Statistical significance for genomewide studies

John D. Storey*† and Robert Tibshirani‡

*Department of Biostatistics, University of Washington, Seattle, WA 98195; and ‡Departments of Health Research and Policy and Statistics, Stanford University, Stanford, CA 94305

Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA, and approved May 30, 2003 (received for review January 28, 2003)

With the increase in genomewide 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 genomewide data set are tested against some null hypothesis, where a number of features are expected to be significant. Here we propose an approach to measuring statistical significance in these genomewide studies based on the concept of the false discovery rate. This approach offers a sensible balance between the number of true and 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. The *q* value is similar to the well known *p* value, except it is a measure of significance in terms of the false discovery rate rather than the false positive rate. 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.

false discovery rates | genomics | multiple hypothesis testing | *q* values

Some of the earliest genomewide studies involved testing for linkage at loci spanning a large portion of the genome. Because a separate statistical test is performed at each locus, traditional *p*-value cutoffs of 0.01 or 0.05 had to be made stricter to avoid an abundance of false positive results. The threshold for significance in linkage analysis is usually chosen so that the probability of any single false positive among all loci tested is ≤ 0.05. This strict criterion is used mainly because one or very few loci are expected to show linkage in any given study (1, 2). Because of the recent surge in high-throughput technologies and genome projects, many more types of genomewide studies are now underway. The analyses of these data also involve performing statistical tests on thousands of features in a genome. As opposed to the linkage case, it is expected that many more than one or two of the tested features are statistically significant. Guarding against any single false positive occurring is often going to be much too strict and will lead to many missed findings. The goal is therefore to identify as many significant features in the genome as possible, while incurring a relatively low proportion of false positives.

We are specifically concerned with situations in which a well defined statistical hypothesis test is performed on each of thousands of features represented in a genome. These "features" can be genes, all nucleotide words of a certain length, single-nucleotide polymorphism markers, etc. Several motivating examples are given below. For each feature, a null hypothesis is tested against an alternative hypothesis. In this work, we say that a feature is truly null if the null hypothesis is true, and a feature is truly alternative if the alternative hypothesis is true. If a feature is called significant, then the null hypothesis is rejected in favor of the alternative hypothesis. The goal is to propose and estimate a measure of significance for each feature that meets the practical goals of the genomewide study and that is easily interpreted in terms of the simultaneous testing of thousands of features.

We propose that the recently introduced *q* value (3, 4) is a well suited measure of significance for this growing class of genomewide tests of significance. The *q* value is an extension of a quantity called the "false discovery rate" (FDR) (5), which has received much recent attention in the statistics literature (6–11). A FDR method has been used in detecting differential gene expression in DNA microarray experiments (12), which can be shown to be equivalent to the method in ref. 5 under certain assumptions. Also, ideas similar to FDRs have appeared in the genetics literature (1, 13).

Similarly to the *p* value, the *q* value gives each feature its own individual measure of significance. 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 FDR. The false positive rate and FDR are often mistakenly equated, but their difference is actually very important. Given a rule for calling features significant, the false positive rate is the rate that truly null features are called significant. The FDR is the rate that significant features are truly null. For example, a false positive rate of 5% means that on average 5% of the truly null features in the study will be called significant. A FDR of 5% means that among all features called significant, 5% of these are truly null on average.

The *q* value provides a measure of each feature's significance, automatically taking into account the fact that thousands are simultaneously being tested. Suppose that features with *q* values $\leq 5\%$ are called significant in some genomewide test of significance. This results in a FDR of 5% among the significant features. A *p*-value threshold of 5% yields a false positive rate of 5% among all null features in the data set. In light of the definition of the false positive rate, a *p*-value cutoff says little about the content of the features actually called significant. The *q* values directly provide a meaningful measure among the features called significant. Because significant features will likely undergo some subsequent biological verification, a *q*-value threshold can be phrased in practical terms as the proportion of significant features that turn out to be false leads.

Here we show that the FDR is a sensible measure of the balance between the number of true positives and false positives in many genomewide studies. We motivate our proposed approach in the context of several recent and prominent papers in which awkwardly chosen *p*-value cutoffs were used in an attempt to achieve at least qualitatively what the *q* value directly achieves. We also introduce a fully automated method for estimating *q* values, with an initial treatment of dependence issues between the features and guidelines as to when the estimates are accurate. The proposed methodology is applied to some gene expression data taken from cancer tumors (14), supporting previously shown results and providing some additional information.

Motivating Examples

Consider the following four recent articles in which thousands of features from a genomewide data set were tested against a null hypothesis. In each case, *p*-value thresholds were used to decide which features to call significant, the ultimate goal being to identify many truly alternative features without including too many false positives.

Example 1: Detecting Differentially Expressed Genes. A common goal in DNA microarray experiments is to detect genes that show differential expression across two or more biological conditions (15). In this scenario, the "features" are the genes, and they are tested against the null hypothesis that there is no differential

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: FDR, false discovery rate; pFDR, positive FDR.

†To whom correspondence should be addressed. E-mail: jstorey@u.washington.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1530509100

gene expression. One of the goals of Hedenfalk *et al.* (14) was to find genes that are differentially expressed between *BRCA1*- and *BRCA2*-mutation-positive tumors by obtaining several microarrays from each cell type. In their analysis they computed a modified F statistic and used it to assign a p value to each gene. A p-value cutoff of 0.001 was selected to find 51 genes of 3,226 that show differential gene expression. A rough calculation shows that about three false positives are expected with this cutoff. These authors later used a threshold of 0.0001 and concluded that 9–11 genes are differentially expressed.

Example 2: Identifying Exonic Splicing Enhancers. Exonic splice enhancers are short oligonucleotide sequences that enhance pre-mRNA splicing when present in exons (16). Fairbrother *et al.* (17) analyzed human genomic DNA to predict exonic splice enhancers based on the statistical analysis of exon–intron and splice-site composition. They assessed the statistical significance of all 4,096 possible hexamers, the null hypothesis being a mathematical formulation of a hexamer not being an exonic splice enhancer. A statistic is formed based on the location of the hexamers in 4,817 human genes where the exon–intron structure has been well characterized. The end product is a p value associated with each of the 4,096 hexamers. A p-value cutoff of 10^{-4} was used based on the rationale that, at most, $4{,}096 \times 10^{-4} < 1$ false positive is expected under this criterion. This cutoff yields 238 significant hexamers, a number of which were subsequently biologically verified.

Example 3: Genetic Dissection of Transcriptional Regulation. Global monitoring of gene expression and large-scale genotyping were recently used to study transcriptional regulation in yeast. Brem *et al.* (18) crossed two strains of yeast, where many genes appeared to be expressed differentially between these two strains. For 40 of the resulting haploid progeny, the expression levels of 6,215 genes were measured by using microarrays. Linkage was tested between 3,312 markers spanning the genome and each of these 6,215 "quantitative traits." A statistically significant linkage between a gene's expression level and a marker indicates that a regulator for that gene is located in the region of the marker. In analyzing these data, one can perform a statistical test for each gene-marker combination, resulting in millions of p values, or one can test each gene for showing linkage to at least one locus, resulting in 6,215 p values. Taking the latter approach and using a p-value cutoff of 8.5×10^{-3}, Brem *et al*. reported that 507 genes show linkage to at least one locus, where 53 are expected by chance. A cutoff of 1.6×10^{-4} yields 205 genes showing linkage to at least one locus, where 1 is expected by chance. The p values are calculated according to a permutation scheme to capture the dependence between adjacent markers (19). The above-mentioned cutoffs correspond to respective thresholds of 5×10^{-5} and 2×10^{-6} when testing every gene–marker combination. Several other p-value cutoffs with similar pieces of information are given throughout ref. 18.

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 genomewide binding analysis has allowed the characterization of genomic sites bound by specific transcriptional regulators. Lee *et al.* (20) used genomewide 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 3,985 interactions found to be significant at this threshold, ≈6–10% are false positives.

Table 1. Possible outcomes from thresholding m features for significance

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

Reasonable p-value thresholds were sought in each of the four examples. Three of them used four or more cutoffs in an attempt to circumvent the inherent difficulty in interpreting a p-value threshold in a genomewide study. The significance of the results is consequently obfuscated by the multiple cutoffs that are applied to the p values. Two pieces of information make such analyses more straightforward and universally interpretable. The first is an estimate of the overall proportion of features that are truly alternative, even if these cannot be precisely identified. For example, what proportion of the 3,226 genes in example 1 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.

Note that, in example 1, one could just as well work with the modified F statistic and threshold it directly. Directly thresholding the F statistic is equivalent to thresholding the p values described above. The proposed methodology described in terms of the original statistics can be intuitively pleasing for certain cases, proving that p values are not a necessary intermediate step. However, in other cases, such as examples 2 and 3, the test statistics and null distributions are much more complicated, and p values provide a convenient numerical measure of the strength of evidence against the null for each feature. For this reason, we describe our proposal in terms in p values rather than test statistics. It is also preferable to present the q-value estimates in terms of p values to make the method widely applicable. However, working with the original test statistics and null distributions will lead to the same q-value estimates (3).

Proposed Method and Results

The dilemma of how to consider, say, m p values is seen more clearly by considering the various outcomes that occur when a significance threshold is applied to them. Table 1 lists these outcomes: specifically, F is the number of false positives, T is the number of true positives, and S is the total number of features called significant. Also, m_0 is the number of truly null features in the study, and $m_1 = m - m_0$ is the number of truly alternative features. These quantities can be used to form an overall error measure for any given p-value cutoff. Regardless of whether the p-value threshold is fixed or data-dependent, the quantities F, T, and S are random variables. Therefore, it is common statistical practice to write the overall error measure in terms of an expected value, which we denote by E[·].

If the false positive rate is the error measure used, then a simple p-value threshold is used. A p-value threshold of 0.05, for example, guarantees only that the expected number of false positives is $E[F] \leq 0.05\,m$. This number is much too large for all of the examples we have considered, and the false positive rate is too liberal. The error measure that is typically controlled in genome scans for linkage is the familywise error rate, which can be written as $\Pr(F \geq 1)$. [Note that we can guarantee that $\Pr(F \geq 1) \leq \alpha$ by calling all features significant with p values $\leq \alpha/m$, which is the well known Bonferroni correction.] Controlling $\Pr(F \geq 1)$ is practical when very few features are expected to be truly alternative (e.g., in the linkage case), because any false positive can lead to a large waste of time. However, the familywise error rate is much too conservative for

many of the genomewide studies currently being performed, including the four examples we considered where many features are expected to be truly alternative.

It is therefore useful to find an error measure in between these, specifically, one that provides a sensible balance between the number of false positive features, F, and the number of true positive features, T. This balance can be achieved efficiently by considering the ratio

$$\frac{\text{no. false positive features}}{\text{no. significant features}} = \frac{F}{F+T} = \frac{F}{S},$$

which can be stated in words as the proportion of false positive features among all of those called significant. We are particularly interested in the FDR, which is defined to be the expected value of this quantity:

$$\text{FDR} = \text{E}\left[\frac{F}{F+T}\right] = \text{E}\left[\frac{F}{S}\right].$$

To be completely rigorous, there is the possibility that $S = 0$, in which case F/S is undefined, so some adjustment has to be made to this definition (see *Remark A* in *Appendix*). The FDR can also be written in terms of the well known specificity, $(m_0 - F)/m_0$, and sensitivity, T/m_1:

$$\text{FDR} = \text{E}\left[\frac{m_0 \cdot [1 - \text{specificity}]}{m_0 \cdot [1 - \text{specificity}] + m_1 \cdot \text{sensitivity}}\right].$$

Clearly, the FDR is a useful measure of the overall accuracy of a set of significant features for the examples we described and many other genomewide studies. But one would also like a measure of significance that can be attached to each individual feature. The q value is a measure designed to reflect this level of attachment.

Suppose that we list the features in order of their evidence against the null hypothesis. It is practical to arrange the features in this way because calling one feature significant means that any other feature with more evidence against the null should also be called significant. Hence, we list the features from smallest to largest p value. If a threshold value is chosen, we call all features significant up through that threshold.

The q value for a particular feature is the expected proportion of false positives incurred when calling that feature significant. Therefore, calculating the q values for each feature and thresholding them at q-value level α produces a set of significant features so that a proportion of α is expected to be false positives. Typically, the p value is described as the probability of a null feature being as or more extreme than the observed one. "As or more extreme" in this setup means that it would appear higher on the list. The q value of a particular feature can be described as the expected proportion of false positives among all features as or more extreme than the observed one. The q value has a special probabilistic relationship to the p value (yielding the origin of its name) that is briefly explained in *Remark A* in *Appendix*.

As a concrete example, we considered the data from ref. 14 to identify genes that are differentially expressed between *BRCA1*- and *BRCA2*-mutation-positive tumors. Using a two-sample t statistic, we calculated a p value for each of 3,170 genes under the null hypothesis of no differential gene expression. See *Remark C* in *Appendix* for specific details. Fig. 1 shows a density histogram of the 3,170 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.

Given the definition of the q value, it makes sense to begin by estimating the FDR when calling all features 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

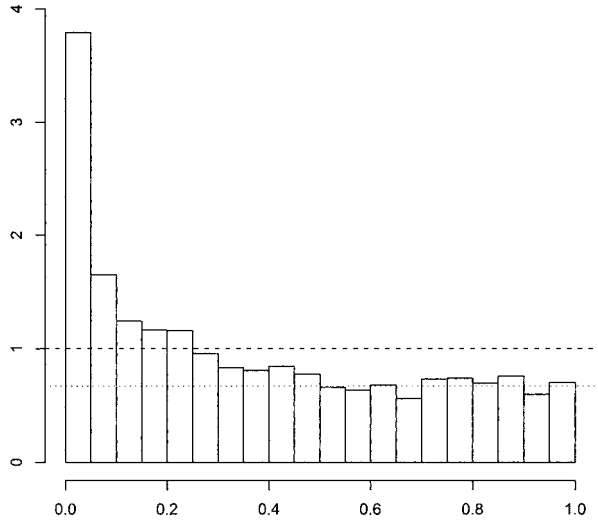

Fig. 1. A density histogram of the 3,170 p values from the Hedenfalk *et al.* (14) data. The dashed line is the density histogram we would expect if all genes were null (not differentially expressed). The dotted line is at the height of our estimate of the proportion of null p values.

$$F(t) = \#\{\text{null } p_i \leq t; i = 1, \ldots, m\} \text{ and}$$

$$S(t) = \#\{p_i \leq t; i = 1, \ldots, m\}.$$

We then want to estimate

$$\text{FDR}(t) = \text{E}\left[\frac{F(t)}{S(t)}\right].$$

Because we are considering many features (i.e., m is very large), it can be shown that

$$\text{FDR}(t) = \text{E}\left[\frac{F(t)}{S(t)}\right] \approx \frac{\text{E}[F(t)]}{\text{E}[S(t)]}. \qquad [1]$$

A simple estimate of $\text{E}[S(t)]$ is the observed $S(t)$; that is, the number of observed p values $\leq t$. In estimating $\text{E}[F(t)]$, recall that p values corresponding to truly null hypotheses should be uniformly distributed. [If the null p values are not uniformly distributed, then one wants to err in the direction of overestimating p values (i.e, underestimating significance). Correctly calculated p values are an important assumption underlying our methodology. See also *Remark D* in *Appendix*.] Thus, the probability a null p value is $\leq t$ is simply t, and it follows from Table 1 that $\text{E}[F(t)] = m_0 \cdot t$. Because the total number of truly null features m_0 is unknown it has to be estimated. Equivalently, one can estimate the (more interpretable) proportion of features that are truly 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 alternative p values. However, exploiting the fact that null p values are uniformly distributed, a reasonable estimate can be formed. From Fig. 1 we can see that the histogram density of p values beyond 0.5 looks fairly flat, which indicates that there are mostly null p values in this region. The height of this flat portion actually gives a conservative estimate of the overall proportion of null p values. This can be quantified with

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \ldots, m\}}{m(1-\lambda)},$$

which involves the tuning parameter λ. Setting $\lambda = 0.5$, we estimate that 67% of the genes in the data from ref. 14 are not differentially expressed. Note that through significance tests, prediction models, and various other techniques, it has been qualitatively argued that

BRCA1- and *BRCA2*-mutation-positive tumors can be distinguished by their genetic profiles (14). 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. Using traditional *p*-value cutoffs, Hedenfalk *et al.* (14) were comfortable only with concluding that 9–11 genes are differentially expressed of >3,000.

The rationale behind the estimate of π_0 is that *p* values of truly alternative 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 $\#\{\text{null } p_i > \lambda\}/m(1-\lambda)$ would be an unbiased estimate of π_0. The inclusion of a few alternative *p* values only makes this estimate conservative. If we take $\lambda = 0$, then $\hat{\pi}_0(\lambda) = 1$, which is usually going to be much too conservative in genomewide data sets, where a sizable proportion of features are expected to be truly alternative. However, as we set λ closer to 1, the variance of $\hat{\pi}_0(\lambda)$ increases, making the estimated *q* values more unreliable. By examining the data in Fig. 1, a common sense choice for λ was $\lambda = 0.5$. In general, it is useful to automate this choice. We introduce a fully automated method in *Remark B* in *Appendix* for estimating π_0 that borrows strength across a range of $\hat{\pi}_0(\lambda)$. This automated method also happens to result in $\hat{\pi}_0 = 0.67$.

By plugging these quantities into the right side of Eq. **1**, FDR(*t*) is estimated by

$$\widehat{\text{FDR}}(t) = \frac{\hat{\pi}_0 m \cdot t}{S(t)} = \frac{\hat{\pi}_0 m \cdot t}{\#\{p_i \le t\}}.$$

The more mathematical definition of the *q* value is the minimum FDR that can be attained when calling that feature significant (see *Remark A* in *Appendix*). Thus, the *q* value of feature *i* is $\min_{t \ge p_i} \text{FDR}(t)$, where we have simply considered all thresholds $t \ge p_i$. We can estimate the *q* value of feature *i* by simply plugging $\widehat{\text{FDR}}(t)$ into the definition above:

$$\hat{q}(p_i) = \min_{t \ge p_i} \widehat{\text{FDR}}(t).$$

Note that this guarantees that the estimated *q* values are increasing in the same order as the *p* values. This method is presented in an easily implemented and fully automated algorithm in *Remark B* in *Appendix*.

We mention two mathematical results concerning the accuracy of the estimated *q* values that hold for large *m* under what we call "weak dependence" of the *p* values (or features). Weak dependence can loosely be described as any form of dependence whose effect becomes negligible as the number of features increases to infinity. (See *Remark D* in *Appendix* and ref. 10.) The first result is that if we call all features significant with *q* values $\le \alpha$, then for large *m* the FDR will be $\le \alpha$. The second result is that the estimated *q* values are simultaneously conservative for the true *q* values. This means that the estimated *q* value of each feature is greater than or equal to its true *q* value, across all features at once. Under this result, one can consider each feature's significance simultaneously without worrying about inducing bias. In a sense, the second result implies that one can consider all α cutoffs simultaneously, which is a much stronger generalization of the first result. These conservative properties are desirable because one does not want to underestimate the true *q* values or the true proportion of false positives. We hypothesize that the most likely form of dependence between features in a genomewide 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 because genes behave dependently in small groups (i.e., pathways), with each group essentially being independent of the others, then the dependence can be modeled in blocks in such a way to satisfy the mathematical conditions. More specific details of these mathematical results can be found in *Remark D* in *Appendix*.

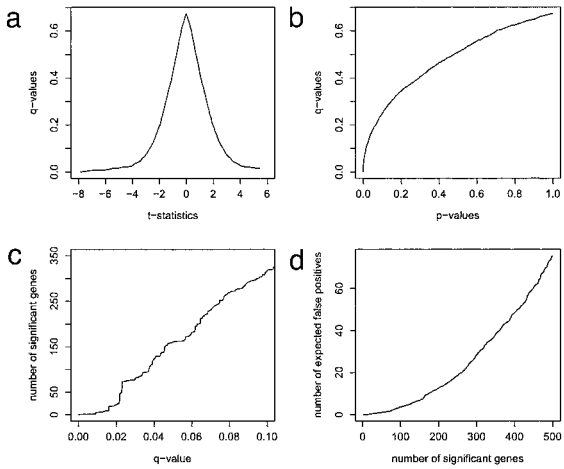

Fig. 2. Results from the Hedenfalk *et al.* (14) data. (*a*) The *q* values of the genes versus their respective *t* statistics. (*b*) The *q* values versus their respective *p* values. (*c*) The number of genes occurring on the list up through each *q* value versus the respective *q* value. (*d*) The expected number of false positive genes versus the total number of significant genes given by the *q* values.

Given this potentially valuable 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 cutoff one would want to apply in a study. (On the other hand, a single cutoff is not always necessary; each feature's estimated *q* value could simply be reported.) Fig. 2*a* shows a plot of the *q* values versus their *t* statistics from the data in ref. 14. Fig. 2*b* is a plot of the *q* values versus their *p* values. One can see the expected proportion of false positives for different *p*-value cutoffs from this plot. Fig. 2*c* shows the number of significant genes for each *q* value. Notice that for estimated *q* values slightly greater than 0.02, a sharp increase occurs in the number of significant genes over a small increase in *q* value. This allows one to easily see that a slightly larger *q*-value cutoff results in many more significant genes. Finally, Fig. 2*d* shows the expected number of false positives as a function of the number of genes called significant. In general, these last three plots can be used concurrently to give the researcher a comprehensive view of what features to examine further.

In our analysis, thresholding genes with *q* values $\le$0.05 yields 160 genes significant for differential expression. This means that $\approx$8 of the 160 genes called significant are expected to be false positives. It has previously been noticed that a large block of genes are overexpressed in *BRCA1*-mutation-positive tumors, in particular, genes involved in DNA repair and apoptosis (14). We find that 117 of the 160 called significant at *q*-value level 0.05 are overexpressed in *BRCA1*-mutation-positive tumors, quantitatively supporting their claim. The 0.05 *q*-value cutoff 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 overexpressed in the *BRCA1*-mutation-positive tumors, indicating increased levels of DNA repair (21).

MSH2's *p* value of 5.05×10^{-5} says that the probability a null (nondifferentially 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 estimated *q* value for *MSH2* is 0.013, meaning that $\approx$0.013 of the genes that are as or more extreme than *MSH2* are false positives. 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 (15), is also overexpressed 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 overexpressed in *BRCA2*-mutation-positive tumors. Activity of this gene is associated with suppressing apoptosis (23), which further supports earlier claims (14). Therefore, our results support the previous observation that many genes are overexpressed 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://genomine.org/qvalue/results.html.

A common mistake is to state that the p value is the probability a feature is a false positive. We stress that the q value is also not the probability that the feature is a false positive. In the example presented above *MSH2* has a q value equal to 0.013. This value does not imply that *MSH2* is a false positive with probability 0.013. Rather, 0.013 is the expected proportion of false positives incurred if we call *MSH2* significant. Because the q-value measure includes genes that are possibly much more significant than *MSH2*, the probability that *MSH2* is itself a false positive may be substantially higher. In terms of the FDR approach, this probability can also be thought of as a "local FDR" (3, 8, 24, 25). Statistical significance involves deciding between null and alternative hypotheses. When assigning multiple measures of statistical significance, it is necessary to account for the fact that decisions are made for m features simultaneously. The q value accomplishes this by conditioning based on the fact that every feature as or more extreme will also be called significant: the probability a feature is a false positive does not. However, the latter quantity clearly provides very useful information, and ideally one would have both estimates available for the analysis of a genomewide study.

Discussion

We have proposed the q value as an FDR-based measure of significance for genomewide studies. The methodology we have proposed is the only methodology theoretically shown to be conservative (over all q values) in situations plausibly encountered in genomics. (See *Remark D* in *Appendix* and ref. 10.) The proposed methodology is easy to implement and interpret, and it is fully automated. The original FDR methodology (5) is too conservative for genomics applications because it assumes $\pi_0 = 1$. For example, controlling the FDR at 0.03, 0.05, or 0.07 in the expression data (14) finds 80, 160, or 231 significant genes, respectively, when our proposed method is used. The methodology in ref. 5 finds only 21, 88, or 153, respectively, indicating that this earlier method's estimates are too conservative and result in a substantial loss of power. The approach in ref. 5 also forces one to choose a single acceptable FDR level before any data are seen, which is often going to be impractical and too restrictive.

The q value of a particular feature in a genomewide data set is the expected proportion of false positives incurred when calling that feature significant. 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 $\leq$ some threshold α to attain a FDR $\leq \alpha$. Most importantly, a systematic use of q values in genomewide tests of significance will yield a clear balance of false positives to true positive results and give a standard measure of significance that can be universally interpreted. The methodology we presented also provides an estimated $\hat{\pi}_0$ of the proportion of features following the null hypothesis. The quantity $\hat{\pi}_1 = 1 - \hat{\pi}_0$ estimates a lower bound on the proportion of truly alternative features. For example, among the 3,170 genes we examined from ref. 14, we found that at least 33% are differentially expressed between *BRCA1*- and *BRCA2*-mutation-positive tumors. Similar estimates from the other examples we considered would be interesting to compute.

The software QVALUE can be downloaded at http://genomine.org/qvalue/. This program takes a list of p values and computes their estimated q values and $\hat{\pi}_0$. A version of Fig. 2 is also generated.

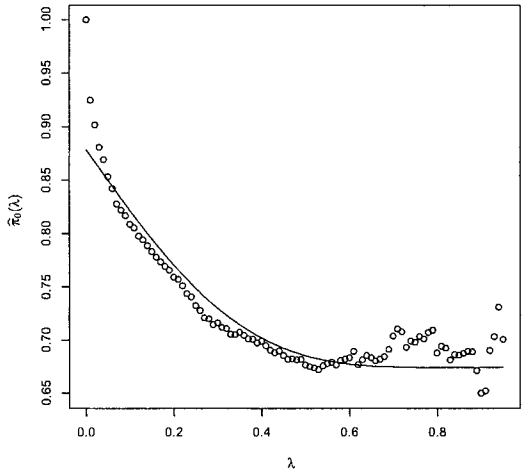

Fig. 3. The $\hat{\pi}_0(\lambda)$ versus λ for the data of Hedenfalk *et al.* (14). The solid line is a natural cubic spline fit to these points to estimate $\hat{\pi}_0(\lambda = 1)$.

Appendix

Remark A: FDR, Positive FDR (pFDR), and the q Value. In this article, we have used FDR and FDR = E[F/S] somewhat loosely. It will almost always be the case that $S = 0$ with positive probability, which implies that E[F/S] is undefined. The quantity E[$F/S|S > 0$]·Pr($S > 0$) was proposed as a solution to this problem (5), which is the result of setting $F/S = 0$ whenever $S = 0$ in the original E[F/S]. This quantity is technically called the FDR in the statistics literature. In our case we want to place a measure of significance on each feature, which is done under the assumption that the feature is called significant. Thus, the inclusion of Pr($S > 0$) is somewhat awkward. An alternative quantity, called the pFDR, was recently proposed (23), which is simply defined as pFDR = E[$F/S|S > 0$]. The q value is most technically defined as the minimum pFDR at which the feature can be called significant (24). Because m is large in genomewide studies, we have that Pr($S > 0$) $\approx$ 1 and FDR $\approx$ pFDR $\approx$ E[F]/E[S], so the distinction is not crucial here. Also, the estimate we use is easily motivated for either quantity (4, 10).

Suppose that each feature's statistic probabilistically follows a random mixture of a null distribution and an alternative distribution. Then under a fixed significance rule, the pFDR can be written as Pr(feature i is truly null|feature i is significant), for any $i = 1, \ldots, m$ (3). Similarly, the false positive rate can be written as Pr(feature i is significant|feature i is truly null), for any $i = 1, 2, \ldots, m$. Notice the similarity between the pFDR and false positive rate: the arguments have simply been swapped in the conditional probabilities. This connection is the motivation for calling our proposed quantity q value. Indeed, the p value of a feature is technically defined to be the minimum possible false positive rate when calling that feature significant (26). Likewise, the q value is based on the minimum possible pFDR.

Remark B: General Algorithm for Estimating q Values. There is a tradeoff between bias and variance in choosing the λ to use in $\hat{\pi}_0(\lambda)$. For well formed p values, it should be the case that the bias of $\hat{\pi}_0(\lambda)$ decreases with increasing λ, the bias being the smallest when $\lambda \rightarrow 1$ (4). Therefore, the method we use here is to estimate $\lim_{\lambda \rightarrow 1} \hat{\pi}_0(\lambda) \equiv \hat{\pi}_0(\lambda = 1)$. In doing so, we will borrow strength across the $\hat{\pi}_0(\lambda)$ over a range of λ, giving an implicit balance between bias and variance.

Consider Fig. 3, where we have plotted $\hat{\pi}_0(\lambda)$ 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 $\hat{\pi}_0(\lambda)$ 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. It can be seen from Fig. 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. For a variety of simulations and forms of dependence (data not shown), this method performed well, often eliminating all bias in $\hat{\pi}_0$.

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 $\lambda = 0, 0.01, 0.02, \ldots, 0.95$, calculate
$$\hat{\pi}_0(\lambda) = \frac{\#\{p_j > \lambda\}}{m(1-\lambda)}.$$
3. Let $\hat{f}$ be the natural cubic spline with 3 df of $\hat{\pi}_0(\lambda)$ on λ.
4. Set the estimate of π_0 to be
$$\hat{\pi}_0 = \hat{f}(1).$$
5. Calculate
$$\hat{q}(p_{(m)}) = \min_{t \geq p_{(m)}} \frac{\hat{\pi}_0 m \cdot t}{\#\{p_j \leq t\}} = \hat{\pi}_0 \cdot p_{(m)}.$$
6. For $i = m-1, m-2, \ldots, 1$, calculate
$$\hat{q}(p_{(i)}) = \min_{t \geq p_{(i)}} \frac{\hat{\pi}_0 m \cdot t}{\#\{p_j \leq t\}} = \min\left(\frac{\hat{\pi}_0 m \cdot p_{(i)}}{i}, \hat{q}(p_{(i+1)})\right).$$
7. The estimated q value for the ith most significant feature is $\hat{q}(p_{(i)})$.

Remark C: Analysis of the Hedenfalk *et al.* Data. The data from ref. 14 can be obtained at http://research.nhgri.nih.gov/microarray/NEJMSupplement. The data consist of 3,226 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 interquartile ranges away from the interquartile range of all of the data and did not seem trustworthy for this example. This left $m = 3{,}170$ genes.

We tested each gene for differential expression between these two tumor types by using a two-sample t statistic. Let the $\log_2$ expression value from the jth array and the ith gene be denoted by x_{ij}. Then $\bar{x}_{i2} = 1/n_2 \Sigma_{j \in BRCA2}\, x_{ij}$ and $s_{i2}^2 = 1/(n_2 - 1)\Sigma_{j \in BRCA2}(x_{ij} - \bar{x}_{i2})^2$ are the sample mean and variance for gene i among the arrays taken from *BRCA2* tumors. We can similarly define $\bar{x}_{i1}$ and s_{i1}^2 to be the sample mean and variance for the ith gene among the *BRCA1* tumor arrays. The two-sample t statistic for the ith gene, allowing for the possibility that the tumors have different variances, is then
$$t_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}}}$$
for $i = 1, 2, \ldots, 3{,}170$.

We next calculated null versions of $t_1, t_2, \ldots, t_{3170}$ when there is no differential gene expression. Because it is not clearly valid 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, $i = 1, 2, \ldots, 3{,}170$ 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 by using the algorithm presented above. All results, including the computer code used to analyze the data, can be found at http://genomine.org/qvalue/results.html.

Remark D: Theoretical Properties. Some mathematical results hold under "weak dependence" of the p values (or features in the genome). These mathematical results indicate that our method yields conservative q-value estimates. The conservative property is desirable because one does not want to underestimate the true q values. (For the same reason one would not want to underestimate a p value.)

Suppose that with probability 1, we have $S(t)/m \to G(t)$ and $F(t)/m_0 \to G_0(t)$ for each $t \in [0, 1]$ as $m \to \infty$, where G and G_0 are continuous functions. In words, this says that the empirical distribution functions of the observed p values and null p values converge pointwise to some continuous functions. Weak dependence is defined as dependence that allows this pointwise convergence. (As a rule of thumb, the more local the dependence is, the more likely it is to meet the weak dependence criterion.) Also suppose that $G_0(t) \leq t$ (i.e., uniform distribution or more conservative), and that m_0/m converges. If we constrain $\hat{\pi}_0 \geq \min_{\ddot{e} \in \mathcal{R}} \hat{\pi}_0(\lambda)$ (which should usually be the case), then it can be shown that for any $\delta > 0$,
$$\lim_{m \to \infty} \min_{p_i \geq \delta} [\hat{q}(p_i) - q\ \text{value}(p_i)] \geq 0.$$
which means that the estimated q values are simultaneously 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
$$\lim_{m \to \infty} \frac{\#\{\text{false positive } \hat{q}(p_i) \leq \alpha\}}{\#\{\hat{q}(p_i) \leq \alpha\}} \leq \alpha,$$
which means that if we call all genes with q values $\leq \alpha$, then in the long run the FDR will be $\leq \alpha$. The proofs of these claims follow from minor modifications to some of the main results in ref. 10.

1. Morton, N. E. (1955) *Am. J. Hum. Gen.* **7,** 277–318.
2. Lander, E. S. & Kruglyak, L. (1995) *Nat. Genet.* **11,** 241–247.
3. Storey, J. D. (2003) *Ann. Stat.*, in press.
4. Storey, J. D. (2002) *J. R. Stat. Soc. B* **64,** 479–498.
5. Benjamini, Y. & Hochberg, Y. (1995) *J. R. Stat. Soc. B* **85,** 289–300.
6. Yekutieli, D. & Benjamini, Y. (1999) *J. Stat. Plan. Inf.* **82,** 171–196.
7. Benjamini, Y. & Hochberg, Y. (2000) *J. Ed. Behav. Stat.* **25,** 60–83.
8. Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. (2001) *J. Am. Stat. Assoc.* **96,** 1151–1160.
9. Genovese, C. & Wasserman, L. (2002) *J. R. Stat. Soc. B* **64,** 499–517.
10. Storey, J. D., Taylor, J. E. & Siegmund, D. (2003) *J. R. Stat. Soc. B*, in press.
11. Tzeng, J. Y., Byerley, W., Devlin, B., Roeder, K. & Wasserman, L. (2003) *J. Am. Stat. Assoc.* **98,** 236–246.
12. Tusher, V., Tibshirani, R. & Chu, C. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 5116–5121.
13. Tatusov, R. L., Altschul, S. F. & Koonin, E. V. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 12091–12095.
14. Hedenfalk, I., Duggan, D., Chen, Y. D., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., *et al.* (2001) *N. Engl. J. Med.* **344,** 539–548.
15. Slonim, D. K. (2002) *Nat. Genet.* **32,** Suppl., 502–508.
16. Blencowe, B. J. (2000) *Trends Biochem. Sci.* **25,** 106–110.
17. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. (2002) *Science* **297,** 1007–1013.
18. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. (2002) *Science* **296,** 752–755.
19. Churchill, G. A. & Doerge, R. W. (1994) *Genetics* **138,** 963–971.
20. 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., *et al.* (2002) *Science* **298,** 799–804.
21. Kolodner, R. (1996) *Genes Dev.* **10,** 1433–1442.
22. Liu, H. T., Wang, Y. G., Zhang, Y. M., Song, Q. S., Di, C. H., Chen, G. H., Tang, J. & Ma, D. L. (1999) *Biochem. Biophys. Res. Commun.* **254,** 203–210.
23. Hishikawa, K., Oemar, B. S., Tanner, F. C., Nakaki, T., Luscher, T. F. & Fujii, T. (1999) *J. Biol. Chem.* **274,** 37461–37466.
24. Efron, B., Storey, J. D. & Tibshirani, R. (2001) *Technical Report 2001-217* (Stanford Univ., Palo Alto, CA).
25. Efron, B. & Tibshirani, R. (2002) *Genet. Epidemiol.* **23,** 70–86.
26. Lehmann, E. L. (1986) *Testing Statistical Hypotheses* (Springer, New York), 2nd Ed.