Bioinformatics, 31(14), 2015, 2303–2309
doi: 10.1093/bioinformatics/btv104
Advance Access Publication Date: 2 March 2015
Original Paper

OXFORD

Gene expression

Estimating the proportion of true null hypotheses when the statistics are discrete

Isaac Dialsingh[1], Stefanie R. Austin[2] and Naomi S. Altman[2,*]

[1]Department of Mathematics and Statistics, The University of the West Indies, St. Augustine Campus, Trinidad and Tobago and [2]Department of Statistics, The Pennsylvania State University, State College, PA 16802-2111, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on August 18, 2014; revised on February 10, 2015; accepted on February 11, 2015

Abstract

Motivation: In high-dimensional testing problems π_0, the proportion of null hypotheses that are true is an important parameter. For discrete test statistics, the *P* values come from a discrete distribution with finite support and the null distribution may depend on an ancillary statistic such as a table margin that varies among the test statistics. Methods for estimating π_0 developed for continuous test statistics, which depend on a uniform or identical null distribution of *P* values, may not perform well when applied to discrete testing problems.

Results: This article introduces a number of π_0 estimators, the regression and 'T' methods that perform well with discrete test statistics and also assesses how well methods developed for or adapted from continuous tests perform with discrete tests. We demonstrate the usefulness of these estimators in the analysis of high-throughput biological RNA-seq and single-nucleotide polymorphism data.

Availability and implementation: implemented in R

Contact: nsa1@psu.edu or naomi@psu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In multiple testing inferential problems, we want to select which among a large number of hypotheses are true. The proportion of truly null hypotheses, π_0, plays a critical role in adjusting for multiple testing, gives a benchmark for the number of statistically significant tests which should be discovered and is an important measure of effect size (Black, 2004).

We assume that m null hypotheses, $H_{01}, \ldots\ldots, H_{0m}$, are being tested corresponding to m parameters or features. The ith hypothesis is associated with an observed test statistic X_i and observed P value p_i. A number of methods are available for estimating π_0 when X_i and hence p_i are continuous (Benjamini and Hochberg, 2000; Markitsis and Lai, 2010; Pounds and Cheng, 2004, 2006; Pounds and Morris, 2003; Wang *et al.*, 2011; Zhang, 2011). These methods rely on modeling the mixture distribution arising from the mix of truly null and non-null tests using various parametric and non-parametric approaches (Langass *et al.*, 2005).

Let $\mathcal{N}$ be the set $\{i|H_{i0}\,\text{is true}\}$. In many cases with continuous response, such as gene expression microarray studies and pixel-wise intensity analysis of images, it is reasonable to believe that the distribution of X_i is the same for all $i \in \mathcal{N}$—for example, we might perform a t-test for each feature and expect the null distribution to be Student's t. Alternatively, we can use p_i as the test statistic, in which case the null distribution is Uniform (0,1). Although it is not known which hypotheses are in $\mathcal{N}$, the empirical distribution of the test statistics is a mixture of m_0 observations from the null and $m_1 = m - m_0$ observations from an alternative distribution, so that deviations of the empirical distribution from the known null can be used to estimate π_0 (Storey, 2003, Strimmer, 2008).

For the discrete case, the situation is more complicated. Each X_i can take on only a finite number of values, which often depend on an ancillary statistic which varies with i. For example, for Fisher's exact test and other tests of independence in two-way tables, the ancillary is a table margin. As a result, even for $i \in \mathcal{N}$, the distribution

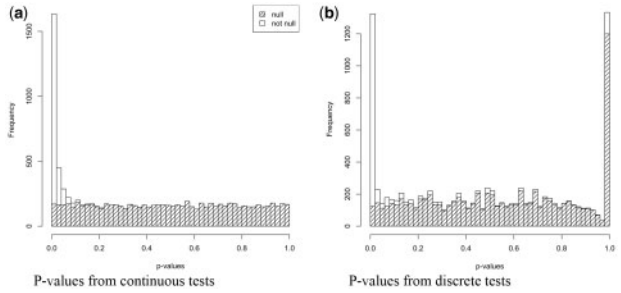

Fig. 1. *P* values from discrete and continuous tests with $\pi_0 = 0.80$. a) P-values for continuous tests b) P-values for discrete tests

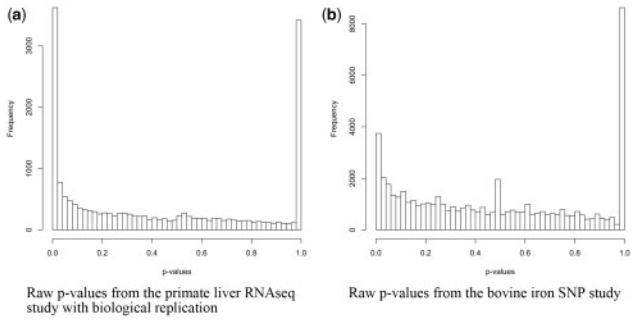

Fig. 2. *P* values from real data. **(a)** Raw *P*-values from the primate liver RNAseq study with biological replication. **(b)** Raw *P*-values from the bovine iron SNP study

of X_i varies with i and the empirical distribution of the observed statistics is a mixture.

For example, Figure 1 displays P values from simulated gene expression data with $\pi_0 = 0.8$. For each feature (gene), there are two treatments and the null hypothesis is no difference in mean expression level. The P values arising from the null distribution are displayed in gray and the P values arising from the non-null features are stacked in white. The two histograms look very different.

In Figure 1a, the P values come from two-sample t-tests, where 20% of the P values come from various non-central t-distributions. Note the relative uniformity of the null (gray) P values versus the non-null (white) P values, which are skewed toward small values. In contrast, Figure 1b displays P values coming from Fisher's exact tests, where 20% of the P values come from various non-central hypergeometric distributions. In this case, both the null and non-null P values are highly non-uniform and there is a non-zero probability of yielding a P value equal to 1 under the null and alternative distributions.

The differences between continuous and discrete tests are apparent with real data. Figure 2 shows P values from real studies: Figure 2a differential expression analysis of the primate liver study using RNA-seq technology with three biological replicates (Blekhman *et al.*, 2010), used in Section 3.1 and Figure 2b the bovine iron single-nucleotide polymorphism (SNP) study (Mateescu *et al.*, 2013), used in Section 3.2.

In this article, we propose new methods for estimating π_0 for discrete tests and compare them to some popular methods developed for continuous data. We also apply these methods to determine the proportion of differentially expressing genes in primate livers and for selecting SNPs associated with bovine muscular iron levels.

2 Background

We assume that after performing the m tests and observing test statistics $X_1 \cdots X_m$, we decide whether to reject each null hypothesis. In a slight abuse of notation, we will write $H_{0i} = 1$ when the ith null hypothesis is true, and $H_{0i} = 0$ when it is false. The number of true null hypotheses is $m_0 = \sum_{i=1}^{m} H_{0i}$ and $\pi_0 = m_0/m$.

In many testing situations, $X_i|H_{0i} = 1$ are identically distributed. In this case, we denote the null density (or mass) function of $X_i|H_{0i} = 1$ as $f_0(x)$. In other cases, $X_i|H_{0i} = 1$ depends on an ancillary statistic A_i, which is observable and independent of the value of H_{0i}. In this case, the conditional null density function is $f(x_i|H_{0i} = 1, A_i = a_i) = f_{i0}(x)$ and we can write the null distribution as $f_0(x) = \sum_{i=1}^{m} f_{i0}(x)\text{Prob}(A_i = a_i)$. The alternative distribution of $X_i|H_{0i} = 0$ usually depends on i. However, in the same spirit, we will write $f_1(x)$ to mean the mixture distribution of alternatives. Then, we can consider the marginal density of X:

$$f(x) = \pi_0 f_0(x) + (1 - \pi_0) f_1(x) \tag{1}$$

This representation of the marginal density is central to estimation of π_0 using the m observed values of the test statistic.

2.1 Continuous tests

When the test statistic is continuous, it is common to assume that $X_i|H_{0i} = 1$ are identically distributed. A number of estimators of π_0 are available but we will focus on three popular estimators that all use the P value as the test statistic. In this case, $f_0(P) = 1$ for $0 \le P \le 1$.

We describe the estimators and also give some heuristics about their use with discrete P values.

2.1.1 Storey's method

Storey (2002) is one of the most popular methods for estimating π_0 and has been shown to estimate π_0 well for continuous test statistics. The estimator is:

$$\hat{\pi}_0(\lambda) = \frac{\#\{P_i | P_i > \lambda\}}{m(1 - \lambda)} \tag{2}$$

where $\lambda \in [0, 1]$ is a tuning parameter and $\#(S)$ is the number of elements in set S. Although λ is sometimes selected adaptively from the data, we used $\lambda = 0.5$. Note that if all the tests are null, $E(\#\{P_i | P_i > \lambda\}) = m(1 - \lambda)$.

The absolute deviance from true π_0 in the Storey estimator is larger for smaller values of π_0 unless there is perfect power to detect the non-null hypotheses at level λ. As well the pile-up of P values at $P = 1$ for discrete tests has the effect of creating further over-estimation and can yield $\hat{\pi}_0(\lambda) > 1$. We actually use min $(1, \hat{\pi}_0(\lambda))$ as the estimator where min (a, b) is the minimum of a and b.

2.1.2 Nettleton's method

Nettleton *et al.* (2006) presents an algorithm for estimating π_0 by estimating the proportion of observed P values that follow the uniform distribution. The idea is to create bins in the interval $[0, 1]$ and use the excess of expected versus observed P values in those bins to iteratively update the estimate of m_0. The algorithm is provided in detail in Nettleton *et al.* (2006).

For the implementation of Nettleton's method, we used the R function from Nettleton's website, http://www.public.iastate.edu/~dnett/microarray/multtest.txt. The number of bins is a tuning parameter. In our simulation studies, we found that partitioning the P values into $B = 25$ bins gave good results. Note that this method relies on the heuristic that the P values of the hypotheses $i \notin \mathcal{N}$ should be skewed toward zero. For discrete data, the pile-up of P values at $P = 1$ depletes the other bins of the histogram. Since the algorithm utilizes the bins with P close to zero, again it seems that

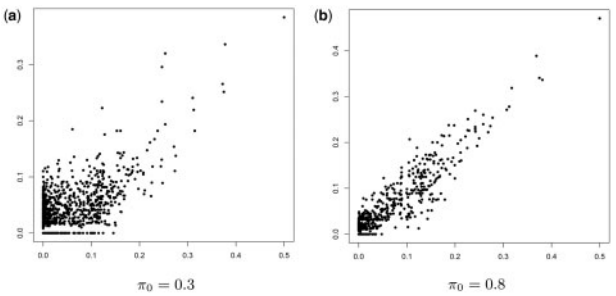

Fig. 3. Plots of $\hat{\phi}_{jt}$ versus ϕ_{0jt} for one random sample of RNA-2, $m = 10\,000$, data for two different π_0 values. (**a**) $\pi_0 = 0.3$. (**b**) $\pi_0 = 0.8$

there may be over-estimation of π_0. On the other hand, since the P value histogram can be erratic with peaks in the center, it is not as clear whether Nettleton's method is prone to inaccuracy, and if it is, whether it will over- or under-estimate π_0.

2.1.3 Pounds and Cheng method

Pounds and Cheng (2006) proposes an estimator of π_0 when the test statistics are continuous and two sided. Their estimator can be summarized as $\hat{\pi}_0 = \min(1, 2\overline{p})$ where $\overline{p} = \frac{1}{m}\sum_{i=1}^{m} p_i$. This estimator tends to over-estimate π_0, but the difference is small when $Pf_1(p)$ is small, that is when $f_1(p)$ has most of its mass at small P. Like the Storey method, the difference is also smaller when π_0 is close to 1. They also proposed this estimator for discrete two-sided P values.

2.2 Discrete tests

Equation (1) readily creates heuristics for estimating π_0 for continuous tests because of the assumption that $X_i, i \in \mathcal{N}$ is an i.i.d. sample from f_0. In most discrete testing situations, the X_i's and the P values have null and alternative distributions that depend on the realization of the ancillary statistic A_i. We will continue to use p_i as our test statistic as it is a transformation of X_i which is on a convenient scale for plotting.

The achievable values of P under the alternative distribution are a subset of those achievable under the null. Usually, this will include mass at $P = 1$ albeit with lower probability than under the null hypothesis. Hence, even if none of the hypotheses are true, we expect some mass at $P = 1$, that is $f_1(1) > 0$.

Since the number of achievable P values is finite, for each value of A_i, there is a minimum achievable P value. For example, for Fisher's exact test for 2×2 tables, an ancillary is A_i, the total of the first row. If $A_i = 1$, the minimum and maximum P value from Fisher's exact test is $P = 1$, while for $A_i = 2$, if the column totals are equal, the maximum is $P = 1$ and the minimum is $P = 0.5$. For $A_i = 1$ or 2 we can never reject the null hypothesis. For discrete tests, the power depends on the ancillary statistic, and in many situations, the region around $P = 0$ is depleted compared with the Uniform(0,1) distribution, even when $\pi_0 < 1$, and regions in the middle of the (0, 1) interval can have greatly increased mass. For simulation studies, we have considered A_i to be random and generated new values with each round of simulation. However, for estimation, we condition on the observed values of A_i in each realization.

We consider several methods for estimating π_0 each of which takes advantage of the observed values of the ancillary statistic. We assume that π_0 does not depend on the ancillary, which has d unique observed values $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_d$ each corresponding to a known null distribution $f_{01}, f_{02}, \ldots, f_{0d}$ respectively. Each null distribution f_{0j} has a finite set of achievable P values $S_j = \{S_{j1}, S_{j2}, \ldots, S_{jT_j}\}$ with $S_{j1} < S_{j2} < \cdots < S_{jT_j} = 1$ and with corresponding probabilities $\phi_{0j1}, \phi_{0j2}, \ldots, \phi_{0jT_j}$. Note that set S_j is the support of f_{0j}, with $\phi_{0j1} = S_{j1}$ and $\phi_{0jk} = S_{jk} - S_{j,k-1}$ for $2 \le k \le T_j$. The hypotheses are partitioned into sets, so that if the null distribution of the ith P value is known to be f_{0j}, then the corresponding support is S_j. The alternative distributions have the same support but unknown probabilities.

2.2.1 Regression method

A new method proposed by Dialsingh (2012), the regression method is applicable when m is large enough, so that $m\text{Prob}(A_i = \mathcal{A}_j) > 1$ for at least one ancillary statistic with at least three achievable P values. Then we have $\text{Prob}(P_i = S_{jt} | A_i = \mathcal{A}_j, H_{0i} = 1) = \phi_{0jt}$, which is known. When the null hypothesis is false, we denote the P value mass function as $\text{Prob}(P_i = S_{jt} | A_i = \mathcal{A}_j, H_{0i} = 0) = \phi_{1jt}$ where the probabilities ϕ_{1jt} come from an unknown mixture of alternatives with $A_i = \mathcal{A}_j$. We assume that the distribution of $P_i = S_{jt} | A_i = \mathcal{A}_j, H_{0i} = 0$ is stochastically smaller than the distribution of $P_i = S_{jt} | A_i = \mathcal{A}_j, H_{0i} = 1$. The alternative distribution of $P_i | H_{0i} = 0$ usually depends on i, but continuing spirit of Equation 1, we will write ϕ_{1jt} to mean the mixture distribution of alternatives:

$$\text{Prob}(P_i = S_{jt} | A_i = \mathcal{A}_j) = \phi_{jt} = \pi_0 \phi_{0jt} + (1 - \pi_0)\phi_{1jt} \qquad (3)$$

When the set $D_j = \{H_{0i} : A_i = \mathcal{A}_j\}$ is sufficiently large, ϕ_{jt} can be estimated from the data. From Equation (3), $\text{Prob}(P_i = S_{jt} | A_i = \mathcal{A}_j)$ is an approximately linear function of ϕ_{0jt}, with slope π_0 (Fig. 3). Of the M_j hypotheses in D_j, we observe that K_{jt} of the hypotheses have P value S_{jt}, so that

$$\hat{\phi}_{jt} = \frac{K_{jt}}{M_j} \qquad (4)$$

where M_j is the cardinality of set D_j and K_{jt} is the number of hypotheses in D_j that have P value S_{jt}. We know that $E(\hat{\phi}_{jt}) = \phi_{jt} = \pi_0 \phi_{0jt} + (1 - \pi_0)\phi_{1jt}$ and that ϕ_{0jt} is known. We regress $\hat{\phi}_{jt}$ on ϕ_{0jt}. The slope is an estimator of π_0.

Figure 3 illustrates the positive linear relationship between $\hat{\phi}_{jt}$ and ϕ_{0jt}. As π_0 increases, the relationship gets stronger: for random samples of the RNA-2, $m = 10\,000$ data of Section 4.1, the correlation increases from 0.51 for $\pi_0 = 0.1$ to 0.97 for $\pi_0 = 1$. The cluster of points near the origin for low values of π_0 due to the non-null distributions increases the variance of the slope and degrades the performance.

Any consistent estimator of the slope is a suitable plug-in estimate in our calculations. We use ordinary least squares. The results in Section 3 of Eicker's paper (Eicker, 1967) prove the consistency of the estimated slope.

To obtain estimates $\hat{\phi}_{jt}$, we need M_j to be sufficiently large. In our simulation study, we arbitrarily set a lower bound and used only the sets D_j with $M_j \ge 10$.

Remark 1: *Since $\hat{\beta}$, the estimate of the slope, can be <0 or >1, we truncate $\hat{\pi}_0$ to the interval* [0, 1].

2.2.2 Bancroft method

The method developed by Bancroft *et al.* (2013) is an adaptation of Nettleton's method to discrete tests. Nettleton's method is applied to each of the d sets of P values corresponding to the d unique null distributions, $f_{01}, f_{02}, \ldots, f_{0d}$ to come up with an estimate of m_{0j}, the number of tests corresponding to true null hypotheses in set D_j using the achievable P values as the 'bins'. Note that the initial estimate of m_{0j} would be M_j, the total number of tests in D_j.

Then, the estimate of π_0 becomes

$$\hat{\pi}_0 = \frac{\hat{m}_{01} + \hat{m}_{02} + \cdots + \hat{m}_{0d}}{m} \quad (5)$$

This estimator can be computationally intense as m, the total number of hypotheses, grows. Additionally, when there exist unique distributions with relatively large number of bins (support values) compared with the number of P values that follow that distribution, the estimator will be larger than the true π_0, especially when very few observed P values fall into the leftmost bins.

Remark 2: *Due to the relative computational intensity of Bancroft's algorithm, for our simulations, we modified the method: If the minimum expected cell count for a table (distribution) is three or fewer, we utilize this discrete method. If all expected cell counts are at least 4, we assume uniformity and combine the P values of those distributions into a single set and utilize Nettleton's continuous method. For example, if d_{disc} of the d unique distributions have a minimum expected cell count of three or fewer, then there will be $(d_{\text{disc}} + 1)$ total estimators of the number of true null hypothesis tests, which are then summed and divided by m to obtain the estimator of π_0 for that simulation. The algorithm is provided in the* Supplementary Materials.

2.2.3 T methods

When S_j is small, f_{0j} yields a component of the null distribution that is far from uniform. The corresponding values of the ancillary statistic typically yield tests with zero power regardless of the effect size, because the smallest achievable P value is larger than the boundary of typical rejection regions, say $P < 0.05$ or $P < 0.01$. Tarone (1990) noted that for improving the power of multiple comparisons adjustments, tests with zero power should be filtered out. We introduce methods that filter tests with zero power and denote them as 'T' methods.

Filtering out zero power tests improves the uniformity of the P values. This is particularly effective when the distribution of the ancillary is highly skewed toward small values. For RNA-seq data, the ancillary is typically the total reads per feature, whereas for SNPs, it is the number of individuals with the rare variant. The ancillary distribution is usually very skewed with most features having small values. Filtering rare RNA-seq features and SNPs dramatically reduces the peak at $P = 1$, as well as intermediate peaks at P values associated with values of the small ancillary statistics.

Let $\mathcal{A}$ be the set of null hypotheses for which the ancillary statistic passes the power threshold. Since our *a priori* assumption is that the ancillary statistic A_i is independent of whether or not $H_{0i} = 1$, the proportion of truly null tests in $\mathcal{A}$ is π_0 and applying an estimator restricted to $\mathcal{A}$ does not bias the estimate, although it does decrease the number of test statistics. We denote the method of restricting an estimator M to the (possibly random) set $\mathcal{A}$ as an $M - T$ estimator. In our simulation study, we use the T estimator for each of the Storey, Pounds and Nettleton estimators.

3 Application to real datasets

3.1 Application of methods to a primate RNA-seq dataset

The estimation methods were applied to RNA-seq data obtained from gene expression levels in livers from three primate species (human, chimpanzee and Rhesus monkey), using three male and three female samples from each species (GEO dataset GSE17274) (Blekhman *et al.*, 2010). We considered two comparisons: male humans versus male chimpanzees and male humans versus male Rhesus monkeys. On the basis of the evolutionary divergence, we expect humans to be more similar to chimpanzees than to Rhesus monkeys and hence expect a larger value for π_0 for this comparison.

Since the data include biological replication, which is likely to induce extra-Poisson variation, we fitted a Negative Binomial Model using the R package `edgeR` (Robinson *et al.*, 2010). `edgeR` uses a sophisticated method of dispersion shrinkage, which increases the power of the tests by using information from all the tests. Because of this, it is not clear how to compute an ancillary statistic to use for the regression and Bancroft methods. We used only procedures developed for continuous tests and their corresponding T-methods.

Since biological variation tends to induce *over*-dispersion in the counts, tests based on a Poisson distribution are anti-conservative. Therefore, removing tests that have zero power under the Poisson assumption will be conservative, while still improving the estimator. Therefore, we remove genes from the analysis that have fewer than six reads summed over all the samples. Even for the 'T' methods, however, there is some question about how to proceed, because the dispersion shrinkage uses all the genes. Therefore, genes removed due to lack of power need to be removed from the entire analysis, not just from the π_0 estimation stage. We take a compromise stance. When using 'T' methods, we first use all the genes that have total reads above the threshold, even if those reads are scattered over multiple treatments. We then do the pairwise comparison using only those genes that are above a more conservative threshold based on only the samples involved in the comparison.

A total of 20 689 features were used in the study, but 2803 were not detected in any sample. The first level of filtering, in which we retained only genes with at least 6 reads over all 18 samples, removes 5273 genes from the analysis. The second level of filtering, in which we conservatively retained genes with at least 4 reads in the 6 samples used in each the pairwise comparison, removed a further 860 genes from the comparison with chimpanzee and 864 from the comparison with Rhesus monkey.

It is common practice to remove low expressing genes from the differential expression analysis. Using the recommendation in Robinson *et al.* (2010), genes with fewer than 25 reads would be removed—for both comparisons the smallest P value among these genes is <0.00005. Many of these genes could be detected as non-null if left in the analysis.

The estimation of π_0 for each comparison is presented in Table 1. We see firstly that as expected from evolutionary history, all the methods estimate a higher value of π_0 for the comparison of humans versus chimpanzee than versus Rhesus monkeys. We also see that for each method, the corresponding 'T' method estimates a lower value of π_0 than the unfiltered method. This is expected both by theory and our simulation results. Another interesting thing to notice, however, is that Nettleton's method gives a much lower estimate of π_0 than the other two methods but that Nettleton's 'T' method is quite similar to the other two 'T' methods.

3.2 Application of methods to a bovine SNP dataset

Samples of muscle tissue from 2285 Angus cattle were obtained for genotyping on 53 367 SNPs. Iron concentration in the tissue was also measured. The study and results for the entire sample are described in Mateescu *et al.* (2013).

One hundred samples were selected at random from the 2285 to use as an example. Animals with more than 5000 missing SNPs and SNPs with more than 20 missing values or just one genotype were removed. The remaining 95 samples were divided into high- and

Table 1. Estimates of π_0 for both comparisons

Method	Humans versus chimpanzees	Humans versus rhesus monkeys
Storey	0.94	0.81
Pounds	0.97	0.85
Nettleton	0.84	0.73
Storey T	0.72	0.57
Pounds T	0.78	0.63
Nettleton T	0.74	0.57

Table 2. Estimates of π_0 for bovine SNP analysis for muscle iron content

Method	Raw *P* values	'T' method
Storey	0.78	0.76
Pounds	0.86	0.83
Nettleton	0.81	0.77

low-iron groups, splitting at the median, giving 47 samples with iron level above the median and 48 below. For each of the 48 816 SNPs, we computed Fisher's exact test.

Because of missing values, it was not possible to find enough tables with the same margin to use the regression or Bancroft methods. Tables for which the sum of the two smallest margins is <5 have zero power at critical value 0.05 when there are no missing values for most animals. This increases mildly for SNPs with many missing values, but for simplicity we used this cut-off for filtering throughout for the 'T' methods.

P values for the detected SNPs are displayed in Figure 2b. Filtering removes much of the mass at $P = 1$ and at $P = 0.5$.

The estimated values of π_0 are listed in Table 2. As in other analyses, the estimate of π_0 is always lower for the 'T' methods. However, despite the dramatic change in the height of the histogram at $P = 1$, the filtered and unfiltered methods do not vary as dramatically as for the RNA-seq data. It seems reasonable to estimate π_0 at around 80% for these data.

4 Simulation studies

Much of the work in multiple testing since the seminal paper by Benjamini and Hochberg (1995) has been motivated by problems in high-throughput biology, in which tens or hundreds of thousands of biological features (mRNAs, proteins and SNPs) are measured simultaneously. Typical questions of interest are whether there are quantitative differences among two or more conditions. We focus on two types of study in which the response for each variable is a count: RNA-seq data for gene expression and SNP data for genotype association. In this article, we focus on discrete data with just two conditions (e.g. the treatment and control scenario).

For the simulation studies, we generated data from two RNA-seq scenarios without biological replication (RNA-seq studies with biological replication was discussed in the real data example) and two SNP scenarios. For each configuration of each type of data, we generated 1000 datasets for both $m = 1000$ and 10 000 hypotheses. The proportion of nulls used in the simulations included $0.10, 0.20, \ldots, 0.90, 0.95, 1.00$. The algorithms for generating the simulation data are in the Supplementary Materials.

For both types of data, we will focus on the hypothesis that for each feature the mean frequency is the same for the treatment and control. We use Fisher's exact test as the test statistic, although the chi-squared test could also be used when the table entries are sufficiently large. Our objective is to estimate π_0, and we compare eight estimators: (i) Nettleton (ii) Storey (iii) Pounds (iv) Nettleton-T (v) Storey-T (vi) Pounds-T (vii) Regression and (viii) Modified Bancroft and Nettleton.

Table 3. Sample 2×2 table from RNA-seq data showing the reads for the ith feature

	Reads from treatment	Reads from control	Total
Reads from feature *i*	l_i	$n_i - l_i$	n_i
Reads from all others	$N_{T+} - l_i$	$N_{C+} - (n_i - l_i)$	$(N_{T+} + N_{C+}) - n_i$
Total reads	N_{T+}	N_{C+}	$N = N_{T+} + N_{C+}$

4.1 Simulated RNA data

The simulated RNA-seq data were based on the real dataset analyzed in Section 3.1. We use a discretized log-Normal(4,2) in condition 'RNA-1' and a discretized log-Normal(3,2) in condition 'RNA-2'. The latter distribution produces a slightly larger number of genes with small total counts. In the main simulation study, the features were simulated independently. In an additional set of simulations suggested by a referee, we generated the features in correlated clusters. An example of data for the *i*th feature is presented in Table 3. *N*, the total count in the two samples, is typically in the tens of millions. The column totals N_{T+} and N_{C+} are the same for each feature but differ from each other. The ancillary statistic is the total of the first row, $A_i = n_i$.

4.1.1 Results of RNA-seq simulations

To investigate the performance of the eight methods in estimating π_0, we obtained the distribution of $\hat{\pi}_0$ for each data configuration. The simulation results are presented in Figure 4. Because over-estimation is generally considered a less serious error than under-estimation, the 25th percentile of the simulations is shown. The plots for $m = 1000$ and $m = 10\,000$ are quite similar except that the under-estimation of the Nettleton methods for large π_0 is more pronounced for smaller m. We show only the results for $m = 10\,000$.

4.1.2 Discussion of the results of RNA-seq simulations

All the methods over-estimate for small values of π_0. The regression and Nettleton-T methods have the smallest deviation from π_0, especially for large values of π_0, whereas Pounds-T and Storey-T have quite similar performance to the regression method for small π_0 but over-estimate for large π_0. The variances of the estimators are not shown, but they are quite similar except near $\pi_0 = 1$ where the truncation of the more positively deviated estimators squeezes them to 1.0. For all combinations of simulation conditions and m, Pounds method produces the highest mean estimate of π_0 and the Nettleton-T produces the lowest. The Nettleton, Nettleton-T and Modified Bancroft/Nettleton methods can under-estimate more egregiously, especially for smaller values of m (not shown), and hence cannot be recommended.

Of the remaining methods, all over-estimate for $\pi_0 < 0.9$. This is not surprising: first, we have already seen that Pounds' and Storey's methods have intrinsic positive deviation not removed by filtering out the zero-power tests. Pounds' and Storey's methods are highly correlated, with even higher correlation after filtering but are less correlated with the regression method.

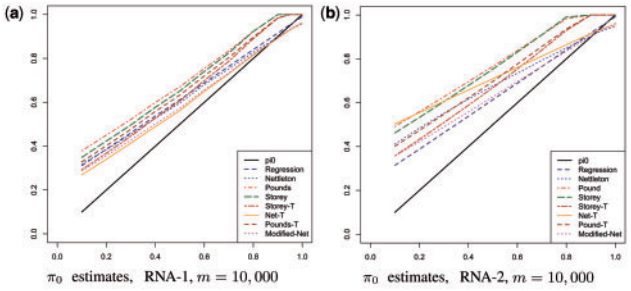

Fig. 4. The 25th percentile of $\hat{\pi}_0$ for different estimators for the RNA-seq simulations. (**a**) π_0 estimates, RNA-1, $m = 10\,000$. (**b**) π_0 estimates, RNA-2, $m = 10\,000$

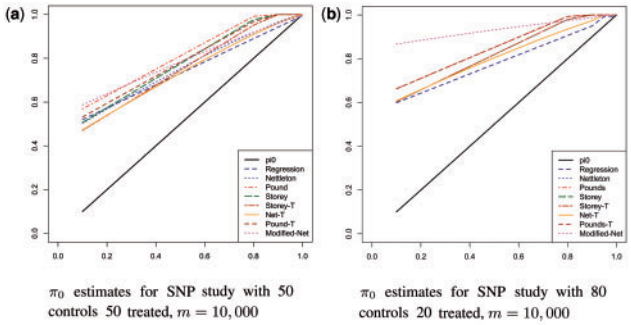

Fig. 5. The 25th percentile of $\hat{\pi}_0$ for different estimators for the SNP simulations. (**a**) π_0 estimates for SNP study with 50 controls 50 treated, $m = 10\,000$. (**b**) π_0 estimates for SNP study with 80 controls 20 treated, $m = 10\,000$

Table 4. Sample SNP 2×3 table for the ith SNP

	GG	Gg	gg	Total
Treatment	n^i_{11}	n^i_{12}	n^i_{13}	N_{T+}
Control	n^i_{21}	n^i_{22}	n^i_{23}	N_{C+}
	A^i_{GG}	A^i_{Gg}	A^i_{gg}	$N = N_{T+} + N_{C+}$

4.2 Simulated SNP data

The simulations for the SNP data are based on a random sample of 100 animals from the bovine SNP dataset (Mateescu *et al.*, 2013) discussed in Section 3.2. Linkage disequilibrium was induced by sampling in sets of five tables. For the simulated data, we considered two scenarios: one with 50 animals in each of the treatment and control groups, and one with 20 animals in the treatment group and 80 in the control group. The data for a feature are summarized as a count of individuals for each genotype for each treatment as in Table 4. The association between genotype and treatment is tested using a Fisher's exact test or chi-squared test for each SNP.

4.2.1 Results of SNP simulations

To investigate the performance of the eight methods in estimating π_0, we obtained the distribution of $\hat{\pi}_0$ for each data configuration. The simulation results are presented in Figure 5. Because over-estimation is generally considered a less serious error than under-estimation, the 25th percentile of the simulations is shown. The plots for $m = 1000$ and $m = 10\,000$ are quite similar except for the regression method. The regression method performs very poorly for $m = 1000$, possibly because there are too few tables with the same margins. We show only the results for $m = 10\,000$.

4.2.2 Discussion of the results of SNP simulations

The performance of the eight estimators in the SNP data varied dramatically, especially considering their relatively comparable efficacy in the RNA-seq data. We see that all methods generally over-estimate until π_0 is close to 1, particularly when the sample sizes for the two conditions are very unequal. Of these, the Modified Bancroft/Nettleton method has the largest absolute difference from π_0. Though not shown here, the regression method did not perform as consistently as the others when the number of tests was $m = 1000$. Among the methods considered, Pounds, Storey and Nettleton, along with their 'T' methods, have the best performance and are very similar.

5 Discussion

There is no clear 'winner' among the π_0 estimators. In part, this is because discrete tests for count data become closer to uniform as the size of the counts increase.

The regression method requires computing f_{0d}, the null distribution conditional on the observed values of the ancillary statistic A_d for each d. These null distributions have more support points for larger values of A_d, adding to the computational burden while providing fewer observations for computing $\hat{\phi}_{jt}$. As well, it may be difficult to compute $\hat{\phi}_{jt}$ in studies in which the number of different values of the ancillary is large. Bancroft *et al.* (2013) is also very computationally complex as the number of different null distributions increases. Thus, it also is not an ideal candidate when there is a large number of unique ancillary statistics. Finally, these methods do not extend readily to continuous tests.

In contrast, the 'T' methods naturally adapt to higher counts because the filtered features are those with small counts. Hence, as frequencies increase and the P values are closer to continuous, fewer features are filtered and the 'T' method is closer to the continuous method on which it is based. The 'T' methods also adapt naturally to missing data—features with missing data are naturally those with smaller counts. As well, in complex studies such as genome-wide association studies with continuous response or RNA-seq studies with biological replication, it is still reasonable to assume that features with small counts will be associated with very low power tests. Hence, the 'T' methods are readily implemented.

For these reasons, we recommend the use of 'T' methods except in simple studies with multiple tests with the same ancillary statistic, for which the regression method may be better. Of the three 'T' methods tested here, Storey-T appears to be most resilient to different types of data, ancillary distributions and sample sizes.

It has become common practice in RNA-seq studies to filter out genes with low total counts; generally, the cut-off is selected arbitrarily. For example, the `edgeR` software (Robinson *et al.*, 2010) recommends the total for genes retained should be about 100 reads per million detected, while the primate liver paper (Blekhman *et al.*, 2010) dropped genes with median count per sample of 1 or less. The 'T' method suggests a more principled approach—select the largest P value for which the investigator would be willing to declare significance and filter out features for which the test statistic has minimum achievable P value which is higher. For 2×2 tables with large margins in one direction such as those generated by RNA-seq studies, a simple rule of thumb is that the test has zero power at $P \leq 0.05$ when the minimum margin (total reads for a gene) is <6. For 2×3 tables with large margins in one direction such as those generated by genome-wide association studies using SNPs, a simple rule of thumb is that the test has zero power if the sum of the two smallest margins is <5.

In many scenarios, tests are not independent. For the RNA-seq scenario, the primary simulation study was done using independently generated features. We also simulated 1000 runs of RNA-seq

data with correlated test statistics and $m = 1000$ tests. The algorithm used to simulate the data is provided in the Supplementary Materials. We found little difference in the performance of the π_0 estimators between the independent tests and the dependent tests; the 'T' methods still performed well overall. For the SNP scenario, the SNPs were simulated in small correlated clusters in the primary study. The results suggest that the π_0 estimators do not require independence of the tests to work well.

6 Conclusion

For discrete tests, the null distribution of the P values is discrete, and thus, estimates of π_0 need to take this into account. For multiple comparisons methods or false discovery rate estimators, it is clear that the statistical significance calls need be adjusted only for tests that have sufficient power to be significant. Hence, the 'T' methods provide the most suitable estimates of π_0 because they estimate the proportion of truly null hypotheses among the hypotheses for which significance could be achieved in the study at hand.

One question not addressed by the analysis is the interpretation of π_0. On occasion, π_0 is used as a measure of overall signal strength - for example, Zhang *et al.* (2005) used the estimated number of differentially expressed genes as an estimate of the proportion of genes involved in organ differentiation. Since genes not expressing in the organs in question cannot be involved in organ differentiation, estimates of π_0 probably should include these non-expressing genes among the null hypotheses. Similarly, SNPs that exhibit no variation in the samples clearly cannot be related to differences among the phenotypes under study. On the other hand, the very low expressing genes or genotypes present at very low frequency might be non-null, even though the low detection levels in the study do not allow enough power to determine this. For this type of study, it is very important to distinguish among those hypotheses for which the null may be true trivially (through a response of zero under all treatments) and those for which even the most extreme observed differences in response do not provide sufficient power to distinguish between the null and alternative hypotheses.

Acknowledgements

The authors thank Dr. James Reecy, Iowa State, for kindly providing data from the bovine iron SNP study. Some of this work was completed by Isaac Dialsingh and Stefanie Austin as part of his Ph.D. dissertation and her Master's thesis, respectively, both completed at The Pennsylvania State University under the direction of Naomi Altman.

Funding

This work was supported in part by NSF DMS 1007801, NSF IOS 0820729 and NIH UL1RR033184 (to N.S.A.) and NSF DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute (to N.S.A.).

Conflict of Interest: none declared.

References

Bancroft,T. *et al.* (2013) Estimation of false discovery rate using sequential permutation p-values. *Biometrics,* **69**, 1–7.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B,* **57**, 289–300.

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

Black,M.A. (2004) A note on the adaptive control of false discovery rates. *J. R. Stat. Soc. B,* **66**, 297–304.

Blekhman,R. *et al.* (2010) Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.,* **20**, 180–189.

Dialsingh,I. (2012) False Discovery Rates When the Statistics are Discrete. PhD thesis, Dept. of Statistics, The Pennsylvania State University, USA.

Eicker,F. (1967) Limit theorems for regressions with unequal and dependent errors. In: Cam,L.L. and Neyman,J. (eds) *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* University of California Press, Berkeley CA, Vol. 1, pp. 59–82.

Langass,M. *et al.* (2005) Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. B,* **67**, 1979–1987.

Markitsis,A. and Lai,Y. (2010) A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bioinformatics,* **26**, 640–646.

Mateescu,R. *et al.* (2013) Genome-wide association study of concentration of iron and other minerals in longissimus muscle of Angus cattle. *Technical report*, Iowa State University. Draft.

Nettleton,D. *et al.* (2006) Estimating the number of true null hypotheses from a histogram of p-values. *J. Agric. Biol. Environ. Stat.,* **11**, 337–356.

Pounds,S. and Cheng,C. (2004) Improving false discovery rate estimation. *Bioinformatics,* **20**, 1737–1745.

Pounds,S. and Cheng,C. (2006) Robust estimation of the false discovery rate. *Bioinformatics,* **22**, 1979–1987.

Pounds,S. and Morris,S. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics,* **19**, 1236–1242.

Robinson,M. *et al.* (2010) edger: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* **26**, 139–140.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B,* **64**, 479–498.

Storey,J.D. (2003) The positive false discovery rate. *Ann. Stat.,* **31**, 2013–2035.

Strimmer,K. (2008) A unified approach to false discovery rate estimation. *BMC Bioinformatics,* **9**, 303.

Tarone,R.E. (1990) A modified Bonferroni method for discrete data. *Biometrics,* **46**, 515–522.

Wang,H.-Q. *et al.* (2011) SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics,* **27**, 225–231.

Zhang,S.-D. (2011) Towards accurate estimation of the proportion of true null hypotheses in multiple testing. *PLoS One,* **6**, e18874.

Zhang,X. *et al.* (2005) Genome-wide expression profiling and identification of gene activities during early flower development in *Arabidopsis*. *Plan Mol. Biol.,* **58**, 401–419.