This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.
BIOINFORMATICS AND COMPUTATIONAL BIOLOGY SOLUTIONS USING R AND BIOCONDUCTOR

Biostatistics 140.688
Rafael A. Irizarry

Advanced Differential Expression Analysis

Outline

• Review of the basic ideas
• Introduction to (Empirical) Bayesian Statistics
• The multiple comparison problem
• SAM
Quantifying Differentially Expression

Two questions

• Can we order genes by interest? One goal is to assign a one number summary and consider large values interesting. We will refer to this number as a score
• How interesting are the most interesting genes? How do their scores compare to the those of genes known not to be interesting?

Example

• Consider a case were we have observed two genes with fold changes of 2

• Is this worth reporting? Are they both as interesting? Some journals require statistical significance. What does this mean?
Repeated Experiment

Review of Statistical Inference

- Let \( Y \) be our measurement representing differential expression.
- What is the typical null hypothesis?
- P-value is \( \text{Prob}(Y \text{ as extreme under null}) \) and is a way to summarize how interesting a gene is.
- Popular assumption: Under the null, \( Y \) follows a normal distribution with mean 0 and standard deviation \( \sigma \).
- Without \( \sigma \) we do not know the p-value.
- We can estimate \( \sigma \) by taking a sample and using the sample standard deviation \( s \).

Note: Different genes have different \( \sigma \).
Sample Summaries

Observations: \(X_1, \ldots, X_M\), \(Y_1, \ldots, Y_N\)

Averages: \(\overline{X} = \frac{1}{M} \sum_{i=1}^{M} X_i\), \(\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i\)

SD\(^2\) or variances:
\[
\text{s}_X^2 = \frac{1}{M-1} \sum_{i=1}^{M} (X_i - \overline{X})^2, \quad \text{s}_Y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \overline{Y})^2
\]

The t-statistic

\[
t - \text{statistic:} \quad \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{\text{s}_Y^2}{N} + \frac{\text{s}_X^2}{M}}}
\]

Properties of t-statistic

- If the number of replicates is very large the t-statistic is normally distributed with mean 0 and SD of 1
- If the observed data, i.e. \(Y-X\), are normally distributed then the t-statistic follows a t distribution regardless of sample size
- With one of these two we can compute p-values with one R command
Problems

- **Problem 1**: T-statistic bigger for genes with smaller standard errors estimates
  - **Implication**: Ranking might not be optimal

- **Problem 2**: T-statistic not t-distributed.
  - **Implication**: p-values/inference incorrect

Problem 1

- With few replicates SD estimates are unstable
- Empirical Bayes methodology and Stein estimators provides a statistically rigorous way of improving this estimate
- SAM, a more ad-hoc procedure, works well in practice

Note: We won’t talk about Stein estimators. See a paper by Gary Churchill for details

Problem 2

- Even if we use a parametric model to improve standard error estimates, the assumptions might not be good enough to provide trustworthy p-values
- We will describe non-parametric approaches for obtaining p-values

Note: We still haven’t discussed the multiple comparison problem. That comes later.
Introduction to Empirical Bayes

Outline

• General Introduction

• Models for relative expression

• Models for absolute expression

BASIC TWO STAGE SAMPLING

\[ \theta \sim G \]

\[ Y \mid \theta \sim f(y \mid \theta) \]

\( G \) is the prior
\( f \) is the sampling distribution
Use the “rules of probability” to get the:

Posterior Distribution

\[ g(\theta \mid Y) = \frac{f(y \mid \theta)g(\theta)}{f_Y(Y)} \]

Marginal Distribution

\[ f_Y(Y) = \int f(y \mid w)g(w)dw \]
THE BASIC GAUSSIAN/GAUSSIAN MODEL

Prior: \( G \sim N(\mu, \tau^2) \)
Sampling distn.: \( f \sim N(\theta, \sigma^2) \)
Marginal distn.: \( U \sim N(\mu, \sigma^2 + \tau^2) \)

Overdispersion

- If \((\mu, \tau^2, \sigma^2)\) are known, the posterior is Gaussian:
  \[
  E(\theta | Y) = B\mu + (1 - B)Y \\
  V(\theta | Y) = (1 - B)\sigma^2 \\
  B = \frac{\sigma^2}{\sigma^2 + \tau^2}
  \]

- The Gaussian prior is conjugate
- Shrinkage and variance reduction
- Increasing \(\sigma^2\) or decreasing \(\tau^2\) produces greater shrinkage
**Borrowing Strength**

- An advantage of having tens of thousands of genes is that we can try to learn about *typical* standard deviations by looking at all genes.
- Empirical Bayes gives us a formal way of doing this.

**Modeling Relative Expression**

Courtesy of Gordon Smyth

**Hierarchical Model**

<table>
<thead>
<tr>
<th>Normal Model</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\mu}_u \sim N(\mu_u, \sigma_u^2)$</td>
<td>$P(\mu_u = 0) = p$</td>
</tr>
<tr>
<td>$\beta_u</td>
<td>\mu_u \sim N(0, \sigma_\beta^2)$</td>
</tr>
</tbody>
</table>

Reparameterization of Lönnstedt and Speed 2002

Normality, independence assumptions are wrong but convenient, resulting methods are useful.
Posterior Statistics

Posterior variance estimators

\[ \hat{s}_g^2 = \frac{s_g^2 d_x + s_0^2 d_0}{d_x + d_0} \]

Moderated t-statistics

\[ \tilde{t}_g = \frac{\hat{\beta}_g}{\hat{s}_g \sqrt{C_{gg}}} \]

Eliminates large t-statistics merely from very small \( s \)

Marginal Distributions

The marginal distributions of the sample variances and moderated t-statistics are mutually independent

\[ s_g^2 \sim s_0^2 F_{d_x, d_0} \]

\[ \tilde{t}_g \sim \left\{ \begin{array}{ll}
\frac{t_{d_x - d}}{\sqrt{1 + s_g^2 / c \tilde{t}_{d_x - d}}} & \text{with prob } 1 - p \\
\frac{t_{d_x - d}}{\sqrt{1 + s_g^2 / C_{gg}}} & \text{with prob } p 
\end{array} \right. \]

Degrees of freedom add!

Shrinkage of Standard Deviations

The data decides whether \( \tilde{t}_g \)

should be closer to \( t_{g, pooled} \) or to \( t_g \)
Posterior Odds

Posterior probability of differential expression for any gene is

\[ \frac{\pi(\theta \neq 0 \mid \hat{\theta}, s^2)}{\pi(\theta = 0 \mid \hat{\theta}, s^2)} = \frac{p}{1 - p} \left( \frac{c}{c + c_0} \right)^{\gamma/\alpha} \left( \frac{P^2 + d + d_0}{P^2 - c - c_0 - d + d_0} \right)^{1 + \gamma/\alpha} \]

Monotonic function of \( \hat{r}^2 \) for constant \( d \)

Reparametrization of Lönnstedt and Speed 2002

Modeling the Absolute Expression

Courtesy of Christina Kendziorski
Hierarchical Model for Expression Data (Two conditions)

- Let \( x = [x_1, x_2] \) denote data (one gene) in conditions C1 and C2.
- Two patterns of expression:
  - \( P_0 \) (EE): \( \mu_{c1} = \mu_{c2} \)
  - \( P_1 \) (DE): \( \mu_{c1} \neq \mu_{c2} \)

  - For \( P_0 \), \( x \sim f(x|\mu)\mu(\mu)\) \( = f_0(x) \)
  - For \( P_1 \), \( x \sim f(x|\mu, \mu)\mu(\mu)\mu(\mu)\) \( = f_1(x) \)

Hierarchical Mixture Model for Expression Data

- Two conditions:
  \[ x \sim p_0 f_0(x) + p_1 f_1(x) \]

- Multiple conditions:
  \[ x \sim \sum_{k=1}^{K} p_k f_k(x) \]

Parameter estimates via EM

- Bayes rule determines threshold here; could target specific FDR.

For every transcript, two conditions \( \Rightarrow \) two patterns (DE, EE)

EE: \( m_1 = m_2 \)

DE: \( m_1 \neq m_2 \)

\[
\text{odds} = \frac{P(\text{DE}|y)}{P(\text{EE}|y)} = \frac{f(y|x \mid \text{DE})P(\text{DE})}{f(y|x \mid \text{EE})P(\text{EE})}
\]

Empirical Bayes methods make use all of the data to make gene specific inferences.
Hierarchical model is used to estimate posterior probabilities of patterns of expression. The model accounts for the measurement error process and for fluctuations in absolute expression levels.

Multiple conditions are handled in the same way as two conditions (no extra work required).

Posterior probabilities of expression patterns are calculated for every transcript.

Threshold can be adjusted to target a specific FDR.

In Bioconductor
Empirical Bayes for Microarrays (EBarrays)

On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes from Microarray Data by M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner, and K.W. Tsui


Inference and the Multiple Comparison Problem
Many slides courtesy of John Storey

Hypothesis testing
- Once you have a given score for each gene, how do you decide on a cut-off?
- p-values are popular.
- But how do we decide on a cut-off?
- Are 0.05 and 0.01 appropriate?
- Are the p-values correct?
P-values by permutation

- It is common for the assumptions used to derive the statistics used to summarize interest are not approximate enough to yield useful p-values

- An alternative is to use permutations

\[ p \text{-values by permutations} \]

We focus on one gene only. For the \( b \)th iteration, \( b = 1, \ldots, B \):

1. Permute the \( n \) data points for the gene \((x)\). The first \( n_1 \) are referred to as “treatments”, the second \( n_2 \) as “controls”.

2. For each gene, calculate the corresponding two sample t-statistic, \( t_b \).

After all the \( B \) permutations are done;

3. Put \( p = \#(b: |t_b| \geq |t_{observed}|)/B \) (p lower if we use \( > \)).

Multiple Comparison Problem

- If we do have useful approximations of our p-values, we still face the multiple comparison problem

- When performing many independent tests p-values no longer have the same interpretation
Hypothesis Testing

• Test for each gene null hypothesis: no differential expression.

• Two types of errors can be committed
  – Type I error or false positive (say that a gene is differentially expressed when it is not, i.e., reject a true null hypothesis).
  – Type II error or false negative (fail to identify a truly differentially expressed gene, i.e., fail to reject a false null hypothesis).

Multiple Hypothesis Testing

• What happens if we call all genes significant with p-values $\leq 0.05$, for example?

<table>
<thead>
<tr>
<th></th>
<th>Called Significant</th>
<th>Not Called Significant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null True</td>
<td>$V$</td>
<td>$m_0 - V$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>Altern. True</td>
<td>$S$</td>
<td>$m_1 - S$</td>
<td>$m_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$R$</td>
<td>$m - R$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

Other ways of thinking of P-values

• A p-value is defined to be the minimum false positive rate at which an observed statistic can be called significant.

• If the null hypothesis is simple, then a null p-value is uniformly distributed.
Multiple Hypothesis Test

Error Controlling Procedure

• Suppose $m$ hypotheses are tested with p-values $p_1, p_2, \ldots, p_m$

• A multiple hypothesis error controlling procedure is a function $T(p; \alpha)$ such that rejecting all nulls with $p_i \leq T(p; \alpha)$ implies that $Error \leq \alpha$

• Error is a population quantity (not random)

Weak and Strong Control

• If $T(p; \alpha)$ is such $Error \leq \alpha$ only when $m_0 = m$, then the procedure provides weak control of the error measure

• If $T(p; \alpha)$ is such $Error \leq \alpha$ for any value of $m_0$, then the procedure provides strong control of the error measure – note that $m_0$ is not an argument of $T(p; \alpha)$!

Error Rates

• Per comparison error rate (PCER): the expected value of the number of Type I errors over the number of hypotheses $PCER = E(V)/m$

• Per family error rate (PFER): the expected number of Type I errors $PFER = E(V)$

• Family-wise error rate: the probability of at least one Type I error $FEWR = Pr(V \geq 1)$

• False discovery rate (FDR) rate that false discoveries occur $FDR = E(V/R; R>0) = E(V/R | R>0)Pr(R>0)$

• Positive false discovery rate (pFDR): rate that discoveries are false $pFDR = E(V/R | R>0)$. 
Bonferroni Procedure

\[ T(p; \alpha) = \max \left\{ p_j : p_j \leq \frac{\alpha}{m} \right\} \]

Provides strong control, ...

\[ \Pr(V \geq 1) \leq \Pr\left( \min_{j=1}^{m} p_j \leq \frac{\alpha}{m} | H_0^C \right) \]
\[ = \sum_{j=1}^{m} \Pr(p_j \leq \frac{\alpha}{m} | H_0^C) \]
\[ = m \cdot \frac{\alpha}{m} \]

Sidak Procedure

\[ T(p; \alpha) = \max \left\{ p_j : p_j \leq 1 - (1 - \alpha)^{1/m} \right\} \]

\[ \Pr(V \geq 1) \leq \Pr\left( \min_{j=1}^{m} p_j \leq 1 - (1 - \alpha)^{1/m} | H_0^C \right) \]
\[ = 1 - \prod_{j=1}^{m} \Pr(p_j > 1 - (1 - \alpha)^{1/m} | H_0^C) \]
\[ = \alpha \]

Requires independence for strong control

Holm Procedure

Order the p-values \( p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)} \)

\[ T(p; \alpha) = \min \left\{ p_{(i)} : p_{(i)} > \frac{\alpha}{m - i + 1} \right\} \]

\[ T(p; \alpha) = \min \left\{ p_{(i)} : p_{(i)} > 1 - \left(1 - \alpha\right)^{(m-i+1)} \right\} \]

Requires independence for strong control
**Hochberg Procedure**

\[ T(p; \alpha) = \max \left\{ p_{(i)} : p_{(i)} \leq \frac{\alpha}{m - i + 1} \right\} \]

...the step-up analogue of Holm

---

**Simes/BH Procedure**

\[ T(p; \alpha) = \max \left\{ p_{(i)} : p_{(i)} \leq \frac{i \cdot \alpha}{m} \right\} \]

- Weak controls the FWER (Simes 1986)
- Strongly controls FDR (Benjamini & Hochberg 1995)
- Both require the null p-values to be independent

---

**False Discovery Rate**

- The "false discovery rate" measures the proportion of false positives among all genes called significant:

\[
\frac{\text{# false positives}}{\text{# called significant}} = \frac{V}{V + S} - \frac{V}{R}
\]

- This is usually appropriate because one wants to find as many truly differentially expressed genes as possible with relatively few false positives
- The false discovery rate gives the rate at which further biological verification will result in dead-ends
False Positive Rate versus False Discovery Rate

- False positive rate is the rate at which truly null genes are called significant
  \[ \text{FPR} = \frac{\text{false positives}}{\text{truly null}} = \frac{V}{m_0} \]
- False discovery rate is the rate at which significant genes are truly null
  \[ \text{FDR} = \frac{\text{false positives}}{\text{called significant}} = \frac{V}{R} \]

False Positive Rate and P-values

- The \textit{p-value} is a measure of significance in terms of the false positive rate (aka Type I error rate)
- P-value is defined to be the minimum false positive rate at which the statistic can be called significant
- Can be described as the probability a truly null statistic is “as or more extreme” than the observed one

False Discovery Rate and Q-values

- The \textit{q-value} is a measure of significance in terms of the false discovery rate
- Q-value is defined to be the minimum false discovery rate at which the statistic can be called significant
- Can be described as the probability a statistic “as or more extreme” is truly null
Bayesian Interpretation

- Suppose \( n \) hypothesis tests are performed with independent statistics \( X_1, \ldots, X_n \) and significance region \( \Gamma \).
- Let \( H_i = 0 \) if null hypothesis \( i \) is true, and \( H_i = 1 \) if it is false.
- Assume \( \Pr(H_i = 0) = \pi_0 \) and \( \Pr(H_i = 1) = \pi_1 \).
- Assume each statistic comes from the mixture distribution, \( X_i \sim (1 - H_i) \cdot F_0 + H_i \cdot F_1 \), where \( F_0 \) is the null and \( F_1 \) is the alternative.

Theorem: (Storey 2001)

\[
pFDR(\Gamma) = E \left[ \frac{Y(\Gamma)}{R(\Gamma)} \right] \mathbb{1}(R(\Gamma) > 0) = \frac{\pi_0 \cdot Pr(X \in \Gamma | H = 0)}{Pr(X \in \Gamma)} = \pi_0 \cdot Pr(H = 0 | X \in \Gamma).
\]

Power / Type I Error Decomposition

- Under the mixture model assumptions ...

\[
pFDR(\Gamma) = \frac{\pi_0 \cdot Pr(X \in \Gamma | H = 0)}{\pi_0 \cdot Pr(X \in \Gamma | H = 0) + \pi_1 \cdot Pr(X \in \Gamma | H = 1)}
\]

\[
= \frac{\pi_0 \cdot \text{Type I error rate}}{\pi_0 \cdot \text{Type I error rate} + \pi_1 \cdot \text{Power}}
\]

q-values

- In general, for a nested set of significance regions \( \{\Gamma\} \), the p-value of an observed statistic \( x \) is defined to be

\[
p\text{-value}(x) = \inf_{x \in \Gamma} Pr(X \in \Gamma | H = 0)
\]

- Likewise, under the independent mixture model,

\[
q\text{-value}(x) = \inf_{x \in \Gamma} pFDR(\Gamma) = \inf_{x \in \Gamma} Pr(H = 0 | X \in \Gamma).
\]
Bayesian Connections

- This allows Bayesians to estimate FDR as well:
  \[ pFDR(\Gamma) = \int \Pr(H = 0 \mid X = x) f(x \mid x \in \Gamma) dx \]
- This motivates the name “q-value” directly:
  \[ p - \text{value}(x_i) = \Pr(\mid X \mid x_i, \mid H = 0) \]
  \[ q - \text{value}(x_i) = \Pr(\mid H = 0 \mid X \mid x_i, \mid) \]
- All the estimation presented below can be viewed as an “empirical Bayes” approach

Possible FDR Goals

1. For some pre-chosen \( \alpha \), estimate a significance cut-off so that on average FDRs \( \alpha \)
2. For some pre-chosen significance cut-off, estimate FDR so that \( E[FDR] \geq FDR \)
3. Estimate FDR so that it’s simultaneously conservative over all significance cut-offs
4. Estimate q-values for all genes that are simultaneously conservative

Universal Goal

1. The q-value, an FDR-based measure of significance, is associated with each gene
2. The estimated q-values are conservative over all genes simultaneously

*In doing so, all four options will be met*
Estimate of FDR

- We begin by estimating FDR when calling all genes significant with p-values ≤ \( t \)
- **Heuristic motivation:**

\[
\text{FDR}(t) = \frac{E[R(t)]}{E[R(t)]} \left[ \frac{E[\# \{ \text{null} P_i \leq t \}]}{E[\# \{ P_i \leq t \}]} \right] = \frac{m_0 \cdot t}{\# \{ P_i \leq t \}}
\]

Estimate of \( \pi_0 \)

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

\[
\hat{\pi}_0(\lambda) = \frac{\# \{ P_i > \lambda \}}{m \cdot (1 - \lambda)}
\]

- Then clearly \( \hat{m}_0 = \hat{\pi}_0 \cdot m \)
Choose $\lambda$ by Balancing Bias and Variance

Overall FDR Estimate

- The overall estimate of $\text{FDR}(\hat{\pi}_0)$ is
  \[
  \hat{\text{FDR}}(t) = \frac{\hat{\pi}_0 \cdot t}{m \cdot t}
  \]
- The implicit estimate used in the original FDR paper is a special case of the above estimate with $\hat{\pi}_0$ estimated as 1.

Numerical Example

- Suppose we call all genes significant with p-values $\leq 0.03$
- The estimate of the FDR is
  \[
  \hat{\text{FDR}} = \frac{0.67 \times 3170 \times 0.03}{462} = \frac{64}{462} = 0.14
  \]
- Could use any threshold $0 \leq t \leq 1$
Q-value Estimate

- The mathematical definition of the q-value of gene $i$ is

\[ q\text{-value}(p_i) = \min_{t \geq p_i} pFDR(t) \]

- Since $pFDR = FDR$, we estimate the q-value of gene $i$ by

\[ \hat{q}(p_i) = \min_{t \geq p_i} \hat{FDR}(t) \]

Q-Plots

Theoretical Results

- Suppose that the empirical distribution functions of the null statistics and of the alternative statistics converge as the number of genes $m$ gets large ...
- The FDR estimates are asymptotically conservative ... simultaneously over all significance regions
- The estimated q-values are simultaneously conservative over all genes
- This is equivalent to controlling the FDR at all levels $\alpha$ simultaneously
**The Estimates**

\[ \text{FDR}_{\lambda}(t) = \frac{\hat{q}(\lambda) \cdot t}{\Pr(P \leq t)} \]

\[ \Pr_H(H = 0 | P \leq t) = \frac{\hat{q}(\lambda) \cdot t}{\Pr(P \leq t)} \]

\[ q_{\alpha}(p_i) = \min_{i \geq n} \Pr_H(H = 0 | P \leq t) \]

- Can define a more robust estimate of q-value based on \( \mu \text{FDR}_{\lambda}(t) \)
- Can get rid of \( \lambda \) by the technique mentioned earlier

---

**Using q-value and FDR in Four Scenarios**

1. Suppose we call all \( p \)-values \( \leq t \) significant, use \( \text{FDR}_{\lambda}(t) \) to estimate \( \text{FDR}(t) \).

2. To control the FDR at level \( \alpha \), reject all null hypothesis with \( q_{\alpha}(p_i) \leq \alpha \).

   Note: This procedure with \( \lambda = 0 \) is equivalent to the Benjamini and Hochberg (1995) threshold \( T_B = \max \{ p(i) : p(i) \leq \frac{i}{m} \cdot \alpha \} \).

   This follows because \( \text{FDR}_{\lambda=0}(p(i)) = \frac{p(i)}{i/m} \).

---

**Using q-value and FDR in Four Scenarios**

3. Suppose we want to estimate \( \text{FDR}(t) \) over all thresholds simultaneously. Examine \( \text{FDR}_{\lambda}(t) \) over \( 0 \leq t \leq 1 \). Estimating the "simultaneous controlling curve."

4. To calculate a measure of significance for each test, form the q-value estimates: \( q_{\alpha}(p_i) \). Estimate minimum FDR at which each test can be called significant (in addition to Bayesian interpretation).
**Finite Sample Results**

- Suppose the null p-values are independent ...(No mixture model or Bayesian assumptions)
- Then
  \[ \mathbb{E}[\text{FDR}_n(t)] \geq \text{FDR}(t) \]
  \[ \mathbb{E}[p\text{FDR}_n(t)] \geq p\text{FDR}(t). \]
  (Storey 2001)
- Strong control:
  \[ \text{FDR}\left(\{q\text{-value}_k(p_k) \leq \alpha, p_k \leq \lambda\}\right) \leq \alpha. \]
  (Storey, Taylor, Siegmund 2002)
- Are the null p-values independent in microarrays?

**Dependence in Microarrays**

- Since measured expression levels of genes are dependent, the statistics (p-values) are dependent:
  1. Genes in the same pathway will be dependent
  2. Genes near each other on the array will be dependent
  3. Genes with sequence similarity will be dependent
- Each of these dependencies is local. Probably occur in finite clumps.

**Empirical Distributions**

- Recall that:
  \[ V(t) \overset{m_0}{=} \#\{\text{null } p_i : p_i \leq t\}, \]
  \[ S(t) \overset{m_1}{=} \#\{\text{alternative } p_i : p_i \leq t\}. \]
- Suppose that with probability 1, we have for each t:
  \[ V(t) \overset{m_0}{\rightarrow} F(t) \leq t, \]
  \[ S(t) \overset{m_1}{\rightarrow} F(t). \]
- Also suppose \( \lim_{m_1 \rightarrow \infty} m_1/m = \tau_0 \) exists.
- Then with probability 1...
Simulation Study

- Performed 3000 hypothesis tests of $H_0: N(0,1)$ versus $H_1: N(2,1)$
- The statistics had correlation 0.40 in blocks of 50
- Two conclusions:
  1. The true q-values under this dependence structure are the same as those given under the independence model
  2. The estimated q-values are simultaneously conservative

Conservative Consistency

- Then for any $\delta > 0$, we have that with probability 1 ...

\[
\begin{align*}
&\lim_{m \to \infty} \inf_{t \geq \delta} FDR(t) \geq 0 \\
&\lim_{m \to \infty} \inf_{t \geq \delta} \left| \text{q-value}(t) - \text{q-value}(t_i) \right| \geq 0
\end{align*}
\]

(Storey, Taylor, Siegmund 2002)

- Plausibly holds for microarray data.

Translations: Given "clumpy microarray dependence" and large $m$ ...

Bayesian interpretation holds

- $FDR(t) \sim pFDR(t) \to \Pr_m(H = 0 | P \leq t)$

Can look at all thresholds simultaneously

- $\sigma FDR(t)$ dominates $FDR(t)$ over all $t$

The FDR is controlled

- Significance rule $\text{q-value}(p_i) \leq \alpha$ controls the FDR at level $\alpha$

The estimated q-values conservatively estimate the true q-values

- $\widehat{\text{q-value}}(t)$ dominates $\text{q-value}(t)$ over all $t$ (even $t = p_i$)
Power Comparison

<table>
<thead>
<tr>
<th>FDR Level</th>
<th># Significant BH</th>
<th># Significant PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>0.02</td>
<td>8</td>
<td>21</td>
</tr>
<tr>
<td>0.03</td>
<td>21</td>
<td>80</td>
</tr>
<tr>
<td>0.04</td>
<td>76</td>
<td>123</td>
</tr>
<tr>
<td>0.05</td>
<td>88</td>
<td>160</td>
</tr>
<tr>
<td>0.10</td>
<td>221</td>
<td>317</td>
</tr>
</tbody>
</table>

$\tilde{\alpha}_0 = 1 \quad \tilde{\alpha}_0 = 0.67$
What should one look for in a multiple testing procedure?

As we will see, there is a bewildering variety of multiple testing procedures. How can we choose which to use? There is no simple answer here, but each can be judged according to a number of criteria:

**Interpretation:** does the procedure answer a relevant question for you?

**Type of control:** strong or weak?

**Validity:** are the assumptions under which the procedure applies clear and definitely or plausibly true, or are they unclear and most probably not true?

**Computability:** are the procedure’s calculations straightforward to calculate accurately, or is there possibly numerical or simulation uncertainty, or discreteness?

---

Selected references


J Storey (2001): 3 papers (some with other authors), www-stat.stanford.edu/~jstorey/

The positive false discovery rate: a Bayesian interpretation and the q-value.

A direct approach to false discovery rates

Estimating false discovery rates under dependence, with applications to microarrays

Significance analysis of microarrays (SAM)

- A clever adaptation of the t-ratio to borrow information across genes
- In Bioconductor, siggenes package is available

SAM-statistic

- For gene i
  \[ d_i = \frac{\bar{y}_i - \bar{y}_0}{s_i + s_0} \]
  \( \bar{y}_i \) = mean of Irradiated samples
  \( \bar{y}_0 \) = mean of Unirradiated samples
  \( s_i \) = Standard deviation of residuals for gene i
  \( s_0 \) = Exchangeability factor estimated using all genes

The exchangeability factor

- Chosen to make signal-to-noise ratios independent of signal
- Computation
  - Let \( q' \) be the \( \alpha \) percentile of the \( q \) values.
  - Let \( \tau_j = q' / (s_j + s') \)
  - Compute the 100 quantiles of the \( \tau \) values, denoted by \( q_1 < q_2 < \cdots < q_{100} \)
  - \( \alpha \in \{0.05, 0.10, \ldots, 1.0\} \)
  - For each \( \alpha \)
    - Compute \( \gamma_j = \min(\tau_j | s_j \in [q_j, q_{j+1}]) \), \( j = 1, 2, \ldots, 99 \),
      where mad is the median absolute deviation from the median, divided by 0.64
    - Compute \( \gamma_j \) as coefficient of variation of the
    - Choose \( \hat{\sigma} = \min(\gamma_j \hat{\sigma}) \), \( s_j = \hat{\sigma}^2 \), and \( \gamma_j \)
Scatter plots of relative difference

The reference distribution

- Order the values of $d_i$ (could be any stat)
  
  $d_{i1} \leq d_{i2} \leq \cdots \leq d_{in}$

- Permute the treatment labels, and compute a new set of ordered values
  
  $d'_{i1} \leq d'_{i2} \leq \cdots \leq d'_{in}$

- Repeat step 2 for, say, 100 permutations:
  
  $d''_{i1} \leq d''_{i2} \leq \cdots \leq d''_{in}$

  $d'''_{i1} \leq d'''_{i2} \leq \cdots \leq d'''_{in}$

  $\vdots$

  $d^{(100)}_{i1} \leq d^{(100)}_{i2} \leq \cdots \leq d^{(100)}_{in}$

- From these, compute the average largest, average second largest etc.

Selected genes
More general versions of SAM

More than two groups
Paired data
Survival data, with censored response
Limitations of SAM

• Solutions for $s_0$ are often at the extremes and sensitive to the resolution of the quantile grid.
• Permutation analysis throws all genes in the same bag
• Requires a monotone signal-to-noise relationship