MICROARRAYS EMPIRICAL BAYES METHODS, AND FALSE DISCOVERY RATES

by

BRADLEY EFRON
JOHN D. STOREY
ROBERT TIBSHIRANI

Technical Report No. 2001-23B/217
September 2001

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305

MICROARRAYS EMPIRICAL BAYES METHODS, AND FALSE DISCOVERY RATES

by

Bradley Efron
Department of Statistics
Stanford University

John D. Storey
Department of Statistics
Stanford University

Robert Tibshirani
Department of Statistics
Stanford University

Technical Report No. 2001-23B/217
September 2001

This research was supported in part by National Institutes of Health grants 2R01 CA72028 and 2R01 CA59039, National Science Foundation grants DMS-9972405 and DMS-0072360

Department of Statistics
Sequoia Hall
STANFORD UNIVERSITY
Stanford, California 94305

http://www-stat.stanford.edu

Microarrays Empirical Bayes Methods, and False Discovery Rates

Bradley Efron *,
John D. Storey, †
Robert Tibshirani ‡

July 24, 2001

Abstract

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.

*Department of Statistics and Division of Biostatistics, Stanford University, Stanford CA 94305; brad@stat.stanford.edu

†Department of Statistics, Stanford University, Stanford CA 94305; jstorey@stat.stanford.edu

‡Division of Biostatistics and Department of Statistics, Stanford University, Stanford CA 94305; tibs@stat.stanford.edu

1 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 employed here, a rather typical microarray experiment, we will have more than three thousand Wilcoxon two-sample tests to consider at once.

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. (2000, 2001). The two approaches are closely related and can be used to support each other, which is the principal point of this paper.

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 3226 genes. Table 1 shows a small portion of the resulting 3226 $\times$ 22 data matrix.

Here is a schematic description of the genetic technology behind the numbers in Table 1. The known DNA base sequences for each of the 3226 genes were printed at known positions on the microarray plates. (There were actually 5361 genes to begin with, only 3226 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 employed as a control. The numbers in Table 1 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 Remark A of Section 6.

Table 1: A 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 3226 genes on 22 microarray plates; 7 from women with BRCA1 mutation, 8 BRCA2, 7 Sporadic (neither). Tabled values are adjusted log(red/green) ratios from spotted cDNA microarrays.

	BRCA1				**BRCA2**				**Sporadic**			
	1	2	$\cdots$	7	1	2	$\cdots$	8	1	2	$\cdots$	7
gene1	-1.29	-1.41	$\cdots$	-0.55	-0.70	1.33	$\cdots$	1.14	-0.44	0.26	$\cdots$	-0.23
gene2	2.03	0.58	$\cdots$	-0.12	0.23	-0.91	$\cdots$	-0.39	0.70	-1.55	$\cdots$	2.17
gene3	0.32	-0.44	$\cdots$	1.25	0.53	-0.96	$\cdots$	-0.51	-1.26	-0.74	$\cdots$	-0.64
gene4	-1.31	-0.98	$\cdots$	0.24	-0.24	0.28	$\cdots$	2.13	0.32	0.42	$\cdots$	-0.65
gene5	-0.66	-0.07	$\cdots$	1.22	-0.41	-0.88	$\cdots$	-0.83	0.25	-0.97	$\cdots$	-0.21

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 1. For this analysis each gene's data was summarized by its Wilcoxon statistic: the 15 expression levels for gene i, 7 BRCA1 and 8 BRCA2, were ranked, giving the rank sum statistic

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

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,

$$36 \leq Y_i \leq 92. \tag{1.2}$$

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 *down-regulation* or *up-regulation.*)

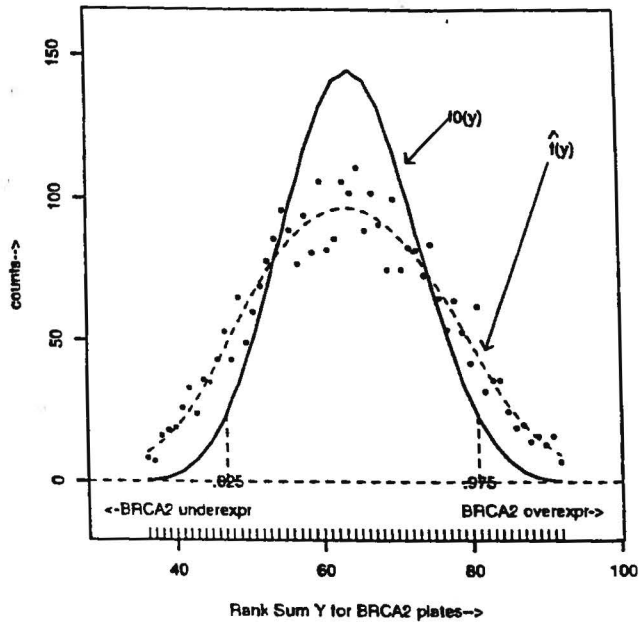

Figure 1: Rank sum statistics comparing BRCA1 vs BRCA2 for the 3226 genes; points are actual counts, solid curve show expected counts under the null hypothesis of no activity differences. Dashed curve is a Poisson regression fit to the actual counts, as explained in Section 2.

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 $1, 2, \ldots, 15$ (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 3226 genes can we confidently label as differently active?" The naive answer would be to run 3226 separate Wilcoxon tests. 614 of the Y_i's lie either below

the .025 point for a standard Wilcoxon(7,8) distribution or above its .975 point. This would give a reasonable criteria for declaring any single prechosen gene differently active, but it leads to an expected 161 false declarations if none of the 3226 genes are actually different.

Efron, Tibshirani et al. (2000) 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 Section 4.

Section 5 returns to the full data set of Table 1, using the empirical Bayes methodology to make a three-way activity comparison between BRCA1, BRCA2, and Sporadic tumors. We close in Section 6 with some notes and remarks.

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. (2001).

2 Empirical Bayes Inferences

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 $p_o = 1 - p_1$, with corresponding prior densities $f_1(y)$ and $f_o(y)$ for the summary statistic Y,

$$\begin{array}{rcll} p_1 & = & \text{Prob\{Different\}} & f_1(y) \text{ density of } Y_i \text{ if gene}_i \text{ "Different"} \\ p_o & = & \text{Prob\{Not Different\}} & f_o(y) \text{ density of } Y_i \text{ if gene}_i \text{ "Not Different".} \end{array} \tag{2.1}$$

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

$$f(y) = p_o f_o(y) + p_1 f_1(y). \tag{2.2}$$

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

$$p_1(y) \equiv \text{Prob\{Different}|Y_i = y\} = 1 - p_o f_o(y)/f(y)$$

and (2.3)

$$p_o(y) \equiv \text{Prob\{Not Different}|Y_i = y\} = p_o f_o(y)/f(y)$$

Full Bayesian analysis would require prior specification of $p_o, p_1, f_o(y)$, 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 Robbins 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: $f_o(y)$ here is the discrete density for a Wilcoxon(7,8) variate, the solid curve in Figure 1 divided by 3226; $f(y)$ has been estimated by a Poisson regression fit to the Y counts. (Specifically by modelling $f(y)$ as a natural spline having 5 degrees of freedom and with offset $\log(f_o(y))$, giving $\widehat{f}(y)$ 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),

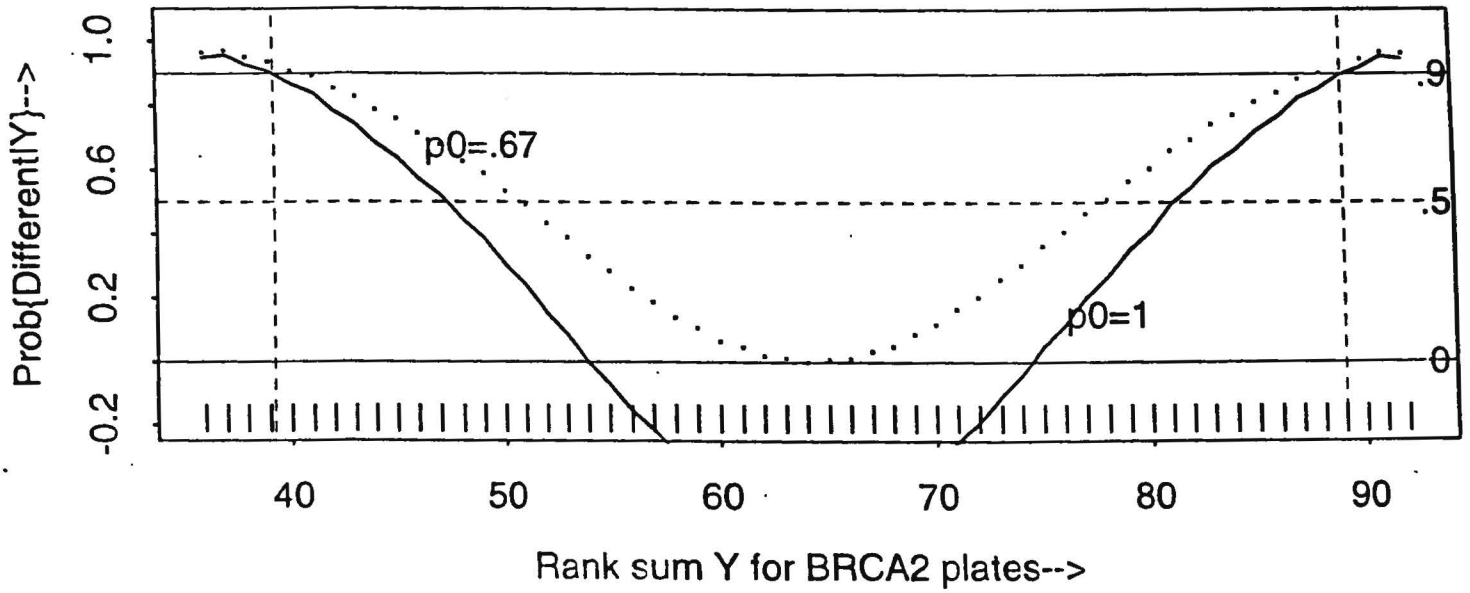

Figure 2: Empirical Bayes estimates (2.3) of $p_1(y) = \text{Prob}\{\text{Different}|Y_i = y\}$ for the comparison of BRCA1 and BRCA2 in Figure 1. *Solid curve*: assuming prior probability p_o of "Not Different" is 1; *Dotted curve*: assuming $p_o = .67$, the largest value of p_o that makes $p_1(y)$ everywhere nonnegative.

$$\widehat{p}_1(y) = 1 - p_o f_o(y)/\widehat{f}(y). \tag{2.4}$$

The prior "Not Different" probability p_o is unidentifiable without strong parametric assumptions, such as normality, on $f_o(y)$ and $f(y)$. However the most conservative possible choice, $p_o = 1$, the choice that minimizes the probability of detecting "Different", still gives interesting results. The solid curve in Figure 2 is $\widehat{p}_1(y)$ for $p_o = 1$ in (2.4). We see that for $p_o = 1$, genes with $Y_i \le 39$ or $Y_i \ge 89$ have $\widehat{\text{Prob}}\{\text{Different}|Y\}$ exceeding 0.90. There are 101 such genes, 49 on the left and 52 on the right.

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

$$p_o \le \widehat{p}_{o,\max} = \min_y\{\widehat{f}(y)/f_o(y)\}. \tag{2.5}$$

The dotted curve in Figure 2 indicates $\widehat{p}_1(y)$ for $p_o = \widehat{p}_{o,\max} = .67$. This raises $\widehat{p}_1(y)$ somewhat in the tails, so that now $\widehat{\text{Prob}}\{\text{Different}|y\}$ exceeds 0.90 for $Y_i \leq 40$ or $Y_i \geq 88$, a total of 134 genes. (Remark F of Efron et al. (2001) suggests a more stable estimate of $p_{o,\max}$.) We will see in Section 3 that the ambiguity in p_o 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 Sections 3 and 4, where (2.4) is related to the frequentist False Discovery Rate algorithm of Banjamini and Hochberg (1995).

3 Connection With False Discovery Rates

The empirical Bayes analysis of Section 2 is closely related to Benjamini and Hochberg's theory of False Discovery Rates (1995). We begin with a brief review of the FDR algorithm. Suppose one wishes to simultaneously test n null hypotheses $H_1, H_2, \ldots, H_n$ on the basis of independent test statistics $Y_1, Y_2, \ldots, Y_n$. From the Y_i we calculate corresponding p-values P_i, denoting the ordered values as

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

$P_{(1)}$ 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 $P_i \leq .01$". Following work by Simes (1986), Bejamini and Hochberg defined the False Discovery Rate 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 proved a useful algorithm for controlling the FDR below a preset value α: let

$$i_\alpha = \underset{i}{\text{argmax}}\left\{P_{(i)} \leq \frac{i}{n}\,\frac{\alpha}{p_o}\right\} \qquad [p_o \equiv \text{proportion of true } H_i]. \tag{3.3}$$

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, as in Benjamini and Wei (1999), but we will concentrate on (3.3, 3.4).

In the context of Figure 1, $n = 3226$ and $H_i = \{\text{gene}_i \text{ Not Different}\}$. Notice that p_o in (2.1) is the expected proportion of true H_i, nearly the same as its definition in (3.3). The 1995 paper took $p_o = 1$, 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 Hochberg (2000), they consider estimating p_o, see also Storey (2001A,B). Empirical Bayes considerations, as in (2.5) and Remark F of Efron et al. (2001), give intuitively appealing bounds for p_o.

Figure 3 applies the FDR-controlling algorithm to the comparison of BRCA1 with BRCA2, using $\alpha = 0.10$ and $p_o = 1.0$. 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 $\mathcal{R}_{.10}$ 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 degrees.)

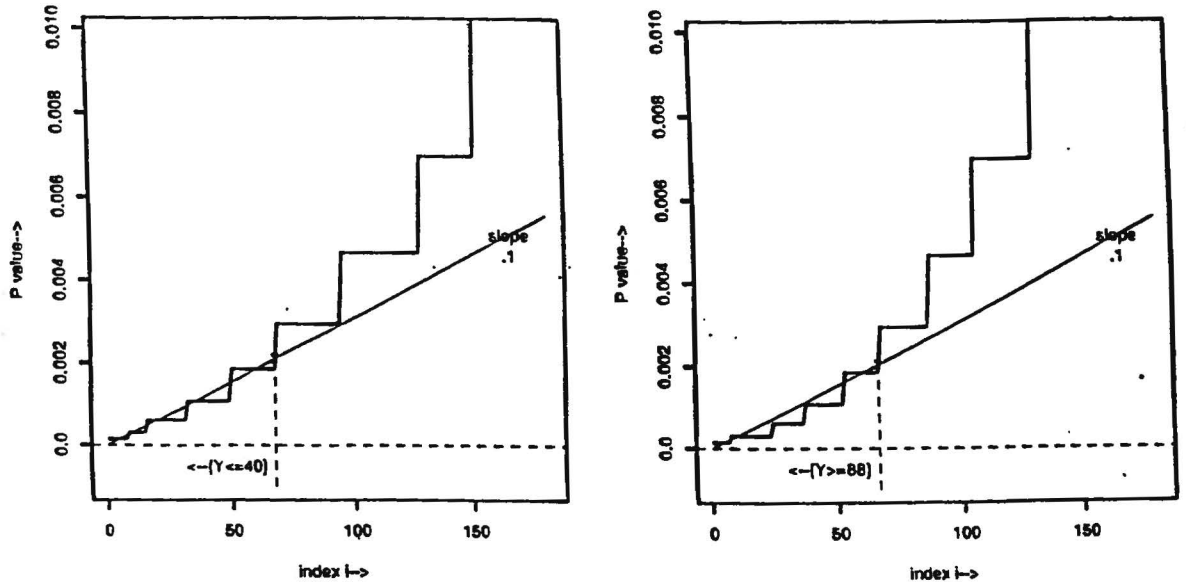

Figure 3: Application of FDR-controlling algorithm to BRCA1/BRCA2 comparison, $\alpha = .10$, $p_o = 1.0$. *Left Panel*: Step function shows ordered p-values for one-sided Wilcoxon tests that reject for small values of rank sum statistic Y_i; $\mathcal{R}_{.10}$ procedure (3.4) rejects for the 68 genes having $Y_i \leq 40$. *Right Panel*: Same, rejecting for large values of Y_i; $\mathcal{R}_{.10}$ 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 Section 2 follows directly from Bayes theorem. Let $F_o(y)$ and $F(y)$ be the cumulative distribution functions (CDF's) corresponding to $f_o(y)$ in (2.1) and $f(y)$ in (2.2), and define the "Bayesian FDR" for $\{Y \leq y\}$ to be

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

as in (2.3), called the "q-value" in Storey (2001A,B). If we have N_i genes with $Y_i \leq y$ then, starting from (2.1) and assuming independence, the number N_{io} of the N_i from the "Not Different" class will be binomially distributed,

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

and for large N_i 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 (2001).

Now let $\bar{F}(y)$ be the usual empirical cdf of the Y_i's, $\bar{F}(y) = \#\{Y_i \le y\}/n$. The obvious nonparametric estimate for $\mathrm{Fdr}(y)$ is

$$\overline{\mathrm{Fdr}}(y) = p_o F_o(y)/\bar{F}(y). \tag{3.8}$$

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

$$y_\alpha = \max_y\{\overline{\mathrm{Fdr}}(y) \le \alpha\}. \tag{3.9}$$

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

Proof Let $Y_{(i)}$ indicate the ith ordered value of $\{Y_1, Y_2, \ldots, Y_n\}$. Then $\bar{F}(Y_{(i)}) = i/n$ and $F_o(Y_{(i)}) = P_{(i)}$. The constraint $\overline{\mathrm{Fdr}}(y) \le \alpha$ is equivalent to

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

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 $\{Y_i \le y\}$ as large as possible subject to the constraint that the estimated empirical Bayes probability $\mathrm{Prob}\{\text{Not Different}|Y \le y\}$ is no greater than α, than our expected proportion of false rejections is also less than α. This is true for any choice of p_o in the two algorithms and in particular for the conservative choice $p_o = 1$. 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 $Y_1, Y_2, \ldots, Y_n$. 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.

4 The Local False Discovery Rate

What we called the Bayesian False Discovery Rate 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

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

with f_o and f defined as in (2.1), (2.2). In Figure 1 for example we might take $\mathcal{Y} = \{Y_i \le 47 \text{ or } Y_i \ge 81\}$, the .05 (actually .054) two-sided Wilcoxon rejection region. Estimating the denominator in (4.1) by the proportion of Y_i's in $\mathcal{Y}$, $614/3226 = 0.190$, gives

$$\overline{\mathrm{Fdr}}(\mathcal{Y}) = p_o \frac{.054}{.190} = p_o \cdot .284. \tag{4.2}$$

Under the conservative assumption $p_o = 1$, we expect about 28% of the ".05 significant" genes to actually be Not Different, while $p_o = \widehat{p}_{o,\max} = .67$ gives 19%.

4.1 Local FDR

Efron et al. (2001) defined the *local false discovery rate* at point y in the Y-space to be the function $p_o(y)$ in (2.3),

$$\begin{aligned} \text{fdr}(y) &= p_o f_o(y)/f(y) = p_o(y) \\ &= \text{Prob}\{\text{Not Different}|Y_i = y\}. \end{aligned} \tag{4.3}$$

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

Averaging Theorem $$\text{Fdr}(\mathcal{Y}) = E_f\{\text{fdr}(y)|y \in \mathcal{Y}\} \quad . \tag{4.4}$$

Proof $E_f\{\text{fdr}(y)|y \in \mathcal{Y}\} = \int_{\mathcal{Y}}[p_o f_o(y)/f(y)]f(y)/\int_{\mathcal{Y}} f(y) = p_o \text{Prob}_{f_o}\{\mathcal{Y}\}/\text{Prob}_f\{\mathcal{Y}\} = \text{Fdr}(\mathcal{Y})$. In 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 $\mathcal{Y} = \{Y_i \le 40\}$, the FDR-controlling set for $\alpha = 0.10$, $p_o = 1.0$, on the left side of Figure 3. It has overall Bayesian $\overline{\text{Fdr}} = .089$, (3.4), but with estimated local values of $\text{fdr}(y)$ ranging from .04 to .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 Section 4.3.

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 $f_o(y)/f(y)$ in situations where $f_o(y)$ is not theoretically determined, see Remark C of Section 6). For example we needed the Poisson regression estimate $\widehat{f}(y)$ in (2.4) to construct the curves $\widehat{p}_1(y) = 1 - \widehat{\text{fdr}}(y)$ of Figure 2.

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

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

with corresponding fdr value $\overline{\text{fdr}}(y) = p_o f_o(y)/\bar{f}(y)$. 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 $Y_i \le y$. 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 $Y_i \le y$ 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 $\widehat{\text{fdr}}(y)$ as in Figure 2, we still might wish to adjust its global average to match $\overline{\text{Fdr}}(y_\alpha)$, by replacing $\widehat{\text{fdr}}(y)$ with

$$\widetilde{\text{fdr}}(y) = c\widehat{\text{fdr}}(y) \tag{4.7}$$

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

4.2 Conservative Estimation Property

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

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

where

$$F_o(\mathcal{Y}) = \int_{\mathcal{Y}} f_o(y) \quad \text{and} \quad \bar{F}(\mathcal{Y}) = N(\mathcal{Y})/n, \tag{4.9}$$

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

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

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

$\phi(\mathcal{Y})$ 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), $\text{FDR}(\mathcal{Y})$. The estimate $\overline{\text{Fdr}}(\mathcal{Y})$, (4.8), amounts to substituting the expectation

$$e_o(\mathcal{Y}) \equiv E_{f_o}\{N_o(\mathcal{Y})\} = np_o F_o(\mathcal{Y}) \tag{4.11}$$

For the unobservable numerator $N_o(\mathcal{Y})$ in (4.10),

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

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

Proof With $N_1(\mathcal{Y})$ fixed, $\overline{Fdr}(\mathcal{Y})$ and $\phi(\mathcal{Y})$ are respectively convex and concave as functions of $N_o(\mathcal{Y})$, as illustrated in Figure 4. Two applications of Jensen's inequality give

$$E\{\overline{Fdr}(\mathcal{Y})|N_1(\mathcal{Y}) = N_1\} \geq \frac{e_o(\mathcal{Y})}{e_o(\mathcal{Y}) + N_1} \geq E\{\phi(\mathcal{Y})|N_1(\mathcal{Y}) = N_1\} \tag{4.13}$$

This proves the stronger result that conditional on any value of $N_1(\mathcal{Y})$, $\overline{Fdr}(\mathcal{Y})$ is biased upward as an estimator of the expectation of $\phi(\mathcal{Y})$; the theorem follows from $E\{\phi(\mathcal{Y})\} = \text{FDR}(\mathcal{Y})$, unconditionally. 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 $N_o(\mathcal{Y})$ among those genes having Y_i in a region of interest $\mathcal{Y}$. To this end we used $e_o(\mathcal{Y})$, (4.11), or $e_o(y) = np_o f_o(y)$ for the local fdr. Overestimates of

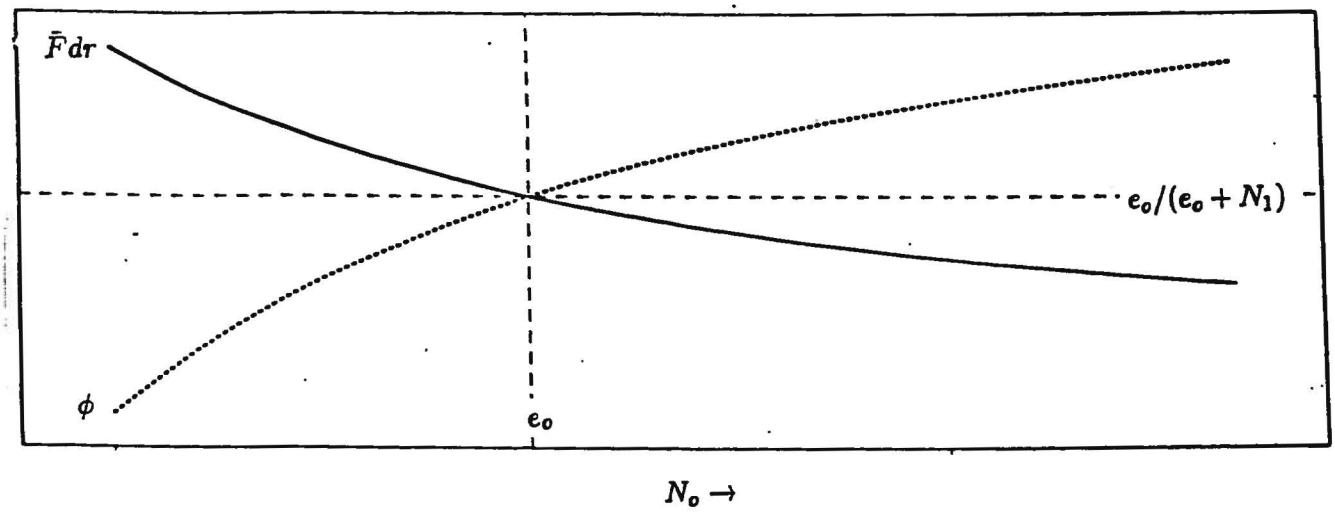

Figure 4: $\overline{Fdr}(\mathcal{Y})$ and $\phi(\mathcal{Y})$ as functions of the unobservable quantity $N_o(\mathcal{Y})$, (4.13).

$E\{N_o(\mathcal{Y})\}$, by taking $p_o = 1$ 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 $E\{N_o(\mathcal{Y})\}$.

The conservative bias theorem applies to a fixed choice of $\mathcal{Y}$, while the original FDR algorithm (3.3), (3.4) selects the rejection set $\mathcal{Y}_\alpha$ 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). Roughly speaking the anticonservative maximization of $\mathcal{Y}_\alpha$ in (3.3) is more than counteracted by the effects at work in Figure 4.

4.3 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 wish to incorporate non-exchangeable 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 $p_o = \widehat{p}_{o,\max} = 0.67$,

$$e_o(y) = E\{N_o(y)\} = np_o f_o(y) = 7.05. \tag{4.14}$$

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

$$\bar{p}_o(y) = \overline{fdr}(y) = \frac{7.05}{36} = .196, \tag{4.15}$$

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 3226. 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 $\mathcal{R}_\alpha$ are considered equally significant.)

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

$$e_o^N(y) \equiv E_o\{N_o(y)|N(y)\} = N(y)p_o(y). \tag{4.16}$$

Replacing $p_o(y)$ with $\widehat{p}_o(y) = p_o f_o(y)/\widehat{f}(y)$ produces the estimate

$$\widehat{e}_o^N(y) = N(y)p_o f_o(y)/\widehat{f}(y). \tag{4.17}$$

The empirical density $\widehat{f} = \bar{f}(y) = N(y)/n$ makes (4.17) identical to (4.14), but smoothed estimates $\widehat{f}(y)$ give different results. The dashed curve in Figure 1 has $\widehat{p}_o(y) = p_o\widehat{f}_o(y)/F(y) = 0.227$ and

$$\widehat{e}_o^N(y) = 36 \cdot 0.227 = 8.18. \tag{4.18}$$

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

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

$$p_{oi} = \text{Prob}\{\text{gene}_i \text{ Not Different}\} \tag{4.19}$$

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

$$f(y) = p_o f_o(y) + p_1 f_1(y) \tag{4.20}$$

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

$$p_{oi}(y) = p_o(y)\frac{r_i}{1-(1-r_i)p_o(y)} \qquad \text{where} \qquad r_i = \frac{p_{oi}}{1-p_{oi}}\Big/\frac{p_o}{1-p_o} \tag{4.21}$$

Given prior probabilities p_{oi}, perhaps obtained from a previous experiment, we could substitute $\widehat{p}_o(y) = p_o f_o(y)/\widehat{f}(y)$ into (4.21) to obtain updated estimates $\widehat{p}_{oi}(y)$. Here $\widehat{f}(y)$ 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 $r_i = .50$, indicating it was roughly half as likely a priori to be Not Different, and modify $\widehat{p}_o(y) = .227$, (4.18), to

$$\widehat{p}_{oi}(y) = .227\frac{.5}{1-.5\cdot.227} = .128 \tag{4.22}$$

5 Three-Way Comparison

The breast cancer data set of Table 1 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". Multi-way 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 1, 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.

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

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

$$\mathcal{S} = \Big\{Y : Y(j) \geq 0 \quad \text{and} \quad \sum_1^3 Y(j) = 1\Big\}. \tag{5.2}$$

We have $n = 3226$ such points, one for each gene. The Y_i's are essentially two-dimensional, since 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 $f_o(y)$, yielding an estimate of $p_1(y) = \text{Prob}\{\text{Different } | Y_i = y\}$, as described in Remark D of Section 6.

Figure 5 shows smoothed contours of $\widehat{p}_1(y)$ 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 5 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 $\widehat{p}_1(y) = .90$, 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 5 points out some strengths and limitations of the non-parametric empirical Bayes model (2.1). A strength is the local nature of $p_o(y)$ and $p_1(y)$ in (2.3). These depend only on the density ratio $f_o(y)/f(y)$ 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 straightforward applications to multi-way comparisons. On the other hand, an inference of "Different" is less definitive for multi-way 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 5. 71 of the 3226 points fall beyond the .90 contours, having posterior probability greater than .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.

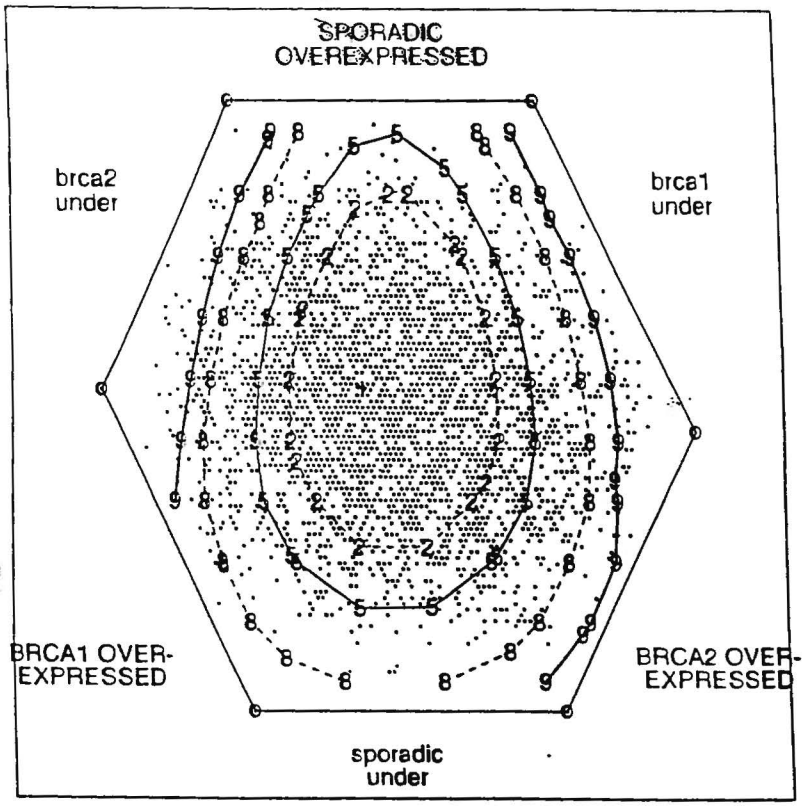

Figure 5: Three-way comparison of the breast cancer microarray data; contours of $\widehat{p}_1(y) =$ Prob{Different $Y_i = y$}; "9" shows $\widehat{p}_1(y) = .90$ etc. The contours are vertically oriented, indicating stronger expression differences between BRCA1 and BRCA2 tumors than between Sporadic and either BRCA. Hexagonal boundary indicates feasible region for rank-sum vectors Y_i, (5.1); Points are the 3226 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.

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.

6 Remarks

A. *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 microarrays data, that is each column of the 3226 × 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 $\widehat{f}(y)$ 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.

B. *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 $f_o(y)$ now the standard t density, 13 degrees of freedom. As in Figure 1, the smooth parametric density $\widehat{f}(y)$ fit to the 3226 Y_i's was substantially wider than $f_o(y)$. The equivalent of the $p_o = 0$ curve in Figure 2 yielded 50 genes having $p_1(Y_i) \geq .90$; $\widehat{p}_{o,\max} = .66$ in (2.5).

C. *Estimating $f_o(y)$* It isn't clear that the t_{13} density is the correct choice for $f_o(y)$ in Remark B. Microarray data structures allow us to estimate $f_o(y)$ 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_o. 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 $f_o(y)$, arising from those genes in the genuinely "Different" class.

20 independent almost balanced permutations were used to estimate $f_o(y)$ in the context of Remark B. The resulting $20 \cdot 3226$ Y's had a distribution that was slightly shorter-tailed than t_{13}. Using this estimate of f_o, the equivalent of the solid curve in Figure 2 gave 112 genes having $p_1(Y_i) \geq .90$, 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_o" 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 $p_1(Y_i) \geq .90$.) In this case permutation methods were essential to the estimation of $f_o(y)$.

D. *The Three-Way Comparison* The contours in Figure 5 were computed using logistic regression: $10 \cdot 3226$ vectors y_i were generated by randomly permuting the integers $1, 2, \ldots, 22$, partitioning them into groups of 7, 8, and 7, and applying definition (5.1). The 3226 actual vectors Y_i and the 32260 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 $\widehat{\pi}(Y)$, 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

$$\widehat{p}_1(Y) = 1 - p_o \frac{1 - \widehat{\pi}(Y)}{10 \cdot \widehat{\pi}(Y)} \,. \tag{6.1}$$

with p_o set equal to 1 in Figure 5. (Formula (6.1) follows from the ratio of Successes to Failures, $\pi(Y) = f(Y)/(f(Y) + 10 \cdot f_o(Y))$.) Notice that the *shape* of the contours does not depend on p_o, while the probability level assigned to the curves does, with $\widehat{p}_o(y) = 1 - \widehat{p}_1(y)$ being directly proportional to p_o.

E. *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 $H_1, H_2, \ldots, H_n$. Only the p-values for the true H_i needed to be independent uniform variates. The same phenomenon occurs for Bayesian False Discovery Rates: 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 $e_o(\mathcal{Y}) = E_{f_o}\{N_o(\mathcal{Y})\}$, the expected number of "true" Y_i in $\mathcal{Y}$. Only $f_o(y)$ plays a computational role in the Bayesian assumptions (2.1-2.2), while $f_1(y)$ is functionally unimportant. However this

doesn't diminish the point of Section 4.3, that the *interpretation* of the FDR results, Bayesian or frequentist, requires some form of exchangeability for application to any particular gene.

F. *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 paper 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 3226 vector of data from a new plate, and suppose we wish to classify it as the basis of a linear discriminant function $Q = \Sigma w_i X_i$. It is intuitively obvious that only genes in the Different class should receive non-zero weights.

Let $\{x_{ij}\}$, $i = 1, 2, \ldots, 3226$, $j = 1, 2, \ldots, 15$, 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

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

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

Acknowledgment This research was supported in part by National Institute of Health grant 2R01 CA72028 and 2R01 CA59039, National Science Foundation grant DMS-9972405 and DMS-0072360.

References

Benjamini, Y. & Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Jour. Royal Stat. Soc., B* **57**, 289-300.

Benjamini, Y. & Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency". To appear *Annals of Statistics.*

Dudoit, S., Yang, Y., Callow, M. & Speed, T. (2000), "Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments". Technical Report, Dept. Statistics, U. Cal. Berkeley.

Efron, B., Tibshirani, R., Storey, J.D., & Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment". Stanford Technical Report #216.

Hedenfalk, I., Duggen, D., Chen, Y., et al. (2001), "Gene Expression Profiles in Hereditary Breast Cancer". *New England Journal of Medicine* **344**, 539-548.

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". To appear, *J. Comp. Biology.*

Robbin, H. (1956), "An empirical Bayes approach to statistics". *Proc. Third Berkeley Symp.* **1**, 157-163, Univ. Calif. Press.

Simes, R. (1986), "An Improved Bonferroni Procedure for Multiple Tests of Significance". *Biometrika* **73**, 751-754.

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, C. (2000), "Significance analysis of microarrays applied to transcriptional responses to ionizing radiation". To appear *Proc. Nat. Acad. Sci.*