Advanced Differential Expression Analysis

Outline

- Review of the basic ideas
- Introduction to (Empirical) Bayesian Statistics
- The multiple comparison problem
- SAM

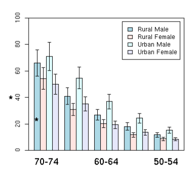
Quantifying Differentially Expression

Two questions

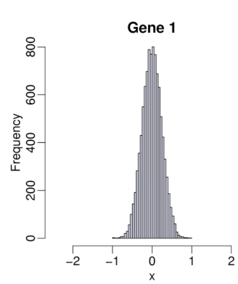
- Can we order genes by interest? One goal is to assign a one number summary and consider large values interesting. We will refer to this number as a *score*
- How interesting are the most interesting genes? How do their scores compare to the those of genes known not to be interesting?

Example

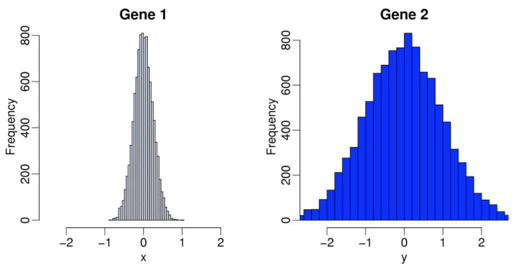
- Consider a case were we have observed two genes with fold changes of 2
- Is this worth reporting? Are they both as interesting? Some journals require *statistical significance*. What does this mean?



Repeated Experiment



Repeated Experiment



Review of Statistical Inference

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

Note: Different genes have different σ,

Sample Summaries

Observations: $X_1,\ldots,X_M \quad Y_1,\ldots,Y_N$

Averages: $\bar{X} = \frac{1}{M}\sum_{i=1}^{M} X_i \quad \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$

SD² or variances:

$$s_X^2 = \frac{1}{M-1}\sum_{i=1}^{M}(X_i - \bar{X})^2 \quad s_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

The t-statistic

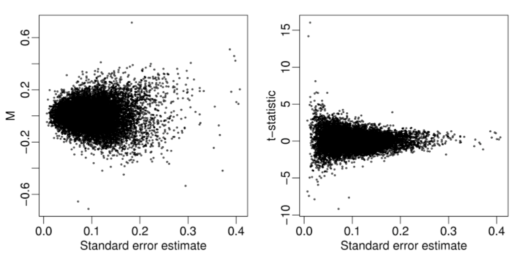
t - statistic:

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{s_Y^2}{N} + \frac{s_X^2}{M}}}$$

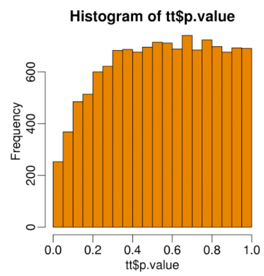
Properties of t-statistic

- **If the number of replicates is very large the t-statistic is normally distributed with mean 0 and 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**

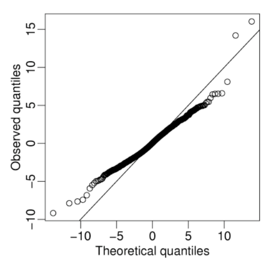
Data Show Problems



Data Show Problems



Data Show Problems



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 trust-worthy 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

$$g(\theta \mid Y) = \frac{f(y|\theta)g(\theta)}{f_G(Y)}$$

Marginal Distribution

$$f_G(Y) = \int f(y \mid u)g(u)du$$

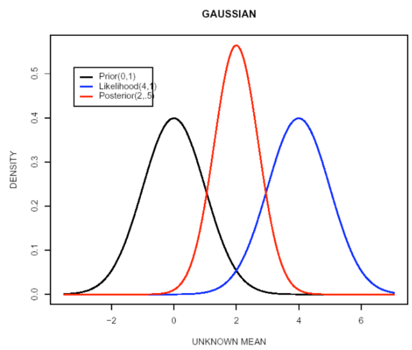
THE BASIC GAUSSIAN/GAUSSIAN MODEL

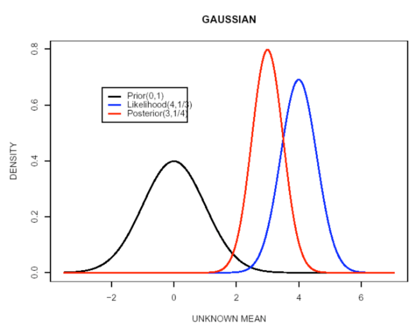
Prior: G $= N(\mu, \tau^2)$
Sampling distn.: f $= N(\theta, \sigma^2)$
Marginal distn.: $f_G = N(\mu, \sigma^2 + \tau^2)$
Overdispersion

- If (μ, τ^2, σ^2) are known, the posterior is Gaussian:

$$\begin{aligned} E(\theta|Y) &= B\mu + (1-B)Y \\ &= \mu + (1-B)(Y-\mu) \\ V(\theta|Y) &= (1-B)\sigma^2 \\ B &= \frac{\sigma^2}{\sigma^2+\tau^2} \end{aligned}$$

- The Gaussian prior is conjugate
- Shrinkage and variance reduction
- Increasing σ^2 or decreasing τ^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

Normal Model

$$\hat{\beta}_{gj} \sim N(\beta_{gj}, c_{gj}\sigma_g^2)$$

$$s_g^2 \sim \sigma_g^2 \chi_{d_g}^2$$

Prior

$$P(\beta_{gj} \neq 0) = p$$

$$\beta_{gj} \mid \beta_{gj} \neq 0 \sim N(0, c_{0j}\sigma_g^2)$$

$$\sigma_g^2 \sim s_0^2 \left(\chi_{d_0}^2 / d_0\right)^{-1}$$

Reparametrization of Lönnstedt and Speed 2002

Normality, independence assumptions are wrong but convenient, resulting methods are useful

Posterior Statistics

Posterior variance estimators

$$\tilde{s}_g^2 = \frac{s_g^2 d_g + s_0^2 d_0}{d_g + d_0}$$

Moderated t-statistics

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{c_{gj}}}$$

Eliminates large t-statistics merely from very small *s*

Marginal Distributions

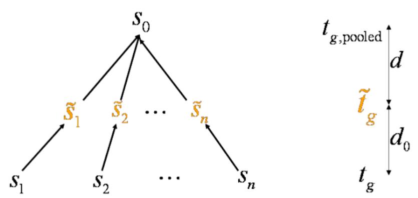
The marginal distributions of the sample variances and moderated t-statistics are mutually independent

$$s_g^2 \sim s_0^2 F_{d,d_0}$$

$$\tilde{t}_g \sim \begin{cases} t_{d_0+d} & \text{with prob } 1-p \\ \sqrt{1+c_0/c}\, t_{d_0+d} & \text{with prob } p \end{cases}$$

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

$$\frac{p(\beta \neq 0 \mid \hat{\beta}, s^2)}{p(\beta = 0 \mid \hat{\beta}, s^2)} = \frac{p}{1-p}\left(\frac{c}{c+c_0}\right)^{1/2}\left\{\frac{\tilde{t}^2 + d + d_0}{\tilde{t}^2 \frac{c}{c-c_0} - d + d_0}\right\}^{\frac{1+d+d_0}{2}}$$

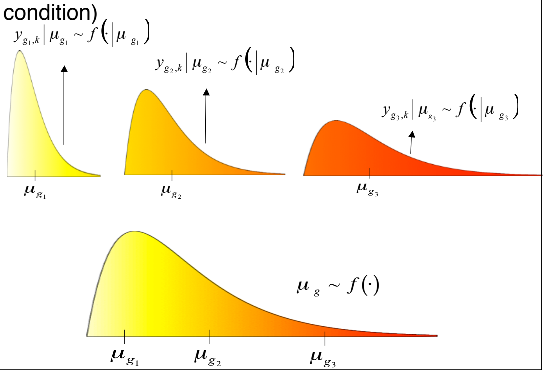
Monotonic function of $\tilde{t}^2$ for constant d

Reparametrization of Lönnstedt and Speed 2002

Modeling the Absolute Expression

Courtesy of Christina Kendziorski

Hierarchical Model for Expression Data (One condition)



Hierarchical Model for Expression Data (Two conditions)

- Let $x = [x_{c1}, x_{c2}]$ denote data (one gene) in conditions C1 and C2.
- Two patterns of expression:

P0 (EE) : $\mu_{c1} = \mu_{c2}$

P1 (DE): $\mu_{c1} \neq \mu_{c2}$

- For P0, $x \sim \int f(x|\mu) f(\mu) d\mu \equiv f_0(x)$
- For P1, $x \sim \int f(x|\mu_{c1}, \mu_{c2}) f(\mu_{c1}, \mu_{c2}) d\mu_{c1} d\mu_{c2}$

$$= \underbrace{\int f(x_{c1}|\mu_{c1}) f(\mu_{c1}) d\mu_{c1}}_{f_0(x_{c1})} \underbrace{\int f(x_{c2}|\mu_{c2}) f(\mu_{c2}) d\mu_{c2}}_{f_0(x_{c2})} \equiv f_1(x)$$

Hierarchical Mixture Model for Expression Data

- Two conditions:

$$x \sim p_0 f_0(x) + p_1 f_1(x) \quad \Rightarrow \quad p(P1|x) = \frac{p_1 f(x|P1)}{p_0 f(x|P0) + p_1 f(x|P1)}$$

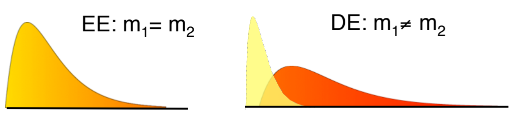
- Multiple conditions:

$$x \sim \sum_{k=1}^{K} p_k f_k(x) \quad \Rightarrow \quad p(Pk'|x) = \frac{p_{k'} f(x|Pk')}{\sum_{k \neq k'} p_k f(x|Pk)}$$

- Parameter estimates via EM

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

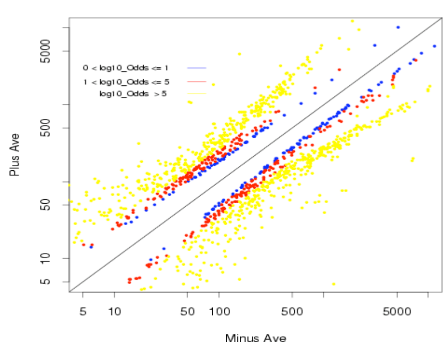
For every transcript, two conditions => two patterns (DE, EE)



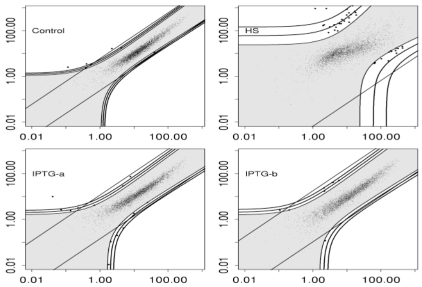
$$\text{odds}_g = \frac{P(DE|y)}{P(EE|y)} = \frac{f(y_g \mid DE)P(DE)}{f(y_g \mid EE)P(EE)}$$

Empirical Bayes methods make use all of the data to make gene specific inferences.

Odds plot: SCD knockout vs. SV129 (Attie lab)



EBarrays: Contour Plots of Odds



Comments on Empirical Bayes Approach(EBarrays)

- 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
Journal of Computational Biology 8: 37-52, 2001.

On Parametric Empirical Bayes Methods for Comparing Multiple Groups
Using Replicated Gene Expression Profiles
by
C.M. Kendziorski, M.A. Newton, H. Lan and M.N. Gould

Statistics in Medicine, to appear, 2003.

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-values by permutations

We focus on one gene only. For the bth iteration, $b = 1, \cdots, B$;

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 ≤ 0.05, for example?

	Called Significant	Not Called Significant	Total
Null True	V	$m_0 - V$	m_0
Altern.True	S	$m_1 - S$	m_1
Total	R	$m - R$	m

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 $T(\mathrm{p}; \alpha)$ such that rejecting all nulls with $p_i \leq T(\mathrm{p}; \alpha)$ implies that $Error \leq \alpha$
- *Error* is a population quantity (not random)

Weak and Strong Control

- **If $T(\mathrm{p}; \alpha)$ is such *Error* ≤ α only when $m_0 = m$, then the procedure provides** *weak control* **of the error measure**
- **If $T(\mathrm{p}; \alpha)$ is such *Error* ≤ α 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 $T(\mathrm{p}; \alpha)$!**

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 ≥ 1)

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

•**Positive false discovery** rate (pFDR): rate that discoveries are false
pFDR = E(V/R I R>0).

Bonferroni Procedure

$$T(\mathbf{p};\alpha) = \max\left\{p_i : p_i \le \frac{\alpha}{m}\right\}$$

Provides strong control…..

$$\begin{aligned}\Pr(V \ge 1) &\le \Pr\left(\min_i p_i \le \frac{\alpha}{m} \mid H_0^C\right)\\ &\le \sum_{i=1}^{m} \Pr\left(p_i \le \frac{\alpha}{m} \mid H_0^i\right)\\ &= m \cdot \frac{\alpha}{m}\end{aligned}$$

Sidak Procedure

$$T(\mathbf{p};\alpha) = \max\{p_i : p_i \le 1-(1-\alpha)^{1/m}\}$$

$$\begin{aligned}\Pr(V \ge 1) &\le \Pr\left(\min_i p_i \le 1-(1-\alpha)^{1/m} \mid H_0^C\right) \\ &= 1-\prod_{i=1}^{m}\Pr\left(p_i > 1-(1-\alpha)^{1/m} \mid H_0^i\right) \\ &= \alpha\end{aligned}$$

Requires independence for strong control…

Holm Procedure

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

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

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

Requires independence for strong control…

Hochberg Procedure

$$T(\mathbf{p};\alpha) = \max\left\{p_{(i)} : p_{(i)} \le \frac{\alpha}{m-i+1}\right\}$$

...the step-up analogue of Holm

Simes/BH Procedure

$$T(\mathbf{p};\alpha) = \max\left\{ p_{(i)} : p_{(i)} \le \frac{i \cdot \alpha}{m} \right\}$$

- **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**

$$\text{FPR} \approx \frac{\#\,\text{false positives}}{\#\,\text{truly null}} = \frac{V}{m_0}$$

- False discovery rate **is the rate at which significant genes are truly null**

$$\text{FDR} \approx \frac{\#\,\text{false positives}}{\#\,\text{called significant}} = \frac{V}{R}$$

False Positive Rate and P-values

- The *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 *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 m hypothesis tests are performed with independent statistics $X_1, \ldots, X_m$ and significance region Γ.
- 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)

$$\mathrm{pFDR}(\Gamma) = \mathrm{E}\left[\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) > 0\right] = \frac{\pi_0 \cdot \Pr(X \in \Gamma | H = 0)}{\Pr(X \in \Gamma)} = \Pr(H = 0 | X \in \Gamma).$$

Power / Type I Error Decomposition

- Under the mixture model assumptions ...

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

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

q-values

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

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

- Likewise, under the independent mixture model,

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

Bayesian Connections

- This allows Bayesians to estimate FDR as well:

$$\text{pFDR}(\Gamma) = \int \Pr(H = 0 \mid X = x) f(x \mid x \in \Gamma) dx$$

- This motivates the name "q-value" directly:

$$\text{p} - \text{value}(x_i) = \Pr\left(|X| \geq |x_i| \,\middle|\, H = 0\right)$$

$$\text{q} - \text{value}(x_i) = \Pr\left(H = 0 \,\middle|\, |X| \geq |x_i|\right)$$

- All the estimation presented below can be viewed as an "empirical Bayes" approach

Possible FDR Goals

1. **For some pre-chosen α, estimate a significance cut-off so that on average FDR≤ α**
2. **For some pre-chosen significance cut-off, estimate FDR so that** $\mathrm{E}\left[\hat{\mathrm{FDR}}\right] \geq \mathrm{FDR}$
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:**

$$\mathrm{FDR}(t) \approx \frac{\mathrm{E}[V(t)]}{\mathrm{E}[R(t)]} = \frac{\overbrace{\mathrm{E}[\#\{\text{null } p_i \leq t\}]}^{=m_0 t}}{\mathrm{E}[\#\{p_i \leq t\}]}$$

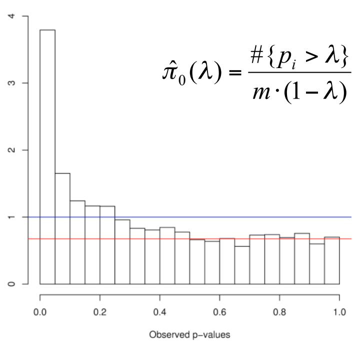
$$\boxed{\hat{\mathrm{FDR}}(t) = \frac{\hat{m}_0 \cdot t}{\#\{p_i \leq t\}}}$$

Estimate of π_0

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

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

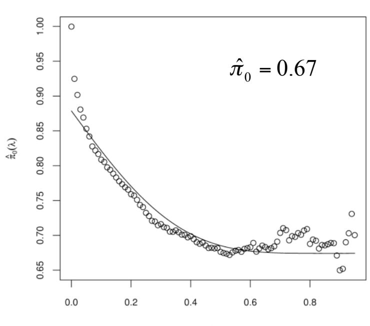
- Then clearly $\hat{m}_0 = \hat{\pi}_0 \cdot m$



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

$\hat{\pi}_0$

Choose λ by Balancing Bias and Variance



$\hat{\pi}_0 = 0.67$

Overall FDR Estimate

- The overall estimate of FDR(t) is

$$\hat{\mathrm{FDR}}(t) = \frac{\hat{\pi}_0 m \cdot t}{\#\{p_i \leq t\}}$$

- The implicit estimate used in the original FDR paper is a special case of the above estimate with π_0 *estimated as 1*.

Numerical Example

- Suppose we call all genes significant with p-values ≤ 0.03
- The estimate of the FDR is

$$\hat{\mathrm{FDR}} = \frac{0.67 \times 3170 \times 0.03}{462} = \frac{64}{462} = 0.14$$

- Could use any threshold $0 \leq t \leq 1$

Q-value Estimate

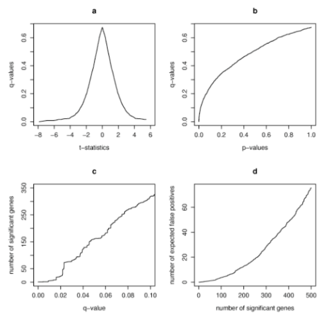
- The mathematical definition of the q-value of gene *i* is

$$\text{q-value}(p_i) = \min_{t \geq p_i} \mathrm{pFDR}(t)$$

- Since pFDR≈FDR, we estimate the q-value of gene *i* by

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

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 α *simultaneously*

The Estimates

$$\widehat{\mathrm{FDR}}_\lambda(t) = \frac{\widehat{\pi}_0(\lambda)\cdot t}{\widehat{\Pr}(P \le t)}$$

$$\widehat{\Pr}_\lambda(H=0|P \le t) = \frac{\widehat{\pi}_0(\lambda)\cdot t}{\widehat{\Pr}(P \le t)}$$

$$\widehat{\text{q-value}}_\lambda(p_i) = \min_{t \ge p_i} \widehat{\Pr}_\lambda(H=0|P \le t)$$

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

Using q-$\widehat{\text{value}}$ and $\widehat{\text{FDR}}$ in Four Scenarios

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

(2) To control the FDR at level α, reject all null hypothesis with q-$\widehat{\text{value}}_\lambda(p_i) \leq \alpha$.

Note: This procedure with $\lambda = 0$ is equivalent to the Benjamini and Hochberg (1995) threshold $T_{BH} = \max\{p_{(i)} : p_{(i)} \leq \frac{i}{m}\alpha\}$. This follows because $\widehat{\text{FDR}}_{\lambda=0}(p_{(i)}) = \frac{p_{(i)}}{i/m}$.

Using q-$\widehat{\text{value}}$ and $\widehat{\text{FDR}}$ in Four Scenarios

(3) Suppose we want to estimate $\text{FDR}(t)$ over all thresholds *simultaneously*. Examine $\widehat{\text{FDR}}_\lambda(t)$ 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-$\widehat{\text{value}}(p_i)$. 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

$$\text{E}[\widehat{\text{FDR}}_\lambda(t)] \geq \text{FDR}(t)$$

$$\text{E}[\text{p}\widehat{\text{FDR}}_\lambda)(t)] \geq \text{pFDR}(t).$$

(Storey 2001)

- Strong control:

$$\text{FDR}\left(\left\{\text{q-}\widehat{\text{value}}_\lambda(p_i) \leq \alpha,\ p_i \leq \lambda\right\}\right) \leq \alpha.$$

(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:

$$\frac{V(t)}{m_0} = \frac{\#\{\text{null } p_i : p_i \leq t\}}{m_0},$$
$$\frac{S(t)}{m_1} = \frac{\#\{\text{alternative } p_i : p_i \leq t\}}{m_1}$$

•Suppose that with probability 1, we have for each t:

$$\frac{V(t)}{m_0} \longrightarrow F_0(t) \leq t,$$
$$\frac{S(t)}{m_1} \longrightarrow F_1(t)$$

•Also suppose $\lim_{m\to\infty} m_0/m = \pi_0$ exists.

•Then with probability 1...

Conservative Consistency

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

(1) $\lim_{m\to\infty} \text{FDR}\left(\left\{\widehat{\text{q-value}}(p_i) \leq \alpha\right\}\right) \leq \alpha.$

(2) $\lim_{m\to\infty} \inf_{p_i \geq \delta} \left[\widehat{\text{q-value}}(p_i) - \text{q-value}(p_i)\right] \geq 0$

(3) $\lim_{m\to\infty} \inf_{t \geq \delta} \left[\widehat{\text{FDR}}(t) - \text{FDR}(t)\right] \geq 0$

(Storey, Taylor, Siegmund 2002)

•*Plausibly holds for microarray data.*

Translation: Given "clumpy microarray dependence" and large m ...

Bayesian interpretation holds

• $\mathrm{FDR}(t) \sim \mathrm{pFDR}(t) \to \Pr^{\infty}(H=0|P \leq t)$

Can look at all thresholds simultaneously

• $\widehat{\mathrm{FDR}}(t)$ dominates $\mathrm{FDR}(t)$ over all t

The FDR is controlled

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

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

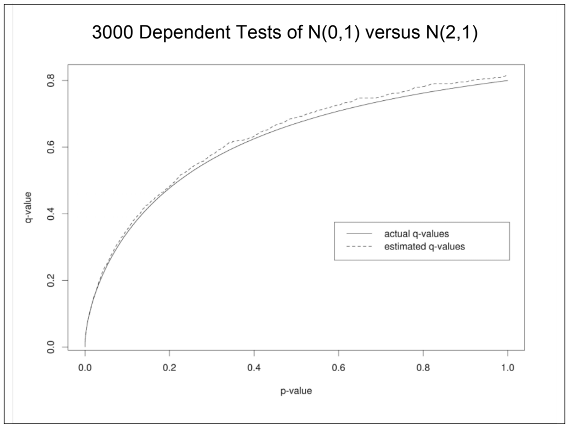
• $\widehat{\text{q-value}}(t)$ dominates q-value(t) over all t (even $t = p_i$!)

Simulation Study

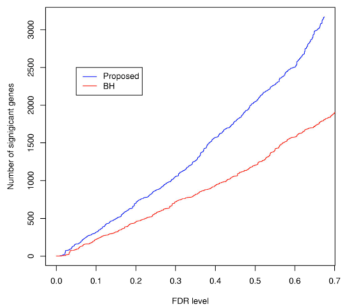
- **Performed 3000 hypothesis tests of H_0:*N*(0,1) versus H_1:*N*(2,1)**
- **The statistics had correlation 0.40 in blocks of 50**
- Two conclusions:

  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



Power Comparison



Power Comparison

FDR Level	# Significant BH	# Significant PP
0.01	**1**	**5**
0.02	**8**	**21**
0.03	21	80
0.04	**76**	**123**
0.05	88	160
0.10	**221**	**317**

$\hat{\pi}_0 = 1$ $\hat{\pi}_0 = 0.67$

SAM Version

$$\hat{\mathrm{FDR}}(\Delta) = \frac{\hat{\pi}_0 \sum_{b=1}^{B} \#\left\{d_i^{0b} : d_i^{0b} \le \ell(\Delta) \text{ or } d_i^{0b} \ge r(\Delta)\right\} / B}{\#\left\{d_i : d_i \le \ell(\Delta) \text{ or } d_i \ge r(\Delta)\right\}}$$

$$= \frac{\hat{\pi}_0 \cdot \text{avg no. nulls called significant}}{\text{no. observed called significant}}$$

$$\hat{\pi}_0(\Delta') = \frac{\#\left\{d_i : d_i > \ell(\Delta') \text{ or } d_i < r(\Delta')\right\}}{\sum_{b=1}^{B} \#\left\{d_i^{0b} : d_i^{0b} > \ell(\Delta') \text{ or } d_i^{0b} < r(\Delta')\right\} / B}$$

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

Westfall, PH and SS Young (1993) *Resampling-based multiple testing: Examples and methods for p-value adjustment*, John Wiley & Sons, Inc

Benjamini, Y & Y Hochberg (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing *JRSS B* 57: 289-300

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
Y Ge et al (2001) Fast algorithm for resampling based p-value adjustment for multiple testing

Significance analysis of microarrays (SAM)

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

SAM-statistic

- For gene i

$$d_i = \frac{\bar{y}_i - \bar{x}_i}{s_i + s_0}$$

$\bar{y}_i =$ mean of Irradiated samples

$\bar{x}_i =$ mean of Unirradiated samples

$s_i =$ Standard deviation of residuals for gene *i a assuming same variance*

$s_0 =$ Exchangeability factor estimated using all genes

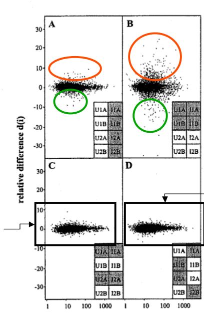
The exchangeability factor

- **Chosen to make signal-to-noise ratios independent of signal**
- **Computation**
  - **Let s^α be the α percentile of the s_i values. Let $d_i^\alpha = r_i/(s_i + s^\alpha)$**
  - **Compute the 100 quantiles of the s_i values, denoted by $q_1 < q_2 < \cdots < q_{100}$**

    $\alpha \in (0, 0.05, 0.10, \ldots, 1.0)$
  - **For**
    - **Compute $v_j = mad(d_i^\alpha \mid s_i \in [q_j, q_{j+1})), j = 1,2,\ldots,99,$ where mad is the median absolute deviation from the median, divided by 0.64**
    - **Compute $cv(\alpha)$= coefficient of variation of the**
    - **Choose $\hat{\alpha} = \arg\min[cv(\alpha)]$. $\hat{s}_0 = s^{\hat{\alpha}}$. and v_j**

Scatter plots of relative difference

Random fluctuations in the data, measured by balanced permutations (for cell line 1 and 2)



Relative difference for a permutation of the data that was balanced betwe en cell lines 1 and 2.

The reference distribution

- Order the values of d_i (could be any stat)

$$d_{(1)} \le d_{(2)} \le \cdots \le d_{(p)}$$

- Permute the treatment labels, and compute a new set of ordered values

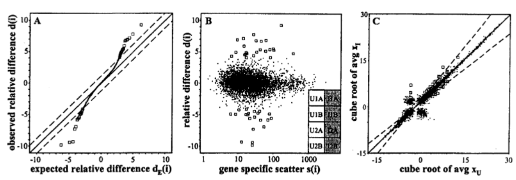
$$d^*_{(1)} \le d^*_{(2)} \le \cdots \le d^*_{(p)}$$

- Repeat step 2 for, say, 100 permutations:

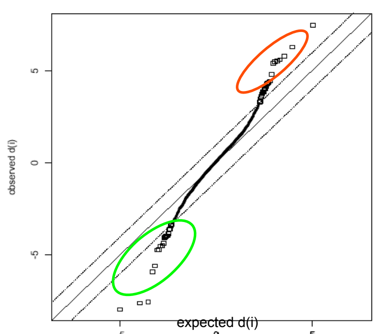
$$\begin{array}{cccc} d^{*1}_{(1)} \le & d^{*1}_{(2)} \le & \cdots & \le d^{*1}_{(p)} \\ d^{*2}_{(1)} \le & d^{*2}_{(2)} \le & \cdots & \le d^{*2}_{(p)} \\ & \vdots & \vdots & \\ d^{*100}_{(1)} \le & d^{*100}_{(2)} \le & \cdots & \le d^{*100}_{(p)} \end{array}$$

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

Selected genes



SAM plot



Delta	Ave # falsely significant	# called significant	False discovery rate
0.3	75.1	294	0.255
0.4	33.6	196	0.171
0.5	19.8	160	0.123
0.7	10.1	94	0.107
1.0	4.0	46	0.086

Delta is the half-width of the bar around the 45-degree line.

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**