# Estimating false discovery rates for contingency tables

Jonathan M. Carlson, David Heckerman and Guy Shani
Microsoft Research
MSR-TR-2009-53
{carlson,heckerma,guyshani}@microsoft.com

May 6, 2009

**Abstract**

When testing a large number of hypotheses, it can be helpful to estimate or control the false discovery rate (FDR), the expected proportion of tests called significant that are truly null. The FDR is intricately linked to probability that a truly null test is significant, and thus a number of methods have been described that estimate or control the FDR by directly using the $p$-values of the hypothesis tests. Most of these methods make the assumption that the $p$-values are uniformly and continuously distributed under the null hypothesis, an assumption that often does not hold for finite data. In this paper, we consider the estimation of FDR for contingency tables. We show how Fisher's exact test can be extended to efficiently calculate the exact null distribution over a set of contingency tables. Using this exact null distribution, we explore the estimation of each of the terms in the FDR estimation, characterize the asymptotic convergence of the estimator, and show how the conservative bias can be reduced by removing certain tests from consideration. The resulting estimator has substantially less conservative bias than traditional approaches.

## 1 Introduction

In modern biomedical applications, researchers often want to test multiple hypotheses at the same time. For example, in an HIV drug resistance study, a researcher may wish to test which observed HIV mutations are correlated with drug resistance. The $p$-value, the probability of getting a result at least as extreme as the observed test assuming the null hypothesis is correct, can be used to filter out low probability correlations. Researchers typically consider the subset of tests with $p$-value below some threshold for a follow-up study. When performing multiple statistical tests, these $p$-value thresholds must be carefully chosen so as to avoid an abundance of false positive results, while at the same time maximizing the number of true positive results. Traditionally, $p$-value thresholds are chosen so as to control the probability of at least one false positive result. More recently, Benjamini & Hochberg (1995) proposed controlling the *false discovery rate* (FDR), which is the expected proportion of false positives among tests that are called significant. Storey (2002, 2003) introduced the FDR-analogue of the $p$-value, called the *q-value*, which estimates the minimum FDR for any given $p$-value threshold. This approach has proven widely applicable in, for example, high throughput biological screens, as it allows a researcher to balance the number of significant associations with the proportion of those tests that are expected to be false positives.

When test statistics are continuous and two-sided, the $p$-values are expected to be uniformly distributed under the null hypothesis. For these cases, Storey (2002, 2003) provides a simple procedure for estimating the FDR. For many emerging biological applications, however, the underlying test statistics are not continuous. For example, tests involving genetic data (such as single nucleotide polymorphisms, or SNPs) often involve a small number of possible outcomes. As we will later illustrate, in such cases the discreteness in the data will cause the $p$-values under the null hypothesis to be heavily skewed toward one, making Storey's FDR method overly conservative (Pounds & Cheng, 2006; Gilbert, 2005). Furthermore, it is often the case that each test will have a different $p$-value distribution under the null hypothesis, further complicating analysis.

Although the discreteness of the data presents unique challenges, we can also leverage it to our advantage. For many discrete tests the exact null distribution can be efficiently computed. We can use these exact computations to provide tighter FDR estimates. In this paper, we consider those cases where the sufficient statistics for a test are a $2 \times 2$ contingency table and Fisher's exact test (FET) is appropriate. In these cases, we show how FET can be leveraged to efficiently compute an FDR estimator that is asymptotically conservative and yields more power than estimations that are based on continuous assumptions.

FET uses the hypergeometric distribution to estimate the null distribution of a single test using all the possible permutations of that test's data. We extend this null distribution to efficiently compute the exact Type I error rate over a large number of tests. We demonstrate it's applicability to discrete data, provide an efficient implementation for exact permutation testing, and prove the asymptotically conservative result. We suggest a family of permutation-based estimators for the $\pi_0$ parameter, estimating the proportion of all tests that are truly null. We explore the convergence properties of our estimators and discuss their convergence properties.

One of the unique advantages of discrete data is the potential to ignore tests than can be proven to be irrelevant. We derive a new filtering criterion that is provably conservative and will, under certain circumstances, provably increase power.

Although this paper focuses on Fisher's exact test over binary contingency tables, our theoretical results will generalize to other exact tests, such as McNamara's test and the exact test for Hardy-Weinberg Equilibrium, for which exact distributions can be efficiently calculated.

## 2 Multiple Hypothesis Examples

To demonstrate the applicability of our methods, we experiment with the following data sets:

**HIV resistance.** A number of effective antiretroviral (ARV) drugs have been developed for HIV. However, for each of these drugs there exists a set of HIV mutations that abrogate the drug's effectiveness. Thus, an active area of research involves the identification and characterization of drug resistance mutations. Harrigan and colleagues (Harrigan et al., 2005) enrolled patients at the start of their ARV therapy and sequenced the infecting HIV genomes. After several years of therapy on multiple classes of ARVs, the researchers resequenced the infecting HIV genomes to identify mutations that were correlated with specific ARVs. They tested whether the presence of a mutation at a given position in the HIV genome was correlated with the patient taking a specific ARV. Three classes of drugs were tested over 1194 variables representing observed amino acids at positions in the HIV Protease and Reverse Transcriptase proteins, resulting in 3582 total tests, each with 281 observations. We will refer to this data set as "Resistance".

**Epitope mapping.** The cellular arm of the immune response identifies and destroys infected cells. Such cells can be identified by the unique strings of viral protein fragments, called *epitopes*, that are displayed on the surface of infected cells. These epitopes are displayed by human leukocyte antigen (HLA) proteins. Thousands of HLA variations have been identified in humans. Thus, a critical component of HIV vaccine design is the identification and characterization of the set of epitopes that each HLA allele can present on the cell surface.

One high-throughput experimental method for epitope mapping is the use of overlapping peptide (OLP) scans, in which a sliding window of 10-15 amino acid peptide fragments are created based on the viral genome of interest. Each OLP is then tested against cells from dozens of patients, each of whom contains six HLA alleles, with an assay that tests if at least one of the HLA proteins binds the peptide (Addo et al., 2003). One then attempts to map HLA alleles to positive assay responses. The test can be performed using FET, comparing the frequency of response to an epitope in patients with or without a given HLA allele. Kiepiela et al. (2007) recently performed this high throughput assay. Here, we reanalyze their data, comparing comparing 219 HLA alleles with 343 OLPs, resulting in 74,774 total tests, with an average of 724 observations per test. This data set is denoted "Epitope".

**Sieve analysis.** In sieve analysis we try to identify positions in the viral genome at which the variability differs between two different populations (Gilbert, 2005). For example, if the position is more variable in

patients from one region than those from another, it may indicate that different forces are acting on that region. Following Gilbert, we have created a data set consisting of 567 HIV clade B sequences (Brumme et al., 2008) and 567 HIV clade C sequences (Kiepiela et al., 2007; Rousseau et al., 2008) and, for 363 positions in the HIV Gag protein, compared the frequency at which sequences match the consensus for their respective clades. We will call this data set "Sieve".

**Linkage disequilibrium mapping.** Linkage disequilibrium (LD) occurs when two sites on a chromosome are not statistically independent of each other. This typically occurs when the sites are nearby on the chromosome, such that inheritance of a given allele at one site is correlated with inheritance of a given allele at the other site. When performing genome wide association studies (GWAS), it is helpful to identify a set of variations (SNPs) that are in LD with each other, so that potentially redundant results can be identified. We have taken the SNP data from a recent GWAS (Chio et al., 2009), binarized the data and tested for independence between each of 401,017 pairs of neighboring SNPs. There was an average of 1085 observations per test. We will call this data set "SNP".

**Synthetic data** We also created various types of synthetic data based on the Epitope data set (see Appendix). Through such synthetic data we can demonstrate each aspect of our proposed approach.

# 3 Background

## 3.1 Contingency Tables and Fisher's Exact Test

Consider an experiment that tests the independence of two random binary variables $X$ and $Y$. Suppose there are $n$ observations, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The results can be summarized in a contingency table $t = (a, b, c, d)$, as defined in Table 1. We denote the marginal counts of $t$ as $\theta^t = (\theta_X, \theta_{\bar{X}}, \theta_Y, \theta_{\bar{Y}})$, which represent the number of times each variable is observed in each state. These counts capture the maximum likelihood estimators for the marginal probabilities $\mathrm{pr}(X = 1)$ and $\mathrm{pr}(Y = 1)$. We will often drop the superscript $t$ when it is clear from context.

Table 1: $2 \times 2$ contingency table

| | $Y = 1$ | $Y = 0$ | $\sum X$ |
|---|---|---|---|
| $X = 1$ | $a$ | $b$ | $\theta_X = a + b$ |
| $X = 0$ | $c$ | $d$ | $\theta_{\bar{X}} = c + d$ |
| $\sum Y$ | $\theta_Y = a + c$ | $\theta_{\bar{Y}} = b + d$ | $n$ |

Our goal is to test the null hypothesis $H = 0$ that $X$ and $Y$ are independent. Let $T$ be the random variable representing a table with marginals $\theta$. Fisher (1922) showed that, if $X$ and $Y$ are independent and each is independent and identically distributed (IID), then the probability $\mathrm{pr}(t)$ of observing $T = t$ is given by the hypergeometric distribution:

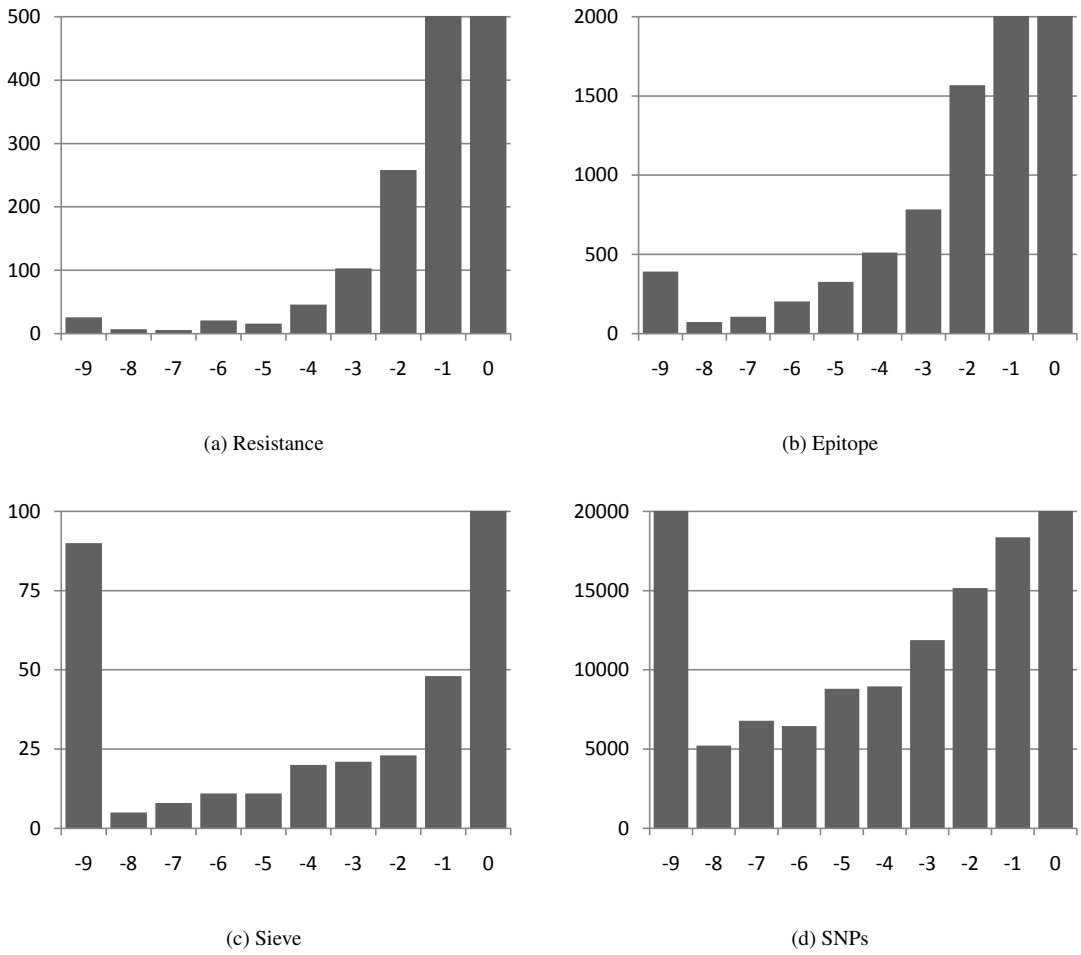
$$\mathrm{pr}(t) \triangleq \frac{\binom{\theta^t_X}{a}\binom{\theta^t_{\bar{X}}}{c}}{\binom{n}{\theta^t_Y}}. \tag{1}$$

From (1), a two-tailed *marginal* $p$-value $\mathrm{p}(t)$ can be computed for $t$ using

$$\mathrm{p}(t) \triangleq \sum_{t' \in \mathrm{perm}(t)} \mathrm{pr}(t')\, \mathbb{1}\{\mathrm{pr}(t') \leq \mathrm{pr}(t)\}\,, \tag{2}$$

where $\mathrm{perm}(t) = \{t' : \theta^{t'} = \theta^t\}$, and $\mathbb{1}\{\cdot\}$ is the indicator function that evaluates to one if the constraints are satisfied and 0 otherwise. We call Equation (2) a *marginal p-value* because it is conditioned on the marginals $\theta$.

Many statistics have a uniform distribution of $p$-values under the null. That is, if $P$ is a continuous random variable representing a $p$-value, $\mathrm{pr}(P \leq \alpha \mid H = 0) = \alpha$. It is evident from (1), however, that the distribution of FET $p$-values under the null may be strongly skewed toward one and therefore $\mathrm{pr}(P \leq \alpha \mid H = 0) \leq \alpha$. For example, the minimum achievable $p$-value for a given $n$ is $1/\binom{n}{n/2}$, and that can only be achieved when the marginals are evenly distributed (i.e., when $\theta_X = \theta_{\bar{X}} = \frac{n}{2}$). Figure 1 shows the distribution of $p$-values from our various data sets.




(a) Resistance

(b) Epitope

(c) Sieve

(d) SNPs

Figure 1: The histogram of marginal $p$-values for the various data sets. $p$-values are transformed using $\log_2 p$.

## 3.2 Positive False Discovery Rates

Suppose that $m$ hypothesis tests over $2 \times 2$ contingency tables $t_1, \ldots, t_m$ are simultaneously tested using Fisher's exact test, with corresponding marginal $p$-values $p_1, p_2, \ldots, p_m$, and we wish to estimate or control the false positive rate in some way. Given the values of $H = 1, H_2, \ldots, H_m$, where $H_i = 0$ if the $i$th null hypothesis is true and $H_i = 1$ if the null hypothesis is false, and a rejection region $\Gamma$, the possible results are summarized in Table 2.

Note that only $m$, $R$ and $W$ are observed. Here, we will focus on nested rejection regions defined by the Type I error rate $\alpha$. That is, given an $\alpha \in [0, 1]$, we will reject all tests with $P \leq \alpha$. For convenience, we will

Table 2: Outcomes when testing $m$ hypotheses

| *Hypothesis* | *Accept* | *Reject* | *Total* |
|---|---|---|---|
| Null true | $U(\Gamma)$ | $V(\Gamma)$ | $m_0$ |
| Alternative true | $T(\Gamma)$ | $S(\Gamma)$ | $m_1$ |
| Total | $W(\Gamma)$ | $R(\Gamma)$ | $m$ |

use $\alpha$ to denote this rejection region.

Benjamini & Hochberg (1995) proposed controlling the False Discovery Rate (FDR), which they defined to be

$$\text{FDR}(\alpha) \triangleq E\left(\frac{V(\alpha)}{R(\alpha)}\right).$$

When $R(\alpha) = 0$, this quantity is undefined. In this case, Benjamini and Hochberg defined $\text{FDR}(\alpha)$ to be 0, which is equivalent to defining

$$\text{FDR}(\alpha) \triangleq E\left(\frac{V(\alpha)}{R(\alpha)} \;\middle|\; R(\alpha) > 0\right) \text{pr}(R(\alpha) > 0).$$

In practice, we are only interested in $p$-value thresholds that result in the rejection of at least one test in our data. Thus, Storey (2002) proposed the *positive false discovery rate* (pFDR), which is conditioned on the rejection of at least one test:

$$\begin{aligned}\text{pFDR}(\alpha) &= E\left(\frac{V(\alpha)}{R(\alpha)} \;\middle|\; R(\alpha) > 0\right) \\ &= \frac{1}{\text{pr}(R(\alpha) > 0)}\text{FDR}(\alpha).\end{aligned}$$

Following Storey (2002), assume that the random variables $H_i$ are IID Bernoulli variables with prior probability $\text{pr}(H_i = 1) = 1 - \pi_0$ and that the $p$-values are IID with

$$P_i \mid H_i \sim (1 - H_i)F_0 + H_i F_1,$$

for some null distribution $F_0$ and some alternative distribution $F_1$. Under these assumptions, Storey (2002) showed that

$$\begin{aligned}\text{pFDR}(\alpha) &= \text{pr}(H = 0 \mid P \leq \alpha) \\ &= \frac{\pi_0 \,\text{pr}(P \leq \alpha \mid H = 0)}{\text{pr}(P \leq \alpha)} \\ &= \frac{E(V(\alpha))}{E(R(\alpha))}.\end{aligned} \tag{3}$$

For small samples, it may be quite likely that no tests are rejected at threshold $\alpha$. Therefore, Storey (2002) proposed the following estimator for pFDR:

$$\widehat{\text{pFDR}}(\alpha) \triangleq \frac{\hat{\pi}_0 \,\hat{\text{pr}}(P \leq \alpha \mid H = 0)}{\hat{\text{pr}}(P \leq \alpha)\,\hat{\text{pr}}(R(\alpha) > 0)}. \tag{4}$$

Equation (4) provides estimators for each term of (3), with an added correction term that estimates $\text{pr}(R(\alpha) > 0)$. (Note that for large $m$, pFDR $\approx$ FDR, because $\text{pr}(R(\alpha) > 0) \approx 1$.)

For $p$-values that are not biased towards 0 under the null distribution, Storey (2002) showed that

$$\widehat{\text{pFDR}}(\alpha) \triangleq \frac{W(\lambda)\,\alpha}{(1 - \lambda)R(\alpha)\,(1 - (1 - \alpha)^m)},$$

for some well-chosen $\lambda$, $0 \leq \lambda < 1$, is conservative both asymptotically and in expectation for finite samples using the intuition that $\text{pr}(P \leq \alpha \mid H = 0) \leq \alpha$ and $\pi_0 \approx \frac{W(\lambda)}{(1-\lambda)m}$. Storey (2002); Storey & Tibshirani (2003) and others (e.g., Langaas et al. 2005) have proposed methods for choosing $\lambda$.

In practice, Storey suggested estimating $\text{pFDR}(\alpha)$ for each observed $p$-value $p_i$ and proved that, under the aforementioned independence assumptions, the estimates are simultaneously conservative for all $p_i$ Storey et al. (2004). When the tests are continuous and uniformly distributed, $\text{pr}(P \leq p_i \mid H = 0) = p_i$, making Storey's estimator quite tight in practice. When the tests are discrete, however, $\text{pr}(P \leq p_i \mid H = 0)$ can be significantly less than $\alpha$. For example, in the case of Fisher's exact test, each observed $p$-value $p_i$ is a marginal $p$-value dependent on the marginals $\theta_i$. Although $\text{pr}(P \leq p_i \mid H = 0, \theta_i) = p_i$, for some test $j \neq i$, $\text{pr}(P \leq p_i \mid H = 0, \theta_j) \leq p_i$. Thus, the overall probability $\text{pr}(P \leq P_i \mid H = 0) \leq p_i$, making Storey's estimate overly conservative.

# 4 Computing pFDR for Fisher's exact test

Fisher's exact test provides an efficient means of computing $\text{pr}(P \leq \alpha \mid H = 0, \theta)$ exactly. Although the resulting marginal $p$-values should not be applied directly to pFDR estimation, the computation can be efficiently leveraged to provide a tight estimate for each of the terms in (4). In this section, we define each of these estimates and show that each estimate is unbiased or conservative (in this context, an estimator is conservative if it is expected to overestimate terms in the numerator of (4) or underestimate terms in the denominator of (4)). We evaluate the asymptotic convergence properties of (4), showing that it results in a conservative FDR estimate and characterize conditions under which the estimates will be closer to the true values. In addition, we discuss ways in which the exact test can be further leveraged to increase power by filtering irrelevant tests.

**Assumption 1.** We assume that $p$-values are IID and the $H_i$ are IID Bernoulli random variables. Because we are using Fisher's exact test, we also need to consider the marginals $\theta$, which we assume are IID. Although we we make no assumptions about the specific distribution of $\theta$, we assume that the $p$-values depend on the marginals according to the mixture model

$$P \mid H, \theta \sim (1 - H)F_0(\theta) + HF_1(\theta) \tag{5}$$

where $F_0(\theta)$ follows the hypergeometric distribution as defined in (2) and $F_1(\theta)$ is the generating function of the alternative model. We further assume that $\theta$ is independent of H, though we will later relax this assumption.

## 4.1 Estimating $\text{pr}(P \leq \alpha \mid H = 0)$ from pooled $p$-values

The typical approach to FDR estimation is to assume for each $p_i$ that $\text{pr}(P \leq p_i \mid H = 0) = p_i$, which is equivalent to assuming $\text{pr}(P \leq \alpha \mid H = 0) = \text{pr}(P \leq \alpha \mid H = 0, \theta_j)$ for $j = 1, \ldots, m$. It is evident from equation (2), however, that the $p$-values are not independent of the marginals. Thus, more accurate estimate is to sum out the marginals:

$$\text{pr}(P \leq \alpha \mid H = 0) = \sum_{\theta'} \text{pr}(P \leq \alpha \mid H = 0, \theta') \, \text{pr}(\theta' \mid H = 0) \tag{6a}$$

$$= \sum_{\theta'} \text{pr}(P \leq \alpha \mid H = 0, \theta') \, \text{pr}(\theta') \,, \tag{6b}$$

where (6b) follows from the assumption that the marginals are independent and identically distributed. Furthermore, this IID assumption allows us to construct an unbiased estimate using *pooled $p$-valued*, calculated

from

$$\widehat{\mathrm{pr}}(P \leq \alpha \mid H = 0) = \sum_{\theta'} \mathrm{pr}(P \leq \alpha \mid H = 0, \theta')\, \widehat{\mathrm{pr}}(\theta') \tag{7a}$$

$$= \sum_{\theta'} \mathrm{pr}(P \leq \alpha \mid H = 0, \theta') \left( \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{\theta_i = \theta'\} \right) \tag{7b}$$

$$= \frac{1}{m} \sum_{i=1}^{m} \sum_{\theta'} \mathrm{pr}(P \leq \alpha \mid H = 0, \theta')\, \mathbb{1}\{\theta_i = \theta'\} \tag{7c}$$

$$= \frac{1}{m} \sum_{i=1}^{m} \mathrm{pr}(P \leq \alpha \mid H = 0, \theta_i) \tag{7d}$$

where we estimate $\mathrm{pr}(\theta)$ using the maximum likelihood estimate

$$\widehat{\mathrm{pr}}(\theta) \triangleq \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{\theta_j = \theta\}\,.$$

Because $\widehat{\mathrm{pr}}(\theta)$ is an unbiased estimator for $\mathrm{pr}(\theta)$, it follows that

$$E(\widehat{\mathrm{pr}}(P \leq \alpha \mid H = 0)) = \mathrm{pr}(P \leq \alpha \mid H = 0)\,. \tag{8}$$

Equation (7d) can be written

$$\widehat{\mathrm{pr}}(P \leq \alpha \mid H = 0) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t' \in \mathrm{perm}(t_i)} \mathrm{pr}(\mathrm{pr}(t') \mid H = 0)\, \mathbb{1}\{\mathrm{p}(t') \leq \alpha\}\,,$$

indicating that the pooled $p$-value is taken from every possible permutation of the data. Note that the rightmost summation is itself a $p$-value and the constraint $\mathbb{1}\{p(T') \leq \alpha\}$ is such that the summation is the maximum achievable $p$-value under $\theta_i$ that is at most $\alpha$. Thus,

$$E(\widehat{\mathrm{pr}}(P \leq \alpha \mid H = 0)) \leq \alpha \tag{9}$$

with equality when all the observed marginals can achieve $\alpha$ exactly. Thus, when not all marginals can achieve $\alpha$ exactly, power can be gained over the marginal $p$-values by using the pooled $p$-values, which yield a tighter estimate of $\widehat{\mathrm{pr}}(P \leq \alpha \mid H = 0)$. The specific reduction in bias can be expressed as the ratio

$$\frac{\widehat{\mathrm{pr}}(P \leq \alpha \mid H = 0)}{\alpha}, \tag{10}$$

which can have quite a large effect in practice, especially for small $\alpha$. Figure 2 plots this ratio as a function of $\alpha$ for each of our data sets.

## 4.2 Estimating $\mathrm{pr}(P \leq \alpha)$ and $\mathrm{pr}(R(\alpha) > 0)$

Given that $R(\alpha)$ is the number of observed tests with $p \leq \alpha$, it follows that

$$E\left(\frac{R(\alpha)}{m}\right) = \mathrm{pr}(P \leq \alpha)\,. \tag{11}$$

We therefore follow Storey (2002) in defining

$$\widehat{\mathrm{pr}}(P \leq \alpha) \triangleq \frac{R(\alpha)}{m}.$$

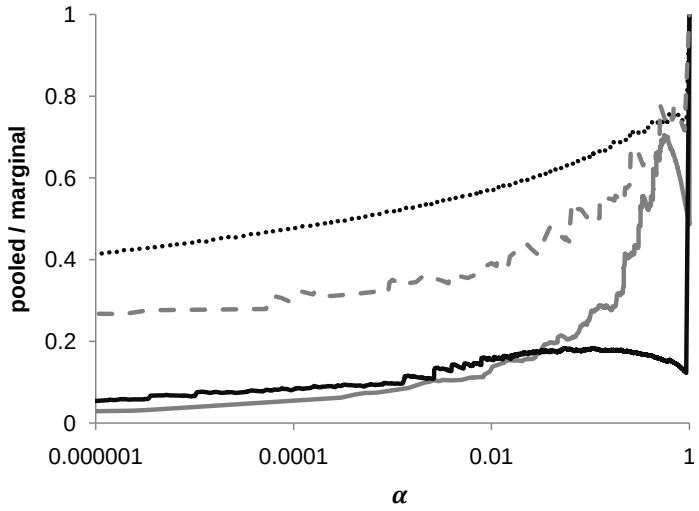



Figure 2: Ratio of pooled to marginal $p$-values as a function of $\alpha$. Solid dark, Epitope; solid gray, Resistance; dotted dark, SNPs; dashed gray, Sieve.

Because (4) would be undefined for $R = 0$, Storey (2002) defines

$$\widehat{\mathrm{pr}}(P \leq \alpha) \triangleq \frac{R(\alpha) \vee 1}{m},$$

where $R(\alpha) \vee 1 = \max(R(\alpha), 1)$, though in practice we are typically only interested in values for $\alpha$ that were observed in our data set.

To determine $\widehat{\mathrm{PFDR}}(\alpha)$, we also require the estimate $\widehat{\mathrm{pr}}(R(\alpha) > 0)$. It follows from our independence assumptions that

$$\begin{aligned} \mathrm{pr}(R(\alpha) > 0) &= 1 - (1 - \mathrm{pr}(P \leq \alpha))^m \\ &\leq 1 - (1 - \mathrm{pr}(P \leq \alpha \mid H = 0))^m , \end{aligned}$$

making

$$\widehat{\mathrm{pr}}(R(\alpha) > 0) \triangleq 1 - (1 - \widehat{\mathrm{pr}}(P \leq \alpha \mid H = 0))^m$$

a conservative estimate of $\mathrm{pr}(R(\alpha) > 0)$. Note that using $1 - (1 - \alpha)^m$ will lead to an anti-conservative (over) estimate of $\mathrm{pr}(R(\alpha) > 0)$ due to inequality (9).

## 4.3 Estimating $\pi_0$

The final step in the PFDR computation is to estimate $\pi_0$. In this section, we use a general framework for conservatively estimating $\pi_0$ and show how existing methods fit within this framework. Given the mixture model (5), we can write

$$\mathrm{pr}(P = p) = \pi_0 \,\mathrm{pr}(P = p \mid H = 0) + \pi_1 \,\mathrm{pr}(P = p \mid H = 1)\,, \tag{12}$$

where $\pi_1 = \mathrm{pr}(H = 1) = 1 - \pi_0$ (Dalmasso et al., 2005; Genovese & Wasserman, 2004; Langaas et al., 2005). Thus, it follows that

$$\mathrm{pr}(P = p) \geq \pi_0 \,\mathrm{pr}(P = p \mid H = 0)$$

and

$$\pi_0 \leq \frac{\mathrm{pr}(P = p)}{\mathrm{pr}(P = p \mid H = 0)}.$$

Moreover, for any non-negative function $\rho(\cdot)$,

$$\pi_0 \leq \frac{\sum_p \rho(p)\mathrm{pr}(P=p)}{\sum_p \rho(p)\mathrm{pr}(P=p \mid H=0)}, \tag{13}$$

which leads to the following lemma.

**Lemma 1.** *Suppose we have $m$ tests that follow Assumption 1. Let $\rho(\cdot)$ be any non-negative function, and*

$$\hat{\pi}_0 \triangleq \frac{\sum_{i=1}^m \rho(p_i)}{\sum_{i=1}^m E(\,\rho(p) \mid \, H=0, \theta_i)}. \tag{14}$$

*Then $E(\hat{\pi}_0) \geq \pi_0$.*

*Proof.* It follows analogously to the derivation of (7) that

$$\mathrm{pr}(P=p \mid H=0) = \frac{1}{m} E\left(\sum_{i=1}^m \mathrm{pr}(P=p \mid H=0, \theta_i)\right) \tag{15}$$

and

$$\mathrm{pr}(P=p) = \frac{1}{m} E\left(\sum_{i=1}^m \mathbb{1}\{p_i=p\}\right). \tag{16}$$

Thus, combining with inequality (13), it follows that

$$\pi_0 \leq \frac{\sum_p \rho(p)\mathrm{pr}(P=p)}{\sum_p \rho(p)\mathrm{pr}(P=p \mid H=0)} = \frac{\sum_p \rho(p)\frac{1}{m} E(\sum_{i=1}^m \mathbb{1}\{p_i=p\})}{\sum_p \rho(p)\frac{1}{m} E(\sum_{i=1}^m \mathrm{pr}(P=p \mid H=0, \theta_i))} \tag{17a}$$

$$= \frac{E(\sum_{i=1}^m \rho(p_i))}{E\Big(\sum_{i=1}^m \sum_p \rho(p)\mathrm{pr}(P=p \mid H=0, \theta_i)\Big)}. \tag{17b}$$

Because $\sum_p \rho(p)\mathrm{pr}(P=p)$ is a linearly increasing function of $\sum_p \rho(p)\mathrm{pr}(P=p \mid H=0)$, it follows from Jensen's inequality that

$$\begin{aligned}
\frac{E(\sum_{i=1}^m \rho(p_i))}{E\Big(\sum_{i=1}^m \sum_p \rho(p)\mathrm{pr}(P=p \mid H=0, \theta_i)\Big)} &\leq E\left(\frac{\sum_{i=1}^m \rho(p_i)}{\sum_{i=1}^m \sum_p \rho(p)\mathrm{pr}(P=p \mid H=0, \theta_i)}\right) \\
&= \frac{\sum_{i=1}^m \rho(p_i)}{\sum_{i=1}^m E(\,\rho(p) \mid \, H=0, \theta_i)} \\
&= \hat{\pi}_0.
\end{aligned}$$

Thus, $E(\hat{\pi}_0) \geq \pi_0$. □

Furthermore, in the limit, estimate (14) asymptotically converges to

$$\pi_0 + \pi_1 \frac{E(\,\rho(p) \mid \, H=1)}{E(\,\rho(p) \mid \, H=0)},$$

which implies that the $\rho(\cdot)$ function minimizing $\frac{E(\,\rho(p) \mid H=1)}{E(\,\rho(p) \mid H=0)}$ will yield the least biased estimator.

**Lemma 2.** *Under the assumptions of Lemma 1,*

$$\lim_{m\to\infty} \hat{\pi}_0 \overset{\text{a.s.}}{=} \pi_0 + \pi_1 \frac{E(\,\rho(p) \mid \, H=1)}{E(\,\rho(p) \mid \, H=0)}.$$

*Proof.* By the strong law of large numbers, equations (15) and (16) imply that $m^{-1}\, E(\sum_{i=1}^{m} \mathrm{pr}(P = p \mid H = 0, \theta_i))$ and $m^{-1}\, E(\sum_{i=1}^{m} \mathbb{1}\{p_i = p\})$ converge almost surely to $\mathrm{pr}(P = p \mid H = 0)$ and $\mathrm{pr}(P = p)$, respectively. Thus, it follows from (17) that

$$\lim_{m\to\infty} \hat{\pi}_0 \overset{\text{a.s.}}{=} \frac{\sum_p \rho(p)\mathrm{pr}(P = p)}{\sum_p \rho(p)\mathrm{pr}(P = p \mid H = 0)}.$$

Furthermore, it follows from the mixture model (12) that

$$\frac{\sum_p \rho(p)\mathrm{pr}(P = p)}{\sum_p \rho(p)\mathrm{pr}(P = p \mid H = 0)} = \pi_0 + \pi_1 \frac{\sum_p \rho(p)\mathrm{pr}(P = p \mid H = 1)}{\sum_p \rho(p)\mathrm{pr}(P = p \mid H = 0)}. \tag{18}$$

Thus,

$$\begin{aligned}\lim_{m\to\infty} \hat{\pi}_0 &\overset{\text{a.s.}}{=} \pi_0 + \pi_1 \frac{\sum_p \rho(p)\mathrm{pr}(P = p \mid H = 1)}{\sum_p \rho(p)\mathrm{pr}(P = p \mid H = 0)} \\ &= \pi_0 + \pi_1 \frac{E(\,\rho(p) \mid H = 1)}{E(\,\rho(p) \mid H = 0)}.\end{aligned}$$

□

Equation (14) gives us great flexibility in computing $\pi_0$ estimates. One such estimation method is given by Storey (2002, 2003)

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{(1-\lambda)m} \tag{19}$$

for some tuning parameter $0 \le \lambda < 1$. For uniformly distributed statistics,

$$(1-\lambda)m = E(\,\#\{\pi > \lambda\} \mid H = 0) = m\,\mathrm{pr}(P > \lambda \mid H = 0)\,.$$

For contingency tables, we can estimate $\mathrm{pr}(P > \lambda \mid H = 0)$ using $\widehat{\mathrm{pr}}(P \le \lambda \mid H = 0)$, which results in an unbiased estimate for $E(\,\#\{\pi > \lambda\} \mid H = 0)$. Therefore, Storey's estimator (19) is a special case of estimator (14) in which

$$\rho(p) = \begin{cases} 0 & \text{if } p \le \lambda, \\ 1 & \text{otherwise} \end{cases}$$

and the P-values are assumed to be continuous and uniformly distributed under the null.

As $\lambda \to 0$, we have increasingly conservative bias, with $\hat{\pi}_0 = 1$ when $\lambda = 0$, whereas the variance of the $\pi_0$ estimate increases as $\lambda \to 1$ due to the decreasing number of observations. Indeed, different heuristic approaches have been proposed to balance the bias-variance tradeoff inherent in picking $\lambda$ (Storey, 2002; Storey & Tibshirani, 2003). Equation (14) suggests an orthogonal heuristic that may be useful in estimating $\pi_0$: choosing a weighting function $\rho(\cdot)$ such that more weight is applied to tests with high $p$-value. A natural choice for a weighting function is $\rho(p) = \mathrm{pr}(P \le p \mid H = 0, \theta_i) = p_i$, which is equivalent to

$$\hat{\pi}_0 \triangleq \frac{E(p)}{E(p \mid H = 0)}. \tag{20}$$

Under this weighting function, tests with low $p$-values will still contribute to the $\pi_0$ estimate, but not as much as tests with high $p$-values. In principle, we could define $\rho(\cdot)$ such that we are effectively summing $p$-values over the range $\lambda < p \le 1$; in practice, however, we have found the $\pi_0$ estimate to be quite stable over a wide range of $\lambda$, and so simply set $\lambda = 0$.

Pounds & Cheng (2006) suggested the estimator $\hat{\pi}_0 \triangleq 2\bar{p}$ for discrete statistics, where $\bar{p}$ is the average observed marginal $p$-value. Estimator (20) is similar to their estimate, except that for contingency tables we can compute $E(p \mid H = 0)$ exactly, rather than assuming $E(p \mid H = 0) = 0.5$, which assumes a unform distribution of $p$-values.

Table 3 compares the true $\pi_0$ for synthetic data against the $\pi_0$ estimators of equation (14), of Storey (2003) (evaluated for $\lambda = 0.5$[1]) and of Pounds & Cheng (2006) computed using marginal $p$-values. Equation (14) provides a tight and conservative estimate, substantially increasing power over methods that assume a uniform $p$-value distribution. Similarly, accounting for the exact $p$-value distribution results in a substantially lower $\pi_0$ estimate on each of the real data sets (Table 3).

Table 3: Comparing $\pi_0$ estimates

| Data set | $\frac{E(P)}{E(P\mid H=0)}$ | $2\bar{p}$ | Storey[a] |
|---|---|---|---|
| 0.001[b] | 0.086 | 0.12 | 0.11 |
| 0.05 | 0.13 | 0.18 | 0.17 |
| 0.1 | 0.18 | 0.25 | 0.24 |
| 0.2 | 0.27 | 0.38 | 0.37 |
| 0.3 | 0.36 | 0.52 | 0.51 |
| 0.4 | 0.45 | 0.64 | 0.63 |
| 0.5 | 0.58 | 0.69 | 0.68 |
| 0.6 | 0.66 | 0.79 | 0.79 |
| 0.7 | 0.75 | 0.91 | 0.90 |
| 0.8 | 0.83 | 1.01 | 1 |
| 0.9 | 0.92 | 1.13 | 1 |
| 1 | 1.003 | 1.24 | 1 |
| Sieve | 0.63 | 0.82 | 0.69 |
| SNPs | 0.34 | 0.44 | 0.41 |
| Epitope | 0.99 | 1.84 | 1 |
| Resistance | 0.97 | 1.38 | 1 |

[a] Evaluated at $\lambda = 0.5$ using pooled $p$-values.
[b] Numeric data sets indicate $\pi_0$ for synthetic data derived from the Epitope data set.

## 4.4 Convergence properties

We have presented estimates for each component of pFDR, showing how each is either unbiased or conservative. It follows that $\widehat{\text{pFDR}}$ is asymptotically conservative as $m \to \infty$. We can, however, be more precise in estimating the convergence properties of our pFDR estimate.

**Theorem 1.** *Given $m$ tests that follow Assumption 1 and a non-negative function $\rho(\cdot)$, we have*

$$\lim_{m\to\infty} \widehat{\text{pFDR}}(\alpha) \overset{\text{a.s.}}{=} \frac{\pi_0 + \pi_1 \frac{E(\rho(p) \mid H=1)}{E(\rho(p) \mid H=0)}}{\pi_0} \text{pFDR}(\alpha).$$

This convergence theorem shows that, for large samples, our pFDR estimate is conservative and becomes tightest when $\hat{\pi}_0$ is computed using a $\rho(\cdot)$ function that is expected to be much lower for true alternative hypotheses than for true null hypotheses.

[1] We used the available R code from http://genomics.princeton.edu/storeylab/qvalue/index.html, using pooled $p$-values as input. Results are truncated at 1. $\lambda = 0.5$ was chosen because the spline-fitting method Storey & Tibshirani (2003) always results in estimates of $\pi_0 = 1$ for these discrete data sets.

*Proof.*

$$\begin{aligned}\lim_{m\to\infty} \widehat{\text{PFDR}}(\alpha) &= \lim_{m\to\infty} \frac{\hat{\pi}_0 \, \widehat{\text{pr}}(P \le \alpha \mid H = 0)}{\widehat{\text{pr}}(P \le \alpha)\, \widehat{\text{pr}}(R(\alpha) > 0)} \\ &= \frac{\lim_{m\to\infty} \hat{\pi}_0 \, \lim_{m\to\infty} \widehat{\text{pr}}(P \le \alpha \mid H = 0)}{\lim_{m\to\infty} \widehat{\text{pr}}(P \le \alpha) \lim_{m\to\infty} \widehat{\text{pr}}(R(\alpha) > 0)}.\end{aligned}$$

By the strong law of large numbers, equation (8) implies that $\widehat{\text{pr}}(P \le \alpha \mid H = 0)$ converges almost surely to $\text{pr}(P \le \alpha \mid H = 0)$, equation (11) implies that $\widehat{\text{pr}}(P \le \alpha)$ converges almost surely to $\text{pr}(P \le \alpha)$, and $\widehat{\text{pr}}(R(\alpha) > 0)$ converges almost surely to 1. Thus

$$\begin{aligned}\lim_{m\to\infty} \widehat{\text{PFDR}}(\alpha) &\overset{\text{a.s.}}{=} \frac{\lim_{m\to\infty} \hat{\pi}_0 \, \text{pr}(P \le \alpha \mid H = 0)}{\text{pr}(P \le \alpha)} \\ &= \frac{\lim_{m\to\infty} \hat{\pi}_0}{\pi_0} \, \text{PFDR}(\alpha).\end{aligned}$$

Finally, it follows from Lemma 2 that

$$\lim_{m\to\infty} \widehat{\text{PFDR}}(\alpha) \overset{\text{a.s.}}{=} \frac{\pi_0 + \pi_1 \frac{E(\rho(p) \mid H=1)}{E(\rho(p) \mid H=0)}}{\pi_0} \, \text{PFDR}(\alpha). \qquad \square$$

## 4.5 Dependent marginals

Until now, we have assumed that $\theta$ is independent of $H$. It is conceivable, however, that true alternative tests will tend to have more balanced marginals that permit lower $p$-values. For example, in the Resistance data, it is possible that positions in which no mutations confer drug resistance may be more conserved due to purifying selection, an evolutionary process that results in less observed variation. Under these conditions, we might expect the PFDR estimate to become more conservative because a substantial proportion of the balanced marginals included in our pooled-$p$-value estimate belong to true alternative tests causing us to overestimate $\text{pr}(P \le \alpha \mid H = 0)$. This conservative bias is analogous to that observed in the microarray community—a bias that arises from permutation testing over alternative data that has a higher variance than the null data (Jiao & Zhang, 2008; Xie et al., 2005). In this section, we formalize the problem for discrete statistics and show that our PFDR estimate becomes asymptotically more conservative as the tendency increases for true alternative tests to have more evenly distributed marginals than null tests. Note that these results hold only for $\pi_0$ estimators in which $\rho(\cdot)$ is non-decreasing, a slightly more restrictive definition than we used in the previous sections. In principle, the conservative bias could be reduced using heuristic measures similar to those proposed in the microarray community (Jiao & Zhang, 2008; Xie et al., 2005); however, these heuristics are not guaranteed to result in a conservative PFDR estimate, so we will not explore them further here.

Let us consider a special case in which $\text{pr}(\theta \mid H = 1) \neq \text{pr}(\theta \mid H = 0)$. Specifically, we shall consider the case where the marginals of true alternative hypotheses tend to have larger $n$ and/or are more balanced, meaning that the marginals of the true alternative hypotheses tend to permit smaller $p$-values than the marginals of the true null hypotheses. This leads to the following set of assumptions:

**Assumption 2.** Assume we have $m$ tests, in which the P-values are independent and identically distributed according to the mixture (5), the $H$ are independent and identically distributed Bernoulli random variables, and for each test $i$, $H_i$ is dependent on $\theta_i$ such that

$$\sum_{\theta} \text{pr}(P \le \alpha \mid H = 0, \theta) \, \text{pr}(\theta \mid H = 0) \le \sum_{\theta} \text{pr}(P \le \alpha \mid H = 0, \theta) \, \text{pr}(\theta \mid H = 1) \, .$$

Under these assumptions, we can derive the following large sample theorem:

**Theorem 2.** *Given $m$ tests that follow Assumption 2, and a non-negative, non-decreasing function $\rho(\cdot)$,*

$$\lim_{m\to\infty} \widehat{\text{PFDR}}(\alpha) \geq \frac{\pi_0 + \pi_1 \frac{E(\,\rho(p)\mid H=1)}{E(\,\rho(p)\mid H=0)}}{\pi_0} \text{PFDR}(\alpha).$$

*Proof.* The proof follows analogously to that of Theorem 1 by noting that the Assumption 2 leads to

$$\lim_{m\to\infty} \widehat{\text{pr}}(P \leq \alpha) \overset{\text{a.s.}}{=} \text{pr}(P \leq \alpha) \tag{21}$$

$$\lim_{m\to\infty} \widehat{\text{pr}}(P \leq \alpha \mid H = 0) \overset{\text{a.s.}}{\geq} \text{pr}(P \leq \alpha \mid H = 0) \tag{22}$$

$$\lim_{m\to\infty} \hat{\pi}_0 \overset{\text{a.s.}}{\geq} \pi_0 + \pi_1 \frac{E(\,\rho(p)\mid\, H=1)}{E(\,\rho(p)\mid\, H=0)}. \tag{23}$$

We shall prove each of these statements in turn.

Equation (21) follows immediately by noting that our estimate $\widehat{\text{pr}}(P \leq \alpha) \triangleq (R \vee 1)/m$ is not affected by the distribution of $\theta$. Equation (22) can be seen by noting that we can no longer use equality (6b) and must instead use equality (6a).

Thus, we have

$$\begin{aligned}
&\lim_{m\to\infty} \widehat{\text{pr}}(P \leq \alpha \mid H = 0) \\
&\quad\overset{\text{a.s.}}{=} \sum_{\theta'} \text{pr}(P \leq \alpha \mid H = 0, \theta')\, \text{pr}(\theta') \\
&\quad= \sum_{\theta'} \text{pr}(P \leq \alpha \mid H = 0, \theta') \times \left\{\text{pr}(\theta' \mid H = 0)\, \pi_0 + \text{pr}(\theta' \mid H = 1)\, \pi_1\right\} \\
&\quad= \pi_0 \sum_{\theta'} \text{pr}(P \leq \alpha \mid H = 0, \theta')\, \text{pr}(\theta' \mid H = 0) + \pi_1 \sum_{\theta'} \text{pr}(P \leq \alpha \mid H = 0, \theta')\, \text{pr}(\theta' \mid H = 1) \\
&\quad\geq \pi_0 \sum_{\theta'} \text{pr}(P \leq \alpha \mid H = 0, \theta')\, \text{pr}(\theta' \mid H = 0) + \pi_1 \sum_{\theta'} \text{pr}(P \leq \alpha \mid H = 0, \theta')\, \text{pr}(\theta' \mid H = 0) \\
&\quad= \text{pr}(P \leq \alpha \mid H = 0)\,,
\end{aligned}$$

where the inequality follows from Assumption 2. Finally, inequality (23) follows from the fact that the added assumptions of Lemma 2 only affect the denominator of our $\pi_0$ estimate (14). Furthermore, inequality (22) implies

$$\lim_{m\to\infty} \widehat{\text{pr}}(P \geq \alpha \mid H = 0) \overset{\text{a.s.}}{\leq} \text{pr}(P \geq \alpha \mid H = 0)\,,$$

from which it follows that

$$\lim_{m\to\infty} \sum_p \rho(p) \widehat{\text{pr}}(P = \alpha \mid H = 0) \overset{\text{a.s.}}{\leq} \sum_p \rho(p) \text{pr}(P = \alpha \mid H = 0)$$

for any non-decreasing function $\rho(p)$. Thus, it follows that

$$\begin{aligned}
\lim_{m\to\infty} \hat{\pi}_0 &= \frac{\sum_p \rho(p) \widehat{\text{pr}}(P \geq \alpha)}{\sum_p \rho(p) \widehat{\text{pr}}(P \geq \alpha \mid H = 0)} \\
&\overset{\text{a.s.}}{\geq} \frac{\sum_p \rho(p) \text{pr}(P \geq \alpha)}{\sum_p \rho(p) \text{pr}(P \geq \alpha \mid H = 0)} \\
&= \pi_0 + \pi_1 \frac{E(\,\rho(p)\mid\, H=1)}{E(\,\rho(p)\mid\, H=0)}.
\end{aligned}$$

□

Hence, under Assumption 2, and provided $\rho(p)$ is non-negative and non-decreasing in $p$, our pFDR and FDR estimates will be asymptotically more conservative than the case where the marginals are independent of $H$. In practice, it is not known whether $H$ and $\theta$ are independent. Consequently, we recommend using the more restricted class of $\rho(\cdot)$ functions allowed by Theorem 2.

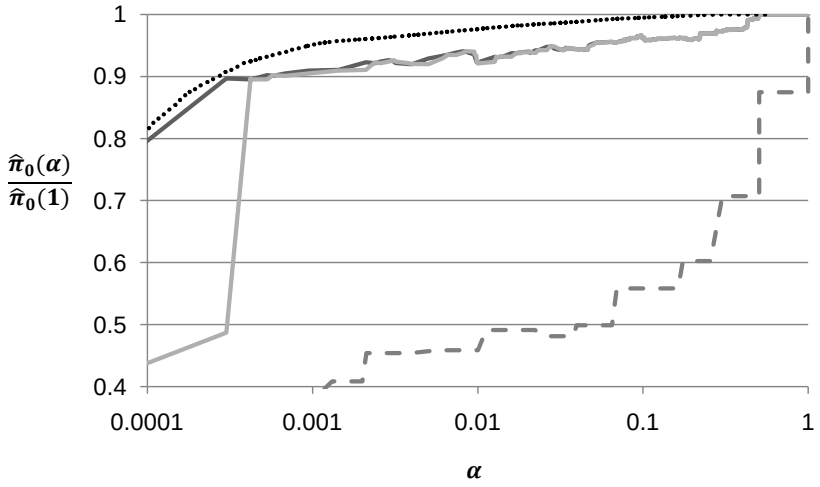



Figure 3: The proportional reduction in the pFDR estimate from using $\hat{\pi}_0(\alpha)$ (filtering) over $\hat{\pi}_0(1)$ (not filtering) on synthetic data derived from: solid dark, Epitope; solid gray, Resistance; dotted dark, SNPs; dashed gray, Sieve.

### 4.6 Filtering irrelevant tests

The discreteness of the data provides a unique opportunity to further increase power. For any discrete test, there exists an $\alpha$ such that $\text{pr}(P \leq \alpha) = 0$. Including such tests in our pFDR estimate will typically increase the conservative bias. Thus, power can often be improved by first *filtering* out all all tests that couldn't possibly achieve the significance threshold $\alpha$. Such filtering will still result in an asymptotically conservative estimate and, in some cases, will provably increase power.

Because the range of the data is finite, computation of the most significant achievable $p$-value given the marginals is possible. We can define the minimum achievable $p$-value for fixed marginals as $p^*(\theta) \triangleq \min_{t \in \text{perm}(\theta)} p(t)$. When computing $\text{pFDR}(\alpha)$, we can ignore (filter) all tests $i$ such that $p^*(\theta_i) > \alpha$. For a set of contingency tables $\mathbf{T}$, we can now write $\mathbf{T} = \mathbf{T}_\alpha^+ \cup \mathbf{T}_\alpha^-$ for the disjoint sets

$$\mathbf{T}_\alpha^+ = \{t_i : t_i \in \mathbf{T} \wedge p^*(\theta_i) > \alpha\},$$
$$\mathbf{T}_\alpha^- = \{t_i : t_i \in \mathbf{T} \wedge p^*(\theta_i) \leq \alpha\}.$$

Filtering on $p^*(\theta) > \alpha$ can be seen as estimating pFDR over $\mathbf{T}_\alpha^-$. Let $\text{pFDR}^*(\alpha)$ and $\widehat{\text{pFDR}}^*(\alpha)$ be the true and estimated pFDR, respectively, for $\mathbf{T}_\alpha^-$. Then the following theorem holds.

**Theorem 3.** *Given $m$ tests that follow Assumptions 1 or 2,* $\text{pFDR}^*(\alpha) = \text{pFDR}(\alpha)$.

*Proof.* Recall that, under either set of assumptions,

$$\text{pFDR}(\alpha) = \frac{E(V(\alpha))}{E(R(\alpha))}.$$

Because $E(\, R(\alpha) \mid \text{pr}(p(T) \leq \alpha) = 0) = 0$, it follows that $\text{pFDR}(\alpha)$ will be the same for $\mathbf{T}$ and $\mathbf{T}_\alpha^-$. Therefore, $\text{pFDR}^*(\alpha) = \text{pFDR}(\alpha)$. □

Therefore, filtering on $p^*(\theta_i) > \alpha)$ does not change the true pFDR. Furthermore, under the assumptions of Lemma 4, we have $\text{pFDR}^*(\alpha) \leq \lim_{m \to \infty} \widehat{\text{pFDR}}^*(\alpha)$, almost surely. Consequently, we can compute $\widehat{\text{pFDR}}^*(\alpha)$ instead of $\widehat{\text{pFDR}}(\alpha)$. Moreover, in certain cases filtering will provably increase power.

**Lemma 3.** *Under assumptions 1 or 2, if* $\frac{\mathrm{pr}(P\le\alpha|H=1)}{\mathrm{pr}(P\le\alpha|H=0)}$ *is non-increasing in* $\alpha$*, then* $\lim_{m\to\infty}\widehat{\mathrm{PFDR}}^{*}(\alpha) \overset{\text{a.s.}}{\le} \lim_{m\to\infty}\widehat{\mathrm{PFDR}}(\alpha)$.

*Proof.* Recall our large sample estimate

$$\begin{aligned}\widehat{\mathrm{PFDR}}(\alpha) &= \frac{\hat{\pi}_0\,\widehat{\mathrm{pr}}(P\le\alpha \mid H=0)}{\widehat{\mathrm{pr}}(P\le\alpha)}\\ &= \frac{\hat{\pi}_0\, m\,\frac{1}{m}\sum_{i=1}^{m}\mathrm{pr}(P\le\alpha \mid H=0,\theta_i)}{R(\alpha)\vee 1}\\ &= \frac{\hat{\pi}_0\sum_{i=1}^{m}\mathrm{pr}(P\le\alpha \mid H=0,\theta_i)}{R(\alpha)\vee 1}\end{aligned}$$

Removing $k$ tests with $p^*(\theta)>\alpha$ will have no effect on $(R(\alpha)\vee 1)$ or on

$$\sum_{i=1}^{m}\mathrm{pr}(P\le\alpha \mid H=0,\theta_i)\,.$$

We will show, however, that, under the present assumptions, our $\pi_0$ estimate under filtering will almost surely be lower than our $\pi_0$ estimate without filtering. Let $p^+$ denote the event $p^*(\theta)>\alpha$ and $p^-$ denote the event $p^*(\theta)\le\alpha$. From equation (18)

$$\lim_{m\to\infty}\hat{\pi}_0 \overset{\text{a.s.}}{=} \pi_0+\pi_1\frac{E(\rho(p)\mid H=1)}{E(\rho(p)\mid H=0)} \tag{24a}$$

$$= \pi_0+\pi_1\frac{E(\rho(p)\mid H=1,p^+)\,\mathrm{pr}(p^+)+E(\rho(p)\mid H=1,p^-)\,\mathrm{pr}(p^-)}{E(\rho(p)\mid H=0,p^+)\,\mathrm{pr}(p^+)+E(\rho(p)\mid H=0,p^-)\,\mathrm{pr}(p^-)} \tag{24b}$$

Let

$$\hat{\pi}_0(\alpha)\triangleq\frac{E(\rho(p)\mid p^-)}{E(\rho(p)\mid H=0,p^-)}$$

be the estimated $\pi_0$ over $\mathbf{T}_\alpha^-$. We wish to show that

$$\lim_{m\to\infty}\hat{\pi}_0(\alpha)\le\lim_{m\to\infty}\hat{\pi}_0(1), \tag{25}$$

which, by (24b) is true if an only if

$$\frac{E(\rho(p)\mid H=1,p^+)\,\mathrm{pr}(p^+)+E(\rho(p)\mid H=1,p^-)\,\mathrm{pr}(p^-)}{E(\rho(p)\mid H=0,p^+)\,\mathrm{pr}(p^+)+E(\rho(p)\mid H=0,p^-)\,\mathrm{pr}(p^-)}\ge\frac{E(\rho(p)\mid H=1,p^-)}{E(\rho(p)\mid H=0,p^-)}.$$

Thus, it follows that (25) is true if and only if

$$\frac{E(\rho(p)\mid H=1,p^+)}{E(\rho(p)\mid H=0,p^+)}\ge\frac{E(\rho(p)\mid H=1,p^-)}{E(\rho(p)\mid H=0,p^-)}. \tag{26}$$

The assumption that $\frac{\mathrm{pr}(P\le\alpha|H=1)}{\mathrm{pr}(P\le\alpha|H=0)}$ is non-increasing in $\alpha$ implies that

$$\frac{\mathrm{pr}(P>\alpha\mid H=1,p^+)}{\mathrm{pr}(P>\alpha\mid H=0,p^+)}\ge\frac{\mathrm{pr}(P>\alpha\mid H=1,p^-)}{\mathrm{pr}(P>\alpha\mid H=0,p^-)},$$

from which inequality (26), and hence Lemma 3, follows from the constraint that $\rho(\cdot)$ is non-decreasing. □

Thus, under the conditions of Lemma 3 filtering is asymptotically guaranteed to provide a tighter estimate and therefore increase power. Although this condition is often not met for finite data, it is often approximately met and provides a good rationale for filtering. In addition, because both $\widehat{\text{pFDR}}(\alpha)$ and $\widehat{\text{pFDR}}^*(\alpha)$ are asymptotically greater than $\text{pFDR}(\alpha)$, for large samples we can compute both the filtered and unfiltered estimates and choose whichever yields the lower value.

The only estimate that filtering changes is $\hat{\pi}_0$. Let $\hat{\pi}_0(\alpha)$ denote the estimated $\pi_0$ over $\mathbf{T}^-_\alpha(\alpha)$, and $\hat{\pi}_0(1)$ denote the estimated $\pi_0$ over $\mathbf{T}$. It may be that $\hat{\pi}_0(\alpha) \leq \hat{\pi}_0(1)$, and $\hat{\pi}_0(\alpha)$ is not a conservative estimate of the true $\pi_0$ of the original set of tests $\mathbf{T}$. $\hat{\pi}_0(\alpha)$ is, however, a conservative estimate of $\pi_0$ among the filtered set of tests $\mathbf{T}^-_\alpha(\alpha)$. That is, $\hat{\pi}_0(\alpha)$ is a conservative estimate of the proportion of tests that are truly null among those tests that could achieve significance level $\alpha$. In addition to providing increased power, $\hat{\pi}_0(\alpha)$ may provide valuable information in cases were a large proportion of tests could not achieve $\alpha$. In such cases, the overall $\pi_0$ may be quite high, but the $\pi_0$ among tests that could achieve $\alpha$ (those that we are interested in) may be much lower. Figure 3 plots the reduction in conservative bias afforded by $\hat{\pi}_0(\alpha)$ over $\hat{\pi}_0(1)$.

# 5 Efficient computations

Our FDR method iterates over all the permutations of the tables, computing hypergeometric probabilities. As a straightforward implementation may be prohibitively expensive, in this section we illustrate several approaches that make this computation extremely efficient. Table 4 shows the execution time on a standard desktop computer of the complete pFDR method with filtering over the real data sets, given these algorithmic improvements.

Table 4: pFDR runtime over the real data sets.

| Data set | $m$ | $n$ | Runtime (seconds) |
|---|---|---|---|
| Resistance | 3582 | 208 | 0.46 |
| Epitope | 74774 | 759 | 6.24 |
| Sieve | 363 | 1134 | 0.50 |
| SNPs | 401017 | 1085 | 416.61 |

## 5.1 Efficient computation of the hypergeometric probability

A key component of the pFDR computation is the probability of a table, which, for independent variables, follows the hypergeometric distribution (equation 1). There are two difficulties with this equation. First, computing factorials for large numbers causes numerical overflow. Second, a direct computation requires $2n$ operations. As we compute $\text{pr}(T = t \mid H = 0)$ many times, reducing computation time is crucial.

We solve the overflow problem by selectively multiplying and dividing terms from the equation. That is, whenever the partial result grows beyond 1, we divide by a term in the denominator, and wherever the partial result shrinks below 1, we multiply by a term in the numerator.

To reduce the overall runtime we factorize the factorials. As we have argued above, in many cases the marginals are not balanced. In that case there is a considerable difference between the minimal and the maximal marginal. If $\theta_Y = c + d$ is the minimal marginal we factorize the computation as follows:

$$\frac{\binom{\theta_X}{a}\binom{\theta_{\bar{X}}}{c}}{\binom{n}{\theta_Y}} = \frac{(\prod_{i=a+1}^{a+c} i)(\prod_{i=c+1}^{c+d} i)(\prod_{i=b+1}^{b+d} i)}{c!(\prod_{i=a+b+1}^{n} i)},$$

which requires $3(c + d)$ operations. In many cases, one of the entries of the table significantly dominates the rest, making the above computation a dramatic improvement over the straight forward implementation. Algorithm 1 provides pseudo-code for our approach.

**Algorithm 1** Efficiently computing the hypergeometric probability of a table $t$.

**Function** **HypergeometricProbability**
**Input:** Contingency table $t = (a, b, c, d)$, $n = a + b + c + d$
**Output:** $\mathrm{pr}(T = t \mid H = 0, \theta^t)$ — the hypergeometric probability of $t$

```
Let p = 1
Let j = n
for i = a + 1 to a + c do
    p = p · i
    if p > 1 then
        p = p/j
        j = j − 1
for i = c + 1 to c + d do
    p = p · i
    if p > 1 then
        p = p/j
        j = j − 1
for i = b + 1 to b + d do
    p = p · i
    if p > 1 then
        p = p/j
        j = j − 1
while j > a + b + 1 do
    p = p/j
    j = j − 1
for i = 1 to c do
    p = p/i
return p
```

## 5.2 Computing FET $p$-values

Computing FET (equation 2) requires an iteration over all tables that are consistent with the marginals and which are at least as significant as the table that is observed. If we order the permutations by increasing value of $a$, we can see that, for two tables $t_a = (a, b, c, d)$ and $\mathrm{t}_{a+1} = (a + 1, b - 1, c - 1, d + 1)$, we have

$$\frac{\mathrm{pr}(t_{a+1})}{\mathrm{pr}(t_a)} = \frac{b\, c}{(a + 1)\,(d + 1)}. \tag{27}$$

Therefore, computing the probabilities of all the permutations can be done by computing the probability of the table with the minimal possible $a$, and then iteratively computing successor probabilities until reaching the maximal possible $a$.

## 5.3 Computing pooled $p$-values for PFDR estimation

Expanding Equation 7 we can write:

$$\widehat{\mathrm{pr}}(P \le \alpha \mid H = 0) \triangleq \frac{1}{m} \sum_{i=1}^{m} \sum_{t' \in perm(t_i)} \mathrm{pr}(t')\, \mathbb{1}\{\mathrm{pr}(t') \le \alpha\}\,.$$

We typically want to estimate PFDR for every observed marginal $p$-value $p_i$. We can, however, iterate only once over the permutations of each table. Assuming that the tables are sorted such that $\mathrm{pr}(t_i) \ge \mathrm{pr}(t_{i-1})$, Algorithm 2 computes all the pooled $p$-values while computing the permutations of each table only once.

**Algorithm 2** Computing pooled $p$-values.

**Input:** Set of contingency table $\tau$
**Output:** Mapping of Fisher scores to FDR
  Let $M$ be a dictionary sorted by increasing key value
  **for each** $t \in \tau$ **do**
    $M[\mathrm{pr}(t)] = 0$
  **for each** $t \in \tau$ **do**
    $M' = AllFisherScores(t)$
    **for each** $k' \in M'$ **do**
      Let $k$ be the first key in $M$ that is no smaller than $k'$
      $M[k] = M[k] + M'[k']$
  **for** $i = 2$ to $|M|$ **do**
    $M[k_i] = M[k_{i-1}] + M[k_i]$
    $M[k_{i-1}] = \frac{M[k_{i-1}]}{i}$
  **return** $M$

## 5.4 Computing $p$-values for all table permutations in a single pass

Algorithm 2 requires that FET $p$-values for all permutations of each table will be computed. For a single table, we can compute $p$-values for all its permutations in a single pass, with the same time complexity as computing a single $p$-value. To achieve that, while executing the incremental hypergeometric computation in Equation 27, we keep all the probabilities that were computed in a list. After we finish iterating through the permutations, we need to sort the list of probabilities, and then sum them in increasing order. Thus, we get a list of the $p$-values of all the permutations of the table. Detailed pseudo-code for this process can be found in Algorithm 3.

**Algorithm 3** Computing Fisher scores for all permutations of a table.

**Function AllFisherScores**
**Input:** Contingency table $t = (a, b, c, d)$
**Output:** Mapping of Fisher scores to hypergeometric probabilities
  Let $\Delta = min(a, d)$
  Let $t_0 = (a - \Delta, b + \Delta, c + \Delta, d - \Delta)$
  Let $L$ be the empty list
  Compute $pr(T = t_0 \mid H = 0)$
  **for** $i = 0$ to $min(a, d) + min(c + d)$ **do**
    Add $pr(T = t_i \mid H = 0)$ to $L$
    $t_{i+1} = (a_i + 1, b_i - 1, c_i - 1, d_i)$
    $pr(t_{i+1} \mid H = 0) = pr(t_i \mid H = 0) \cdot \frac{b_i \cdot c_i}{(a_i+1)\cdot(d_i+1)}$
  Sort $L$ in increasing order
  $FisherScore = 0$
  **for** $j = 0$ to $|L|$ **do**
    $FisherScore = FisherScore + L[i]$
    $M[FisherScore] = L[i]$
  return $M$

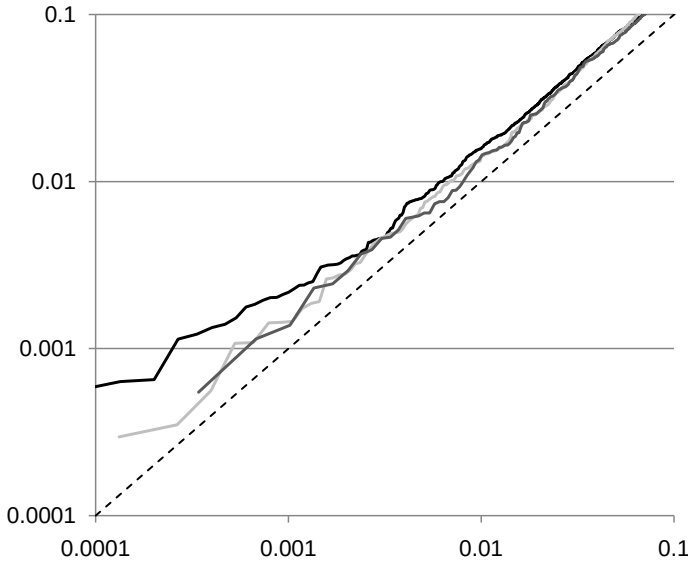



Figure 4: Estimated pFDR vs. true false discovery proportion for synthetic data generated from the Epitope data set. Estimates above the dashed line (representing an unbiased estimate) are conservative. Solid lines, from dark to light, indicate synthetic data with 70, 35 or 10 thousand tables, respectively.

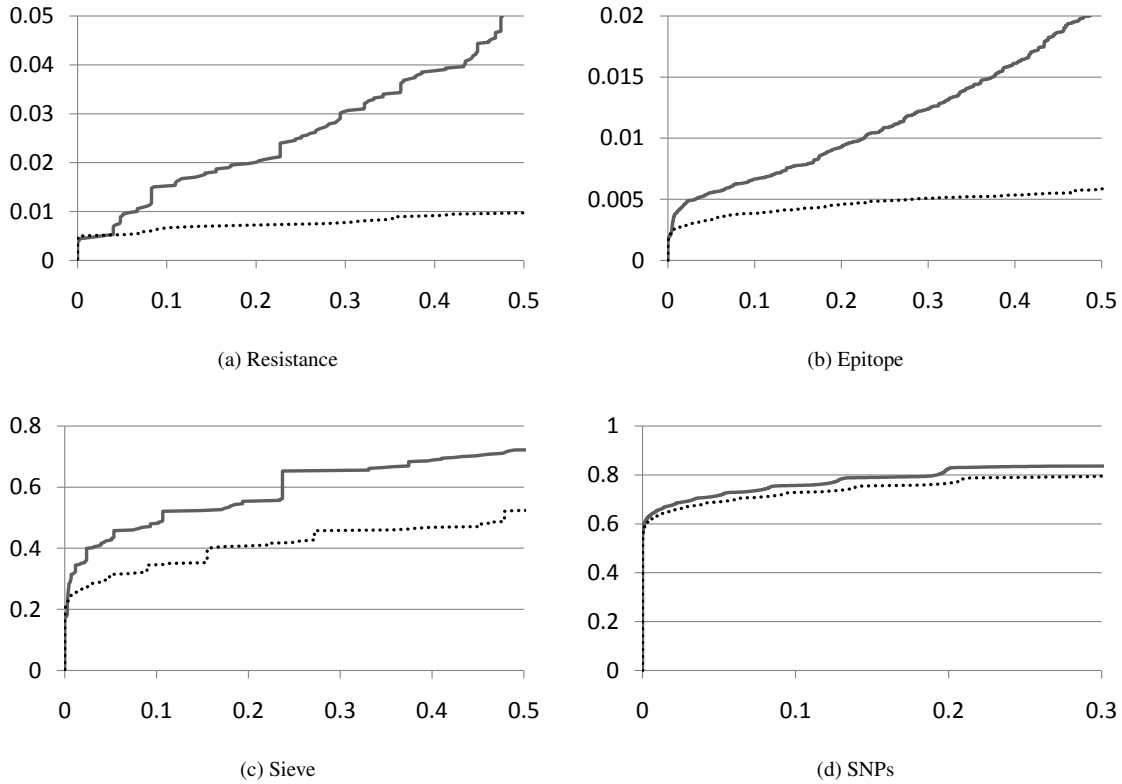
# 6 Numerical results

To explore the applicability of our proposed pFDR estimator, we created a number of Epitope-derived synthetic data sets with different number of tables that follow the mixture model assumptions above, allowing for an unequal distribution of marginals (see the Appendix for details). For each of these data sets, we plotted the estimated $\text{pFDR}(\alpha)$ against the true proportion of false discoveries using $p < \alpha$ as the threshold (Figure 4).

In practice, it is often the case that $\text{pFDR}(\alpha) > \text{pFDR}(\beta)$ for some $\beta > \alpha$. Therefore, there is no reason to choose $\alpha$ as the rejection region, because choosing $\beta$ will result in more rejected tests and a lower proportion of false positives among those rejected tests. For this reason, Storey (2002) proposed the q-value, defined to be $q(\alpha) \triangleq \min_{\beta \geq \alpha} \widehat{\text{pFDR}}(\beta)$. To demonstrate the power gains of our method in practice, we conclude by comparing the number of significant results for each of our example data sets as a function of the q-value threshold (Figure 5). As can be seen, our conservative estimates result in a substantial increase in the number of tests called significant at a variety of thresholds.

# 7 Discussion

The false discovery rate has proven to be an extremely useful tool when testing large numbers of tests, as it allows the researcher to balance the number of significant results with an estimate of the proportion of those results that are truly null. Storey (2002, 2003) presented novel methods for estimating pFDR and $q$-values for general test statistics. He factored the pFDR computation into several components and suggested estimators for each component. Perhaps the most discussed component is the $\pi_0$—the proportion of tests that are expected to be null over the entire data set. For example, Dalmasso et al. (2005) derived a class of $\pi_0$ estimators for continuous distributions that take the same form as equation (14) and explored properties of $\rho(\cdot)$. They proved that a certain class of convex $\rho(\cdot)$ functions yielded provably less biased $\pi_0$ estimators than $\rho(p) = p$. Similarly, Genovese & Wasserman (2004) explore several estimators under a mixture model framework that assumes a uniform continuous null distribution and provide estimates of confidence intervals,




(a) Resistance

(b) Epitope

(c) Sieve

(d) SNPs

Figure 5: Plotting the portion of rejected cases vs. q-values for the real data sets. The solid line is the proposed method for discrete data and the dotted line is the method of Storey & Tibshirani (2003) using marginal $p$-values.

and Langaas et al. (2005) use the mixture model to define $\pi_0$ estimators that perform particularly well under certain continuous convexity assumptions.

When the data are finite, however, some of the underlying assumptions used by the above methods, such as the uniform distribution of $p$-values under the null and the convexity and monotone distribution of $p$-values under the alternative, are violated (Pounds & Cheng, 2006). In such cases, some of the methods developed for general statistics become overly conservative, and some may provide anti-conservative estimates. For example, the estimators of Dalmasso et al. (2005) assume that the null distribution is non-increasing in $p$. As we have seen, contingency tables provide a common example where these assumptions are grossly violated, even when the number of observations in each table is quite high. In these cases, the use of marginal $p$-values leads to severe conservative bias in the FDR estimation.

Pounds & Cheng (2006) addressed the conservative bias of FDR estimation on finite data by proposing a new $\pi_0$ estimator. This estimator avoids the extreme conservative bias of Storey's spline-fitting method on finite data, in which $\pi_0$ estimates at $\lambda = 1$ may have more bias rather than less. On our data sets, the method of Pounds and Chang was comparable to Storey's estimator at $\lambda = 0.5$. A key assumption in the method of Pounds and Cheng is that the expected $p$-value under the null hypothesis is $0.5$, which was grossly violated in all of our contingency table data sets. Replacing this assumption with the exact null distribution substantially decreased the bias in all our tests. Our theoretical results indicate that optimal $\rho(\cdot)$ is that which minimizes the ratio of the expected $\rho(\cdot)$ under the alternative hypothesis to the expected $\rho(\cdot)$ under the null hypothesis. Other $\rho(\cdot)$ functions than those described here may thus yield less biased estimates.

Several authors have proposed randomization testing as a means of dealing with non-uniform or unknown

$p$-values distributions, with a focus on non-uniform continuous distributions (see Cheng & Pounds 2007 for review). Focusing on Fisher's exact test allows us to implement exact permutation tests efficiently even for very large data sets, resulting in exact estimation of the pooled null distribution, a straightforward analysis of the convergence properties, and the removal of numerical error from the estimation.

Furthermore, the exact null distribution allows us to identify and remove tests that cannot be called significant, thereby increasing power. This approach was first proposed by Gilbert (2005), who proposed choosing a $p$-value threshold $p_0$ and removing *a priori* all tests for which no permutation of the contingency table results in $p \leq p_0$. To choose $p_0$, Gilbert suggested using a derivative of the Bonferroni adjusted $p$-value. Unfortunately, it can be shown that this threshold is too aggressive and will often remove tests that should be considered significant. In contrast, choosing $p_0 = \alpha$ leaves the true PFDR unchanged while often achieving an increase in statistical power.

This paper provides estimators for the various components of the PFDR, based on a permutation testing approach. We combine here several ideas that were previously suggested, adapting them to the important case of contingency tables. As we have shown above, our methods can rapidly provide tight estimates of PFDR and $q$-values for very large data sets. Although we have chosen to focus on Fisher's exact test, analogous results can be derived for any discrete test for which all permutations of the data can be efficiently computed.

In addition to describing the theoretical implications of our method, we have derived efficient algorithms for its computation. We have implemented these algorithms as a freely available .NET library and web service, which are available at http://research.microsoft.com/en-us/um/redmond/projects/MSCompBio/FalseDiscoveryRate.

# Appendix

## Creating null and alternative tables from given marginals

To create a null table given a set of marginals $\theta = \{\theta_X, \theta_{\bar{X}}, \theta_Y, \theta_{\bar{Y}}\}$, we draw $n$ tests such that $\mathrm{pr}(X = 1 \mid H = 0) = \theta_X/n$ and $\mathrm{pr}(Y = 1 \mid H = 0) = \theta_Y/n$. To create an alternative tables from $\theta$, we draw $n$ tests such that $\mathrm{pr}(X = 1 \mid H = 0) = \theta_X/n$ and $\mathrm{pr}(Y = 1 \mid H = 1, X = 0) = c/\theta_{\bar{X}}$.

## Selecting marginals

We have created two different types of data sets, one where all the marginals come from the same distribution, and one where the marginals distribution depends on whether the table is null or alternative.

In the case of a single distribution of marginals, we create 10 exponential bins $[1, 1/2], \ldots, (1/512, 1/1024]$ and place each marginal $\theta$ into a bin according to $\min\{\theta_X, \theta_{\bar{X}}, \theta_Y, \theta_{\bar{Y}}\}/\max\{\theta_X, \theta_{\bar{X}}, \theta_Y, \theta_{\bar{Y}}\}$. We then choose a bin uniformly, and select a set of marginals uniformly from the bin. We then designate the selected marginal as null with probability $\pi_0$ and generate the table accordingly. This approach biases us towards choosing marginals that permit lower p-values, which enables us to generate interesting alternative tables, even when we force $\pi_0$ to be much lower than it is in the real data.

When the distribution of marginals depends on the whether the table is null or alternative, we draw the $\theta$ from bin $b \in 1, \ldots, 10$ with probability $1/2^{10-b+1}$ for a null table and with probability $1/2^b$ for an alternative table.

# References

ADDO, M. M., YU, X. G., RATHOD, A., COHEN, D., ELDRIDGE, R. L., STRICK, D., JOHNSTON, M. N., CORCORAN, C., WURCEL, A. G., FITZPATRICK, C. A., FEENEY, M. E., RODRIGUEZ, W. R., BASGOZ, N., DRAENERT, R., STONE, D. R., BRANDER, C., GOULDER, P. J. R., ROSENBERG, E. S., ALTFELD, M. & WALKER, B. D. (2003). Comprehensive epitope analysis of human immunodeficiency virus type 1

(HIV-1)-specific T-cell responses directed against the entire expressed HIV-1 genome demonstrate broadly directed responses, but no correlation to viral load. *J Virol* **77**, 2081–2092.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B Stat Meth* **57**, 289–300.

BRUMME, Z. L., TAO, I., SZETO, S., BRUMME, C. J., CARLSON, J. M., CHAN, D., KADIE, C., FRAHM, N., BRANDER, C., WALKER, B., HECKERMAN, D. & HARRIGAN, P. R. (2008). Human leukocyte antigen-specific polymorphisms in HIV-1 Gag and their association with viral load in chronic untreated infection. *AIDS* **22**, 1277–1286.

CHENG, C. & POUNDS, S. (2007). False discovery rate paradigms for statistical analyses of microarray gene expression data. *Bioinformation* **1**, 436–446.

CHIO, A., SCHYMICK, J. C., RESTAGNO, G., SCHOLZ, S. W., LOMBARDO, F., LAI, S.-L., MORA, G., FUNG, H.-C., BRITTON, A., AREPALLI, S., GIBBS, J. R., NALLS, M., BERGER, S., KWEE, L. C., ODDONE, E. Z., DING, J., CREWS, C., RAFFERTY, I., WASHECKA, N., HERNANDEZ, D., FERRUCCI, L., BANDINELLI, S., GURALNIK, J., MACCIARDI, F., TORRI, F., LUPOLI, S., CHANOCK, S. J., THOMAS, G., HUNTER, D. J., GIEGER, C., WICHMANN, H.-E., CALVO, A., MUTANI, R., BATTISTINI, S., GIANNINI, F., CAPONNETTO, C., MANCARDI, G. L., LA BELLA, V., VALENTINO, F., MONSURRO, M. R., TEDESCHI, G., MARINOU, K., SABATELLI, M., CONTE, A., MANDRIOLI, J., SOLA, P., SALVI, F., BARTOLOMEI, I., SICILIANO, G., CARLESI, C., ORRELL, R. W., TALBOT, K., SIMMONS, Z., CONNOR, J., PIORO, E. P., DUNKLEY, T., STEPHAN, D. A., KASPERAVICIUTE, D., FISHER, E. M., JABONKA, S., SENDTNER, M., BECK, M., BRUIJN, L., ROTHSTEIN, J., SCHMIDT, S., SINGLETON, A., HARDY, J. & TRAYNOR, B. J. (2009). A two-stage genome-wide association study of sporadic amyotrophic lateral sclerosis. *Hum Mol Genet* , ddp059.

DALMASSO, C., BROET, P. & MOREAU, T. (2005). A simple procedure for estimating the false discovery rate. *Bioinformatics* **21**, 660–668.

FISHER, R. A. (1922). On the interpretation of $\chi^2$ from contingency tables, and the calculation of P. *J Roy Stat Soc* **85**, 87–94.

GENOVESE, C. & WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann Stat* **32**, 1035–1061.

GILBERT, P. B. (2005). A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *J Roy Stat Soc C Appl Stat* **54**, 143–158.

HARRIGAN, P. R., HOGG, R. S., DONG, W. W. Y., YIP, B., WYNHOVEN, B., WOODWARD, J., BRUMME, C. J., BRUMME, Z. L., MO, T., ALEXANDER, C. S. & MONTANER, J. S. G. (2005). Predictors of HIV drug-resistance mutations in a large antiretroviral-naive cohort initiating triple antiretroviral therapy. *J Infect Dis* **191**, 339–347. PMID: 15633092.

JIAO, S. & ZHANG, S. (2008). On correcting the overestimation of the permutation-based false discovery rate estimator. *Bioinformatics* **24**, 1655.

KIEPIELA, P., NGUMBELA, K., THOBAKGALE, C., RAMDUTH, D., HONEYBORNE, I., MOODLEY, E., REDDY, S., DE PIERRES, C., MNCUBE, Z., MKHWANAZI, N., BISHOP, K., VAN DER STOK, M., NAIR, K., KHAN, N., CRAWFORD, H., PAYNE, R., LESLIE, A., PRADO, J., PRENDERGAST, A., FRATER, J., MCCARTHY, N., BRANDER, C., LEARN, G. H., NICKLE, D., ROUSSEAU, C., COOVADIA, H., MULLINS, J. I., HECKERMAN, D., WALKER, B. D. & GOULDER, P. (2007). CD8 T-cell responses to different HIV proteins have discordant associations with viral load. *Nat Med* **13**, 46–53.

LANGAAS, M., LINDQVIST, B. H. & FERKINGSTAD, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J Roy Stat Soc B Stat Meth* **67**, 555–572.

POUNDS, S. & CHENG, C. (2006). Robust estimation of the false discovery rate. *Bioinformatics* **22**, 1979–1987.

ROUSSEAU, C. M., DANIELS, M. G., CARLSON, J. M., KADIE, C., CRAWFORD, H., PRENDERGAST, A., MATTHEWS, P., PAYNE, R., ROLLAND, M., RAUGI, D. N., MAUST, B. S., LEARN, G. H., NICKLE, D. C., COOVADIA, H., NDUNG'U, T., FRAHM, N., BRANDER, C., WALKER, B. D., GOULDER, P. J. R., BHATTACHARYA, T., HECKERMAN, D. E., KORBER, B. T. & MULLINS, J. I. (2008). HLA class I-driven evolution of human immunodeficiency virus type 1 subtype C proteome: immune escape and viral load. *J Virol* **82**, 6434–6446.

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

STOREY, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the $q$-value. *Ann Stat* **31**, 2013–2035.

STOREY, J. D., TAYLOR, J. E. & SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J Roy Stat Soc B Stat Meth* **66**, 187–205.

STOREY, J. D. & TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100**, 9440–9445.

XIE, Y., PAN, W. & KHODURSKY, A. B. (2005). A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics* **21**, 4280–4288.