in due course decide between these and other methods that will present themselves. This will take a long time to accumulate. While we accumulate such experience, the technology will continue to change, placing in question the usefulness of any except very recent experience.

Both for boostrap and for permutation methods, the distribution is a poor estimator of the population distribution for small sample sizes. Is it possible to build in parametric assumptions, perhaps assuming that the distribution is normal except in the tails, that will reduce the problem of loss of power relative to normal theory methods?

It would be interesting to do the same comparisons under conditions where the discreteness of the permutation distributions is more of an issue, for example with four or six mice per treatment. Also, why not use the permutation distribution to calibrate p-values that are derived from normal theory assumptions, interpolating between the discrete probabilities from the permutation distribution?.

Variation in the denominators of the sample t-statistics, and in the t-statistics themselves, will be more extreme than variation in the "true" unknown variances. Better ways than at present available are needed to use information on the distribution of variances across different genes to improve the crude variance estimates. A complication is that these variances can be, and in these data are, a function of hybridization intensity, of specific print-tip effects, and of order of printing effects. While most of the methods do not change the rank order of the genes (the sequential methods may change the order), changing the variance estimator will change the ranking.

John D. Storey
Department of Statistics
University of California, Berkeley, USA

Yongchao Ge, Sandrine Dudoit, and Terry Speed have written a lucid article on multiple testing in the context of DNA microarray studies. I have greatly benefitted from the interaction I have had with them, and it has greatly influenced my work in this area. Their presentation of the issues is thoughtful and careful, both in this article and in their previous work (Dudoit et al. (2002b)). This particular article will undoubtedly serve as a standard reference for those wanting to become acquainted with the research area.

Ge, Dudoit, and Speed (henceforth abbreviated by GDS), discuss both FWER and FDR. It seems that most microarray experiments will involve conditions where a reasonable fraction of the genes are differentially expressed. In such cases, the FDR is likely the more appropriate quantity. (Of course, one can create examples where FWER would be more applicable.) Therefore, my comments will be limited to false discovery rates and consist of four points. First, I will show that many of the different approaches to FDR they have presented do in fact become equivalent if one views p-value calculations from a "pooling" point of view. Second, I will review some recent results I have completed with Jonathan Taylor and David Siegmund that directly address some of the concerns they raise. Third, I will discuss where I think dependence is an issue in DNA microarray experiments, where it is not an issue, and how this relates to some of the methods they discuss. Fourth, I will argue that q-values provide a good gene-specific measure of significance as long as one considers them simultaneously in the appropriate way.

Connections Between Procedures

Suppose that m hypothesis tests are simultaneously tested with corresponding p-values $p_1, p_2, \ldots, p_m$. Benjamini and Hochberg (1995) propose the following algorithm for controlling the FDR at level α. Let $T_{BH} = \max\{p_i : p_i \leq \frac{i}{m}\alpha\}$. Then reject all null hypothesis corresponding to $p_i \leq T_{BH}$. When the null p-values are independent and uniformly distributed, this procedure strongly controls the FDR at level α. In Storey (2002a), I suggest the following estimate of FDR for a fixed p-value threshold t:

$$\widehat{FDR}_\lambda(t) = \frac{\widehat{\pi}_0(\lambda) \cdot t}{\frac{1}{m}\sum_{i=1}^{m} I(p_i \leq t)}, \tag{1}$$

where $\widehat{\pi}_0(\lambda)$ is an estimate of π_0, the proportion of true null hypotheses, with tuning parameter $0 \leq \lambda < 1$. The form of $\widehat{\pi}_0(\lambda)$ is

$$\widehat{\pi}_0(\lambda) = \frac{\sum_{i=1}^{m} I(p_i > \lambda)}{m(1-\lambda)}. \tag{2}$$

It is shown in Storey (2002a) under an i.i.d. mixture model that $\mathrm{E}[\widehat{FDR}_\lambda(t)] \geq FDR(t)$, where $FDR(t)$ is the false discovery rate attained when thresholding the p-values for significance at t. This inequality holds

when the null p-values are independent and uniformly distributed (Storey et al. (2002)), the same conditions as in Benjamini and Hochberg (1995).

Even though the above form of "strong control" is from the opposite viewpoint of Benjamini and Hochberg (1995), it is tempting to form the threshold

$$T_\lambda = \max\{t : \widehat{FDR}_\lambda(t) \leq \alpha\} \tag{3}$$

in order to provide strong control of the FDR. It follows that $\widehat{\pi}_0(\lambda = 0) = 1$ so that $\widehat{FDR}_{\lambda=0}(t) = mt/\sum_{i=1}^{m} I(p_i \leq t)$. From this, it easily follows that $T_{BH} = T_{\lambda=0}$. Therefore, if one takes certain liberties with the procedure proposed in Storey (2002a), it can be viewed as a generalization of the BH procedure. In fact, we have shown in Storey et al. (2002) that T_λ strongly controls the FDR at level α (again under the same conditions as in Benjamini and Hochberg (1995)), under the constraint that $T_\lambda \leq \lambda$. The fact that the threshold occurs $\leq \lambda$ is a bit of a nuisance, but makes little difference in practice for wisely chosen λ. This constraint is unnecessary for large m, which I discuss later.

There has been much confusion in the literature recently over the differences between controlling the FDR via p-values or through permutation methods. In fact, GDS quickly dismiss the FDR method used in SAM (Tusher et al. (2001)) as being unconventional and not even worth discussing. For the case of detecting differential gene expression between two conditions, Tusher et al. (2001) define an asymmetric, data-dependent thresholding rule for significance, based on modified t-statistics and a quantile-quantile plot. The thresholding rule is indexed by $0 \leq \Delta < \infty$, where the larger Δ is, the fewer the number of significant genes there are. For a fixed Δ, Tusher et al. (2001) estimate the FDR by $\mathrm{E}[V^*(\Delta)]/R(\Delta)$, where $R(\Delta)$ is the number of significant genes at this threshold. $\mathrm{E}[V^*(\Delta)]$ is the average number of genes called significant under the permutation distribution obtained by scrambling the group labels, using the same asymmetric thresholding rule.

The following result shows that this method is in fact equivalent to the Benjamini and Hochberg (1995) method in the sense described above, as long as one calculates p-values by pooling across genes. Let $\widetilde{\Delta}_i$ be the largest Δ so that gene i is called significant, for $i = 1, 2, \ldots, m$. Then the p-value of gene i, when pooling across genes (i.e., assuming their null distributions are the same), is $p_i = \mathrm{E}[V^*(\widetilde{\Delta}_i)]/m$. This easily follows by the definition given in Lehmann (1986) and by considering the nested set

of significance regions indexed by Δ.

Theorem 1. Let $p_i = \mathrm{E}[V^*(\widetilde{\Delta}_i)]/m$, $\mathrm{E}[V^*(\Delta)]$, and $R(\Delta)$ be defined as above. Then the BH algorithm applied to $p_1, p_2, \ldots, p_m$ is equivalent to calling all genes significant by $\widehat{\Delta}$ in SAM where

$$\widehat{\Delta} = \min \left\{ \Delta : \frac{\mathrm{E}[V^*(\Delta)]}{R(\Delta)} \leq \alpha \right\}.$$

Therefore, one can use the SAM software in the above way to perform the BH method. We indirectly state this fact in Storey and Tibshirani (2001), but we do not explain it as thoroughly. Given this equivalence, I think that GDS have overlooked one potentially greater drawback of SAM. The rule defined by the quantile-quantile plot and Δ is determined from the same set of data on which the FDR estimates are made. It is clear that this can result in "over-fitting" and anti-conservative biases in FDR calculations. As an extreme example, suppose that we apply SAM to detecting differential gene expression in a single gene. It then uses right-sided or left-sided significance regions, depending on whether the observed statistic is respectively positive or negative. It is not hard to show that this results in a p-value that is $1/2$ of its actual size, and therefore the FDR estimates will be two times too small. As the number of genes increases, this bias decreases, but it is always present for the most significant genes.

By noting that any use of averaging over the number of statistics called significant under some simulated null distribution is equivalent to calculating p-values by pooling across genes (or tests), it can be seen that many of the re-sampling based FDR methods are simply p-value based methods with globally defined p-values. Moreover, because the expectation of the sum of indicator random variables is the same regardless of the dependence present between them, it is difficult to see how the re-sampling approach captures dependence in the FDR case. Because of this, I am slightly skeptical about how useful and novel the current re-sampling approaches are in false discovery rates (Of course the scenario is quite different for FWER where one is concerned with $\Pr(V \geq 1)$ (Westfall and Young, 1993)). The "Storey" and "ST" methods employed in GDS would have been completely equivalent if they had pooled across genes to form p-values. GDS argue that there is no reason to suspect that each gene has the same null distribution. Perhaps this is true, but their argument for "subset pivotality" also requires assumptions. Both sets of assumptions can be met with arguments

based on a large number of arrays. Finally, note that it takes B permutations when calculating p-values across genes to get the same resolution as mB permutations when calculating p-values within genes. Recall that m is usually on the order of 3000 to 30,000.

Next, I review several very recent results that do not depend on an independence assumption, nor on the assumption that each gene has the same null or alternative distributions.

Recent Results with Applicability to DNA Microarrays

For a large number of genes m, several results about $\widehat{FDR}_\lambda(t)$ and T_λ (see equations 1-3) have been shown that increase their applicability. Note that $V(t)/m_0 = \frac{\#\{\text{null } P_i \leq t\}}{m_0}$ and $S(t)/m_1 = \frac{\#\{\text{alt. } P_i \leq t\}}{m_1}$ are the empirical distribution functions of the null and alternative p-values, respectively. Almost sure convergence in the point-wise sense as $m \to \infty$ means that with probability 1:

$$\begin{aligned} \frac{V(t)}{m_0} &\to G_0(t) \text{ for each } t \in [0,1], \\ \frac{S(t)}{m_1} &\to G_1(t) \text{ for each } t \in [0,1], \end{aligned} \tag{4}$$

for some functions G_0 and G_1. The following results are proven in Storey et al. (2002). These are closely related to several results in Genovese and Wasserman (2001).

Theorem 2 (*Storey et al. 2002*) Suppose that $V(t)/m_0 = \frac{\#\{\text{null } P_i \leq t\}}{m_0}$ and $S(t)/m_1 = \frac{\#\{\text{alt. } P_i \leq t\}}{m_1}$ converge almost surely point-wise to continuous G_0 and G_1, respectively, where $G_0(t) \leq t$. Also suppose that $\lim_{m\to\infty} m_0/m = \pi_0$ exists. Then for each $\delta > 0$,

$$\lim_{m\to\infty} \inf_{t \geq \delta} \left[\widehat{FDR}_\lambda(t) - FDR(t) \right] \geq 0 \tag{5}$$

with probability 1. Also,

$$\lim_{m\to\infty} FDR(T_0) \leq \lim_{m\to\infty} FDR(T_\lambda) \leq \alpha. \tag{6}$$

Therefore, the estimate $\widehat{FDR}_\lambda(t)$ simultaneously dominates $FDR(t)$ over all thresholds t for large m. Also, the generalized thresholding proce-

dure T_λ asymptotically controls the FDR at level α. We have $\lim_{m\to\infty} FDR(T_0) < \lim_{m\to\infty} FDR(T_\lambda)$ for $\lambda > 0$, when G_0 and G_1 are strictly monotone. Under these conditions, the generalized procedure is more powerful than the BH procedure ($T_0 = T_{BH}$).

Estimates of the q-values for each p_i were given in Storey (2002a). We have also shown that under these conditions the q-values are simultaneously conservatively consistent. Therefore, for large m, one can examine all genes and their q-values simultaneously without inducing bias. This is explicitly stated in the following result.

Corollary 1 (*Storey et al. 2002*) For a given p-value p_i, let $\widehat{q}_\lambda(p_i)$ be its estimated q-value as defined in Storey (2002a). Then under the conditions of Theorem 8.3,

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

for each $\delta > 0$.

These asymptotic results hold under the point-wise convergence of the empirical distribution functions. Note that we did not require each test to have the same null distribution, but rather the null distributions have to converge to some function. Many forms of weak dependence allow point-wise convergence of empirical distribution functions, for example ergodic dependence, blocks of dependent tests, and certain mixing distributions. This is a useful fact for certain applications, for example, when dealing with the dependence encountered in DNA microarrays.

Dependence in DNA Microarrays

I hypothesize that the most likely form of dependence between the genes encountered in DNA microarrays is weak dependence, and more specifically, "clumpy dependence"; that is, the measurements on the genes are dependent in small groups, each group being independent of the others. There are two reasons that make clumpy dependence likely. The first is that genes tend to work in pathways, that is, small groups of genes interact to produce some overall process. This can involve just a few to 50 or more genes. This would lead to a clumpy dependence in the pathway-specific noise in the data. The second reason is that there tends to be cross-hybridization in DNA microarrays. In other words, the signals between two genes can cross because of molecular similarity at the sequence level. Cross-hybridization

would only occur in small groups, and each group would be independent of the others. Typically microarrays measure the expression levels on 3000 to 30,000 genes, and each gene makes up a p-value. Therefore, given the clumpy dependence and large number of genes, I expect Theorem 2 and Corollary 1 to be relevant for the problem of detecting differential gene expression.

Many assumptions that have been made for modeling microarray data have yet to be verified. Hopefully evidence either for or against these assumptions will emerge. I have given a plausibility argument for the assumptions in Theorem 8.3 and Corollary 1. I have also provided numerical evidence in Storey et al. (2002) and Storey (2002b). GDS have stressed the dependence between the genes, not only in this article but in Dudoit et al. (2002b) as well. I leave it as a challenge to them to provide evidence from real microarray data that the aforementioned assumptions do not hold. I have not been able to find it myself. Keep in mind that one can cluster microarray data and see that many genes are in fact related, but this is very different than *stochastic* dependence, especially the type that would violate the assumptions I have argued are true.

Q-values Give Gene-specific Significance ... from a Global View

Given these results and arguments, I would like to suggest a useful way to use the information among all q-values. Since we can essentially consider all estimated q-values simultaneously, one can make various plots in order to find a useful q-value cut-off. The estimated q-values also give a gene-specific measures of significance. In Efron et al. (2001), we approached the problem of detecting differential gene expression from a Bayesian framework, where the posterior probability that a gene is not differentially expressed conditional on its observed statistic can be calculated. We called this quantity a "local false discovery rate" in the sense that it gives proportion of false positives among genes in a small neighborhood around the observed gene. We also used the main result from Storey (2001) to relate these posterior probabilities to the pFDR. This posterior probability is also a gene-specific measure of significance. Therefore, in a sense the qs-value and the traditional posterior probability are natural competitors for gene-specific measures of significance.

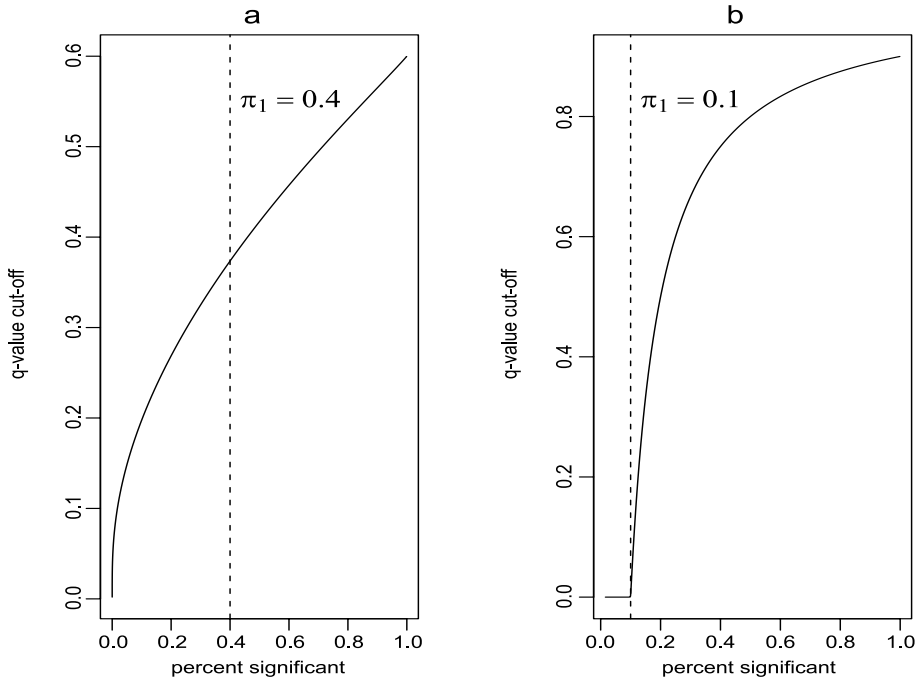

Figure 1: *A plot of the q-value cut-off versus the percent of data called significant with this cut-off when (a)* $\pi_1 = 0.4$, $\mu = 1$ *and (b)* $\pi_1 = 0.1$, $\mu = 6$.

Suppose we observe independent statistics $Z_1, Z_2, \ldots, Z_m$, distributed as a null $N(0,1)$ with probability π_0 and an alternative $N(\mu,1)$ with probability $\pi_1 = 1 - \pi_0$. Also suppose that we use a right-sided significance rule. Suppose $\pi_1 = 0.4$, $\mu = 1$, and we observe $z_i = 2$. Then its local false discovery rate is $\Pr(H_i = 0|Z_i = 2) = 0.25$ and its q-value is $\Pr(H_i = 0|Z_i \geq 2) = 0.18$. Now if we consider $\pi_1 = 0.1$ and $\mu = 6$ with the same observed statistic $z_i = 2$, we get a local false discovery rate of $\Pr(H_i = 0|Z_i = 2) = 0.9997$ and q-value of $\Pr(H_i = 0|Z_i \geq 2) = 0.17$. Therefore, by changing two important parameters we end up with totally different local false discovery rates, but very similar q-values. Clearly, we would not want to call $z_i = 2$ significant in the latter case, but perhaps it is reasonable to in the former case.

With this limited information, it appears that the q-value is not a very good gene-specific measure of significance. Is this a fair assessment? In

the context of looking at a single gene in the marginal sense, then this is a fair assessment. But in the context of the global problem of detecting differentially expressed genes, then this is absolutely not a fair assessment. It also appears to be useful to consider both the q-value and the posterior probability at the same time, but this is both difficult and it requires one to adopt the Bayesian framework.

We show a global use of the q-values does not require one to incorporate these Bayesian quantities. Consider Figure 1 where the percentage of statistics rejected has been plotted versus its corresponding q-value cut-off for both of the above scenarios. The π_1 values are denoted in each plot. From these plots, one can see the local information contained in the q-values is quite different in the two scenarios. Specifically, it can be seen in panel **b** that a q-value cut-off of 0.17 is completely unreasonable, whereas in panel **a** this is not as clear. In panel **b**, one can see that the q-value is virtually zero when 10% of the data have been rejected panel **b**. This information used in conjunction with that fact that $\pi_1 = 10\%$ makes it immediately clear that about 10% of the data being rejected is most reasonable. From panel **a**, this is not the case. Therefore, by considering all q-values simultaneously as well as π_1 in the spirit of Figure 1, one can see which cut-offs make sense. We can do this without being Bayesian and without introducing a totally new quantity.

In the methodology of Storey (2002a), one can obtain estimates of π_1. By generating only 3000 observations from each of these cases, I estimate $\widehat{\pi}_1 = 0.36$ when $\pi_1 = 0.4$ and $\mu = 1$. When $\pi_1 = 0.1$ and $\mu = 6$, I estimate $\widehat{\pi}_1 = 0.10$. The q-value plots are also very similar to the idealized versions in Figure 1. Therefore, it is not clear that the posterior probability (i.e., local false discovery rate) is always necessary.

Peter H. Westfall
Texas Tech University
Lubbock, Texas, USA

The Influence of John Tukey

Among his many other notable contributions to statistics, the late John Tukey also deserves credit for some of the ideas behind this article of Ge, Dudoit and Speed (hereafter GDS). In the late 1980's, several pharmaceu-