

*Statistical Applications in Genetics
and Molecular Biology*

Volume 10, Issue 1

2011

Article 28

The Joint Null Criterion for Multiple
Hypothesis Tests

Jeffrey T. Leek, *Johns Hopkins Bloomberg School of Public
Health*

John D. Storey, *Princeton University*

Recommended Citation:

Leek, Jeffrey T. and Storey, John D. (2011) "The Joint Null Criterion for Multiple Hypothesis Tests," *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 28.

DOI: 10.2202/1544-6115.1673

Available at: <http://www.bepress.com/sagmb/vol10/iss1/art28>

©2011 Berkeley Electronic Press. All rights reserved.

The Joint Null Criterion for Multiple Hypothesis Tests

Jeffrey T. Leek and John D. Storey

Abstract

Simultaneously performing many hypothesis tests is a problem commonly encountered in high-dimensional biology. In this setting, a large set of p-values is calculated from many related features measured simultaneously. Classical statistics provides a criterion for defining what a “correct” p-value is when performing a single hypothesis test. We show here that even when each p-value is marginally correct under this single hypothesis criterion, it may be the case that the joint behavior of the entire set of p-values is problematic. On the other hand, there are cases where each p-value is marginally incorrect, yet the joint distribution of the set of p-values is satisfactory. Here, we propose a criterion defining a well behaved set of simultaneously calculated p-values that provides precise control of common error rates and we introduce diagnostic procedures for assessing whether the criterion is satisfied with simulations. Multiple testing p-values that satisfy our new criterion avoid potentially large study specific errors, but also satisfy the usual assumptions for strong control of false discovery rates and family-wise error rates. We utilize the new criterion and proposed diagnostics to investigate two common issues in high-dimensional multiple testing for genomics: dependent multiple hypothesis tests and pooled versus test-specific null distributions.

KEYWORDS: false discovery rate, multiple testing dependence, pooled null statistics

1 Introduction

Simultaneously performing thousands or more hypothesis tests is one of the main data analytic procedures applied in high-dimensional biology (Storey and Tibshirani, 2003). In hypothesis testing, a test statistic is formed based on the observed data and then it is compared to a null distribution to form a p-value. A fundamental property of a statistical hypothesis test is that correctly formed p-values follow the Uniform(0,1) distribution for continuous data when the null hypothesis is true and simple. (We hereafter abbreviate this distribution by $U(0,1)$.) This property allows for precise, unbiased evaluation of error rates and statistical evidence in favor of the alternative. Until now there has been no analogous criterion when performing thousands to millions of tests simultaneously.

Just as with a single hypothesis test, the behavior under true null hypotheses is the primary consideration in defining well behaved p-values. However, when performing multiple tests, the situation is more complicated for several reasons: (1) among the entire set of hypothesis tests, a subset are true null hypotheses and the remaining subset are true alternative hypotheses, and the behavior of the p-values may depend on this configuration; (2) the data from each true null hypothesis may follow a different null distribution; (3) the data across hypothesis tests may be dependent; and (4) the entire set of p-values is typically utilized to make a decision about significance, some of which will come from true alternative hypotheses. Because of this, it is not possible to simply extrapolate the definition of a correct p-value in a single hypothesis test to that of multiple hypothesis tests. We provide two key examples to illustrate this point in the following section, both of which are commonly encountered in high-dimensional biology applications.

The first major point of this paper is that the joint distribution of the true null p-values is a highly informative property to consider, whereas verifying that each null p-value has a marginal $U(0,1)$ distribution is not as directly informative. We propose a new criterion for null p-values from multiple hypothesis tests that guarantees a well behaved joint distribution, called the Joint Null Criterion (JNC). The criterion is that the ordered null p-values are equivalent in distribution to the corresponding order statistics of a sample of the same size from independent $U(0,1)$ distributions. We show that multiple testing p-values that satisfy our new criterion can be used to more precisely estimate error rates and rank tests for significance. We illustrate with simple examples how this criterion avoids potentially unacceptable levels of inter-study variation that is possible even for multiple testing procedures that guarantee strong control.

The second major point of this paper is that new diagnostics are needed to objectively compare various approaches to multiple testing, specifically those that evaluate properties beyond control of expected error rate estimates over repeated

studies. These new diagnostics should also be concerned with potentially large study specific effects that manifest over repeated studies in terms of the variance of realized error rates (e.g., the false discovery proportion) and the variance of error rate estimates. This has been recognized as a particularly problematic in the case of dependent hypothesis tests where unacceptable levels of variability in error rate estimates may be obtained even though the false discovery rate may be controlled (Owen, 2005). The need for this type of diagnostic is illustrated in an example presented in the next section, where the analysis of gene expression utilizing three different approaches yields drastically different answers. We propose Bayesian and frequentist diagnostic procedures that provide an unbiased standard for null p-values from multiple testing procedures for complex data. When applied to these methods, the reasons for their differing answers are made clearer.

We apply our diagnostics to the null p-values from multiple simulated studies, to capture the potential for study specific errors. We use the diagnostics to evaluate methods in two major areas of current research in multiple testing: testing multiple dependent hypotheses and pooled versus test-specific null distributions. Surprisingly, some popular multiple testing procedures do not produce p-values with a well behaved joint null distribution, leading directly to imprecise estimates of common error rates such as the false discovery rate.

2 Motivating Examples

Here we motivate the need for the JNC and diagnostic tests by providing two general examples and a real data example from a gene expression study. The first general example describes a situation where every p-value has a $U(0,1)$ distribution marginally over repeated studies. However, the joint distribution of study-specific sets of null p-values deviate strongly from that of independent $U(0,1)$ components. The second general example illustrates the opposite scenario: here none of the p-values has a $U(0,1)$ distribution marginally, but the set of study-specific null p-values appear to have a joint distribution equivalent to independent $U(0,1)$ components of the same size. Together, these examples suggest the need for a gold standard for evaluating multiple testing procedures in practice. Finally, we show that different methods for addressing multiple testing dependence give dramatically different results for the same microarray analysis. This indicates that an objective criterion is needed for evaluating such methods in realistic simulations where the correct answer is known.

2.1 Problematic Joint Distribution from Correct Marginal Distributions

In this example, the goal is to test each feature for a mean difference between two groups of equal size. The first 300 features are simulated to have a true mean difference. There is also a second, randomized unmodeled binary variable that affects the data. Features 200-700 are simulated to have a mean difference between the groups defined by the unmodeled variable. The exact model and parameters for this simulation are detailed in Section 5. The result of performing these tests is a $1,000 \times 100$ matrix of p-values, where the p-values for a single study appear in columns and the p-values for a single test across repeated studies appear in rows.

Using these p-values we examine both their marginal distributions as well as the joint distribution of the null p-values. First, we look at a single p-value affected by the unmodeled variable over the 100 repeated studies. The top two histograms in Figure 1 show the behavior of two specific p-values over the 100 simulated studies. Marginally, each is $U(0,1)$ distributed as would be expected. The randomization of the unmodeled variable results in correct marginal distributions for each null p-value.

Next we consider the null p-values from tests 301-700 for a single study, which is a sample from their joint distribution. The bottom two histograms in Figure 1 show two such examples. In one case the null p-values appear smaller than expected from an i.i.d. $U(0,1)$ sample, in the other case they appear to be larger than expected. This is because in the first study, the unmodeled variable is correlated with the group difference and the signal from the unmodeled variable is detected by the test between groups. In the second study, the unmodeled variable is uncorrelated with the group difference and a consistent source of noise is added to the data, resulting in null p-values that are too large. The result is that each null p-value is $U(0,1)$ marginally, but the joint distribution deviates strongly from a sample of i.i.d. $U(0,1)$ random variables.

Of particular interest is the lower left histogram of Figure 1, which shows only the null p-values from a single simulated study with dependence. The p-values appear to follow the usual pattern of differential expression, with some p-values near zero (ostensibly corresponding to differential expressed genes) and some p-values that appear to be drawn from a $U(0,1)$ distribution (ostensibly the null genes). However, in this example all of the genes are true nulls, so ideally their joint distribution would reflect the $U(0,1)$. Inspection of this histogram would lead to the mistaken conclusion that the method had performed accurately and true differential expression had been detected. This strongly motivates the need for new diagnostic tools that consider the joint behavior of null p-values.

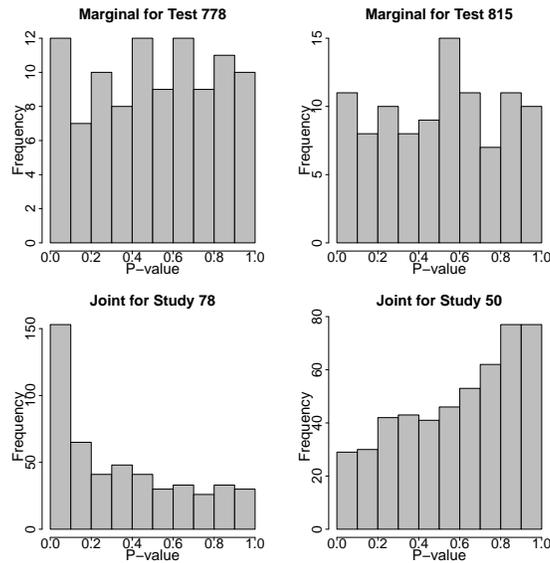


Figure 1: Uniform marginal p -value distribution, but JNC violating joint distribution. Each histogram in the top panel shows the p -value distribution of a single hypothesis test across 100 simulated studies; for each, the marginal distribution is approximately $U(0,1)$ even though each test was subject to a randomized unmodeled variable. Each histogram in the bottom panel shows a sample from the joint distribution of the set of null p -values from a specific realized study. Here, the p -values deviate from the distribution of an i.i.d. $U(0,1)$ sample, depending on the correlation between the randomized unmodeled variable and the group.

2.2 Well Behaved Joint Distribution from Incorrect Marginal Distributions

The second general example also consists of 1,000 tests for mean differences between two groups. The first 300 features are again simulated to have a mean difference between groups. We simulated each feature to have a different variance. The test statistic is a modified t -statistic with a shrinkage constant added to the denominator: $t = \frac{\bar{x}_{i1} - \bar{x}_{i0}}{s_i + a_0}$ where \bar{x}_{ij} is the mean for feature i and group j , s_i is the standard error of $\bar{x}_{i1} - \bar{x}_{i0}$, and a_0 is a single fixed shrinkage constant for all tests. This type of shrinkage statistic is common in the field of multiple testing, where a_0 is frequently estimated from the distribution of the s_i (Tusher, Tibshirani, and Chu, 2001, Cui, Hwang, Qiu, Blades, and Churchill, 2005, Efron, Tibshirani, Storey, and Tusher, 2001). The null statistics are calculated via the bootstrap and pooled across

features (Storey and Tibshirani, 2003). The top row of Figure 2 shows the distribution of two specific p-values across the 100 studies. In this case, since the standard errors vary across tests, the impact a_0 has on the test's null distribution depends on the relative size of a_0 to the s_i .

When the null statistics are pooled, there are individual tests whose p-value follows an incorrect marginal distribution across repeated studies. The reason is that the bootstrap null statistics are pooled across 1,000 different null distributions. The bottom row of Figure 2 shows a sample from the joint distribution of the null p-values for specific studies. The joint distribution behaves like an i.i.d. $U(0,1)$ sample because pooling the bootstrap null statistics captures the overall impact of different variances on the joint distribution of the test statistics coming from true null hypotheses.

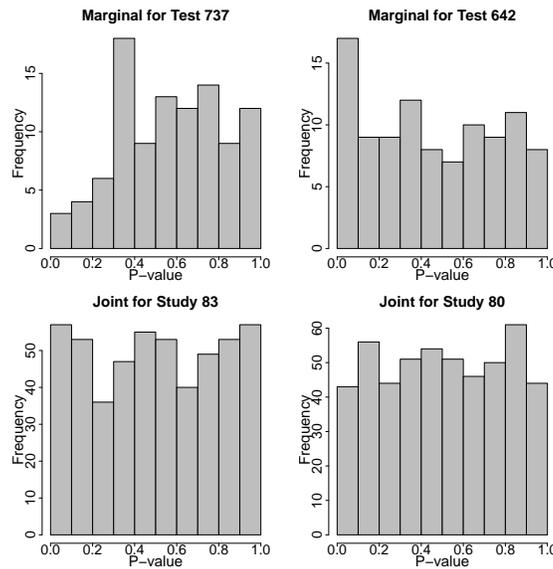


Figure 2: *Non-uniform marginal p-value distribution, but JNC satisfying joint distribution. Each histogram in the top panel shows the p-value distribution of a single hypothesis test using a shrunken t-statistic and pooled null statistics across 100 simulated studies. It can be seen in each that the marginal distribution deviates from $U(0,1)$. Each histogram in the bottom panel shows a sample from the joint distribution of the set of null p-values from two specific realizations of the study. Here, the p-values satisfy the JNC, since pooling the null statistics accounts for the distribution of variances across tests.*

2.3 Microarray Significance Analysis

Idaghdour, Storey, Jadallah, and Gibson (2008) performed a study of 46 desert nomadic, mountain agrarian, and coastal urban Moroccan Amazigh individuals to identify differentially expressed genes across geographic populations. Due to the heterogeneity of these groups and the observational nature of the study, there is likely to be latent structure present in the data, leading to multiple testing dependence. This can be easily verified by examining the residual data after regressing out the variables of interest (Idaghdour et al., 2008). As an example we present two differential expression analyses in Figure 3: agrarian versus desert nomadic, and desert nomadic versus village.

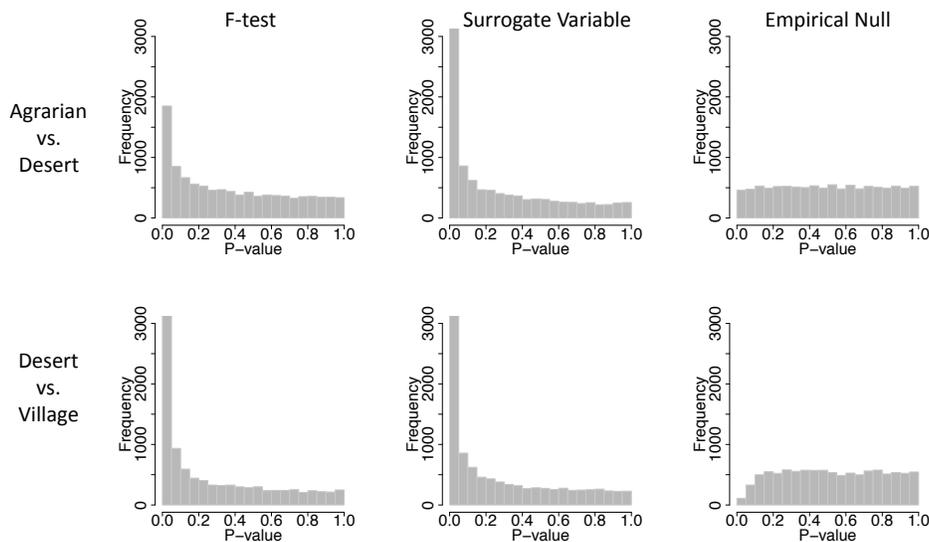


Figure 3: *P*-value histograms from the differential expression analysis comparing agrarian versus desert nomadic (top row) and desert nomadic versus village (bottom row). For each comparison, three different analysis strategies are used: a standard *F*-statistic significance (first column), a surrogate variable adjusted approach (second column), and an empirical null adjusted approach (third column). The last two are methods for adjusting for multiple testing dependence. Both comparisons show wildly different results depending on the analysis technique used.

We perform each analysis in three ways, (1) a simple *F*-test for comparing group means, (2) a surrogate variable adjusted analysis (Leek and Storey, 2007), and (3) an empirical null (Efron, 2004, 2007) adjusted analysis. These last two approaches are different methods for adjusting for multiple testing dependence and

latent structure in microarray data. Figure 3 shows that each analysis strategy results in a very different distribution for the resulting p-values. Idaghdour et al. (2008) found coherent and reproducible biology among the various comparisons only when applying the surrogate variable analysis technique. However, how do we know in a more general sense which, if any, of these analysis strategies is more well behaved since they give such different results? This question motivates a need for a criterion and diagnostic test for evaluating the operating characteristics multiple testing procedures, where the criterion is applied to realistically simulated data where the correct answers are known.

3 A Criterion for the Joint Null Distribution

The examples from the previous section illustrate that it is possible for p-values from multiple tests to have proper marginal distributions, but together form a problematic joint distribution. It is also possible to have a well behaved joint distribution, but composed of p-values with incorrect marginal distributions. In practice only a single study is performed and the statistical significance is assessed from the entire set of p-values from that study. The real data example shows that different methods may yield notably different p-value distributions in a given study.

Thus, utilizing a procedure that produces a well behaved joint distribution of null p-values is critical to reduce deviation from uniformity of p-values within a study, and large variances of statistical significance across studies. A single hypothesis test p-value is correctly specified if its distribution is $U(0,1)$ under the null (Lehmann, 1997). In other words, p is correctly specified if for $\alpha \in (0, 1)$, $\Pr(p < \alpha) = \Pr(U < \alpha) = \alpha$, where $U \sim U(0, 1)$. We would like to ensure that the null p-values from a given experiment have a joint distribution that is stochastically equivalent to an independent sample from the $U(0,1)$ distribution of the same size. Based on this intuition, we propose the following criterion for the joint null p-value distribution.

Definition (Joint Null Criterion, JNC). *Suppose that m hypothesis tests are performed where tests $1, 2, \dots, m_0$ are true nulls and $m_0 + 1, \dots, m$ are true alternatives. Let p_i be the p-value for test i and let $p_{(n_i)}$ be the order statistic corresponding to p_i among all p-values, so that $n_i = \#\{p_j \leq p_i\}$. The set of null p-values satisfy the Joint Null Criterion if and only if the joint distribution of $p_{(n_i)}, i = 1, \dots, m_0$ is equal to the joint distribution of $p_{(n_i)}^*, i = 1, \dots, m_0$, where $p_1^*, \dots, p_{m_0}^*$ are an i.i.d. sample from the $U(0, 1)$ distribution and $p_i^* \stackrel{a.s.}{=} p_i$, for $i = m_0 + 1, \dots, m$.*

Remark 1. If all the p-values correspond to true nulls, the JNC is equivalent to saying that the ordered p-values have the same distribution as the order statistics from an i.i.d. sample of size m from the $U(0,1)$ distribution. \square

Intuitively, when the JNC is satisfied and a large number of hypothesis tests is performed, the set of null p-values from these tests should appear to be equivalent to an i.i.d. sample from the $U(0,1)$ distribution when plotted together in a histogram or quantile-quantile plot. Figure 4 illustrates the conceptual difference between the JNC and the univariate criterion. The p-values from multiple tests for a single study appear in columns and the p-values from a single test across studies appear in rows. The standard univariate criterion is concerned with the behavior of single p-values across multiple studies, represented as rows in Figure 4. In contrast, the JNC is concerned with the joint distribution of the set of study-specific p-values, represented by columns in Figure 4. When only a single test is performed, each column has only a single p-value so the JNC is simply the standard single test criterion.

Remark 2. In the case that the null hypotheses are composite, the distributional equality in the above criterion can be replaced with a stochastic ordering of the two distributions. \square

Remark 3. The JNC is not equivalent to the trivial case where the null p-values are each marginally $U(0,1)$ and they are jointly independent. Let $U_{(1)} \leq U_{(2)} \leq \dots \leq U_{(m_0)}$ be the order statistics from an i.i.d. sample of size m_0 from the $U(0,1)$ distribution. Set $p_i = U_{(i)}$ for $i = 1, \dots, m_0$. It then follows that the null p-values are highly dependent (since $p_i < p_j$ for all $i < j$), none are marginally $U(0,1)$, but their joint distribution is valid. Example 2 from Section 2 provides another scenario where the JNC is not equivalent to the trivial case. \square

Remark 4. The JNC is not a necessary condition for the control of the false discovery rate, as it has been shown that the false discovery rate may be controlled for certain types of dependence that may violate the JNC (Benjamini and Yekutieli, 2001, Storey, Taylor, and Siegmund, 2004). \square

The JNC places a condition on the joint behavior of the set of null p-values. This joint behavior is critical, since error estimates and significance calculation are performed on the set of p-values from a single study (e.g., false discovery rates estimates Storey (2002)). To make this concrete, consider the examples from the previous section. In Example 1, the joint distribution of the null p-values is much more variable than a sample from the $U(0,1)$ distribution, resulting in unreliable error rate estimates and significance calculations (Owen, 2005). The joint p-values in this example fail to meet the JNC. In Example 2, the joint distribution of the

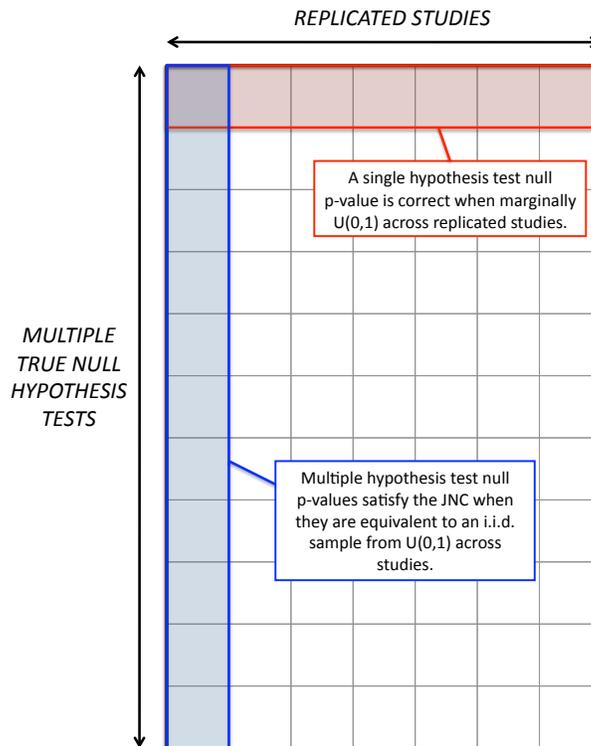


Figure 4: An illustration of the Joint Null Criterion. The p -values from multiple tests for a single study appear in columns and the p -value from a single test across replicated studies compose each row. The JNC evaluates the joint distribution of the set of null p -values, whereas the single test criterion is concerned with the distribution of a single p -value across replicated studies.

p -values satisfies the JNC, resulting in well behaved error rate estimates and significance calculations, even though the marginal behavior of each p -value is not $U(0,1)$.

When the JNC is met, then estimation of experiment-wide error rates and significance cutoffs behaves similarly to the well behaved situation where the true null p -values are i.i.d. $U(0,1)$. Lemma 1 makes these ideas concrete (see Supplementary Information for the proof).

Lemma 1 Suppose that p_1, p_2, \dots, p_m are m p -values resulting from m hypothesis tests; without loss of generality, suppose that p_1, \dots, p_{m_0} correspond to true null hypotheses and p_{m_0+1}, \dots, p_m to true alternative hypotheses. If (1) the JNC is satisfied for p_1, \dots, p_{m_0} and (2) the conditional distribution $\{p_{(n_i)}\}_{i=m_0+1}^m | \{p_{(n_i)}\}_{i=1}^{m_0}$

is equal to the conditional distribution $\{p_{(n_i)}^*\}_{i=m_0+1}^m | \{p_{(n_i)}^*\}_{i=1}^{m_0}$, then any multiple hypothesis testing procedure based on the order statistics $p_{(1)}, \dots, p_{(m)}$ has properties equivalent to those in the case where the true null hypotheses' p-values are i.i.d. $Uniform(0,1)$.

Corollary. When conditions (1) and (2) of Lemma 1 are satisfied, the multiple testing procedures of Shaffer (1995), Benjamini and Hochberg (1995), Storey et al. (2004) provide strong control of the false discovery rate. Furthermore, the controlling and estimation properties of any multiple testing procedure requiring the null p-values to be i.i.d. $Uniform(0,1)$ continue to hold true when the JNC is satisfied.

The Joint Null Criterion is related to two well-known concepts in multiple testing, the marginal determine joint (MDJ) condition (Xu and Hsu, 2007, Calian, Li, and Hsu, 2008) and the joint null domination (jtNDT) condition (Dudoit and van der Laan, 2008). The MDJ is a condition on the observations, which is sufficient to guarantee a permutation distribution is the same as the true distribution (Calian et al., 2008). Meanwhile, the jtNDT condition is concerned with Type I errors being stochastically greater under the test statistics null distribution than under their true distribution. From this, Dudoit and van der Laan (2008) show that two main types of null distributions for test statistics can be constructed that satisfy this null domination property. The difference between these criteria and the JNC is that the JNC focuses not just one Type I error control, but also controlling the study-to-study variability in Type I errors.

4 Statistical Methods for Evaluating the Joint Null Criterion

Several new multiple testing statistics for the analysis of gene expression data have recently been proposed and evaluated in the literature (Tusher et al., 2001, Newton, Noueiry, Sarkar, and Ahlquist, 2004, Storey, 2007). A standard evaluation of the accuracy of a new procedure is to apply it to simulated data and determine whether a particular error rate, such as the false discovery rate, is conservatively biased at specific thresholds, typically 5% and 10%. The JNC suggests a need for methods to evaluate the joint distribution of null p-values from multiple testing procedures. We propose a three step approach for evaluating whether the joint distribution of null p-values satisfies the JNC.

1. Simulate multiple high-dimensional data sets from a common data generating mechanism that captures the expected cross study variation in signal and noise, and includes any dependence or latent structure that may be present.

2. Apply the method(s) in question to each study individually to produce a set of p-values for each study.
3. Compare the set of null p-values from each specific study to the U(0,1) distribution, and quantify differences between the two distributions across all studies.

The first two steps of our approach involve simulating data and applying the method in question to generate p-values, which we carry out in the next section in the context of multiple testing dependence and pooling null distributions across tests. When the joint null distribution can be characterized directly (Huang, Xu, Calian, and Hsu, 2006), analytic evaluation of the JNC may be possible. A key component of evaluating the JNC is the ability to simulate from a realistic joint distribution for the observed data. Application of these diagnostic criteria requires careful examination of the potential properties, artifacts, and sources of dependence that exist in high-dimensional data. In the remainder of the current section, we propose methods for the third step: summarizing and evaluating null p-values relative to the U(0,1) distribution.

We propose one non-parametric approach based on the Kolmogorov-Smirnov (KS) test and a second approach based on a Bayesian posterior probability for the joint distribution. When applying these diagnostics to evaluate multiple testing procedures that produce a small number of observed p-values ($m < 100$) the asymptotic properties of the KS test may not hold. For these scenarios, the Bayesian diagnostic may be more appropriate. In the more general case, when a large number of tests are performed, the diagnostics are both appropriate.

4.1 Double Kolmogorov-Smirnov Test

In this step we start with m p-values from B simulated studies, p_{1j}, \dots, p_{mj} , $j = 1, \dots, B$. Assume that the first m_0 p-values correspond to the null tests and the last $m - m_0$ correspond to the alternative tests. To directly compare the behavior of the p-values from any study to the U(0,1) distribution, we consider the study-specific empirical distribution function, defined for study j as $F_{m_0}^j(x) = \frac{1}{m_0} \sum_{i=1}^{m_0} 1(p_{ij} < x)$. The empirical distribution is an estimate of the unknown true distribution of the null p-values $F^j(x)$. If the null p-values are U(0,1) distributed then $F_{m_0}^j(x)$ will be close to the U(0,1) distribution function, $F(x) = x$. In practice, none of the empirical distribution functions will exactly match the U(0,1) distribution due to random variation.

One approach to determine if the p-values are “close enough” to the U(0,1) distribution is to perform a KS test (Shorack and Wellner, 1986) using the statistic, $D_{m_0}^j = \sup_x |F_{m_0}^j(x) - x|$ (see also Supplementary Figure S1). Based on this statistic

we can calculate a KS test p-value for each simulated study. Under the null hypothesis the KS tests' p-values will also be U(0,1) distributed. We can then calculate a second KS test statistic based on the empirical distribution of the first stage KS test p-values. If the original test-specific null p-values are U(0,1) distributed, then this double KS test p-value will be large and if not then it will be small. Repeating the KS test across a range of simulated data sets permits us to quantify variation around the U(0,1) distribution. Replication also reduces the potential for getting lucky and picking a single simulated study where the method in question excels.

Note that it is possible to consider metrics less stringent than the supremum norm on which the KS test is based. There are variety of ways in which a metric based on $|F_{m_0}^j(x) - x|$ over the range $0 \leq x \leq 1$ can be calculated.

4.2 Bayesian Posterior Probability

A second approach we propose for evaluating the joint distribution of the null p-values is to estimate the posterior probability that the JNC holds given the sets of m p-values across the B simulated studies. To calculate this posterior probability, we assume that the observed null p-values are drawn from a flexible class of distributions. For example, we assume the null p-values are a sample from a Beta(α, β) distribution, where $(\alpha, \beta) \in [0, A] \times [0, B]$. Supplementary Figure S2 shows examples of the density functions for a range of values of (α, β) . The Beta family is used because Beta distributions closely mimic the behavior of non-null p-values observed in practice (Pounds and Morris, 2003). For example, if $\alpha = 1$ and $\beta > 1$ then the corresponding Beta density function is strictly decreasing between 0 and 1, which is typical of the distribution of p-values from differentially expressed genes in a microarray experiment.

Our approach assigns prior probability 1/2 that the p-values are jointly U(0,1) (i.e., the JNC holds), equivalent to a Beta distribution with $\alpha = \beta = 1$, and prior probability 1/2 that the p-values follow a Beta distribution where either $\alpha \neq 1$ or $\beta \neq 1$. We write $\{p_{ij}\}$ as shorthand for the entire set of simulated null p-values, $\{p_{ij}; i = 1, \dots, m_0, j = 1, \dots, B\}$. From Bayes Theorem we can calculate the posterior probability the JNC holds as follows:

$$\begin{aligned} & \Pr(\text{JNC holds} | \{p_{ij}\}) \\ = & \frac{\frac{1}{2} \Pr(\{p_{ij}\} | \text{JNC holds})}{\frac{1}{2} \Pr(\{p_{ij}\} | \text{JNC holds}) + \frac{1}{2} \Pr(\{p_{ij}\} | \text{JNC does not hold})}. \end{aligned}$$

The first component is calculated as:

$$\Pr(\{p_{ij}\}|\text{JNC holds}) = \Pr(\{p_{ij}\} | (\alpha, \beta) = 1) = \prod_{i=1}^{m_0} \prod_{j=1}^B 1(0 \leq p_{ji} \leq 1) = 1.$$

The second component can be calculated by integrating over the other values of (α, β) :

$$\begin{aligned} & \Pr(\{p_{ij}\}|\text{JNC does not hold}) \\ &= \int_0^A \int_0^B \prod_{i=1}^{m_0} \prod_{j=1}^B \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_{ij}^{\alpha-1} (1 - p_{ij})^{\beta-1} \pi_0(\alpha, \beta) d\alpha d\beta \end{aligned}$$

where $\pi_0(\alpha, \beta)$ is the prior distribution for specific values of (α, β) . In the examples that follow, we utilize independent $U(0,1)$ priors on both α and β , but more informative prior choices could be used to emphasize specific potential alternatives. For example, weighting the prior toward values with $\alpha < 1$ and $\beta > 1$ would emphasize distributions that are stochastically smaller than the $U(0,1)$ distribution and typically occur under the alternative.

5 Applications of the Joint Null Criterion

We apply the proposed JNC and diagnostic tests to assess the behavior of methods or two important challenges in multiple hypothesis testing: (1) addressing multiple testing dependence and (2) determining the validity pooled null distributions. Methods have been developed for both of these issues in multiple testing, but there has not been a standard approach for evaluating whether the resulting significance measures have desirable variability properties.

5.1 Multiple Testing Dependence

Multiple testing dependence is a common problem in the analysis of high-dimensional data such as those obtained from genomics (Leek and Storey, 2007) or imaging experiments (Schwartzman, Dougherty, and Taylor, 2008). Multiple testing dependence has frequently been defined as a type of stochastic dependence among p-values or one-dimensional test-statistics when performing multiple tests (Yekutieli and Benjamini, 1999, Benjamini and Yekutieli, 2001, Efron, 2004, 2007). More recently, the root source of this type of dependence has been identified and addressed as dependence among the data for the tests (Leek and Storey, 2008). It has also been shown that regardless of the dependence structure, dependence in the feature

level data can always be parameterized by a low dimensional set of variables (or factors) called a dependence kernel (Leek and Storey, 2008).

Three different approaches for addressing multiple testing dependence are: surrogate variable analysis (Leek and Storey, 2007, 2008), residual factor analysis for multiple testing dependence (Friguet, Kloareg, and Causer, 2009), and the empirical null (Efron, 2004) as applied to multiple testing dependence (Efron, 2007). Surrogate variable analysis is an approach that performs a supervised factor analysis of the data during the modeling process, before one dimensional summaries such as p-values have been calculated. Residual factor analysis for multiple testing dependence is a reformulation of this approach where the estimated factors are required to be orthogonal to the class variable. The empirical null distribution is calculated based on the observed values of the test statistics. The basic idea is to estimate a null distribution based on the “null part” of the observed distribution where the null statistics are assumed to lie. We note that the empirical null method as a general approach (Efron, 2004, 2007) has not been subjected to simulations where the correct answer is known, so its accuracy and general operating characteristics are heretofore unexplored.

It is often the case that the data for multiple tests from high-throughput experiments are dependent. One example of this type of dependence which is common in both microarray and imaging experiments is dependence due to latent or unmodeled factors (Leek and Storey, 2007, 2008). To mimic this type of dependence in our simulated data, we generate the observations for test i from the model $\mathbf{x}_i = b_{0i} + b_{1i}\mathbf{y} + b_{2i}\mathbf{z} + \varepsilon_i$, where \mathbf{z} is a second latent variable that affects the data for multiple tests, and z_j is Bernoulli with probability 0.5. Under this model we let $b_{1i} \neq 0$ for $i = 1, \dots, 500$ and $b_{1i} = 0$ for $i = 501, \dots, 1000$ as before, but $b_{2i} \neq 0$ for $i = 300, \dots, 800$ and $b_{2i} = 0$ for $i = 1, \dots, 200; 801, \dots, 1000$. We first test the null hypothesis that $b_{1i} = 0$ including the variable \mathbf{z} , even though in general it will not be known to the researcher. In Figure 5a the quantile-quantile plots for the null p-values indicate that the p-values approximately follow the U(0,1) distribution. Correspondingly, the double KS p-value is 0.446 and the median posterior probability of the JNC holding (25th-75th percentile) is 0.967 (0.928,0.978).

Next we apply each of the methods for addressing dependence based on the default R code provided by the authors. The surrogate variable analysis (Leek and Storey, 2007, 2008) and residual factor analysis for multiple testing dependence (Friguet et al., 2009) methods result in additional covariates that are included in the model when testing $b_{1i} = 0$. The empirical null approach adjusts the p-values directly based on the observed test statistics. Figure 5 shows the quantile-quantile plots for the adjusted null p-values using each of these methods and Table 1 gives the resulting double KS p-values and posterior probabilities of the JNC holding.

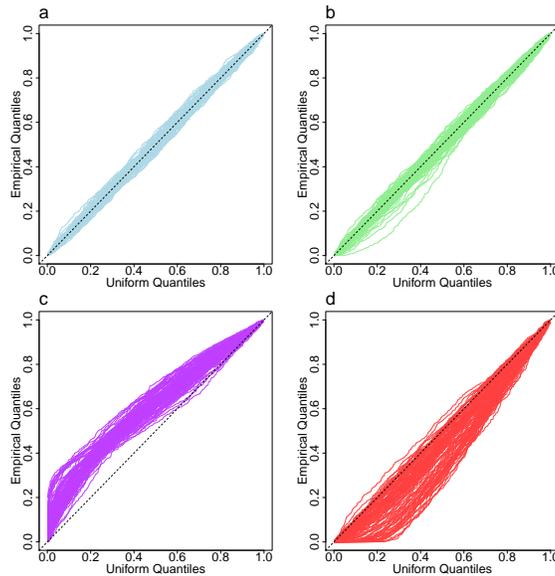


Figure 5: Quantile-quantile plots of the joint distribution of null p-values from 100 simulated studies when the hypothesis tests are dependent. Results when utilizing: **a.** the true latent variable adjustment, **b.** surrogate variable analysis, **c.** empirical null adjustment, and **d.** residual factor analysis.

The surrogate variable adjusted p-values (Figure 5b) behave for the most part like the correctly adjusted p-values in Figure 5a, with the exception of a small number of cases, where the unmodeled variable is nearly perfectly correlated with the group difference. The resulting posterior probability estimates are consistently near 1; however, the double KS p-value is sensitive to the small number of outlying observations.

The empirical null adjustment shows a strong conservative bias, which results in a loss of power (Figure 5c). The reason appears to be that the estimated empirical null is often too wide due to the extreme statistics from the dependence structure. Since the one-dimensional summary statistics conflate signal and noise, it is generally impossible to estimate the null distribution well in the case of dependent data. It has been recommended that the empirical null be employed only when the proportion of truly null hypotheses is greater than 0.90, potentially because of this behavior. Under this assumption, the null p-values are somewhat closer to $U(0,1)$, but still show strong deviations in many cases (Table 1). This indicates the empirical null may be appropriate in limited scenarios when only a small number of tests are truly alternative, such as in genome-wide association studies as originally sug-

gested by Devlin and Roeder (1999) – but not for typical microarray, sequencing, or brain imaging studies.

The residual factor analysis adjusted null p-values, where the factors are required to be orthogonal to the group difference, show strong anti-conservative bias (Figure 5d). The reason is that the orthogonally estimated factors do not account for potential confounding between the tested variable and the unmodeled variable. However, when the unmodeled variable is nearly orthogonal to the group variable by chance, this approach behaves reasonably well and so the 75th percentile of the posterior probability estimates is 0.810.

Table 1: *The posterior probability distribution and the double KS test p-value assessing whether the JNC holds for each method adjusting for multiple testing dependence. Correctly Adjusted = adjusted for the true underlying latent variable, SV = surrogate variable analysis, EN = empirical null, and RF = residual factor analysis.*

Method	Post. Prob. (IQR)	dKS P-value
Correctly Adjusted	0.967 (0.928,0.978)	0.446
SV Adjusted	0.961 (0.918,0.975)	0.132
EN Adjusted	0.000 (0.000,0.000)	< 2e-16
EN Adjusted (95% Null)	0.685 (0.081,0.961)	1.443e-14
RF Adjusted	0.000 (0.000,0.810)	< 2e-16

Supplementary Figures S3 and S4 show the estimates of the FDR and the proportion of true nulls calculated for the same simulated studies. Again, the estimates using the correct model and surrogate variable analysis perform similarly, while the empirical null estimates are conservatively biased and the residual factor analysis p-values are anti-conservatively biased. For comparison purposes, Supplementary Figure S5 shows the behavior of the unadjusted p-values and their corresponding false discovery rate estimates. It can be seen that since surrogate variables analysis satisfies the JNC, it produces false discovery rate estimates with a variance and a conservative bias close to the correct adjustment. However, the empirical null adjustment and residual factor analysis produce substantially biased estimates. The unadjusted analysis produces estimates with a similar expected value to the correct adjustment, although the variances are very large.

Another way to view this analysis is to consider the sensitivity and specificity of each approach. The ROC curves for each of the four proposed methods are shown in Supplementary Figure S6. The approaches that pass the JNC criteria - the correctly adjusted analysis and the surrogate variable adjusted analysis - have

similarly high AUC values, while the approaches that do not pass the JNC (residual factor analysis and empirical null) have much lower AUC values. This suggests that another property of the JNC is increased sensitivity and specificity of multiple testing procedures.

5.2 Pooled Null Distributions

A second challenge encountered in large-scale multiple testing in genomics is in determining whether it is valid to form an averaged (called “pooled”) null distribution across multiple tests. Bootstrap and permutation null distributions are common for high-dimensional data, where parametric assumptions may be difficult to verify. It is often computationally expensive to generate enough null statistics to make test-specific empirical p-values at a fine enough resolution. (This requires at least as many resampling iterations as there are tests.) One proposed solution is to pool the resampling based null statistics across tests when forming p-values or estimating other error rates (Tusher et al., 2001, Storey and Tibshirani, 2003). By pooling the null statistics, fewer bootstrap or permutation samples are required to achieve a fixed level of precision in estimating the null distribution. The underlying assumption here is that averaging across all tests’ null distributions yields a valid overall null distribution. This approach has been criticized based on the fact that each p-value’s marginal null distribution may not be $U(0,1)$ (Dudoit, Shaffer, and Boldrick, 2003). However, the JNC allows for this criticism to be reconsidered by considering the joint distribution of pooled p-values.

Consider the simulated data from the previous subsection, where $\mathbf{x}_i = b_{0i} + b_{1i}\mathbf{y} + \varepsilon_i$. Suppose that $b_{1i} \neq 0$ for $i = 1, \dots, 500$, $b_{1i} = 0$ for $i = 501, \dots, 1000$, and $\text{Var}(\varepsilon_{ij}) \sim \text{InverseGamma}(10, 9)$. Suppose $y_j = 1$ for $j = 1, 2, \dots, n/2$ and $y_j = 0$ for $j = n/2 + 1, \dots, n$. We can apply the t-statistic to quantify the difference between the two groups for each test. We compute p-values in one of two ways. First we permute the labels of the samples and recalculate null statistics based on the permuted labels. The p-value is the proportion of permutation statistics that is larger in absolute value than the observed statistic.

A second approach to calculating the null statistics is with the bootstrap. To calculate bootstrap null statistics, we fit the model $\mathbf{x}_i = b_{0i} + b_{1i}\mathbf{y} + \varepsilon_i$ by least squares and calculate residuals $\mathbf{r}_i = \mathbf{x}_i - \hat{b}_{0i} - \hat{b}_{1i}\mathbf{y}$. We calculate a null model fit using the model $\mathbf{x}_i = b_{0i}^0 + \varepsilon_i$, sample with replacement from the residuals \mathbf{r} to obtain bootstrapped residuals \mathbf{r}_i^* , rescale the bootstrapped residuals to have the same variance as the original residuals, and add the bootstrapped residuals to the null model fit to obtain null data $\mathbf{x}_i^* = \hat{b}_{0i}^0 + \mathbf{r}_i^*$. The p-value is the proportion of bootstrap

Method	Post. Prob. (IQR)	dKS P-value
T-Statistic/Perm./Test-Specific	0.436 (0.006,0.904)	0.171
T-statistic/Perm./Pooled	0.966 (0.946,0.979)	0.850
T-statistic/Boot./Test-Specific	0.508 (0.002,0.946)	0.181
T-statistic/Boot./Pooled	0.967 (0.942,0.977)	0.759
ODP/Perm./Test-Specific	0.748 (0.024,0.955)	0.068
ODP/Perm./Pooled	0.000 (0.000,0.000)	0.000
ODP/Boot./Test-Specific	0.000 (0.000,0.178)	0.127
ODP/Boot./Pooled	0.971 (0.946,0.980)	0.121

Table 2: The posterior probability distribution and the double KS test p-value assessing whether the JNC holds for each type of permutation or bootstrap analysis.

statistics that is larger in absolute value than the observed statistic. This is the bootstrap approach employed in Storey, Dai, and Leek (2007).

We considered two approaches to forming resampling based p-values: (1) a pooled null distribution, where the resampling based null statistics from *all* tests are used in calculating the p-value for test *i* and (2) a test-specific null distribution, where only the resampling based null statistics from test *i* are used in calculating the p-value for test *i*. Table 2 shows the results of these analyses for all four scenarios with the number of resampling iterations set to $B = 200$. The pooled null outperforms the marginal null because the marginal null is granular, due to the relatively small number of resampling iterations. The pooling strategy is effective because the t-statistic is a pivotal quantity, so its distribution does not depend on unknown parameters. In this case, the test-specific null distribution can reasonably be approximated by the joint null distribution that comes from pooling all of the null statistics.

Many statistics developed for high-dimensional testing that borrow information across tests are not pivotal. Examples of non-pivotal statistics include those from SAM (Tusher et al., 2001), the optimal discovery procedure (Storey et al., 2007), variance shrinkage (Cui et al., 2005), empirical Bayes methods (Efron et al., 2001), limma (Smyth, 2004), and Bayes methods (Gottardo, Pannucci, Kuske, and Brettin, 2003). As an example, to illustrate the behavior of non-pivotal statistics under the four types of null distributions we focus on the optimal discovery procedure (ODP) statistics. The ODP is an extension of the Neyman-Pearson paradigm to tests of multiple hypotheses (Storey, 2007). If m tests are being performed, of which m_0 are null, the ODP statistic for the data \mathbf{x}_i for test i is given by: $S_{odp}(\mathbf{x}_i) = \frac{\sum_{m_0+1}^m f_{1i}(\mathbf{x}_i)}{\sum_{i=1}^{m_0} f_{0i}(\mathbf{x}_i)}$, where f_{1i} is the density under the alternative and f_{0i} is

the density under the null for test i . When testing for differences between group A and group B, an estimate for the ODP test statistic can be formed using the Normal probability density function, $\phi(\cdot; \mu, \sigma^2)$:

$$\widehat{S}_{odp}(\mathbf{x}_j) = \frac{\sum_{i=1}^m \phi(\mathbf{x}_{Aj}; \widehat{\mu}_{Ai}, \widehat{\sigma}_{Ai}^2) \phi(\mathbf{x}_{Bj}; \widehat{\mu}_{Bi}, \widehat{\sigma}_{Bi}^2)}{\sum_{i=1}^m \phi(\mathbf{x}_j; \widehat{\mu}_{0i}, \widehat{\sigma}_{0i}^2)}$$

The ODP statistic is based on the estimates of the mean and variance for each test under the null hypothesis model restrictions $(\widehat{\mu}_{0i}, \widehat{\sigma}_{0i}^2)$ and unrestricted $(\widehat{\mu}_{Ai}, \widehat{\sigma}_{Ai}^2, \widehat{\mu}_{Bi}, \widehat{\sigma}_{Bi}^2)$. The data for each test \mathbf{x}_j is substituted into the density estimated from each of the other tests. Like variance shrinkage, empirical Bayes, and Bayesian statistics, the ODP statistic is not pivotal since the distribution of the statistic depends on the parameters for all of the tests being performed.

We used the ODP statistics instead of the t-statistics under the four types of null distributions; the results appear in Table 2. With a non-pivotal statistic, pooling the permutation statistics results in non-uniform null p-values. The variance of the permuted data for the truly alternative tests is much larger than the variance for the null tests, resulting in bias. The test-specific null works reasonably well under permutation, since the null statistics for the alternative tests are not compared to the observed statistics for the null tests. The bootstrap corrects the bias, since the residuals are resampled under the alternative and adjusted to have the same residual variance as the original data. The bootstrap test-specific null distribution yields granular p-values causing the Bayesian diagnostic to be unfavorable, but yielding a favorable result from the double KS test. The pooled bootstrap null distribution meets the JNC in terms of both diagnostic criteria. These results suggest that non-pivotal high-dimensional statistics that employ permutations for calculating null statistics may result in non-uniform p-values when the null statistics are pooled, but those that employ variance adjusted bootstrap pooled distributions meet the JNC. It should be noted that Storey et al. (2007) prescribe using the pooled bootstrap null distribution as implemented here and the permutation null distribution is not advocated.

Our results suggest that the double KS test may be somewhat sensitive to outliers, suggesting that it may be most useful when strict adherence to the JNC is required from a multiple testing procedure. Meanwhile, the Bayesian approach is sensitive to granular p-value distributions commonly encountered with permutation tests using a small sample, suggesting it may be more appropriate for evaluating parametric tests or high-dimensional procedures that pool null statistics.

6 Discussion

Biological data sets are rapidly growing in size and the field of multiple testing is experiencing a coordinated burst of activity. Existing criteria for evaluating these procedures were developed in the context of single hypothesis testing. Here we have proposed a new criterion based on evaluating the joint distribution of the null statistics or p-values. Our criterion is more stringent than requiring strong control of specific error rates, but flexible enough to deal with the type of multiple testing procedures encountered in practice. When the Joint Null Criterion is met, we have shown that standard error rates can be precisely and accurately controlled. We have proposed frequentist and Bayesian diagnostics for evaluating whether the Joint Null Criterion has been satisfied in simulated examples. Although these diagnostics can not be applied in real examples, they can be a useful tool to diagnose multiple testing procedures when they are proposed and evaluated in simulated data. Here we focused on two common problems in multiple testing that arise in genomics, however our criterion and diagnostic tests can be used to evaluate any multiple testing procedure to ensure p-values satisfy the JNC and result in precise error rate estimates.

References

- Benjamini, Y. and Y. Hochberg (1995): “Controlling the false discovery rate—a practical and powerful approach to multiple testing,” *J Roy Stat Soc B*, 57, 289–300.
- Benjamini, Y. and D. Yekutieli (2001): “The control of the false discovery rate in multiple testing under dependency,” *Ann Stat*, 29, 1165–88.
- Calian, V., D. Li, and J. Hsu (2008): “Partitioning to uncover conditions for permutation test to control multiple testing error rate,” *Biometrical Journal*, 50, 756–766.
- Cui, X., J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill (2005): “Improved statistical tests for differential gene expression by shrinking variance components estimates,” *Biostatistics*, 6, 59–75.
- Devlin, B. and K. Roeder (1999): “Genomic control for association studies,” *Biometrics*, 55, 997–1004.
- Dudoit, S., J. P. Shaffer, and J. C. Boldrick (2003): “Multiple hypothesis testing in microarray experiments,” *Statistical Science*, 18, 71–103.
- Dudoit, S. and M. J. van der Laan (2008): *Multiple Testing Procedures with Applications to Genomics*, Springer.
- Efron, B. (2004): “Large-scale simultaneous hypothesis testing: The choice of a null hypothesis,” *J Am Stat Assoc*, 99, 96–104.

- Efron, B. (2007): “Correlation and large-scale simultaneous significance testing,” *J Am Stat Assoc*, 102, 93–103.
- Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001): “Empirical bayes analysis of a microarray experiment,” *Journal of Computational Biology*, 96, 1151–60.
- Friguet, C., M. Kloareg, and D. Causer (2009): “A factor model approach to multiple testing under dependence.” *Journal of the American Statistical Association*, to appear.
- Gottardo, R., J. A. Pannucci, C. R. Kuske, and T. Brettin (2003): “Statistical analysis of microarray data: a bayesian approach,” *Biostatistics*, 4, 597–620.
- Huang, Y., H. Xu, V. Calian, and J. Hsu (2006): “To permute or not to permute,” *Bioinformatics*, 22, 2244–2248.
- Idaghdour, Y., J. D. Storey, S. Jadallah, and G. Gibson (2008): “A genome-wide gene expression signature of lifestyle in peripheral blood of moroccan amazighs,” *PLoS Genetics*, 4, e1000052.
- Leek, J. T. and J. D. Storey (2007): “Capturing heterogeneity in gene expression studies by surrogate variable analysis,” *PLoS Genetics*, 3, e161.
- Leek, J. T. and J. D. Storey (2008): “A general framework for multiple testing dependence.” *Proc. Nat. Acad. Sci. U.S.A.*, 105, 18718–18723.
- Lehmann, E. L. (1997): *Testing Statistical Hypotheses*, Springer.
- Newton, M. A., A. Noueir, D. Sarkar, and P. Ahlquist (2004): “Detecting differential gene expression with a semiparametric hierarchical mixture method.” *Biostatistics*, 5, 155–76.
- Owen, A. (2005): “Variance of the number of false discoveries,” *J Roy Stat Soc B*, 67, 411–26.
- Pounds, S. and S. W. Morris (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.
- Schwartzman, A., R. F. Dougherty, and J. Taylor (2008): “False discovery rate analysis of brain diffusion direction maps.” *Ann Appl Stat*, 2, 153–175.
- Shaffer, J. P. (1995): “Multiple hypothesis testing,” *Annu. Rev. Psychol.*, 46, 561–84.
- Shorack, G. R. and J. A. Wellner (1986): *Empirical Processes with Applications to Statistics*, Wiley.
- Smyth, G. K. (2004): “Linear models and empirical bayes methods for assessing differential expression in microarray experiments.” *Statistical Applications in Genetics and Molecular Biology*, 1, 3.
- Storey, J. D. (2002): “A direct approach to false discovery rates,” *J Roy Stat Soc B*, 64, 479–98.
- Storey, J. D. (2007): “The optimal discovery procedure: A new approach to simultaneous significance testing.” *J Roy Stat Soc B*, 69, 347–68.

- Storey, J. D., J. Y. Dai, and J. T. Leek (2007): “The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments,” *Biostatistics*, 8, 414–32.
- Storey, J. D., J. E. Taylor, and D. Siegmund (2004): “Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach,” *J Roy Stat Soc B*, 66, 187–205.
- Storey, J. D. and R. Tibshirani (2003): “Statistical significance for genome-wide studies,” *Proc Natl Acad Sci USA*, 100, 9440–9445.
- Tusher, V. G., R. Tibshirani, and G. Chu (2001): “Significance analysis of microarrays applied to the ionizing radiation response,” *Proc Natl Acad Sci, U.S.A.*, 98, 5116–21.
- Xu, H. and J. Hsu (2007): “Using the partitioning principle to control the generalized family error rate,” *Biometrical Journal*, 49, 52–67.
- Yekutieli, D. and Y. Benjamini (1999): “Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics,” *J Statist Plan Inf*, 82, 171–96.