly delivers more power than the Bonferroni correction by testing only the most extreme p value against the strictest criterion, and the others against progressively less strict criteria.
- The Holm adjusted p-value is

$$p_j^{Holm} = \min\{m - j + 1 \times p_j, 1\}$$

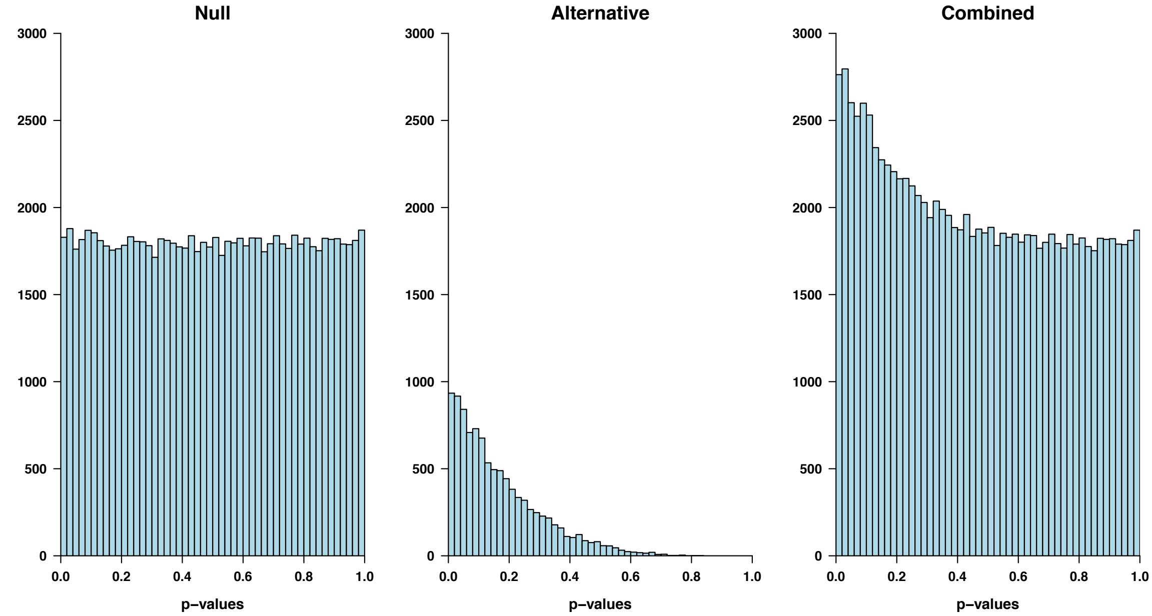
- The point here is that we do not multiply every p_j by the same factor m.

- The FWER is appropriate when you want to guard against **any** false positives.
- For example, this is usually done in genome-wide association studies.



- The general null hypothesis (that all the null hypotheses are true) is rarely of interest.
- There is a high probability of type 2 errors, i.e. of not rejecting the general null hypothesis when important effects exist.

- In many cases (particularly in genomics) we can live with a certain number of false positives.
- This is for example the case in gene expression studies, when we suspect a fair number of genes to be differentially expressed.
- In these cases, the more relevant quantity to control is the False Discovery Rate (FDR).
- The FDR is designed to control the proportion of false positives *among the set of rejected hypotheses.*



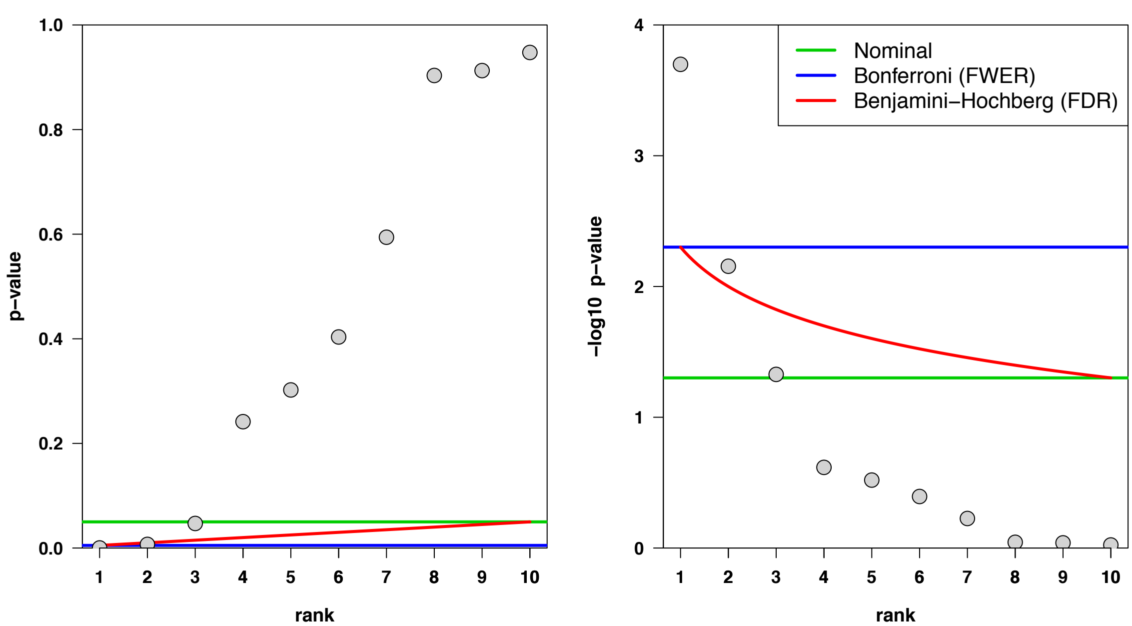
Declared ↓	H_0 is true	H_a is true	Total
Significant	V	S	R
Non-significant	U	T	m – R
Total	m_0	m - m_0	m

- Bonferroni and such control the family-wise error rate.
  → V/(V+U).
- The FDR controls the false positive rate.
  → V/(V+S).

Benjamini and Hochberg FDR.

To control FDR at level δ:

- Order the unadjusted p-values: $p_1 \leq p_2 \leq \cdots \leq p_m$.
- Find the test with the highest rank j for which the p-value p_j is less than or equal to (j / m) $\times$ δ.
- Declare the tests of rank 1, 2, . . ., j as significant.



Difference in interpretation:

Suppose 550 out of 10,000 genes are significant at the 0.05 level.

- False Discovery Rate < 0.05:

  Expect $0.05 \times 550 = 27.5$ false positives.
- Family Wise Error Rate < 0.05:

  The probability of at least 1 false positive < 0.05.
- In most settings, the latter is extremely unlikely, unless the sample size is huge!

John Storey's positive FDR (pFDR):

$$\text{FDR} = \text{E}\left[\frac{V}{R} \mid R > 0\right] \times P(R > 0)$$

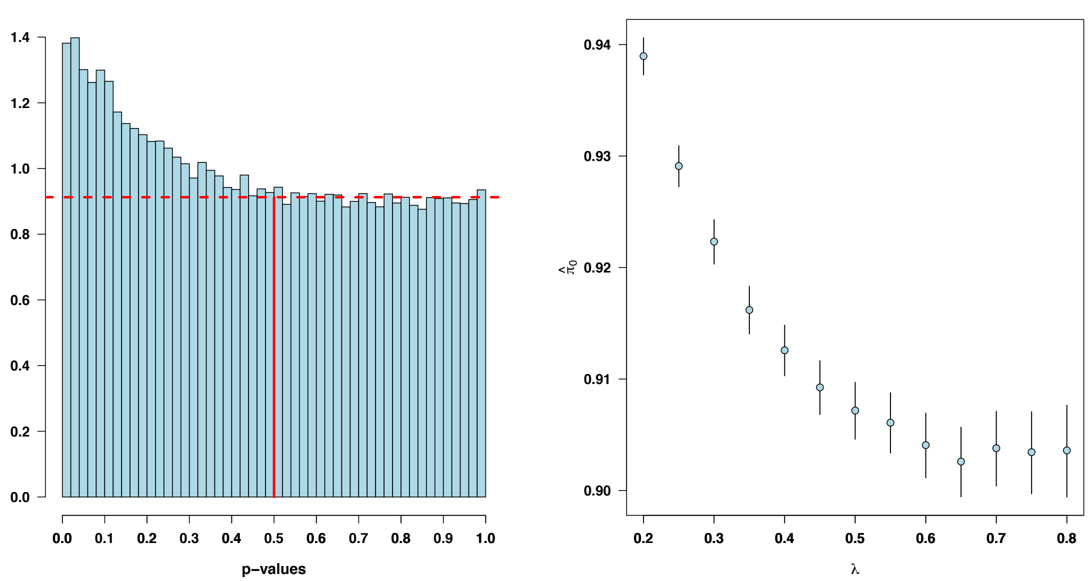
$$\text{pFDR} = \text{E}\left[\frac{V}{R} \mid R > 0\right]$$

- Since $P(R > 0)$ is ~ 1 in most genomics experiments, the FDR and the pFDR are very similar.
- Omitting $P(R > 0)$ facilitated the development of a measure of significance in terms of the FDR for each hypothesis.

Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

Q-values:

- The q-value is defined as the minimum FDR that can be attained when calling that gene significant (i.e., expected proportion of false positives incurred when calling that gene significant).
- The estimated q-value is a function of the p-value for that test and the distribution of the entire set of p-values from the family of tests being considered.
- In testing for differential expression, if a gene has a q-value of 0.10 it means that we can expect 10% of genes that show p-values at least as small as this gene to be false positives.



We begin by estimating the FDR when calling all genes significant with p-values $\leq$ t.

A heuristic motivation:

$$\text{FDR}(t) \approx \frac{\text{E}\left[V(t)\right]}{\text{E}\left[R(t)\right]} = \frac{\text{E}\left[\#\{\text{null } p_i \leq t\}\right]}{\text{E}\left[\#\{p_i \leq t\}\right]} = \frac{m_0 \times t}{\text{E}\left[\#\{p_i \leq t\}\right]}$$

Thus:

$$\widehat{\text{FDR}}(t) = \frac{\hat{m}_0 \times t}{\#\{p_i \leq t\}}$$

PMID 12883005

- We first estimate the more easily interpreted term $\pi_0 = m_0/m$, the proportion of truly null (non-differentially expressed) genes.

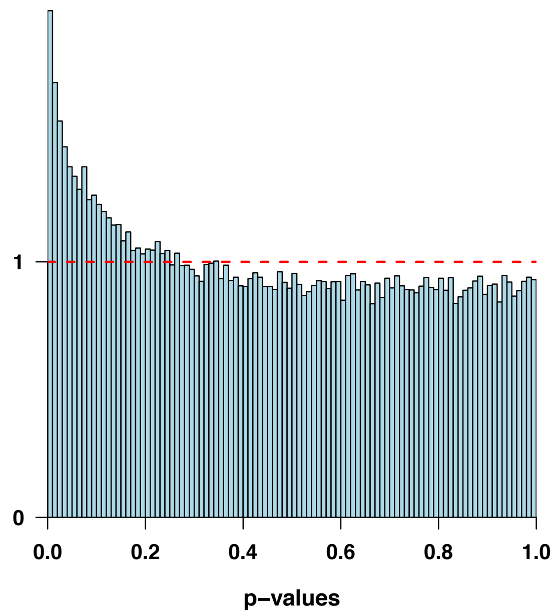
$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{m \times (1 - \lambda)}$$

- We then use $\hat{m}_0 = \hat{\pi}_0 \times m$.
- Note that $1 - \hat{\pi}_0$ is the estimated proportion of differentially expressed genes.
- The q-value is formally defined as the minimum FDR that can be attained when calling that gene significant:

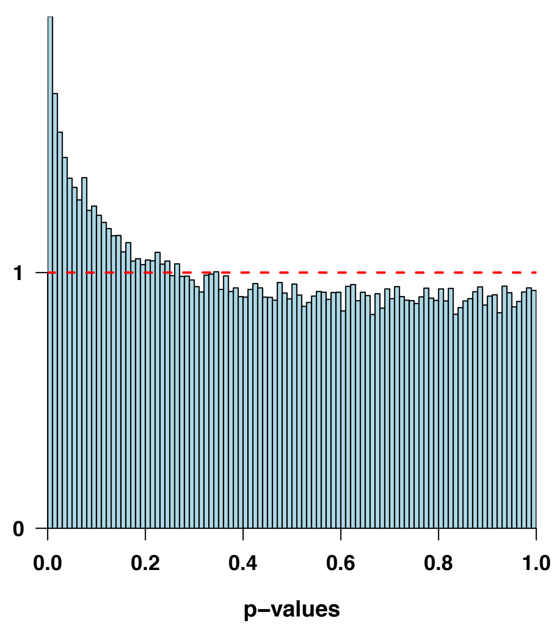
$$\hat{q}(p_i) = \min_{t \geq p_i} \widehat{\text{FDR}}(t)$$

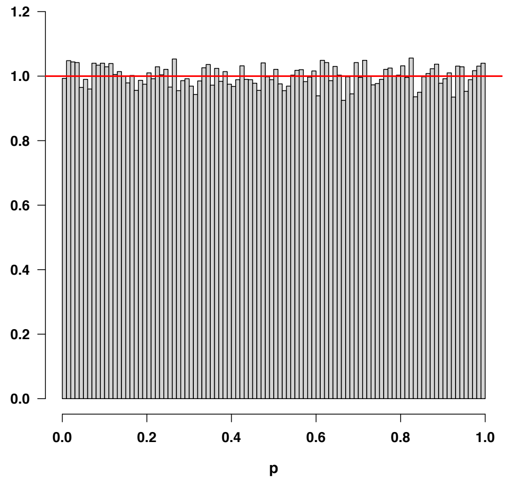
PMID 12883005

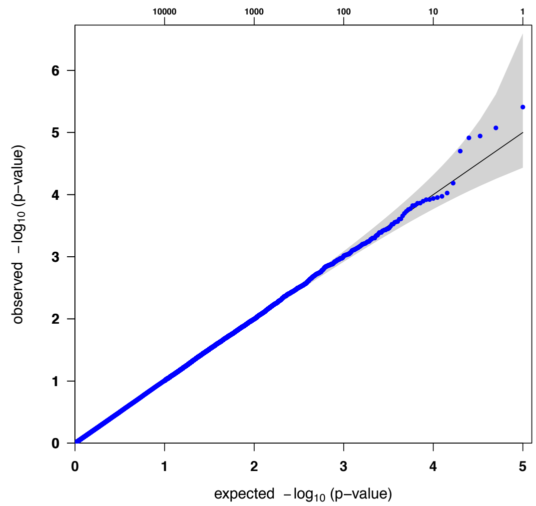
	p	q	p.bonf
1	1.92e-06	0.11814	0.19188
2	3.90e-06	0.11814	0.38955
3	8.43e-06	0.11814	0.84317
4	9.35e-06	0.11814	0.93497
5	1.11e-05	0.11814	1.00000
6	1.14e-05	0.11814	1.00000
7	1.14e-05	0.11814	1.00000
8	1.22e-05	0.11814	1.00000
9	1.24e-05	0.11814	1.00000
10	1.44e-05	0.11814	1.00000
11	1.44e-05	0.11814	1.00000
12	1.64e-05	0.11997	1.00000
13	1.76e-05	0.11997	1.00000
14	1.97e-05	0.11997	1.00000
15	1.99e-05	0.11997	1.00000
16	2.53e-05	0.14035	1.00000
17	2.64e-05	0.14035	1.00000
18	3.03e-05	0.14795	1.00000
19	3.11e-05	0.14795	1.00000
20	3.67e-05	0.15919	1.00000

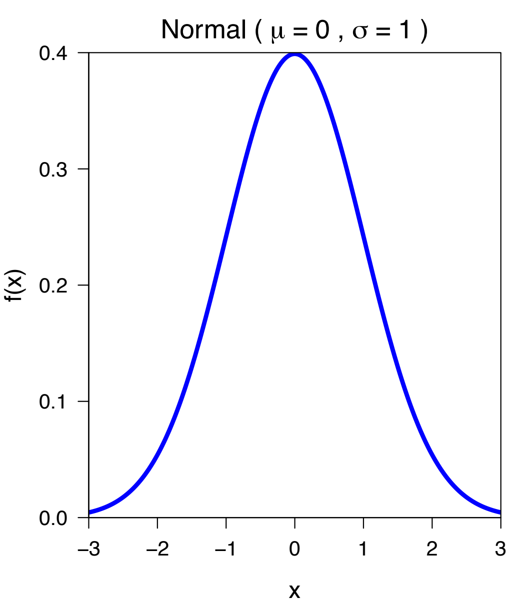


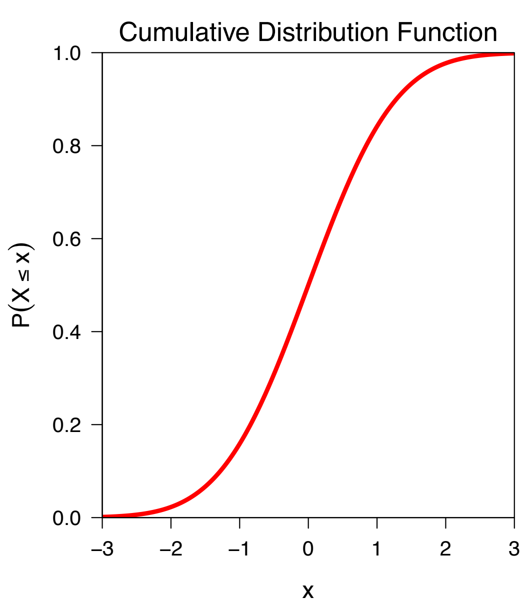
	p	q	p.bonf
1	2.27e-31	2.05e-26	2.27e-26
2	9.05e-13	4.09e-08	9.05e-08
3	3.37e-12	1.02e-07	3.37e-07
4	2.57e-11	5.80e-07	2.57e-06
5	2.20e-09	3.97e-05	2.20e-04
6	1.92e-06	2.89e-02	1.92e-01
7	3.90e-06	5.03e-02	3.90e-01
8	5.26e-06	5.94e-02	5.26e-01
9	8.43e-06	6.99e-02	8.43e-01
10	9.35e-06	6.99e-02	9.35e-01
11	9.55e-06	6.99e-02	9.55e-01
12	1.11e-05	6.99e-02	1.00e+00
13	1.14e-05	6.99e-02	1.00e+00
14	1.14e-05	6.99e-02	1.00e+00
15	1.22e-05	6.99e-02	1.00e+00
16	1.24e-05	6.99e-02	1.00e+00
17	1.44e-05	7.22e-02	1.00e+00
18	1.44e-05	7.22e-02	1.00e+00
19	1.64e-05	7.77e-02	1.00e+00
20	1.76e-05	7.95e-02	1.00e+00

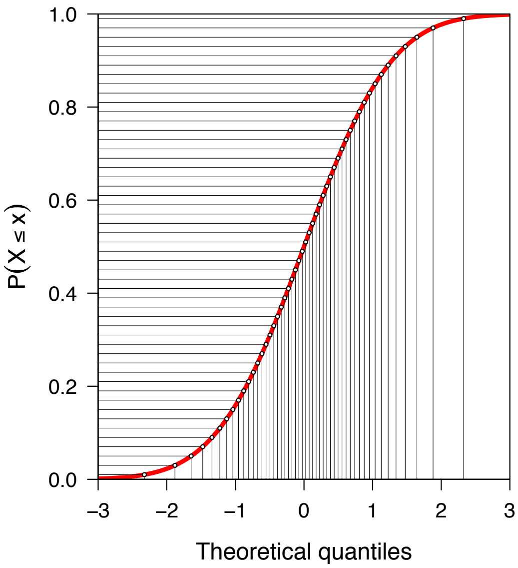


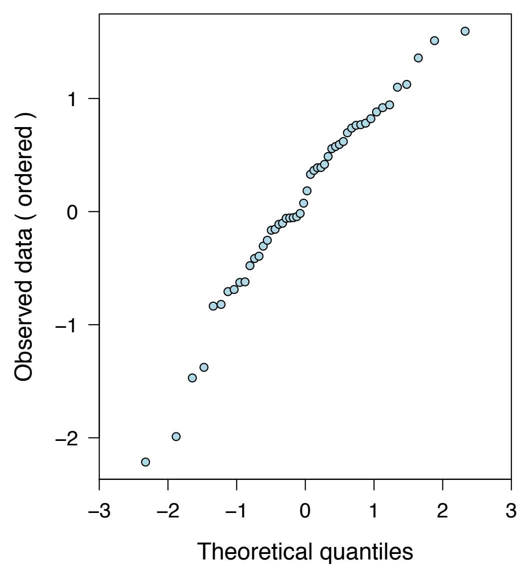

p

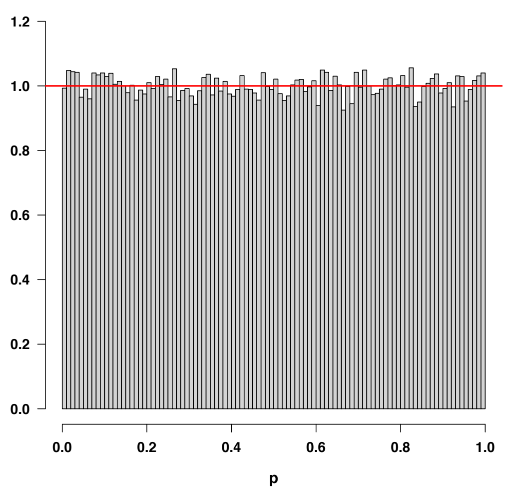

observed $-\log_{10}$ (p−value)
expected $-\log_{10}$ (p−value)
10000
1000
100
10
1

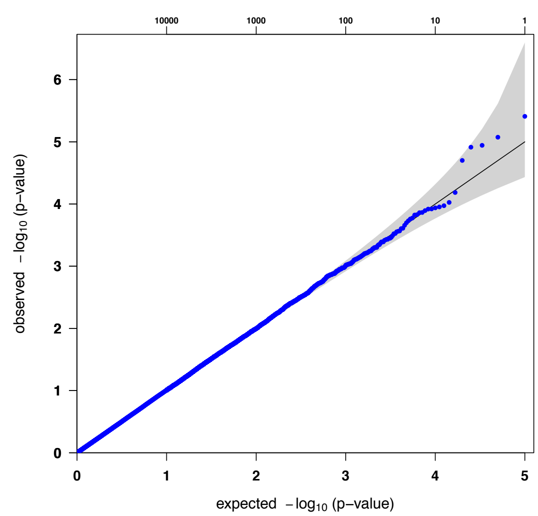

Normal (μ = 0 , σ = 1)
f(x)
x

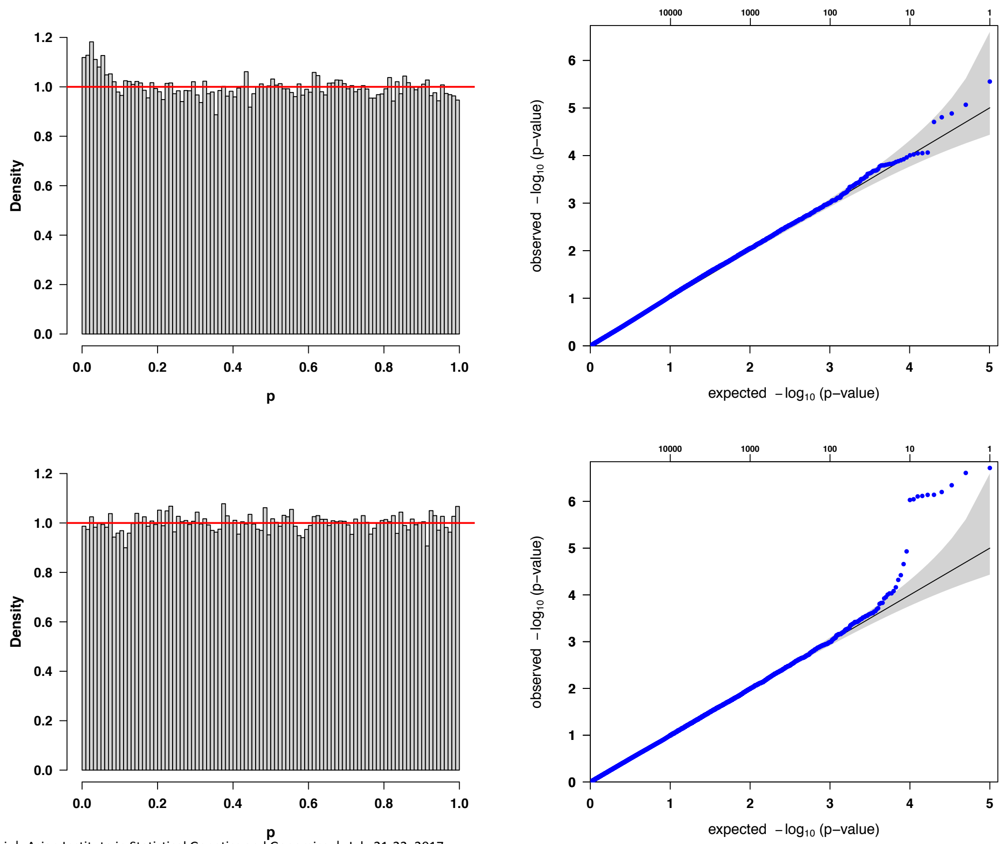

Cumulative Distribution Function
P(X ≤ x)
x

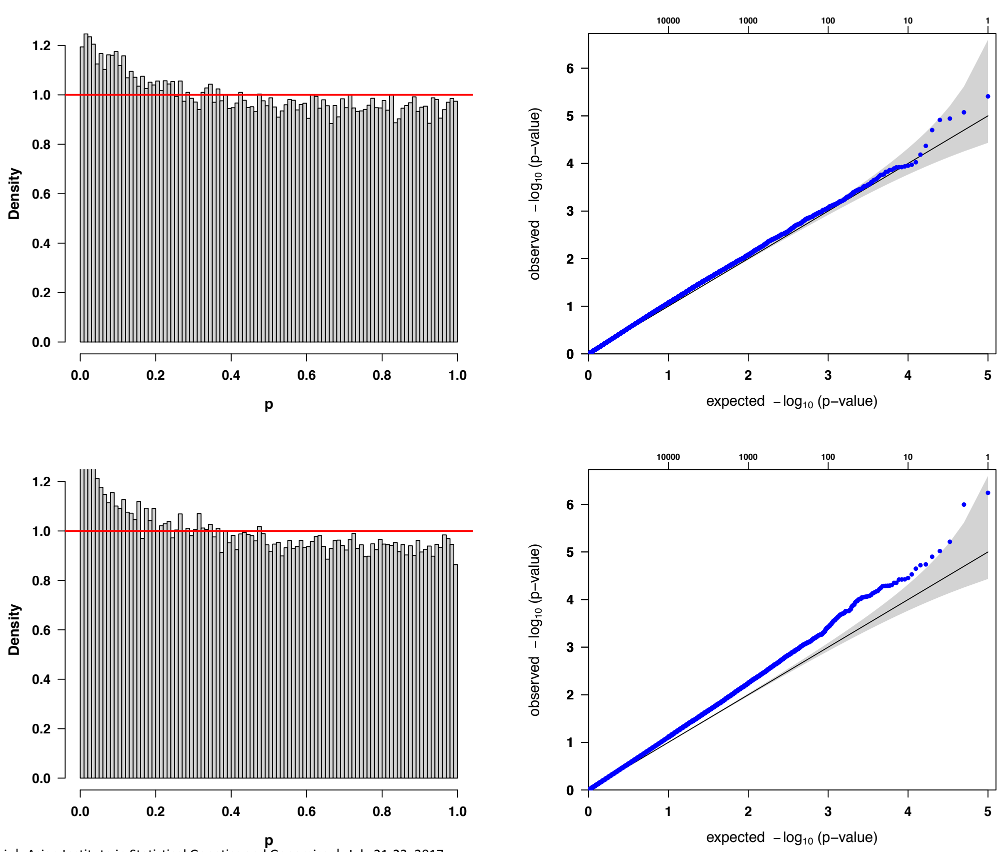

1.0
0.8
0.6
0.4
0.2
0.0
P(X ≤ x)
−3
−2
−1
0
1
2
3
Theoretical quantiles


1
0
−1
−2
Observed data (ordered)
−3
−2
−1
0
1
2
3
Theoretical quantiles


1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
p


10000
1000
100
10
1
6
5
4
3
2
1
0
observed −log10 (p−value)
0
1
2
3
4
5
expected −log10 (p−value)


Density
p
observed −log10 (p−value)
expected −log10 (p−value)


Density
p
observed −log10 (p−value)
expected −log10 (p−value)

Iterson *et al. Genome Biology* (2017) 18:19
DOI 10.1186/s13059-016-1131-9

Genome Biology

METHOD Open Access

Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution

Maarten van Iterson[1*], Erik W. van Zwet[2], the BIOS Consortium and Bastiaan T. Heijmans[1]

Abstract

We show that epigenome- and transcriptome-wide association studies (EWAS and TWAS) are prone to significant inflation and bias of test statistics, an unrecognized phenomenon introducing spurious findings if left unaddressed. Neither GWAS-based methodology nor state-of-the-art confounder adjustment methods completely remove bias and inflation. We propose a Bayesian method to control bias and inflation in EWAS and TWAS based on estimation of the empirical null distribution. Using simulations and real data, we demonstrate that our method maximizes power while properly controlling the false positive rate. We illustrate the utility of our method in large-scale EWAS and TWAS meta-analyses of age and smoking.

Keywords: Epigenome- and transcriptome-wide association studies, Bias, Inflation, Empirical null distribution, Gibbs sampler, Meta-analysis

PMID 28129774

Dependency in expression data

Since measured gene expression levels are not independent, the statistics (p-values) are not independent.

Genes in the same pathway, near each other, with sequence similarity, might be dependent.

Each of these dependencies is **local**. They probably occur in finite clumps.

Given "clumpy microarray dependence" and a large number of hypothesis tests, Storey et al showed that
1) the FDR is controlled, and
2) the estimated q-values conservatively estimate the true q-values.

PMID 12883005

STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:1960–1972
Published online 12 March 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3237

Practical guidelines for assessing power and false discovery rate for a fixed sample size in microarray experiments

Tiejun Tong[1] and Hongyu Zhao[2,3,*,†]

[1]*Department of Applied Mathematics, University of Colorado, Boulder, CO 80309, U.S.A.*
[2]*Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, U.S.A.*
[3]*Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, U.S.A.*

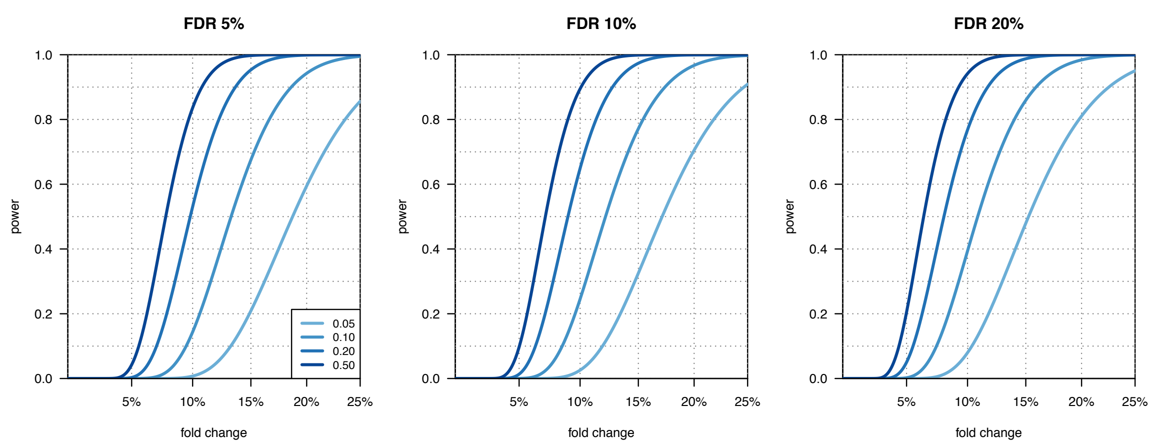
SUMMARY

One major goal in microarray studies is to identify genes having different expression levels across different classes/conditions. In order to achieve this goal, a study needs to have an adequate sample size to ensure the desired power. Owing to the importance of this topic, a number of approaches to sample size calculation have been developed. However, due to the cost and/or experimental difficulties in obtaining sufficient biological materials, it might be difficult to attain the required sample size. In this article, we address more practical questions for assessing power and false discovery rate (FDR) for a fixed sample size. The relationships between power, sample size and FDR are explored. We also conduct simulations and a real data study to evaluate the proposed findings.

KEY WORDS: false discovery rate; gene expression data; power; sample size; T-statistic

PMID 18338314

FDR based sample size justification



Data-driven hypothesis weighting increases detection power in genome-scale multiple testing

Nikolaos Ignatiadis, Bernd Klaus, Judith B Zaugg & Wolfgang Huber

Hypothesis weighting improves the power of large-scale multiple testing. We describe independent hypothesis weighting (IHW), a method that assigns weights using covariates independent of the *P*-values under the null hypothesis but informative of each test's power or prior probability of the null hypothesis (http://www.bioconductor.org/packages/IHW). IHW increases power while controlling the false discovery rate and is a practical approach to discovering associations in genomics, high-throughput biology and other large data sets.

PMID 27240256

Table 1 | Examples of covariates

Application	Covariate
Differential expression	Sum of read counts per gene across all samples[12]
GWAS	Minor allele frequency
eQTL, chromatin immunoprecipitation–QTL	Distance between genetic variant and locus of expression, or comembership in a topologically associated domain[16]
t-test	Overall variance[9]
Two-sided tests	Sign of the effect
Various applications	Signal quality, sample size

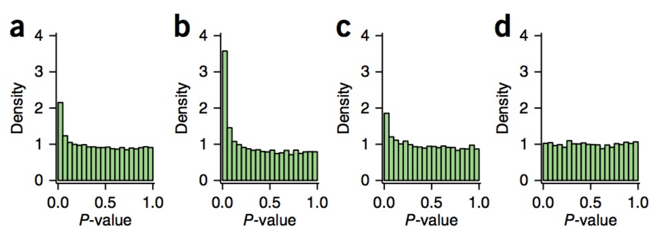


Figure 1 | Histograms stratified by the covariate as a diagnostic plot. (**a**) The histogram of all *P*-values shows a mixture of a uniform distribution and an enrichment of small *P*-values to the left. Such a well-calibrated histogram is the starting point for most multiple testing methods. (**b**–**d**) Histograms after splitting the hypotheses into three groups based on the values of the covariate.

PMID 27240256

Reporting and interpretation in genome-wide association studies

Jon Wakefield

Accepted 4 December 2007

Background In the context of genome-wide association studies we critique a number of methods that have been suggested for flagging associations for further investigation.

Methods The *P*-value is by far the most commonly used measure, but requires careful calibration when the *a priori* probability of an association is small, and discards information by not considering the power associated with each test. The *q*-value is a frequentist method by which the false discovery rate (FDR) may be controlled.

Results We advocate the use of the Bayes factor as a summary of the information in the data with respect to the comparison of the null and alternative hypotheses, and describe a recently-proposed approach to the calculation of the Bayes factor that is easily implemented. The combination of data across studies is straightforward using the Bayes factor approach, as are power calculations.

Conclusions The Bayes factor and the *q*-value provide complementary information and when used in addition to the *P*-value may be used to reduce the number of reported findings that are subsequently not reproduced.

Keywords Bayes theorem, epidemiologic methods, genetic polymorphism, testing

PMID 18270206