Statistical Significance for Genome-Wide Experiments

John D. Storey[*] and Robert Tibshirani[†]

January 2003

Abstract: With the increase in genome-wide experiments and the sequencing of multiple genomes, the analysis of large data sets has become commonplace in biology. It is often the case that thousands of features in a genome-wide data set are tested against some null hypothesis, where many features are expected to be significant. Here we propose an approach to statistical significance in the analysis of genome-wide data sets, based on the concept of the false discovery rate. This approach offers a sensible balance between the number of true findings and the number of false positives that is automatically calibrated and easily interpreted. In doing so, a measure of statistical significance called the *q*-value is associated with each tested feature in addition to the traditional *p*-value. Our approach avoids a flood of false positive results, while offering a more liberal criterion than what has been used in genome scans for linkage.

Keywords: false discovery rates, genomics, multiple hypothesis testing, p-values, q-values

[*] Department of Statistics, University of California, Berkeley CA 94720. Email: `storey@stat.berkeley.edu`.

[†] Department of Health Research & Policy and Department of Statistics, Stanford University, Stanford CA 94305. Email: `tibs@stat.stanford.edu`.

Introduction

Some of the earliest genome-wide analyses involved testing for linkage at loci spanning a large portion of the genome. Since a separate statistical test is performed at each locus, traditional p-value cut-offs of 0.01 or 0.05 had to be made stricter to avoid an abundance of false positive results. The criterion for statistical significance controls the probability that any single false positive occurs among all loci tested. This strict criterion is used mainly because one or very few loci are expected to show linkage in any given experiment (Lander & Kruglyak 1995). Due to the recent surge in high-throughput technologies and genome projects, many more types of genome-wide data sets are available. The analysis of these data sets also involves tests on thousands of features in the genome, except many more than one or two of them are are significant. In these genome-wide tests of significance, guarding against one or more false positives is typically going to be too strict and will lead to many missed findings. The goal is therefore to identify as many significant features as possible, while incurring a relatively low proportion of false positives.

We propose that the recently introduced q-value (Storey 2002) is a well suited measure of significance for this growing class of genome-wide tests of significance. The q-value is an extension of a quantity in statistics called the "false discovery rate" (Benjamini & Hochberg 1995). The q-value gives each feature its own individual measure of significance, similarly to the p-value. Whereas the p-value is a measure of significance in terms of the *false positive rate*, the q-value is a measure in terms of the *false discovery rate*. The difference in the definitions of the false positive rate and false discovery rate is subtle in theory, yet very important in practice. For a given significance threshold, the false positive rate is the probability a *null statistic* is significant. The false discovery rate, on the other hand, is the expected *proportion of null statistics* among *all statistics* called significant (i.e., the proportion of false positives among all significant features). The false discovery rate can also be written as the probability a significant feature is a false positive (Storey 2001). Note that the false discovery rate definition is often confused with the false positive rate, so it is important to keep the distinction in mind.

The q-value does not force one to choose a hard and often arbitrary cut-off for significance, rather it provides an exploratory guide for each feature's significance while taking into account the fact that thousands of features are simultaneously being tested. Suppose that features with q-values less than or equal to 5% are called significant in some genome-wide test of significance. This implies that among all significant features, about 5% of them are false positives. A p-value threshold of 5% on the other hand implies that among all the null features present in the data set, about 5% will meet this threshold – it says little about the content of the features actually called significant. The information from a p-value threshold is not helpful for making the significance decision, however, the q-value threshold is helpful since it measures what proportion of significant features are false positives. Because features called significant will likely undergo some subsequent biological verification, the q-value threshold indicates how many of these will lead to a waste of time.

Here we argue that the q-value is a sensible measure of the balance between the number of truly significant findings and the number of false positives in many genome-wide tests of significance. We motivate our proposed approach in the context of several recent and prominent papers in which awkwardly chosen p-value cut-offs were used in an attempt to at least qualitatively achieve what the q-value directly achieves. We also introduce a fully automated method for estimating q-values, with a careful treatment of dependence issues between the features and guidelines as to when the estimates are accurate. The proposed methodology is applied to data from Hedenfalk et al. (2001), supporting their results as well as providing some new information.

Motivating Examples

Consider the following four recent articles in which thousands of features from a genome-wide data set were tested against a null hypothesis. In each case, p-values thresholds were employed to decide which features to call significant, although the ultimate goal was to find many truly significant features without including too many false positives.

Example 1: Detection of differentially expressed genes. A common goal in DNA microarray experiments is to detect genes that show differential expression across two or more biological conditions (Slonim 2002). This is an important question to answer since it allows one to discover genes involved in differentiating complex biological states. In this situation, the "features" are the genes, and they are tested against the null hypothesis that there is no differential gene expression. One of the goals of Hedenfalk et al. (2001) is to find genes that are differentially expressed between BRCA1-mutation-positive tumors and BRCA2-mutation-positive tumors by obtaining several microarrays from each cell type. In their analysis they assign a p-value to each gene by standard methods and use a p-value cut-off of 0.001 to find 51 genes out of 3226 that show differential gene expression. A rough calculation says that we would have expected about 3 false positives with this cut-off. They later use a threshold of 0.0001 and conclude that 9 to 11 genes are differentially expressed.

Example 2: Identification of exonic splicing enhancers. Exonic splice enhancers (ESEs) are short oligonucleotide sequences that enhance pre-mRNA splicing when present in exons. In a recent study, Fairbrother et al. (2002) analyzed human genomic DNA in order to predict ESEs based on the statistical analysis of exon-intron and splice site composition. They assess the statistical significance of all 4096 possible hexamers, the null hypothesis being that the hexamer is not an ESE. A statistic was formed based on the location of the hexamers in 4817 human genes where the exon-intron structure has been well characterized. The end product is a p-value associated with each of the 4096 hexamers. A p-value cut-off of 10^{-4} is used because about $4096 \times 10^{-4} \ll 1$ false positive is expected under this criterion. This cut-off yielded 238 significant results.

Example 3: Genetic dissection of transcriptional regulation. In linkage analysis the goal is to find a statistically significant association between a phenotype and a marker locus. In a recent study, Brem et al. (2002) cross two strains of yeast that show much differential gene expression. In 40 of the resulting haploid progeny, the expression level of each gene is treated as a quantitative trait. This essentially results in 6215 simultaneous genome scans for linkage, where a positive result for a particular gene indicates that a regulator for the gene's expression is located in the region showing linkage. A p-value is calculated for each gene under the null hypothesis that the gene's expression is not linked to any locus. It is reasonable to expect that out of the 6215 genes, many of them are going to show linkage. In fact, their results indicate that this is the case: with a p-value cut-off of 5×10^{-5}, 507 genes show linkage with 53 expected by chance, and with a cut-off of 2×10^{-6}, 205 genes are significant with less than 1 expected from chance. Several other cut-offs with similar bits of information are given throughout the article.

Example 4: Finding binding sites of transcriptional regulators. Transcriptional regulatory proteins bind to specific promoter sequences to participate in the regulation of gene expression. The availability of complete genome sequences and the development of a method for genome-wide binding analysis has allowed the characterization of genomic sites bound by specific transcriptional regulators. Recently Lee et al. (2002) used genome-wide location analysis to investigate how yeast transcriptional regulators bind to promoter sequences across the genome. Specifically, binding of 106 transcriptional factors was measured across the genome; at each genomic location, a p-value was calculated under the null hypothesis that no binding occurs resulting in the consideration of thousands of p-values. Lee et al. "generally describe results obtained at a p-value threshold of 0.001 because [their] analysis indicates that this threshold maximizes inclusion of legitimate regulator-DNA interactions and minimizes false positives." They estimate that among the 3985 interactions found to be significant at this threshold, about 6% to 10% are false positives.

In the four above examples, the researchers tried to decide on a sensible cut-off for their p-values. Three articles used four or more cut-offs in an attempt to circumvent the difficulty in interpreting a p-value threshold in a genome-wide test of significance. The results are consequently obfuscated by the multiple cut-offs that are applied to the p-values. Two pieces of information would have made their analyses more straightforward and universally interpretable. The first is an estimate of the proportion of features that are truly significant, even if these cannot be precisely identified. For

Table 1: *Possible outcomes from thresholding m features for significance.*

	Called Significant	Called Not Significant	Total
Feature *Not* Significant	V	$m_0 - V$	m_0
Feature Truly Significant	S	$m_1 - S$	m_1
Total	R	$m - R$	m

example, what proportion of the 3218 genes in Hedenfalk et al. (2001) are differentially expressed? The second is a measure of significance that can be associated with each feature so that thresholding these numbers at a particular value has an easy interpretation. We provide both of these in our proposed approach. We also apply the method to the data from *Example 1* (Hedenfalk et al. 2001).

Proposed Method and Results

The dilemma of how to threshold, say, m p-values can be seen more clearly by considering Table 1, which lists the various outcomes that occur when testing m features in a genome-wide data set. For a given threshold, V is the total number of false positives, and R is the total number of features called significant. If we use a p-value threshold of 0.05, then we have only guaranteed that $\mathrm{E}[V] \leq 0.05 \cdot m$, which is a number much too large for all the examples we have considered. The error measure that is typically controlled in genome scans for linkage is the family-wise error rate, which can be written as $\Pr(V \geq 1)$ and is the probability of committing *any* false positives. (Note that we can guarantee that $\Pr(V \geq 1) \leq \alpha$ by calling all genes significant with p-values less that or equal to α/m, which is the well known Bonferroni correction.) Controlling $\Pr(V \geq 1)$ makes sense in the linkage case since any false positives can lead to a large waste of time, but it is way too conservative for many of the genome-wide tests of significance currently being performed. Many recent papers involving a genome-wide analysis find a cut-off so that $\mathrm{E}[V] \leq 1$, including *Example 2*. This seems too restrictive in general since one should have the flexibility to let $\mathrm{E}[V] = 2$, for example, if many more truly significant features are found.

It is therefore useful to find a sensible balance between the number of false positive features V and the number of truly significant features S. This balance can efficiently be achieved by considering the ratio

$$\frac{\#\text{false positive features}}{\#\text{significant features}} = \frac{V}{V+S} = \frac{V}{R},$$

which can be stated in words as the proportion of false positive features among all of those called significant. We are particulary interested in the false discovery rate, which is defined to be[1]

$$FDR = \mathrm{E}\left[\frac{V}{R}\middle| R > 0\right].$$

In taking the expectation of V/R, we have simply conditioned on $R > 0$, since we want to prevent $R = 0$ in which case V/R is undefined.

In each of the above motivating examples, the researchers had to deal with thousands of p-values simultaneously, where it is expected that many of these p-values are significant. A qualitative attempt is made to balance the number of p-values found significant for a given cut-off with the number expected to be false positives. The false discovery measures this trade-off precisely in a way that takes into account the joint behavior of all the p-values. The false discovery rate is therefore a useful measure of the overall accuracy of a set of significant features, but we would also like a number that can be attached to each individual feature. The q-value is a measure designed to reflect this.

Let us now precisely define the q-value. Suppose that we list the features in order of their evidence against the null hypothesis (e.g., in ascending order of their p-values). It is practical to arrange the features in this way since calling one feature significant means that any other feature with more evidence against the null should also be called significant. Therefore, one could think of listing the features in ascending order of their p-values. A threshold would be determined by calling all features significant up to some point on the list.

> *The q-value for a particular feature is the expected proportion of false positives occurring up through that feature on the list.*

Therefore, if we calculate the q-values for each feature, then thresholding them at q-value level α produces a set of significant features so that a proportion of about α are false positives. The precise definition of the q-value for a particular feature is the following.

> *The q-value for a particular feature is the minimum false discovery rate that can be attained when calling all features up through that one on the list significant.*

The q-value has a special relationship to the p-value (yielding the origin of its name) that is beyond the scope of this work. The interested reader is therefore referred to Storey (2001) for a mathematical definition of the q-value and its statistical relationship to the p-value.

As a concrete example, we considered the data from Hedenfalk et al. (2001) to test for differentially expressed genes between BRCA1-mutation-positive tumors and BRCA2-mutation-positive tumors. Using a two-sample t-statistic, we calculated a p-value for each of 3170 genes under the null

[1] This quantity is more technically defined to be the positive false discovery rate (pFDR).

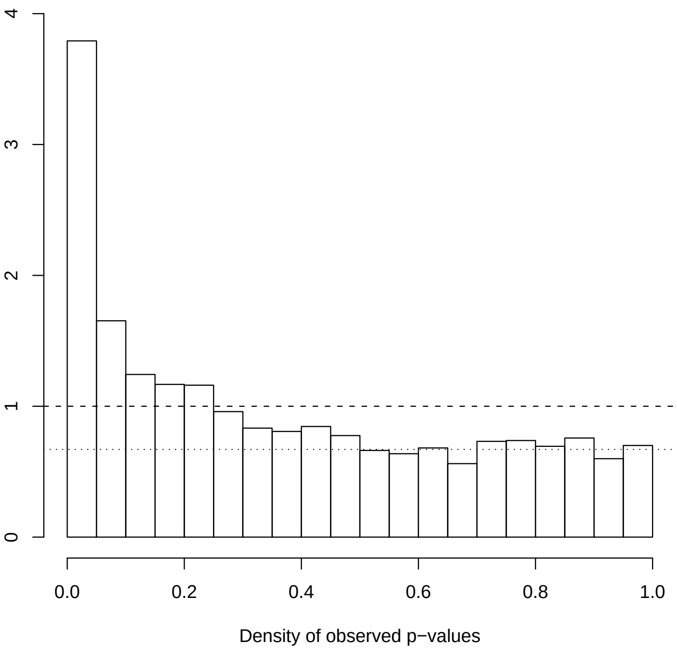

Figure 1: *A density histogram of the 3170 p-values from the Hedenfalk et al. data. The dashed line is the density histogram we would expect if all genes were null (not differentially expressed). The dotted line is at the height of our estimate of the proportion of null p-values.*

hypothesis of no differential gene expression. See Remark B in the Appendix for specific details. Figure 1 shows a density histogram of the 3170 p-values. The dashed line is the density we would expect if all genes were null (not differentially expressed), so it can be seen that many genes are differentially expressed. The p-value of a gene is the probability a statistic is as extreme or more extreme than the observed statistic, given the gene is not differentially expressed. In other words, it is the probability of seeing a gene with that much evidence for differential gene expression by chance under the null hypothesis. The p-value is a measure of significance for a single hypothesis test, and the smaller it is, the more evidence there is for differential gene expression.

Fortunately, q-values can be directly estimated from p-values. Given the definition of the q-value, it makes sense to begin by estimating the FDR when calling all genes significant whose p-value is less than or equal to some threshold t, where $0 < t \leq 1$. Denote the m p-values by $p_1, p_2, \ldots, p_m$, and let

$$V(t) = \#\{\text{false positive } p_i \leq t;\ i = 1, \ldots, m\} \text{ and } R(t) = \#\{p_i \leq t;\ i = 1, \ldots, m\}.$$

We then want to estimate

$$FDR(t) = \mathrm{E}\left[\left.\frac{V(t)}{R(t)}\right| R(t) > 0\right].$$

Since we are considering many features (i.e., m is very large), it follows that

$$FDR(t) = \mathrm{E}\left[\left.\frac{V(t)}{R(t)}\right| R(t) > 0\right] \approx \frac{\mathrm{E}[V(t)]}{\mathrm{E}[R(t)]}. \tag{1}$$

A simple estimate of $\mathrm{E}[R(t)]$ is the observed $R(t)$; that is, the number of observed p-values less than or equal to t. For estimating $\mathrm{E}[V(t)]$, recall that the null p-values are uniformly distributed. Therefore, the probability a null p-value is less than or equal to t is simply t, and it follows from Table 1 that $\mathrm{E}[V(t)] = m_0 \cdot t$. Since m_0 is unknown, we have to estimate it, or equivalently estimate the (more interpretable) proportion of features that are null, which we denote by $\pi_0 \equiv m_0/m$.

It is difficult to estimate π_0 without specifying the distribution of the truly significant p-values. However, exploiting the fact that null p-values are uniformly distributed, a reasonable estimate can be formed. From Figure 1 we can see that the histogram density of p-values beyond 0.5 looks fairly flat, which indicates that there are mostly uniformly distributed null p-values. The height of this flat portion actually gives us a conservative estimate of the overall proportion of null p-values. This can be quantified with

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

which involves the tuning parameter λ. Setting $\lambda = 0.5$, we estimate that 67% of the genes in the Hedenfalk et al. (2001) data are not differentially expressed. Note that through significance tests, prediction models, and various other techniques, Hedenfalk et al. (2001) qualitatively argue that BRCA1-mutation-positive tumors and BRCA2-mutation-positive tumors can be distinguished by their genetic profiles. Our estimate of 67% provides a direct measurement of this: we estimate that at least 33% of the examined genes are differentially expressed between these two tumor types.

The rationale for the estimate of π_0 is that p-values of truly significant features will tend to be close to zero, whereas p-values of null features will be uniformly distributed among $[0, 1]$. "Most" of the p-values we observe near 1 will be null then. If we were able to count only null p-values, then $\frac{\#\{\text{null } p_j > \lambda\}}{m(1-\lambda)}$ would be an unbiased estimate of π_0. The inclusion of a few alternative p-values will only make this estimate conservative. If we take $\lambda = 0$, then $\widehat{\pi}_0(\lambda) = 1$, which is usually going to be much too conservative in genome-wide data sets where a sizable proportion of features are suspected to be significant. However, as we set λ closer to 1, the variance of $\widehat{\pi}_0(\lambda)$ increases, making our estimated q-values more unreliable. By examining the data in Figure 1, a common sense choice for λ was $\lambda = 0.5$, however, it is useful to automate this choice. We introduce a novel and fully automated method in Remark A of the Appendix for estimating π_0 that borrows strength across a range of $\widehat{\pi}_0(\lambda)$. This automated method also happens to result in $\widehat{\pi}_0 = 0.67$.

By plugging these quantities into the right hand side of 1, the FDR when calling all p-values less than or equal to t significant can be estimated by

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

Recalling the definition of the q-value for a particular feature, we can estimate the q-value of feature i, $i = 1, 2, \ldots, m$, by

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

This method is presented in an easily implemented and fully automated algorithm in Remark A of the Appendix.

Several mathematical results about the accuracy of the estimated q-values hold under what we call "weak dependence" of the p-values (or features). First, the estimated q-values are *simultaneously* conservative for the true q-values. This means we can consider all q-values at once, without worrying about inducing bias. Second, if we call all features significant with q-values less than or equal to α, then in the long run the false discovery rate will be less than or equal to α. These conservative properties are desirable because we don't want to be underestimating the true q-values or the true proportion of false positives. We hypothesize that the most likely form of dependence between features in a genome-wide data set will meet the weak dependence requirement, although this has to be considered for each application. Specifically for DNA microarray data, we argue that "clumpy dependence" holds, which is a special case of weak dependence; that is, the genes behave dependently in small groups (i.e., pathways), each group being independent of the others. Specific details of these mathematical results and the definition of weak dependence can be found in Remark C of the Appendix.

Given this theoretical justification for considering all q-values simultaneously, even in the presence of weak dependence, it is possible to use several plots to calibrate the q-value cut-off one would want to apply in a study. (On the other hand, we recommend making use of the information contained in each feature's q-value.) Figure 2a shows a plot of the q-values versus their t-statistics from the Hedenfalk et al. (2001) data. This gives one a picture of significance according to the two-sample t-statistic. Several other plots only depending on the p-values are useful for understanding the behavior of the q-values as we adjust the cut-off. These are shown in b-d of Figure 2. Figure 2b is a plot of the q-values versus their p-values. One can see the expected proportion of false positives for different p-value cut-offs from this plot. We can then examine Figure 2c which shows the number of significant genes for each q-value. Finally, Figure 2d shows the expected number of false positives as a function of the number of genes called significant. These last three plots can be used simultaneously to give the researcher a comprehensive view of what features to examine further.

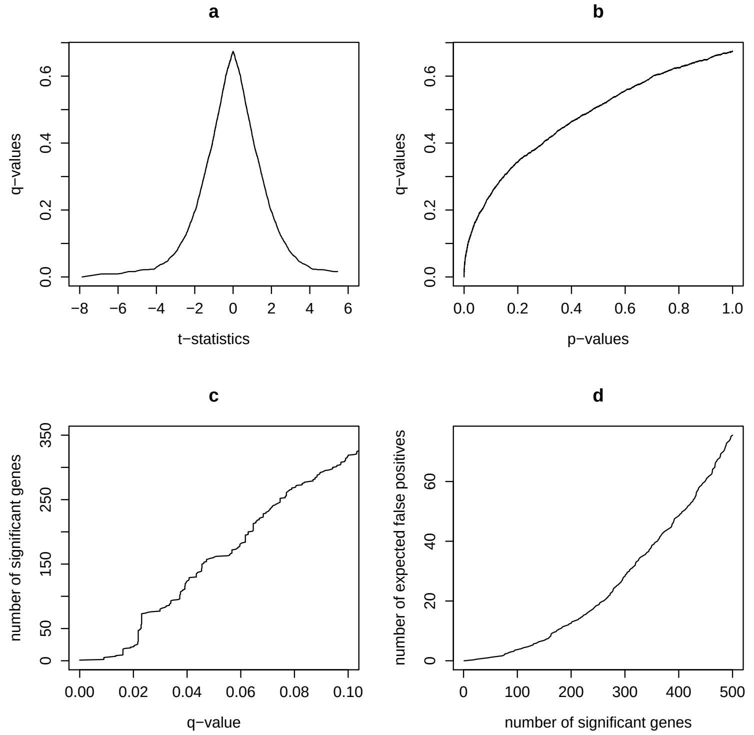

Figure 2: *Results from the Hedenfalk et al. data. (a) The q-values of the genes versus their respective t-statistics. (b) The q-values versus their respective p-values. (c) The number of genes occurring on the list up through each q-value versus the respective q-value. (d) The expected number of false positive genes versus the total number of significant genes given by the q-values.*

In our analysis, thresholding genes with q-values less than 0.05 yields 160 genes significant for differential expression. This means that about 8 of the 160 genes called significant are expected to be false positives. Hedenfalk et al. (2001) notice that a large block of genes are over-expressed in BRCA1-mutation-positive tumors, particularly genes involved in DNA repair and apoptosis. We find that 117 of the 160 called significant at q-value level 0.05 are over-expressed in BRCA1-mutation-positive tumors, quantitatively supporting their claim. The 0.05 q-value cut-off is arbitrary, and we do not recommend that this value necessarily be used. Considering particular genes allows us to examine their individual q-values. For example, the MSH2 gene (clone 32790) is the eighth most significant gene for differential expression with a q-value of 0.013 and a p-value of 5.05×10^{-5}. This gene is over-expressed in the BRCA1-mutation-positive tumors indicating increased levels of DNA repair (Kolodner 1996).

MSH2's p-value of 5.05×10^{-5} says that the probability a null (non-differentially expressed) gene would be as or more extreme than MSH2 is 5.05×10^{-5}. But MSH2's statistic could also be unlikely for a differentially expressed gene. The q-value allows a quantification of this: the q-value for MSH2 is 0.013, meaning that about 0.013 of the genes as or more extreme than MSH2 are false positives. One can also state that the probability a gene as or more extreme than MSH2 is a false positive is 0.013 (Storey 2001). Notice this latter interpretation's comparableness to the p-value's definition. (In fact, a common mistake is to state the p-value's definition as being the probability a gene is a false positive given its statistic is as or more extreme than the observed one.) The PDCD5 gene (clone 502369) is the 47th most significant gene with a q-value of 0.022 and p-value of 4.79×10^{-4}. This gene, associated with inducing apoptosis (Liu et al. 1999), is also over-expressed in BRCA1-mutation-positive tumors. The CTGF gene (clone 38393) is the 159th most significant gene for differential expression (q-value = 0.049, p-value=0.0036), and is over-expressed in BRCA2-mutation-positive. Activity of this gene is associated with suppressing apoptosis (Hishikawa et al. 1999), further supporting Hedenfalk et al.'s (2001) claims. Therefore our results support Hedenfalk et al.'s (2001) observation that many genes are over-expressed in BRCA1-mutation-positive tumors, particularly genes involved in DNA repair and apoptosis. A full list of genes with their q-values, p-values and fold-change is available at `http://www.stat.berkeley.edu/~storey/qvalue/`.

Discussion

False discovery rates have received much recent attention in the statistics literature (Yekutieli & Benjamini 1999, Benjamini & Hochberg 2000, Storey 2001, Genovese & Wasserman 2001, Efron et al. 2001, Storey 2002). Tusher et al. (2001) use a false discovery rate method in detecting differential gene expression, which can be shown to be equivalent to Benjamini & Hochberg (1995) upon close examination. The methodology we have proposed is superior for a number of reasons. First, it is the only methodology theoretically shown to be conservative (over all q-values) in likely situations encountered in genomics (Remark C of the Appendix, Storey et al. 2002). Second, it has been shown to be nearly optimal, if not sometimes completely optimal (Storey 2002). (That is, the goal is to find an upper bound on the true false discovery rate(s), but at the same time one wants to be as close to the true rate as possible to avoid a loss in statistical power.) Third, the proposed methodology is easy to implement and interpret, and it is fully automated. The original FDR methodology (Benjamini & Hochberg 1995) is too conservative for genomics applications since it assumes $\pi_0 = 1$. For example, controlling the FDR at 0.03, 0.05, or 0.07 in the Hedenfalk et al. (2001) data finds 80, 160, or 231 significant genes, respectively, using our proposed method. The Benjamini & Hochberg (1995) methodology only finds 21, 88, or 153, respectively, indicating a huge loss of power. The Benjamini & Hochberg (1995) methodology also forces one to choose an acceptable FDR level before any data are seen, which is often going to be impractical.

We have proposed a useful measure of significance when performing thousands of significance tests on a genome-wide data set. Given a ranking of the tested features in order of their evidence against the null hypothesis, the q-value for a particular feature is the minimum false discovery rate that can be attained when calling all features significant up through that one on the list. One may use the q-values as an exploratory guide for which features to investigate further. One may also take all features with q-values less than or equal to some threshold α to attain a false discovery rate less than or equal to α. Most importantly, a systematic use of q-values in exploratory genome-wide tests of significance will yield a clear balance of false positives to truly significant results, and give a standard measure of significance that can be universally interpreted.

We recommend reporting the p-value of each feature along with the estimated q-value. The p-value is widely used and its interpretation is well understood. It also only depends on the null distribution, and serves as a good counterpart to the q-value. The methodology we presented also provides an estimate $\widehat{\pi}_0$ of the proportion of features following the null hypothesis. The quantity $\widehat{\pi}_1 = 1 - \widehat{\pi}_0$ estimates a lower bound on the proportion of truly significant features. For example, among the 3170 genes we examined from Hedenfalk et al. (2001), we found that at least 33% of them are differentially expressed between BRCA1-mutation-positive tumors and BRCA2-mutation-positive tumors. Similar estimates from the other examples we considered would be interesting to compute.

Software

The freely available software QVALUE can be downloaded at `http://www.stat.berkeley.edu/~storey/qvalue/`. This program takes a list of p-values and computes their q-values and $\widehat{\pi}_0$. A version of Figure 2 is also generated.

Appendix

Remark A. General Algorithm for Estimating Q-values

Recall that there is a trade-off between bias and variance in choosing the λ to use in $\widehat{\pi}_0(\lambda)$. Therefore, λ could be chosen to minimize the sum of the variance and squared bias of $\widehat{\pi}_0(\lambda)$ relative to π_0: $\mathrm{E}\left[\widehat{\pi}_0(\lambda) - \pi_0\right]^2 = \mathrm{Bias}^2\left[\widehat{\pi}_0(\lambda)\right] + \mathrm{Var}\left[\widehat{\pi}_0(\lambda)\right].$ For wisely chosen significance regions, it should be the case that the bias of $\widehat{\pi}_0(\lambda)$ decreases with increasing λ (Storey 2002). Moreover in many cases, the bias goes to zero when $\lambda \rightarrow 1$. Therefore, the method we use here is to estimate

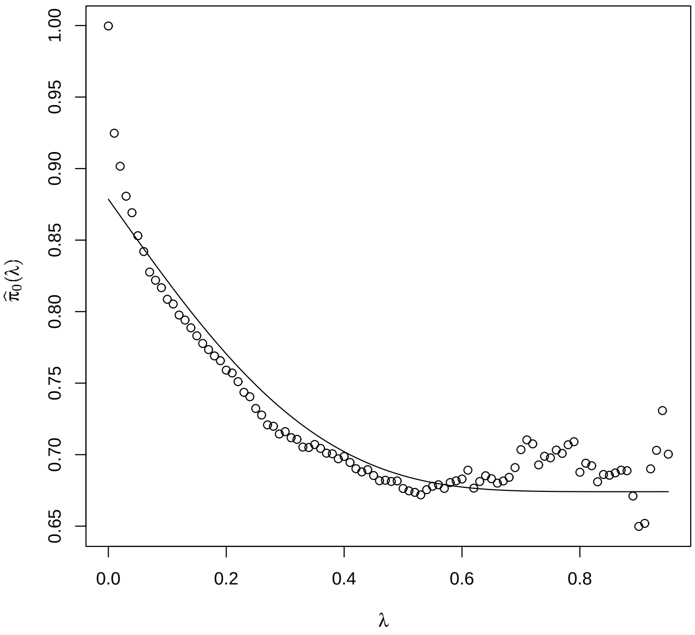

Figure 3: *The $\widehat{\pi}_0(\lambda)$ versus λ for the Hedenfalk et al. data. The solid line is a natural cubic spline fit to these points to estimate $\widehat{\pi}_0(\lambda = 1)$.*

$\lim_{\lambda \to 1} \widehat{\pi}_0(\lambda) = \widehat{\pi}_0(1)$. In doing so, we will borrow strength across the $\widehat{\pi}_0(\lambda)$ over a range of λ, giving an implicit balance between bias and variance.

Consider Figure 3, where we have plotted $\widehat{\pi}_0(\lambda)$ versus λ for $\lambda = 0, 0.01, 0.02, \ldots, 0.95$. By fitting a natural cubic spline to these data (solid line), we have estimated the overall trend of $\widehat{\pi}_0(\lambda)$ as λ increases. We purposely set the degrees of freedom of the natural cubic spline to 3; this means we limit its curvature to be like a quadratic function, which is suitable for our purposes. We also weight the observation $(\lambda, \widehat{\pi}_0(\lambda))$ by $1 - \lambda$ in the natural cubic spline, which means that we trust the $\widehat{\pi}_0(\lambda)$ with small λ to be more accurate. It can be seen from Figure 3 that the natural cubic spline fits the points quite well. The natural cubic spline evaluated at $\lambda = 1$ is our final estimate of π_0. It is normally dangerous to extrapolate from a model, but the natural cubic spline makes linear extrapolations that are smooth extensions of the model, which is exactly what we want. For a variety of simulations and forms of dependence (data not shown), this method performed well, often eliminating all bias in $\widehat{\pi}_0$.

The following is the general algorithm for estimating q-values from a list of p-values.

1. Let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ be the ordered p-values. This also denotes the ordering of the features in terms of their evidence against the null hypothesis.

2. For a range of λ, say $\mathcal{R} = \{0, 0.01, 0.02, \ldots, 0.95\}$, calculate

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

3. Let f be the natural cubic spline with 3 degrees of freedom of $\widehat{\pi}_0(\lambda)$ on λ. Moreover, we recommend weighting each "observation" $(\lambda, \widehat{\pi}_0(\lambda))$ by $1-\lambda$ in the natural cubic spline.

4. Set the estimate of π_0 to be

$$\widehat{\pi}_0 = f(1).$$

5. Estimate

$$\widehat{q}(p_{(m)}) = \min_{t \geq p_{(m)}} \frac{\widehat{\pi}_0 m \cdot t}{\#\{p_j \leq t\}} = \widehat{\pi}_0 \cdot p_{(m)}.$$

6. For $i = m-1, m-2, \ldots, 1$, estimate

$$\widehat{q}(p_{(i)}) = \min_{t \geq p_{(i)}} \frac{\widehat{\pi}_0 m \cdot t}{\#\{p_j \leq t\}} = \min\left(\frac{\widehat{\pi}_0 m \cdot p_{(i)}}{i}, \widehat{q}(p_{(i+1)})\right).$$

7. The estimated q-value for the i^{th} gene on the list is $\widehat{q}(p_{(i)})$.

Remark B. Analysis of the Hedenfalk et al. Data

We downloaded the data for Hedenfalk et al. (2001) from `http://research.nhgri.nih.gov/microarray/NEJM_Supplement/`. The file contained 3226 genes on $n_1 = 7$ BRCA1 arrays and $n_2 = 8$ BRCA2 arrays, along with some arrays from sporadic breast cancer which we did not use. If any gene had one or more measurement exceeding 20, then this gene was eliminated. A value of 20 is several IQR's (interquartile range) away from the IQR of all the data, and did not seem trustworthy for this example. This left $m = 3170$ genes.

We tested whether each gene has a difference in average gene expression between these two tumor types with a two-sample t-statistic. Let the expression value from the j^{th} array and the i^{th} gene be denoted by x_{ij}. Then $\overline{x}_{i2} = \frac{1}{n_2}\sum_{j \in \text{BRCA2}} x_{ij}$ and $s_{i2}^2 = \frac{1}{n_2-1}\sum_{j \in \text{BRCA2}} (x_{ij} - \overline{x}_{i2})^2$ are the sample mean and variance for gene i among the arrays taken from BRCA2 tumors. We can similarly define $\overline{x}_{i1}$ and s_{i1}^2 to be the sample mean and variance for the i^{th} gene among the BRCA1 tumor arrays. The two sample t-statistic for the i^{th} gene, allowing for the possibility that the tumors have different variances, is then

$$t_i = \frac{\overline{x}_{i2} - \overline{x}_{i1}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}}}$$

for $i = 1, 2, \ldots, 3170$.

We next calculated null versions of $t_1, t_2, \ldots, t_{3170}$ when there is no differential gene expression. Since it is not possible to assume that the t_i follow a t distribution, we calculate these by a permutation method. Consider all possible ways to assign $n = 15$ arrays to $n_1 = 7$ arrays from BRCA1 and $n_2 = 8$ arrays from BRCA2. Under the assumption that there is no differential gene expression, the t-statistic should have the same distribution regardless of how we make these assignments. Specifically, the labels on the arrays are randomly scrambled, and the t-statistics are recomputed. Therefore, for $B = 100$ permutations of the array labels we get a set of null statistics $t_1^{0b}, \ldots, t_{3170}^{0b}$, $b = 1, \ldots B$. The p-value for gene i, $i = 1, 2, \ldots, 3170$ was calculated by

$$p_i = \sum_{b=1}^{B} \frac{\#\{j : |t_j^{0b}| \geq |t_i|, j = 1, \ldots, 3170\}}{3170 \cdot B}.$$

We estimated the q-values for differential gene expression between the BRCA1 and BRCA2 tumors using the above algorithm. All results, including the computer code used to analyze the data can be found at `http://www.stat.berkeley.edu/~storey/qvalue/`.

Remark C. Theoretical Properties

Several mathematical results hold under "weak dependence" of the p-values (or features in the genome). These mathematical results say when our method yields conservative q-value estimates. The conservative property is necessary because we don't want to be underestimating the true q-values (for the same reason we would not want to underestimate a p-value).

Suppose that the empirical distribution function of the observed p-values converges point-wise to some function. Also suppose that the null p-values are uniformly distributed and that m_0/m converges as the number of features tested m gets large. Then it can be shown that for any $\delta > 0$,

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

which means that the estimated q-values are conservative for the true q-values, even when taking the worst case scenario over $[\delta, 1]$ for arbitrarily small δ. Also, we can conclude that

$$\lim_{m \to \infty} \frac{\#\{\text{false positive } \widehat{q}(p_i) \leq \alpha\}}{\#\{\widehat{q}(p_i) \leq \alpha\}} \leq \alpha$$

which means that if we call all genes with q-values less than or equal to α, then in the long run the false discovery rate will be less than or equal to α. The proofs of these claims follow from minor modifications to some of the main results in Storey et al. (2002).

References

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* **85**: 289–300.

Benjamini, Y. & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics, *J. Edu. and Behav. Stat.* **25**: 60–83.

Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast, *Science* **296**: 752–755.

Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association* **96**: 1151–1160.

Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. (2002). Predictive identification of exonic splicing enhancers in human genes, *Science* **297**: 1007–1013.

Genovese, C. & Wasserman, L. (2001). Operating characteristics and extensions of the FDR procedure, *Journal of the Royal Statistical Society, Series B* **64**: 499–517.

Hedenfalk, I., Duggan, D., Chen, Y. D., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A. & Trent, J. (2001). Gene-expression profiles in hereditary breast cancer, *New England Journal of Medicine* **344**: 539–548.

Hishikawa, K., Oemar, B. S., Tanner, F. C., Nakaki, T., Luscher, T. F. & Fujii, T. (1999). Connective tissue growth factor induces apoptosis in human breast cancer cell line MCF-7, *Journal of Biological Chemistry* **274**: 37461–37466.

Kolodner, R. (1996). Biochemistry and genetics of eukaryotic mismatch repair, *Genes and Development* **10**: 1433–1442.

Lander, E. S. & Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results, *Nature Genetics* **11**: 241–247.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K. & Young, R. A. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science* **298**: 799–804.

Liu, H. T., Wang, Y. G., Zhang, Y. M., Song, Q. S., Di, C. H., Chen, G. H., Tang, J. & Ma, D. L. (1999). TFAR19, a novel apoptosis-related gene cloned from human leukemia cell line TF-1, could enhance apoptosis of some tumor cells induced by growth factor withdrawal, *Biochemical and Biophysical Research Communications* **254**: 203–210.

Slonim, D. K. (2002). From patterns to pathways: Gene expression data analysis comes of age, *Nature Genetics* **32**: 502–508 (supplement).

Storey, J. D. (2001). The positive false discovery rate: A Bayesian interpretation and the q-value. Submitted. Available at `http://www.stat.berkeley.edu/~storey/`.

Storey, J. D. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society, Series B* **64**: 479–498.

Storey, J. D., Taylor, J. E. & Siegmund, D. (2002). A unified estimation approach to false discovery rates. Submitted. Available at `http://www.stat.berkeley.edu/~storey/`.

Tusher, V., Tibshirani, R. & Chu, C. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation, *Proceedings of the National Academy of Sciences* **98**: 5116–5121.

Yekutieli, D. & Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics, *J. Stat. Plan. and Inference* **82**: 171–196.