Using False Discovery Rates in DNA Microarrays

John D. Storey
Department of Statistics
University of California, Berkeley
Email: storey@stat.berkeley.edu
Web: http://www.stat.berkeley.edu/~storey/

This research was done while at Stanford University, partly in collaboration with David Siegmund, Jonathan Taylor, and Rob Tibshirani. Also, thanks to Pat Brown, Brad Efron, and the Kruglyak Lab.

The Topic

How should one deal with false positives when testing thousands of genes for differential gene expression?

We will present some recent results in false discovery rates, and argue that these methods are well suited for the above task.

Microarray Data

n arrays, m genes

	array 1	array 2	array 3	array 4	$\cdots$	array n
gene 1	1.23	-2.61	-3.57	4.22	$\cdots$	5.12
gene 2	3.98	-0.294	1.73	2.97	$\cdots$	-2.43
$\vdots$			$\vdots$			$\vdots$
gene m	0.846	3.72	1.83	-1.10	$\cdots$	-2.94

Detecting Differential Gene Expression

- Suppose that we have n_1 microarrays taken from untreated cells and n_2 microarrays taken from treated cells (e.g., untreated=normal, treated=cancer). $n_1 + n_2 = n$.
- Which genes show a *statistically significant difference in gene expression* between these two types of cells?
- Answering this question helps to narrow down the search for genes involved in differentiating these cell types.
- For example, in the normal versus cancer case, *finding differentially expressed genes helps to identify genes involved in cancer*.

Some Recent Work

Ideker et al. (2000)

Newton et al. (2001)

Tusher, Tibshirani, Chu (2001) – SAM Software

Efron, Tibshirani, Storey, Tusher (2001)

Dudoit et al. (2002)

Example Data

- **Two condition microarray data set – Brem, Yvert, Clinton, Kruglyak, *Science* (2002).**

–Two strains of *S. cerevisiae* were considered. One is from the wild, the other from a lab.

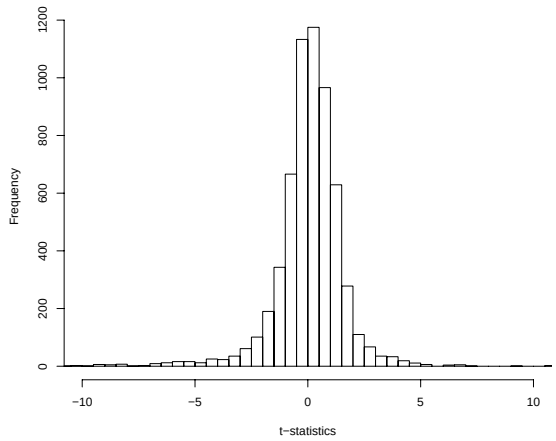
–Want to identify genes that are differentially expressed between the two strains.

–Each strain hybridized to a 6200+ cDNA microarray 6 times, for a total of 12 arrays.

–A two sample t-statistic t_i was calculated for each gene.

- ***Now what do we do?***

6200 t-statistics



The Statistical Approach

(1) Form a statistic for each gene

(2) Calculate the null distribution(s)

(3) Choose the rejection regions to use

(4) Assess the number of false positives given these statistics and rejection regions

For example, if we reject all $t_i < c_1$ or $t_i > c_2$, what can we say about the false positives?

The Intuitive Approach

(A) Rank the genes according to their significance for differential gene expression

(B) Associate a number with each gene that tells us how significant it is

Step (A) involves calculating a statistic for each gene and defining a nested set of significance regions. The significance regions implicitly determine the ranking and vice versa.

Step (B) essentially involves determining the null distribution of the genes, as well as something about false positives and true positives.

Returning to the Example

(A) We can rank the genes in order of their evidence for differential gene expression by their $|t_i|$ values

(B) Let H_i indicate whether gene i differentially expressed or not. Some options for associating a significance measure with gene i:

p-value$(t_i) = \Pr\left(|T_i| \geq |t_i| \,|\, H_i = 0\right)$ ***... marginal***

posterior prob $= \Pr\left(H_i = 0 \,|\, |T_i| = |t_i|\right)$ ***... marginal***

q-value$(t_i) \doteq \Pr\left(H_i = 0 \,|\, |T_i| \geq |t_i|\right)$ ***... both marginal and multivariate!***

Multiple Hypothesis Testing

•Outcomes when testing m hypotheses:

	Accept	Reject	Total
Null True	U	V	m_0
Alternative True	T	S	m_1
	W	R	m

•Error measures:

–FWER = $\Pr(V \geq 1)$ ***Family Wise Error Rate***

–FDR = $\mathrm{E}\left[\frac{V}{R} \,\middle|\, R > 0\right] \Pr(R > 0)$ ***False Discovery Rate***

–pFDR = $\mathrm{E}\left[\frac{V}{R} \,\middle|\, R > 0\right]$ ***positive False Discovery Rate***

FDR – Benjamini and Hochberg (1995) **pFDR – Storey (2001)**

Frequentist Interpretation

•False discovery rates measure the expected proportion of false positives among all significant hypotheses.

•The FDR includes cases where no hypotheses are significant – the "proportion" is set to zero.

•The pFDR only considers cases where at least one significant hypothesis is found.

•If a procedure is applied to call hypotheses significant, then a pFDR of 5%, for example, says that on average the proportion of false positives among significant hypotheses is 5%.

•Loosely ... if we find 100 significant genes under some method with a pFDR of 5%, then we expect about 5 false positive genes.

Bayesian Interpretation

- Suppose m hypothesis tests are performed with independent statistics $X_1, \ldots, X_m$ and rejection 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)

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

q-values

- In general, for a nested set of rejection 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,

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

- We want to estimate the q-value for each gene, and use this number to measure the significance of each gene.

Why use FDR's in DNA Microarrays?

- Microarray experiments tend to be exploratory
- False discovery rates have an easy and useful interpretation – both frequentist and Bayesian
- They are robust against microarray dependence
- The more genes, the better

Recent Work on FDR's

Abramovich, Benjamini, Donoho, Johnstone (2000)
Benjamini and Hochberg (1995)
Benjamini and Hochberg (2000)
Benjamini and Liu (1999)
Benjamini and Yekutieli (2001)
Efron, Tibshirani, Storey, Tusher (2001)
Genovese and Wasserman (2001)
Storey (2001a)
Storey (2001b)
Storey and Tibshirani (2001)
Storey, Taylor, and Siegmund (2002)
Tusher, Tibshirani, Chu (2001)
Yekutieli and Benjamini (1999)

A Frequentist Estimate

- **Suppose m hypothesis tests are performed with p-values $P_1, \ldots, P_m$. The rejection region is $\Gamma = [0, t]$ for some t.**
- **We can re-write:**

$$pFDR(t) = \frac{\pi_0 \cdot t}{\Pr(P \le t)}$$

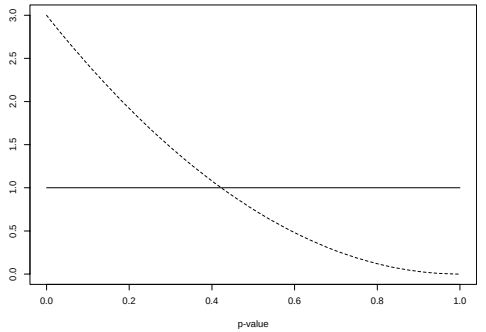
$$FDR(t) = \frac{\pi_0 \cdot t}{\Pr(P \le t | R(t) > 0)}$$

- **Estimates:**

$$\begin{aligned} \widehat{\Pr}(P \le t | R(t) > 0) &= \frac{\#\{p_i : p_i \le t\} \vee 1}{m} \\ \widehat{\Pr}(P \le t) &= \widehat{\Pr}(P \le t | R(t) > 0) \cdot \Pr_0(R(t) > 0) \\ \widehat{\pi}_0(\lambda) &= \frac{\#\{p_i > \lambda\}}{(1-\lambda)m} \end{aligned}$$

Estimate of π_0 (cont'd)

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



The Estimates

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

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

$$\widehat{\Pr}_\lambda(H = 0 | P \le t) = \widehat{pFDR}_\lambda(t)$$

Using $\widehat{pFDR}$ and $\widehat{FDR}$ in Four Scenarios

(1) To estimate for a fixed $[0, t]$, use $\widehat{FDR}_\lambda(t)$ and $\widehat{pFDR}_\lambda(t)$.

(2) To control the FDR at level α, form

$$\widehat{t}^\alpha_\lambda = \max\{t : \widehat{FDR}_\lambda(t) \le \alpha\}.$$

Reject all $p_i \le \widehat{t}^\alpha_\lambda$. Note: when $\lambda = 0$, this is equivalent to the Benjamini and Hochberg (1995) algorithm.

(3) To estimate over all rejection regions simultaneously, simply use $\widehat{FDR}_\lambda(\cdot)$ and $\widehat{pFDR}_\lambda(\cdot)$.

Four Scenarios (cont'd)

(4a) To calculate "FDR adjusted p-values", form:

$$\widehat{\alpha}_{FDR,\lambda}(p_i) = \min_{s \geq p_i} \widehat{FDR}_\lambda(p_i).$$

These estimate the *simultaneous FDR controlling curve.*

(4b) To estimate the q-values, form:

$$\widehat{q}_\lambda(p_i) = \min_{s \geq p_i} p\widehat{FDR}_\lambda(p_i).$$

Note this is also:

$$\widehat{q}_\lambda(p_i) = \min_{s \geq p_i} \widehat{\Pr}_\lambda(H = 0 | P \leq s).$$

Finite Sample Results

- **Suppose the null p-values are independent ...**
- **Then**

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

(Storey 2001)

- **Also, if we limit $\widehat{t}^\alpha_\lambda$ to the interval $[0, \lambda]$, then**

$$FDR\left(\widehat{t}^\alpha_\lambda\right) \equiv \mathrm{E}\left[\left.\frac{V(\widehat{t}^\alpha_\lambda)}{R(\widehat{t}^\alpha_\lambda)}\right| R(\widehat{t}^\alpha_\lambda) > 0\right] \Pr(R(\widehat{t}^\alpha_\lambda) > 0) \leq \alpha.$$

(Storey, Taylor, Siegmund 2002)

- **But does this assumption hold for 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 ...**

(0) $\lim_{m\to\infty} \sup_{t\geq\delta} \left| \widehat{pFDR}_\lambda(t) - \widehat{FDR}_\lambda(t) \right| = 0$

(1) $\lim_{m\to\infty} \inf_{t\geq\delta} \left[\widehat{pFDR}_\lambda(t) - pFDR(t) \right] \geq 0$

(2) $\lim_{m\to\infty} \sup_{t\geq\delta} |pFDR(t) - \Pr^\infty(H = 0|P \leq t)| = 0$**, where**

$$\Pr^\infty(H = 0|P \leq t) \equiv \frac{\pi_0 F_0(t)}{\pi_0 F_0(t) + \pi_1 F_1(t)}$$

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

(4) $\lim_{m\to\infty} FDR\left(\widehat{t}^\alpha_\lambda\right) \leq \alpha$

(Storey, Taylor, Siegmund 2002)

- ***Likely holds for microarray data.***

Translation ...

- **Given "clumpy microarray dependence" ...**

We can look at all rejection regions simultaneously

- $\widehat{pFDR}_\lambda(t) = \widehat{\Pr}_\lambda(H = 0|P \leq t)$ **asymptotically dominates** $pFDR(t)$ **over all** t **simultaneously**
- $FDR(t)$ **and** $pFDR(t) \to \Pr^\infty(H = 0|P \leq t)$

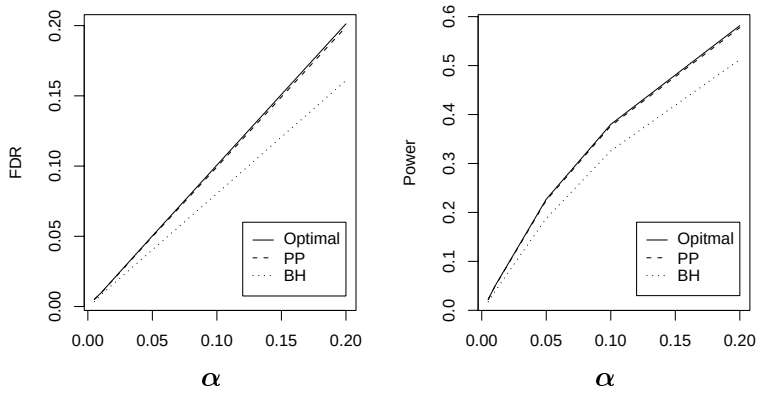
The estimated q-values conservatively estimate the true q-values

- $\widehat{q}_\lambda(t)$ **asymptotically dominates q-value**(t) **over all** t **simultaneously**

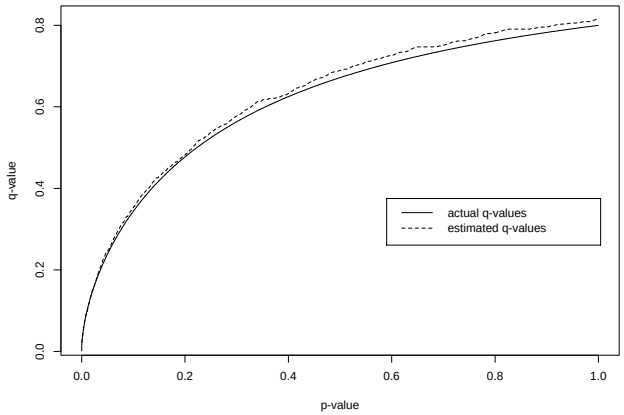
The FDR is controlled

- $\widehat{t}^\alpha_\lambda = \max\{t : \widehat{FDR}_\lambda(t) \leq \alpha\}$ **asymptotically controls the FDR at level** α**.**

3000 Tests of $N(0,1)$ **versus** $N(2,1)$**,** $\pi_0 = 0.8$**,** $\rho = 0.4$



3000 Tests of $N(0,1)$ **versus** $N(2,1)$**,** $\pi_0 = 0.8$**,** $\rho = 0.4$

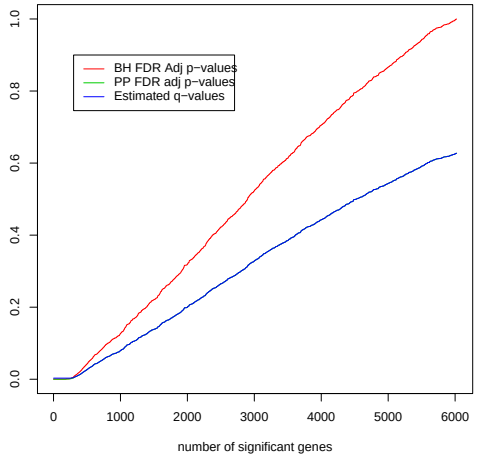


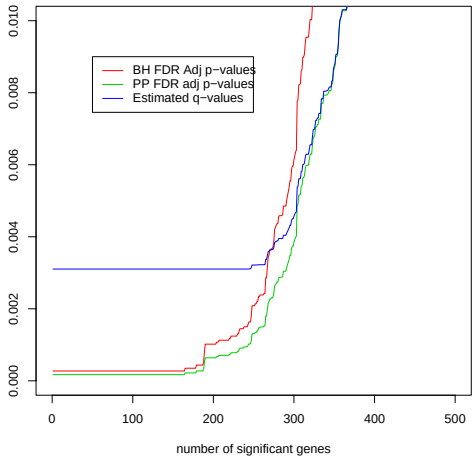
Yeast Data Results

- **Steps 1-3:**

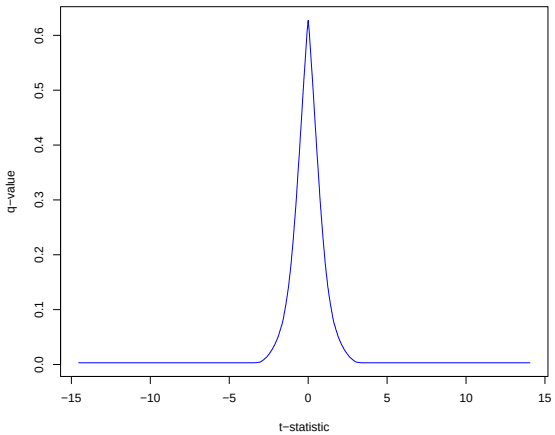
–Two sample t-statistics were formed for each gene.
–Null dist'n calculated by randomly permuting strain labels.
–Symmetric rejection regions used.

- **With $\lambda = 0.5$, $\widehat{\pi}_0 = 0.63$!!!**
- **At $\alpha = 0.001$, BH method finds 189 significant genes. Our method ($\lambda = 0.5$) finds 243 significant genes.**
- **At $\alpha = 0.05$, BH method finds 531 significant genes. Our method finds 697 significant genes.**

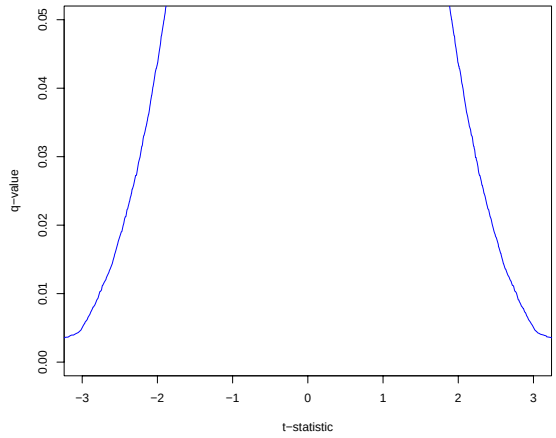




Yeast Data q-values



Yeast Data q-values



Concluding Remarks

- **False discovery rates are a natural false positive measure to use for the problem of detecting differential gene expression**
- **The estimated q-value should be reported for each gene in such an experiment because:**
  - **It takes the multiple comparisons into account [e.g., p_i and $\Pr(H = 0|P = p_i)$ do not]**
  - **It is robust near the origin and against dependence**
  - **It does not force the researcher to make a decision, but rather serves as an exploratory guide**
  - **It has a straightforward posterior probability interpretation**
- **Papers and talk at** `http://www.stat.berkeley.edu/~storey/`