

Empirical Bayes Methods and False Discovery Rates for Microarrays

Bradley Efron^{1,*} and Robert Tibshirani²

¹*Department of Statistics and Division of Biostatistics, Stanford University, Stanford, California*

²*Division of Biostatistics and Department of Statistics, Stanford University, Stanford, California*

In a classic two-sample problem, one might use Wilcoxon's statistic to test for a difference between treatment and control subjects. The analogous microarray experiment yields thousands of Wilcoxon statistics, one for each gene on the array, and confronts the statistician with a difficult simultaneous inference situation. We will discuss two inferential approaches to this problem: an empirical Bayes method that requires very little a priori Bayesian modeling, and the frequentist method of "false discovery rates" proposed by Benjamini and Hochberg in 1995. It turns out that the two methods are closely related and can be used together to produce sensible simultaneous inferences. *Genet. Epidemiol.* 23: 70–86, 2002. © 2002 Wiley-Liss, Inc.

Key words: multiple comparisons; simultaneous hypothesis tests; a posteriori probability of gene significance

INTRODUCTION

Microarrays epitomize the high-throughput devices that are revolutionizing biomedical research. They are also enlivening statistics. When applied in a comparative experiment, for example, comparing gene activity in tumor and normal cells, microarrays produce intriguing but difficult simultaneous inference problems. In the main example used here, a rather typical microarray experiment, we will have more than 3,000 Wilcoxon two-sample tests to consider at once.

*Correspondence to: Robert Tibshirani, Department of Health Research and Policy, and Statistics, HRP Redwood Building, Room T101C, Stanford University, Stanford, CA 94305.

E-mail: tibs@stat.stanford.edu

Received for publication 11 December 2001; Revision accepted 13 March 2002

Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/gepi.01124

Two analyses will be discussed: a frequentist approach based on Benjamini and Hochberg’s [1995] false discovery rate procedure, and an empirical Bayes methodology developed in Efron et al. [2001A, 2001B]. The two approaches are closely related and can be used to support each other, which is the principal point of this article.

Hedenfalk et al. [2001] report on a microarray experiment concerning the genetic basis of breast cancer. It is known that unfavorable mutations of two different genes, *BRCA1* and *BRCA2*, lead to greatly increased breast cancer risk. How do the tumors resulting from the two different mutations differ in their genetic activity? To answer this question, tumors from 22 women were analyzed, with seven of the women known to have the *BRCA1* mutation, eight known to have *BRCA2*, and seven, labeled “sporadics,” having neither mutation. Each woman’s tumor cells were analyzed on a separate microarray plate that measured expression levels for 3,226 genes. Table I shows a small portion of the resulting 3,226×22 data matrix.

Here is a schematic description of the genetic technology behind the numbers in Table I. The known DNA base sequences for each of the 3,226 genes were printed at known positions on the microarray plates. (There were actually 5,361 genes to begin with, only 3,226 of which produced accurately readable results.) When the tumor cells were hybridized on a plate, they generated messenger RNA in proportion to each gene’s activity, producing a measurable expression level at its corresponding DNA plate location. The expression levels were optically read using a red dye for the effect of interest and a green dye for a background measurement used as a control. The numbers in Table I are the logarithms of the ratio of red to green intensities measured at each gene location as described in detail in Figure 1 of Hedenfalk et al. [2001]. Some adjustments were made to the raw ratios (see Remarks section, Data Adjustments).

Figure 1 concerns the comparison of gene activity in *BRCA1* tumors versus *BRCA2* tumors, and so involves only the first 15 columns of the matrix begun in Table I. For this analysis, each gene’s data were summarized by its Wilcoxon statistic: the 15 expression levels for gene *i*, 7 *BRCA1* and 8 *BRCA2*, were ranked, giving the rank sum statistic

$$Y_i = \text{sum of BRCA2 ranks, } (i = 1, 2, \dots, n = 3226). \tag{1.1}$$

TABLE I. Small Portion of the Data from a Microarray Experiment by Hedenfalk et al. [2001] Concerning Genetic Activity Differences in Breast Cancer Cells; Expression Levels for 3,226 Genes on 22 Microarray Plates; 7 from Women with *BRCA1* Mutation, 8 *BRCA2*, 7 Sporadic (Neither).^a

	BRCA1				BRCA2				Sporadic			
	1	2	—	7	1	2	—	8	1	2	—	7
Gene1	-1.29	-1.41	—	-0.55	-0.70	1.33	—	1.14	-0.44	0.26	—	-0.23
Gene2	2.03	0.58	—	-0.12	0.23	-0.91	—	-0.39	0.70	-1.55	—	2.17
Gene3	0.32	-0.44	—	1.25	0.53	-0.96	—	-0.51	-1.26	-0.74	—	-0.64
Gene4	-1.31	-0.98	—	0.24	-0.24	0.28	—	2.13	0.32	0.42	—	-0.65
Gene5	-0.66	-0.07	—	1.22	-0.41	-0.88	—	-0.83	0.25	-0.97	—	-0.21

^a Tabled values are adjusted log(red/green) ratios from spotted cDNA microarrays.

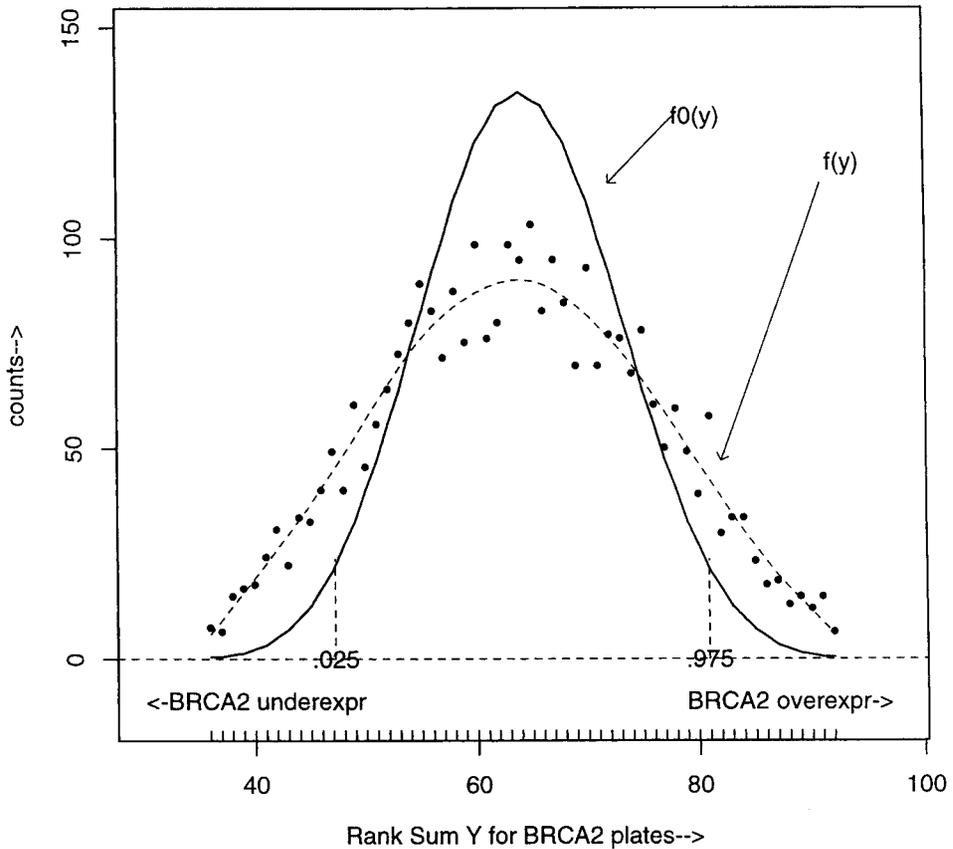


Fig. 1. Rank sum statistics comparing BRCA1 versus BRCA2 for the 3,226 genes; points are actual counts, solid curve shows expected counts under the null hypothesis of no activity differences. Dashed curve is a Poisson regression fit to the actual counts, as explained in the Empirical Bayes Inferences section.

The Y_i range from a low of 36, if the BRCA2 numbers were the 8 smallest among the 15, to a high of 92 if they were the 8 largest,

$$36 \leq Y_i \leq 92. \quad (1.2)$$

In the usual terminology, small or large values of Y_i correspond respectively to *underexpression* or *overexpression* of gene i for BRCA2 compared to BRCA1 tumors (or equivalently *downregulation* or *upregulation*).

The points in Figure 1 are the actual Y counts. For example, the leftmost point, plotted at (36,8), represents the 8 genes for which Y_i equaled 36. The solid curve shows the expected counts assuming no difference between BRCA1 and BRCA2 expression levels, i.e., under the permutation distribution of the numbers 1, 2, ..., 15 [called the “Wilcoxon (7,8)” distribution in what follows]. The expected count is only 0.501 for $Y = 36$, so there are 16 times as many genes with $Y_i = 36$ as we

would expect if there were no expression differences between BRCA1 and BRCA2 tumors.

The dashed line, a smooth Poisson regression fit to the points, is much wider than the expected curve, clearly indicating substantial genetic activity differences for at least some of the genes. The question of interest is “which of the 3,226 genes can we confidently label as differently active?” The naive answer would be to run 3,226 separate Wilcoxon tests. Six hundred fourteen of the Y_i 's lie either below the 0.025 point for a standard Wilcoxon(7,8) distribution or above its 0.975 point. This would give a reasonable criterion for declaring any single prechosen gene differently active, but it leads to an expected 161 false declarations if none of the 3,226 genes are actually different.

Efron et al. [2001B] developed a simple empirical Bayes approach to this kind of simultaneous inference problem. As described in the next section, the approach produces believable a posteriori probabilities of activity differences for each gene, starting with a minimum of a priori assumptions. In Figure 1's case, we will see that the estimated values of probability{different| Y_i } for the 614 “rejected” genes range from a low of 0.50 near the rejection thresholds to a high of nearly 0.95 at the extremes of the Y scale.

The downside of the empirical Bayes approach is its ad hoc appearance compared to the mathematical certitudes of standard hypothesis testing theory. Benjamini and Hochberg [1995], beginning with an algorithm of Simes [1986], developed an attractive new multiple comparison technique that produces exact frequentist inferences for what they call the “false discovery rate” (FDR). Section 3 discusses the FDR algorithm and shows that in an important sense it exactly matches the empirical Bayes methodology, perhaps strengthening belief in both techniques. We can use the two approaches in a complementary way to answer the kind of simultaneous inference problems raised in Figure 1. A useful variant called the “local false discovery rate” is introduced in the section on Local False Discovery Rate.

The Three-Way Comparison section returns to the full data set of Table I, using the empirical Bayes methodology to make a three-way activity comparison between BRCA1, BRCA2, and Sporadic tumors. We close in the Remarks section with some notes and comments.

The statistics literature for microarrays is quite recent, with much of it unpublished. Useful references for simultaneous testing situations include Newton et al. [2000], Dudoit et al. [2000], Tusher et al. [2000], as well as Efron et al. [2001B].

EMPIRICAL BAYES INFERENCE

We assume that there are two classes of genes, “Different” and “Not Different,” in our example meaning that the gene is either differently or not differently expressed in BRCA1 and BRCA2 tumors. Let the prior probabilities of the two classes be p_1 and $p_0 = 1 - p_1$, with corresponding prior densities $f_1(y)$ and $f_0(y)$ for the summary statistic Y ,

$$\begin{aligned}
 p_1 &= \text{Prob}\{\text{Different}\} & f_1(y) & \text{density of } Y_i \text{ if gene}_i \text{ “Different”} \\
 p_0 &= \text{Prob}\{\text{Not Different}\} & f_0(y) & \text{density of } Y_i \text{ if gene}_i \text{ “Not Different”}.
 \end{aligned}
 \tag{2.1}$$

Finally let $f(y)$ be the mixture density

$$f(y) = p_0 f_0(y) + p_1 f_1(y). \tag{2.2}$$

A direct application of Bayes' theorem gives *a posteriori* probabilities

$$p_1(y) \equiv \text{Prob}\{\text{Different} | Y_i = y\} = 1 - p_0 f_0(y)/f(y) \tag{2.3}$$

and

$$p_0(y) \equiv \text{Prob}\{\text{Not Different} | Y_i = y\} = p_0 f_0(y)/f(y).$$

Full Bayesian analysis would require prior specification of $p_0, p_1, f_0(y)$, and $f_1(y)$, but we can use the massively parallel structure of microarray data to estimate an empirical Bayes version of (2.3). In doing so we will be carrying out the kind of empirical Bayes or compound Bayes analysis suggested by Robbin nearly 50 years ago, for instance in Robbins [1956], but rarely practical in traditional biometric settings.

Figure 2 shows an empirical Bayes analysis for the situation of Figure 1: $f_0(y)$ here is the discrete density for a Wilcoxon(7,8) variate, the solid curve in Figure 1 divided by 3,226; $f(y)$ has been estimated by a Poisson regression fit to the Y counts. (Specifically by modeling $f(y)$ as a natural spline having 5 degrees of freedom and with offset $\log[f_0(y)]$, giving $\hat{f}(y)$ proportional to the dashed curve in Figure 1.) Together these give an estimate of $p_1(y) = \text{prob}\{\text{different} | y\}$ in (2.3),

$$\hat{p}_1(y) = 1 - p_0 f_0(y)/\hat{f}(y). \tag{2.4}$$

The prior “not different” probability p_0 is unidentifiable without strong parametric assumptions, such as normality, on $f_0(y)$ and $f(y)$. However, the most conservative possible choice, $p_0 = 1$, the choice that minimizes the probability of detecting “different,” still gives interesting results. The solid curve in Figure 2 is $\hat{p}_1(y)$ for $p_0 = 1$ in (2.4). We see that for $p_0 = 1$, genes with $Y_i \leq 39$ or $Y_i \geq 89$ have $\text{prob}\{\text{different} | Y\}$ exceeding 0.90. There are 101 such genes, 49 on the left and 52 on the right.

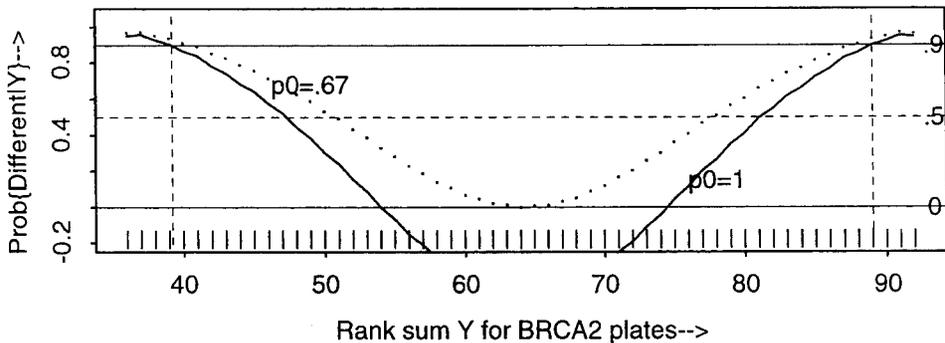


Fig. 2. Empirical Bayes estimates (2.3) of $p_1(y) = \text{Prob}\{\text{Different} | Y_i = y\}$ for the comparison of BRCA1 and BRCA2 in Figure 1. Solid curve: Assuming prior probability p_0 of “Not Different” is 1; Dotted curve: assuming $p_0 = 0.67$, the largest value of p_0 that makes $p_1(y)$ everywhere nonnegative.

An obvious objection to setting $p_0 = 1$ is that $p_1(y)$ then becomes negative near the middle of the Y scale. Expression (2.3) shows that in order for $\hat{p}_1(y)$ to be always nonnegative, we must have

$$p_0 \leq \hat{p}_{0,\max} = \min_y \{ \hat{f}(y)/f_0(y) \}. \tag{2.5}$$

The dotted curve in Figure 2 indicates $\hat{p}_1(y)$ for $p_0 = \hat{p}_{0,\max} = 0.67$. This raises $\hat{p}_1(y)$ somewhat in the tails, so that now $\text{prob}\{\text{different}|y\}$ exceeds 0.90 for $Y_i \leq 40$ or $Y_i \geq 88$, a total of 134 genes. (Remark F of Efron et al. [2001B] suggests a more stable estimate of $p_{0,\max}$.) We will see in the next section that the ambiguity in p_0 plays the same role in the FDR theory as here.

The argument leading to Figure 2 has a strong heuristic foundation but no formal basis. To this end, the asymptotic accuracy of (2.4) as the number of genes goes to infinity is established under some restrictions in Storey [2001A]. We take another approach in the next two sections, where (2.4) is related to the frequentist FDR algorithm of Benjamini and Hochberg [1995].

CONNECTION WITH FALSE DISCOVERY RATES

The empirical Bayes analysis of Section 2 is closely related to Benjamini and Hochberg’s theory of FDRs [1995]. We begin with a brief review of the FDR algorithm. Suppose one wants to simultaneously test n null hypotheses H_1, H_2, \dots, H_n on the basis of independent test statistics Y_1, Y_2, \dots, Y_n . From the Y_i we calculate corresponding P -values P_i , denoting the ordered values as

$$P_{(1)} \leq P_{(2)} \leq \dots P_{(n)}, \tag{3.1}$$

$P_{(1)}$ being the most significant and $P_{(n)}$ the least significant in the usual terminology.

Let $\mathcal{R}(\mathbf{Y})$ be a proposed rule for selecting which of the null hypotheses to reject, e.g., “Reject H_i if P_i is among the smallest 5% of the P -values and $P_i \leq 0.01$.” Following work by Simes [1986], Benjamini and Hochberg defined the FDR of \mathcal{R} to be its expected proportion of false rejections,

$$\text{FDR}(\mathcal{R}) = E\{\text{proportion of rejected } H_i \text{ that are actually true}\}, \tag{3.2}$$

(with the proportion equaling zero if nothing is rejected) and provided a useful algorithm for controlling the FDR below a preset value α : let

$$i_\alpha = \operatorname{argmax}_i P_{(i)} \leq \frac{i}{n} \alpha \quad [p_0 \equiv \text{proportion of true } H_i]. \tag{3.3}$$

Then the rejection rule

$$\mathcal{R}_\alpha = \{\text{Reject all } H_i \text{ with } P_i \leq P_{(i_\alpha)}\} \tag{3.4}$$

has

$$\text{FDR}(\mathcal{R}_\alpha) \leq \alpha; \tag{3.5}$$

(3.5) becomes an equality if the Y_i are continuous as well as independent, theorem (5.1) of Benjamini and Yekutieli [2001]. Other FDR-controlling rules are available, but we will concentrate on (3.3, 3.4).

In the context of Figure 1, $n = 3,226$ and $H_i = \{\text{gene}_i \text{ Not Different}\}$. Notice that p_0 in (2.1) is the expected proportion of true H_i , nearly the same as its definition

in (3.3). The 1995 paper took $p_0 = 1$, which here as in (2.3) is the most conservative choice, minimizing i_α and making inequality (3.5) least sharp. In more recent work, Benjamini and Yekutieli [2001], they consider estimating p_0 , see also Storey [2001A,B]. Empirical Bayes considerations, as in (2.5) and remark F of Efron et al. [2001B], give intuitively appealing bounds for p_0 .

Figure 3 applies the FDR-controlling algorithm to the comparison of BRCA1 with BRCA2, using $\alpha = 0.10$ and $p_0 = 1.0$. The step function in the left panel shows the ordered P values (3.1) for one-sided Wilcoxon tests of H_i versus the alternative that gene $_i$ underexpresses BRCA2; that is, P_i is the probability that a Wilcoxon(7,8) variable is equal or less than the observed value Y_i . The right panel shows $\mathcal{R}_{0.10}$ applied to the overexpression of BRCA2, now with $P_i = \text{Prob}\{\text{Wilcoxon}(7,8) \geq Y_i\}$. (Notice that the step functions are empirical cdf's of the P values, rotated 90° .)

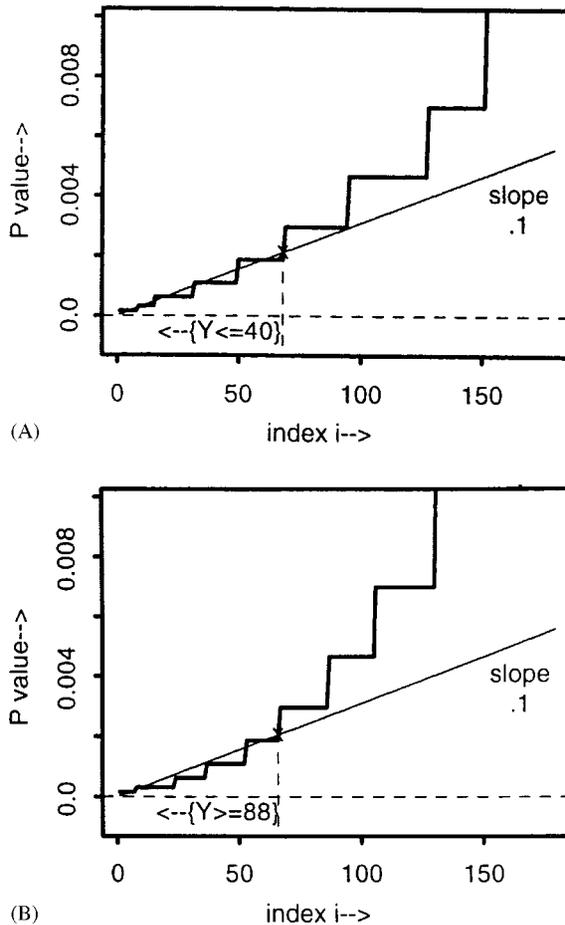


Fig. 3. Application of false discovery rate-controlling algorithm to BRCA1/BRCA2 comparison, $\alpha = 0.10$, $p_0 = 1.0$. A: Step function shows ordered p-values for one-sided Wilcoxon tests that reject for small values of rank sum statistic Y_i ; $\mathcal{R}_{0.10}$ procedure (3.4) rejects for the 68 genes having $Y_i \leq 40$. B: Same, rejecting for large values of Y_i ; $\mathcal{R}_{0.10}$ rejects for the 66 genes with $Y_i \geq 88$.

The close connection of Benjamini and Hochberg’s FDR procedure with the empirical Bayes methodology of the section on Empirical Bayes Inferences follows directly from Bayes theorem. Let $F_0(y)$ and $F(y)$ be the cumulative distribution functions (CDFs) corresponding to $f_0(y)$ in (2.1) and $f(y)$ in (2.2), and define the “Bayesian FDR” for $\{Y \leq y\}$ to be

$$\begin{aligned} \text{Fdr}(y) &\equiv p_0 F_0(y)/F(y) \\ &= \text{Prob}\{\text{gene}_i \text{ Not Different} | Y_i \leq y\} \end{aligned} \tag{3.6}$$

as in (2.3). If we have N_y genes with $Y_i \leq y$ then, starting from (2.1) and assuming independence, the number N_{y0} of the N_y from the “not different” class will be binomially distributed,

$$N_{y0} | N_y \sim \text{Bi}(N_y, \text{Fdr}(y)), \tag{3.7}$$

and for large N_y we can expect $\text{Fdr}(y)$ to be close to $\text{FDR}(Y_i \leq y)$, (3.2). This will be true even if the Y_i are correlated, a mixing condition being enough to ensure asymptotic equivalence, as shown in Genovese and Wasserman [2001] and Storey [2001A].

Now let $\bar{F}(y)$ be the usual empirical CDF of the Y_i ’s, $\bar{F}(y) = \#\{Y_i \leq y\}/n$. The obvious nonparametric estimate for $\text{Fdr}(y)$ is

$$\bar{\text{Fdr}}(y) = p_0 F_0(y)/\bar{F}(y). \tag{3.8}$$

Equivalence Theorem: The Benjamini-Hochberg rule \mathcal{R}_α , (3.4) is equivalent to rejecting all H_i with $Y_i \leq y_\alpha$, where y_α is defined by

$$y_\alpha = \max_y \{\bar{\text{Fdr}}(y) \leq \alpha\}. \tag{3.9}$$

Reversing the y scale, a similar result holds for rejection regions $\{Y_i \geq y\}$.

Proof: Let $Y_{(i)}$ indicate the i th ordered value of $\{Y_1, Y_2, \dots, Y_n\}$. Then $\bar{F}(Y_{(i)}) = i/n$ and $F_0(Y_{(i)}) = P_{(i)}$. The constraint $\bar{\text{Fdr}}(y) \leq \alpha$ is equivalent to

$$p_0 P_{(i)}/(i/n) \leq \alpha \quad \text{or} \quad P_{(i)} \leq \frac{i}{n} \frac{\alpha}{p_0}, \tag{3.10}$$

coinciding with the FDR definitions (3.3), (3.4). Tied values of Y_i can be ordered arbitrarily without affecting this argument, as can be seen from inspection of Figure 3.

The equivalence theorem says that if we choose the rejection region $\{Y_i \leq y\}$ as large as possible subject to the constraint that the estimated empirical Bayes probability $\text{Prob}\{\text{Not Different} | Y \leq y\}$ is no greater than α , then our expected proportion of false rejections is also less than α . This is true for any choice of p_0 in the two algorithms and in particular for the conservative choice $p_0 = 1$. In this situation one can be both a Bayesian and frequentist simultaneously.

The FDR theorem was originally proved under an independence assumption on the test statistics Y_1, Y_2, \dots, Y_n . Recent work by Benjamini and Yekutieli [2001], relaxes this assumption to allow a form of positive dependence. However, independence plays no essential role in the empirical Bayes approach — all we need is $\bar{F}(y)$ in (3.8) to be a reasonable estimator of $F(y)$ — which suggests that the

FDR algorithm should give reasonably accurate results under quite general conditions on the test statistics. The assumptions underlying the empirical Bayes and FDR methods are further discussed in the next section, which provides a further connection between the FDR and empirical Bayes approaches, and illustrates the principal advantage of the latter.

LOCAL FALSE DISCOVERY RATE

What we called the Bayesian FDR in (3.6) can be defined for general rejection regions, including infinitesimally “local” ones. For \mathcal{Y} , a subset of the Y sample space, let

$$\begin{aligned} \text{Fdr}(\mathcal{Y}) &\equiv p_0 \text{Prob}_{f_0} \{ Y \in \mathcal{Y} \} / \text{Prob}_f \{ Y \in \mathcal{Y} \} \\ &= \text{Prob} \{ \text{gene}_i \text{ Not Different} \mid Y_i \in \mathcal{Y} \}, \end{aligned} \tag{4.1}$$

with f_0 and f defined as in (2.1), (2.2). In Figure 1, for example, we might take $\mathcal{Y} = \{ Y_i \leq 47 \text{ or } Y_i \geq 81 \}$, the 0.05 (actually 0.054) two-sided Wilcoxon rejection region. Estimating the denominator in (4.1) by the proportion of Y_i 's in \mathcal{Y} , $614/3,226 = 0.190$, gives

$$\overline{\text{Fdr}}(\mathcal{Y}) = p_0 \frac{0.054}{0.190} = p_0 \cdot 0.284. \tag{4.2}$$

Under the conservative assumption $p_0 = 1$, we expect about 28% of the “0.05 significant” genes to actually be Not Different, whereas $p_0 = \hat{p}_{0,\max} = 0.67$ gives 19%.

Local FDR

Efron et al. [2001A and 2001B] defined the *local FDR* at point y in the Y space to be the function $p_0(y)$ in (2.3),

$$\begin{aligned} \text{fdr}(y) &= p_0 f_0(y) / f(y) = p_0(y) \\ &= \text{Prob} \{ \text{Not Different} \mid Y_i = y \}. \end{aligned} \tag{4.3}$$

There is a simple Bayesian relationship between $\text{Fdr}(\mathcal{Y})$ and $\text{fdr}(y)$:

Averaging Theorem:

$$\text{Fdr}(\mathcal{Y}) = E_f \{ \text{fdr}(y) \mid y \in \mathcal{Y} \}. \tag{4.4}$$

Proof:

$$E_f \{ \text{fdr}(y) \mid y \in \mathcal{Y} \} = \frac{\int_{\mathcal{Y}} [p_0 f_0(y) / f(y)] f(y) dy}{\int_{\mathcal{Y}} f(y) dy} = p_0 \text{Prob}_{f_0} \{ \mathcal{Y} \} / \text{Prob}_f \{ \mathcal{Y} \} = \text{Fdr}(\mathcal{Y}).$$

In other words, $\text{Fdr}(\mathcal{Y})$ is the conditional f -average of $\text{fdr}(y)$ for $y \in \mathcal{Y}$.

The advantage of the local fdr is its specificity: it provides a measure of belief in gene i 's “significance” that depends on Y_i 's exact value, not on its inclusion in a larger set of possible values. Consider $\mathcal{Y} = \{ Y_i \leq 40 \}$, the FDR-controlling set for $\alpha = 0.10$, $p_0 = 1.0$, on the left side of Figure 3. It has overall Bayesian $\overline{\text{Fdr}} = 0.089$,

(3.4), but with estimated local values of $\text{fdr}(y)$ ranging from 0.04 to 0.13. This just says the obvious, that the boundary value $y = 40$ is the most likely point in \mathcal{Y} to yield a false detection, but it is nice to have a quantitative assessment. A biogeneticist could use the observed fdr values quite flexibly, without necessarily declaring a sharp boundary between significant and not significant cases, and perhaps including a priori opinions of differential gene activity as discussed in the section entitled Exchangeability and Prior Beliefs.

The main disadvantage of the local fdr is the need to estimate the density $f(y)$ in (4.3) (or more generally to estimate the ratio $f_0(y)/f(y)$ in situations where $f_0(y)$ is not theoretically determined, see Estimating $f_0(y)$ of the Remarks section). For example, we needed the Poisson regression estimate $\hat{f}(y)$ in (2.4) to construct the curves $\hat{p}_1(y) = 1 - \mathbf{fdr}(y)$ of Figure 2.

In discrete situations like that of Figure 1, the simplest estimate of $f(y)$ is

$$\tilde{f}(y) = \#\{Y_i = y\}/n, \tag{4.5}$$

with corresponding fdr value $\overline{\text{fdr}}(y) = p_0 f_0(y)/\tilde{f}(y)$. The averaging theorem (4.4) then gives

$$\overline{\text{Fdr}}(\mathcal{Y}) = \sum_{Y_i \in \mathcal{Y}} \overline{\text{fdr}}(Y_i)/\#\{Y_i \in \mathcal{Y}\}, \tag{4.6}$$

so that $\overline{\text{Fdr}}(y)$ in (3.8) equals the average of $\overline{\text{fdr}}(Y_i)$ for $Y_i \leq y$. We can restate the equivalence theorem to say that the Benjamini-Hochberg upper limit y_α is the maximum value y such that the average of $\overline{\text{fdr}}(Y_i)$ for $Y_i \leq y$ is no greater than α .

The estimator $\overline{\text{fdr}}(y)$ can be highly variable, even with n very large. Given a smoothed, less variable estimate $\mathbf{fdr}(y)$ as in Figure 2, we still might want to adjust its global average to match $\overline{\text{Fdr}}(y_\alpha)$, by replacing $\mathbf{fdr}(y)$ with

$$c\mathbf{fdr}(y) \tag{4.7}$$

where c is α divided by $\sum_{Y_i \in \mathcal{Y}} \mathbf{fdr}(Y_i)/\#\{Y_i \in \mathcal{Y}\}$. In this way we obtain a global rejection region \mathcal{Y}_α from the Benjamini-Hochberg algorithm with guaranteed FDR control, along with compatible local fdr estimates that differentiate error probabilities within \mathcal{Y}_α . Notice that $\mathbf{fdr}(y)$ in (4.7) is only required for $y \in \mathcal{Y}_\alpha$ so even a rough guess of $\hat{f}(y)$'s tail behavior can be used to approximate $\mathbf{fdr}(y)$.

Conservative Estimation Property

The empirical estimate of the Bayesian False Discovery Rate $\text{Fdr}(\mathcal{Y})$, (4.1), is

$$\overline{\text{Fdr}}(\mathcal{Y}) = p_0 F_0(\mathcal{Y})/\bar{F}(\mathcal{Y}), \tag{4.8}$$

where

$$F_0(\mathcal{Y}) = \sum_{\mathcal{Y}} f_0(y) \quad \text{and} \quad \bar{F}(\mathcal{Y}) = N(\mathcal{Y})/n, \tag{4.9}$$

$N(\mathcal{Y}) \equiv \#\{Y_i \in \mathcal{Y}\}$. We will show that $\overline{\text{Fdr}}(\mathcal{Y})$ is biased upward for estimating the actual FDR, in a strong sense described next.

Let $N_1(\mathcal{Y})$ and $N_0(\mathcal{Y})$ indicate the number of ‘‘Different’’ and ‘‘Not Different’’ genes with $Y_i \in \mathcal{Y}$, so $N(\mathcal{Y}) = N_0(\mathcal{Y}) + N_1(\mathcal{Y})$, and define

$$\phi(\mathcal{Y}) = N_0(\mathcal{Y})/N(\mathcal{Y}); \tag{4.10}$$

$\phi(\mathcal{Y})$ is the actual proportion of false detections if we reject all null hypotheses having $Y_i \in \mathcal{Y}$, while its expectation is Benjamini and Hochberg's definition (3.2), $\overline{\text{FDR}}(\mathcal{Y})$. The estimate $\overline{\text{Fdr}}(\mathcal{Y})$, (4.8), amounts to substituting the expectation

$$e_0(\mathcal{Y}) \equiv E_{f_0} \{N_0(\mathcal{Y})\} = np_0 F_0(\mathcal{Y}) \quad (4.11)$$

for the unobservable numerator $N_0(\mathcal{Y})$ in (4.10),

$$\overline{\text{Fdr}}(\mathcal{Y}) = e_0(\mathcal{Y})/N(\mathcal{Y}). \quad (4.12)$$

Conservative Bias Theorem: The empirical Bayes FDR $\overline{\text{Fdr}}(\mathcal{Y})$ is biased upward as an estimator of the frequentist FDR, $\text{FDR}(\mathcal{Y})$, for the rule that rejects all H_i having $Y_i \in \mathcal{Y}$, (3.2).

The proof is given in Efron, Storey, and Tibshirani [2001A] See Remark E, and also Theorem 2 of Storey [2001B].

A crucial assumption for empirical Bayes estimates like those in Figure 2 is that we can estimate the expected number of true null hypotheses $N_0(\mathcal{Y})$ among those genes having Y_i in a region of interest \mathcal{Y} . To this end we used $e_0(\mathcal{Y})$, (4.11), or $e_0(y) = np_0 f_0(y)$ for the local fdr. Overestimates of $E\{N_0(\mathcal{Y})\}$, by taking $p_0 = 1$ for instance, increase the conservative bias. More aggressive empirical Bayes estimators such as the dotted curve in Figure 2 put more strain on accurately estimating $E\{N_0(\mathcal{Y})\}$.

The conservative bias theorem applies to a fixed choice of \mathcal{Y} , whereas the original FDR algorithm (3.3), (3.4) selects the rejection set \mathcal{Y}_α adaptively, in a "greedy" way that might seem to generate an anticonservative bias. However the sophisticated calculations of Benjamini and Hochberg [1995] and Benjamini and Yekutieli [2001] show it is still true that $E\{\phi(\mathcal{Y}_\alpha)\} \leq \alpha$, (3.5).

Exchangeability and Prior Beliefs

Empirical Bayes estimates like those in Figure 2 tacitly assume some form of exchangeability of prior beliefs among the genes. This section examines the exchangeability assumption, also discussing what happens when we want to incorporate nonexchangeable prior information.

As an example consider the value $y = 84$ on the x axis of Figure 1; $N(y) = 36$ of the genes have Wilcoxon statistic $Y_i = 84$, versus an expected number of about 7 "Not Different" genes if we set $p_0 = \hat{p}_{0,\max} = 0.67$,

$$e_0(y) = E\{N_0(y)\} = np_0 f_0(y) = 7.05. \quad (4.13)$$

This gives an estimate of $p_0(y) = \text{fdr}(y) = \text{Prob}\{\text{Not Different} \mid Y = y\}$,

$$\bar{p}_0(y) = \overline{\text{fdr}}(y) = \frac{7.05}{36} = 0.196, \quad (4.14)$$

as in (4.12) with $\mathcal{Y} = y$, or (2.3) with $f(y)$ estimated by $\bar{f}(y) = 36/n$.

The exchangeability assumption is transparent in this case: we expect about 7 of the 36 genes with $Y_i = 36$ to be "Not Different," and assign a posteriori probability $7/36$ to all 36. Notice that exchangeability is required only among the 36 genes, not among all 3,226. In this sense, the local fdr estimate relies less than global estimates

like (3.8) on exchangeability. (The equivalence theorem suggests exchangeability assumptions also lurking in the Benjamini-Hochberg procedure, in the way that all of the genes in \mathcal{R}_α are considered equally significant.)

In place of $e_0(y)$, (4.14), we would usually prefer the more relevant conditional expectation

$$e_0^N(y) \equiv E_0\{N_0(y)|N(y)\} = N(y)p_0(y). \tag{4.15}$$

Replacing $p_0(y)$ with $\hat{p}_0(y) = p_0f_0(y)/\hat{f}(y)$ produces the estimate

$$\hat{e}_0^N(y) = N(y)p_0f_0(y)/\hat{f}(y). \tag{4.16}$$

The empirical density $\hat{f} = \hat{f}(y) = N(y)/n$ makes (4.16) identical to (4.13), but smoothed estimates $\hat{f}(y)$ give different results. The dashed curve in Figure 1 has $\hat{p}_0(y) = p_0\hat{f}_0(y)/F(y) = 0.227$ and

$$\hat{e}_0^N(y) = 36 \cdot 0.227 = 8.18. \tag{4.17}$$

The exchangeability argument still applies, now assigning non-significance probability $8.18/36 = 0.227 = \hat{p}_0(y)$ to each of the 36 genes.

Suppose now that we have varying a priori beliefs for the genes, with prior probabilities

$$p_{0i} = \text{Prob}\{\text{gene}_i \text{ Not Different}\} \tag{4.18}$$

replacing the constant value p_0 in (2.1). Let p_0 be the average of p_{0i} over the genes, $p_1 = 1 - p_0$, and set

$$f(y) = p_0f_0(y) + p_1f_1(y) \tag{4.19}$$

as in (2.2). Defining $p_0(y) \equiv p_0f_0(y)/f(y)$, Bayes theorem and a little algebra yields an expression for $p_{0i}(y) \equiv \text{Prob}\{\text{gene}_i \text{ Not Different} | Y_i = y\}$:

$$p_{0i}(y) = p_0(y) \frac{r_i}{1 - (1 - r_i)p_0(y)} \quad \text{where} \quad r_i = \frac{p_{0i}}{1 - p_{0i}} \frac{p_0}{1 - p_0} \tag{4.20}$$

Given prior probabilities p_{0i} , perhaps obtained from a previous experiment, we could substitute $\hat{p}_0(y) = p_0f_0(y)/\hat{f}(y)$ into (4.21) to obtain updated estimates $\hat{p}_{0i}(y)$. Here $\hat{f}(y)$ would be estimated by fitting the observed counts as in Figure 1, the justification being that $nf(y) = E\{N(y)\}$ as before. In practice we might have only fragmentary prior information, perhaps a list of a few dozen genes that the researchers believe particularly likely to be important. For example, if one of the 36 genes with $Y_i = 84$ was on the list, we might take $r_i = 0.50$, indicating it was roughly half as likely a priori to be Not Different, and modify $\hat{p}_0(y) = 0.227$, (4.17), to

$$\hat{p}_{0i}(y) = 0.227 \frac{0.5}{1 - 0.5 \cdot 0.227} = 0.128 \tag{4.21}$$

THREE-WAY COMPARISON

The breast cancer data set of Table I comprises three groups, BRCA1, BRCA2, and Sporadic, but so far our examples have only compared BRCA1 with BRCA2. This section makes the three-way comparison, using the same simple empirical Bayes

model as before but now applied to a higher-dimensional summary statistic “ Y_i .” Multiway comparisons illustrate an advantage of our local empirical Bayes approach, but also show its limitations.

Each gene is represented by 22 microarray readings, as in Table I, 7 for BRCA1, 8 for BRCA2, and 7 for Sporadic. After ranking the 22 numbers, gene i 's summary statistic was taken to be the 3-vector.

$$Y_i = (\text{BRCA1 rank sum, BRCA2 rank sum, Sporadic rank sum})/253; \quad (5.1)$$

253 is the total rank sum so Y_i is a point in the simplex

$$\mathcal{S} = \left\{ Y : Y(j) \geq 0 \text{ and } \sum_{j=1}^3 Y(j) = 1 \right\}. \quad (5.2)$$

We have $n=3,226$ such points, one for each gene. The Y_i 's are essentially two-dimensional, because the first two components determine the third, which simplifies the actual numerical calculations.

The empirical Bayes model (2.1) still is applicable, with “Not Different” now meaning that a gene has the same expression score distribution for all three tumor classes. Bayes rule still applies as stated in (2.2), (2.3). Simulation was used to approximate the null density $f_0(y)$, yielding an estimate of $p_1(y) = \text{Prob}\{\text{Different} \mid Y_i = y\}$, as described in The Three-Way Comparison in the Remarks section.

Figure 4 shows smoothed contours of $\hat{p}_1(y)$ plotted in \mathcal{S} . The plot is in barycentric coordinates, meaning that the triangular region \mathcal{S} has been laid flat on the 2-dimensional page, preserving the original 3-dimensional geometry. Because (5.1) deals with rank vectors, the points Y_i are constrained to lie within the indicated hexagon surrounding the central value $(1/3, 1/3, 1/3)$. The corners of the triangle, which are outside the range of the plot, are indicated by the “OVEREXPRESSED” labels. For example the corner $(1, 0, 0)$ lies beyond the edge of the hexagon labeled “BRCA1 OVEREXPRESSED”. Points Y_i lying on that edge would correspond to genes where the 7 BRCA1 expression levels exceed the other 15.

Figure 4 displays a striking feature: the differences between BRCA1 and BRCA2 are sharper than the differences between Sporadic and either of the BRCA's. This is clearest in the contours for $\hat{p}_1(y) = 0.90$, labeled “9.” These are vertically oriented and not closed at the top or bottom of the hexagon, indicating that high or low Sporadic scores are *not* indicative of genuine expression differences. To state things phenomenologically, genes that were BRCA2 overexpressed tended to have BRCA1 underexpression but an intermediate expression level for Sporadic, and vice versa for genes with BRCA2 underexpressed. There were no genes for which we can be reasonably certain that both BRCA's were overexpressed or both underexpressed. It is as if the BRCA1 and BRCA2 mutations had diverged in opposite directions from a baseline Sporadic type.

The three-way comparison of Figure 4 points out some strengths and limitations of the nonparametric empirical Bayes model (2.1). A strength is the local nature of $p_0(y)$ and $p_1(y)$ in (2.3). These depend only on the density ratio $f_0(y)/f(y)$ at y , not on an ordering of the Y space, which is why we are able to deal with multi-dimensional Y_i vectors such as (5.1). The original FDR algorithm (3.1)–(3.4) is based on P values, implying on ordering of outcomes and less

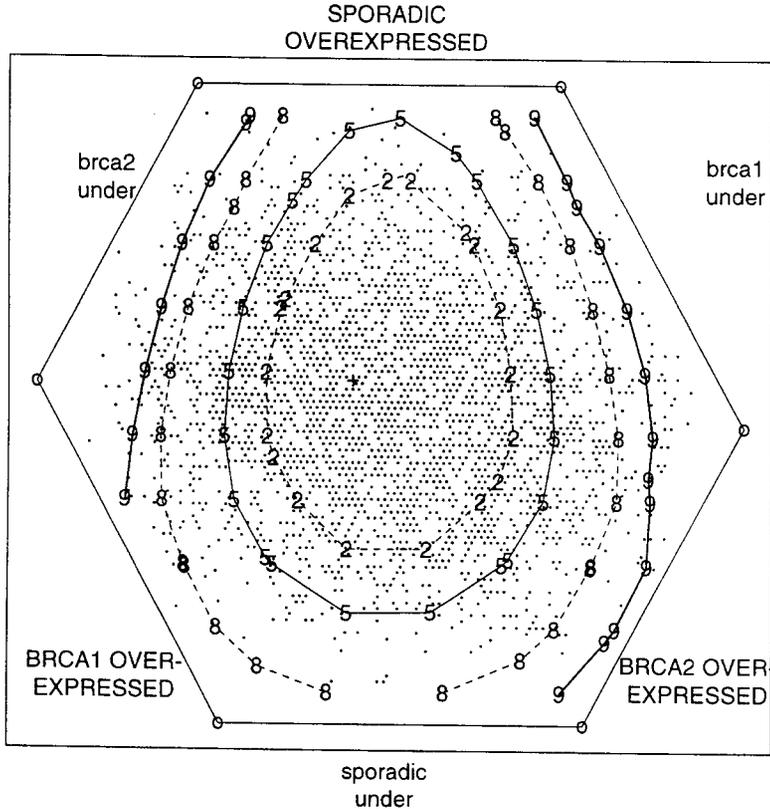


Fig. 4. Three-way comparison of the breast cancer microarray data; contours of $\hat{p}_1(y) = \text{Prob}\{\text{Different } Y_i = y\}$; “9” shows $\hat{p}_1(y) = 0.90$ etc. The contours are vertically oriented, indicating stronger expression differences between BRCA1 and BRCA2 tumors than between Sporadic and either BRCA. Hexagonal boundary indicates feasible region for rank-sum vectors Y_i , (5.1); Points are the 3,226 Y_i vectors “+” is center $(1/3, 1/3, 1/3)$ of simplex \mathcal{S} , (5.2). The three corners of \mathcal{S} lie outside the range of this figure, beyond the “OVEREXPRESSED” legends.

straightforward applications to multiway comparisons. On the other hand, an inference of “Different” is less definitive for multiway comparisons. In the two-way comparison of Figure 2, genes that were significantly Different fell into two clear categories: “Different with BRCA2 expression greater than BRCA1” on the right, and the reverse on the left. Things are less clearcut in Figure 4. Seventy-one of the 3,226 points fall beyond the 0.90 contours, having posterior probability greater than 0.90 of being Different. These are located toward the right or left extremes of the hexagon, with right again indicating BRCA2 expression greater than BRCA1.

However, the status of the Sporadic response for these points is less clear, the choices “BRCA1 < Sporadic < BRCA2,” “BRCA1 < BRCA2 < Sporadic,” etc. remaining ambiguous. Further information is available, by separately examining versions of Figure 2 that apply to the Sporadic-BRCA1 comparison and the Sporadic-BRCA2 comparison, but this tactic was only moderately helpful here.

REMARKS**Data Adjustments**

Processing differences, for example in the treatment of the green-dyed background reference material, can easily produce systematic errors in the readings on any one microarray, making some “brighter” than others. Hedenfalk et al. [2001] adjusted their raw optical measurements for a variety of such factors. We made a final adjustment: each microarray’s data, that is, each column of the $3,226 \times 22$ data matrix, was linearly transformed to have mean 0 and variance 1. Doing so nullifies plate effects, at the expense of possibly reducing the magnitude of genuine expression differences.

Alternatively we might have adjusted each microarray’s mean to its group average, (BRCA1, BRCA2, or Sporadic) rather than to zero. Doing so shifts $\hat{p}(y)$ in Figure 1 roughly 3 units rightward. Making no adjustment at all gave results more like Figure 1. A t-test comparing the 7 BRCA1 plate averages with the 8 BRCA2 averages indicated no systematic differences, and in this case we preferred adjusting all means to zero. We also tried an even more conservative approach, replacing each column of the data matrix with its normal scores vector, but this gave almost the same results.

Continuous Cases

Instead of the discrete Wilcoxon rank-sum statistic (1.1), we might have taken Y_i to be the two-sample t-statistic. Doing so produced results very much like Figure 1, with the solid curve $f_0(y)$ now the standard t density, 13 degrees of freedom. As in Figure 1, the smooth parametric density $\hat{f}(y)$ fit to the 3,226 Y_i s was substantially wider than $f_0(y)$. The equivalent of the $p_0 = 0$ curve in Figure 2 yielded 50 genes having $p_1(Y_i) \geq 0.90$; $\hat{p}_{0,\max} = 0.66$ in (2.5).

Estimating $f_0(y)$

It isn’t clear that the t_{13} density is the correct choice for $f_0(y)$ in the Continuous Cases Remark. Microarray data structures allow us to estimate $f_0(y)$ by permutation methods rather than just accepting the normal-theory answer. Permuting the 15 BRCA1, BRCA2 plates and recalculating the Y statistics gives a direct estimate of f_0 . It can be shown that the permutations should be as *balanced* as possible. For example, if there were 8 plates in each group, each permutation should transfer 4 plates from group to group. Unbalanced permutations add a spurious component of variance to the estimation of $f_0(y)$, arising from those genes in the genuinely “Different” class.

Twenty independent almost balanced permutations were used to estimate $f_0(y)$ in the context of the Continuous Cases section. The resulting $20 \cdot 3,226$ Y s had a distribution that was slightly shorter-tailed than t_{13} . Using this estimate of f_0 , the equivalent of the solid curve in Figure 2 gave 112 genes having $p_1(Y_i) \geq 0.90$, about the same as in the Wilcoxon analysis.

The comparative experiment discussed in Efron et al. [2001] had only four plates for each of the two treatments. There it proved more efficient to add a constant “ a_0 ” to the denominator of the usual two-sample t-statistic when computing the gene

score Y_i . (“More efficient” was defined in terms of the number of genes with $p_1(Y_i) \geq 0.90$.) In this case, permutation methods were essential to the estimation of $f_0(y)$.

The Three-Way Comparison

The contours in Figure 4 were computed using logistic regression: $10 \cdot 3,226$ vectors y_i were generated by randomly permuting the integers $1, 2, \dots, 22$, partitioning them into groups of 7, 8, and 7, and applying definition (5.1). The 3,226 actual vectors Y_i and the 32,260 vectors y_i were plotted in the simplex \mathcal{S} . Thinking of the Y_i s as Successes and the y_i s as Failures, a logistic regression was run to estimate the probability of success, say $\hat{\pi}(Y)$, as a mixed quadratic function of the coordinates of the point Y in \mathcal{S} . Finally, $p_1(Y) = \text{Prob}\{\text{Different} \mid Y\}$ was estimated to be

$$\hat{p}_1(Y) = 1 - p_0 \frac{1 - \hat{\pi}(Y)}{10 \cdot \hat{\pi}(Y)}, \tag{6.1}$$

with p_0 set equal to 1 in Figure 4. [Formula (6.1) follows from the ratio of Successes to Failures, $\pi(Y) = f(Y)/(f(Y) + 10 \cdot f_0(Y))$.] Notice that the *shape* of the contours does not depend on p_0 , while the probability level assigned to the curves does, with $\hat{p}_0(y) = 1 - \hat{p}_1(y)$ being directly proportional to p_0 .

True and Untrue Null Hypotheses

A pleasant surprise of the original FDR algorithm (3.3–3.5) was that its proof required no probabilistic assumptions about the untrue null hypotheses among H_1, H_2, \dots, H_n . Only the P values for the true H_i needed to be independent uniform variates. The same phenomenon occurs for Bayesian FDR: the Conservative Bias Theorem (4.13) holds true *conditionally* on $N_1(\mathcal{Y})$, the number of “Different” genes having $Y_i = \mathcal{Y}$, Different equaling untrue in our terminology.

In fact, as pointed out in (4.11, 4.12), the only quantity required for the estimation of $\text{Fdr}(\mathcal{Y})$ is $e_0(\mathcal{Y}) = E_{f_0}\{N_0(\mathcal{Y})\}$, the expected number of “true” Y_i in \mathcal{Y} . Only $f_0(y)$ plays a computational role in the Bayesian assumptions (2.1–2.2), whereas $f_1(y)$ is functionally unimportant. However, this does not diminish the point of the Exchangeability and Prior Beliefs section, that the *interpretation* of the FDR results, Bayesian or frequentist, requires some form of exchangeability for application to any particular gene.

Prediction

Hedenfalk et al. [2001] were interested in the prediction problem: given a new unclassified microarray plate, how should it be assigned to one of the three categories BRCA1, BRCA2, or Sporadic? The empirical Bayes methodology of this article bears on the prediction problem.

Consider the situation of Figure 1 where we are only interested in the two categories BRCA1 versus BRCA2. Let \mathbf{X} be the 3,226 vector of data from a new plate, and suppose we want to classify it as the basis of a linear discriminant function

$Q = \sum w_i X_i$. It is intuitively obvious that only genes in the Different class should receive nonzero weights.

Let $\{x_{ij}\}$, $i = 1, 2, \dots, 3, 226$, $j = 1, 2, \dots, 15$, represent the Hedenfalk data. It can be deduced, using further empirical Bayes analyses, that the x_{ij} s are roughly uncorrelated and have constant variance across the different genes. Without going into details, it can then be shown that a reasonable estimate for the ideal discriminant function is

$$\hat{Q} = \sum \hat{w}_i X_i \quad \text{where} \quad \hat{w}_i = \hat{p}_1(Y_i) \cdot (\bar{x}_{i2} - \bar{x}_{i1}). \quad (6.2)$$

Here \bar{x}_{i1} and \bar{x}_{i2} are the means for gene i s BRCA1 and BRCA2 expression levels, whereas $\hat{p}_1(Y_i)$ is the estimate (2.4) for $\text{Prob}\{\text{gene}_i \text{ Different} \mid Y_i\}$. Our current work concerns the efficacy of (6.2) in practical prediction problems.

REFERENCES

- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B* 57:289–300.
- Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29:1165–8.
- Dudoit S, Yang Y, Callow M, Speed T. 2000. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Technical Report, Department of Statistics, University of California, Berkeley.
- Efron B, Storey J, Tibshirani R. 2001A. Microarrays, empirical Bayes methods, and false discovery rates. Stanford Technical Report, July 2001.
- Efron B, Tibshirani R, Storey JD, Tusher V. 2001B. Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 96:1151–60.
- Genovese C, Wasserman L. 2001. Operating characteristics and extensions of the FDR procedure. Technical report, Department of Statistics, Carnegie Mellon University.
- Hedenfalk I, Duggen D, Chen Y, et al. 2001. Gene expression profiles in hereditary breast cancer. *N Engl J Med* 344:539–48.
- Newton M, Kendziorski C, Richmond C, Blatter F, Tsui K. 2000. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comp Biol* 8:37–52.
- Robbins H. 1956. An empirical Bayes approach to statistics. Proceedings of the Third Berkeley Symposium 1: Berkeley: University of California Press, 157–63.
- Simes R. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73:751–4.
- Storey J. 2001A. The false discovery rate: a Bayesian interpretation and the q -value. Stanford Technical Report, jstorey@stat.stanford.edu.
- Storey J. 2001B. A new approach to false discovery rates and multiple hypotheses testing. Stanford Technical Report, jstorey@stat.stanford.edu.
- Tusher V, Tibshirani R, Chu G. 2000. Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc Natl Acad Sci USA*.