Abstract

Gene expression studies produce data with which inferences can be made for thousands of genes simultaneously. Standard practice in multiple testing with gene expression data is to use t-statistics as test statistics and to control the error rate under the permutation distribution. In this paper, we revisit the rationale behind such choices and suggest situations in which alternatives are more sensible. We propose an optimal null distribution for testing hypotheses about gene-specific...
parameters which is a Kullback-Leibler projection parameter of the true data generating distribution on the submodel defined by the overall null hypothesis. We propose to estimate this null distribution with maximum likelihood estimation. One can control error rates by resampling from this estimated null distribution. We contrast the optimal null distribution with the null distribution estimated by permutation methods and illustrate that the two are very different in the two sample problem whenever sample sizes are not equal. With real and simulated gene expression data, we evaluate the finite sample performance of different choices of test statistics, estimated null distributions and multiplicity adjustments. Parametric bootstrap methods perform best when the model for the data is known, but since this is rarely true in practice, non-parametric bootstrap methods are recommended. We also find that the standardized t-statistic’s null distribution is harder to estimate than that of the difference in means.

Key words: multiple testing, permutation, bootstrap, gene expression.

Abbreviated title: Multiple testing for gene expression data.

AMS classification: primary 62H15, secondary 62H10

1 Motivation

A gene expression study results in an observed data matrix $X$ whose columns are $n$ copies of a $p$-dimensional vector of measurements, where $n$ is the number of observations, $p$ is the number of variables (e.g: genes), and typically $p > 5000$ while $n < 50$. Researchers frequently want to perform a statistical test for every gene in order to answer questions such as “Which genes
are significantly differently expressed between two (or more) conditions?” or “Which genes have a significant association with an outcome or covariate?”. Ordering the genes based on statistical significance helps one to formulate biological hypotheses and allows one to prioritize the genes for follow-up experiments, such as drug target validation or in situ studies. In addition, identifying such statistically significant subsets of genes reduces the dimension of the data before further analysis, such as clustering or classification. In order to make statements about the statistical significance of thousands of genes simultaneously, however, it is essential to appropriately account for multiple tests.

Traditional approaches to the multiplicity problem are reviewed by Hochberg and Tamhane (1987). More recent developments in the field include resampling methods (Westfall and Young (1993)), step-wise procedures, and less conservative error rate control, such as control of the false discovery rate (Benjamini and Hochberg (2002)).

Multiple testing in the gene expression context must account for

1. a multiplicity problem of much larger scale than in more classical multiple testing situations;

2. the dimension of the data being much larger than the sample size;

3. the correlation structure between the genes;

4. the fact that genes do not share a common marginal null distribution; and

5. the knowledge that some proportion of the genes (frequently 10-30%)
should have significant affects (e.g.: different expression between groups or associations between expression and outcomes).

Dudoit et al. (2002) discuss several of these issues, conduct a comparison study, and review some recent applications of multiple testing to expression data, noting where the methods have been used incorrectly or the assumptions under which they work have been violated. We have investigated these problems further. The goal of this paper is to discuss the problem of multiple testing for gene expression data in a statistical framework, thereby illustrating which methods are suitable for this context. Our results have important implications for multiple testing with gene expression data as well as other types of high dimensional data.

2 Data and Null Hypotheses

Let $X_1, \ldots, X_n$ be i.i.d. $X \sim P \in \mathcal{M}$, where $\mathcal{M}$ is a model and $X$ is a $p$-dimensional vector that may include gene expression, and possibly covariates and outcomes. One outcome of interest is survival, which may be subject to right censoring. Consider real valued parameters $\mu_1, \ldots, \mu_p$, that is $\mu_j(P) \in \mathbb{R}$. The gene-specific parameters $\mu_1, \ldots, \mu_p$ could be, for example, location parameters (e.g.: means/medians or differences between two population means/medians) or regression parameters (e.g.: association between $x_j$ and $Y$ in a linear/logistic model). Typically, we are interested in simultaneously testing the null hypotheses:

$$H_{0,j} : \mu_j(P) = \mu_j^0, j = 1, \ldots, p,$$  \hspace{1cm} (1)

where the $\mu_j^0$ are hypothesized null values, frequently zero.
3 Multiple Testing Procedures

Given i.i.d. $X_1, \ldots, X_n$ and a family of null hypotheses $\{H_{0,j}, j = 1, \ldots, p\}$, we define a large class of multiple testing procedures (excluding FDR controlling procedures) by specific choices of (i) test statistics, (ii) type of error control, (iii) null distribution, and (iv) testing method.

3.1 Test Statistics

Let $\hat{\mu}_j$ be an efficient or locally efficient estimate of $\mu_j$. We refer to van der Laan and Robins (2002) for locally efficient estimators dealing with the curse of dimensionality. Then, an obvious choice of test statistics is:

$$T_j = \hat{\mu}_j - \mu_j^0, j = 1, \ldots, p.$$  \hspace{1cm} (2)

A statistic $T_j$ that is sufficiently large in absolute value (or in a certain direction for a one-sided test) represents significant evidence against the null hypothesis $H_{0,j}$. Note that it is important to efficiently estimate the parameters of interest $\mu_j$, since doing so provides test statistics $T_j$ with maximal power.

Other choices of test statistics include $T_j = \sqrt{n}(\hat{\mu}_j - \mu_j^0)/sd(\hat{\mu}_j)$. If $\hat{\mu}_j$ is asymptotically linear with influence curve $IC_j(X)$, that is, $\hat{\mu}_j - \mu_j = \frac{1}{n} \sum_{i=1}^n IC_j(X_i) + op(\frac{1}{\sqrt{n}})$, and $sd(\hat{\mu}_j)$ is an estimate of $\sigma_j = \sqrt{VAR(IC_j(X))}/n$, then

$$T_j = \frac{\hat{\mu}_j - \mu_j^0}{sd(\hat{\mu}_j)} \xrightarrow{n \to \infty} N(0,1).$$

Standardizing test statistics so that the asymptotic marginal distributions of all $T_j$ are $N(0,1)$ is a useful tool when one wishes to use tabled null
distributions. Figure 1 shows that in the gene expression context, however, marginal null distributions are far from being equal to $N(0,1)$, even for standardized test statistics and reasonably large sample sizes. In particular, estimation of $sd(\hat{\mu}_j)$ is known to be difficult in the gene expression context (Tusher et al. (2001), Rocke and Durbin (2001)). Thus, we do not recommend using tabled distributions or assuming a common null distribution for all genes. This also eliminates the need to use standardized test statistics. We revisit this issue in the simulations of Section 4.5, where we compare the statistics in Equation 2 to their standardized counterparts and show that it is easier to estimate the null distribution of non-standardized statistics (See Table 1).

### 3.2 Error Control

Multiple testing procedures can be assessed based on estimates of how many erroneous testing decisions they make.

#### 3.2.1 Type I Error Rates

We assume the reader is familiar with the distinction between type I (false positive) and type II (false negative) errors in the standard univariate setting, where the typical approach is to control the type I error rate at a pre-specified level $\alpha$ and compare different procedures with type I error rate $\alpha$ based on their type II error rates (or power). Dudoit et al. (2002) compare different generalizations of type I error control to the multiple testing setting. Let $R$ be the total number of rejected hypotheses, let $V$ be the (unobservable) number of false rejections, and let $k$ be a user supplied constant. Some error
Figure 1: Histograms of null distributions of standardized t-statistics for four genes from the DLBCL data set of Alizadeh et al. (2000) computed by bootstrapping the centered empirical null distribution. The value of the 0.975 quantile of each distribution is given in the title. The Student’s T distribution with appropriate degrees of freedom ($df = 38$) is superimposed on each histogram, showing that the distributions can be heavy/light in the tails or quite skewed. The 0.975 quantile of the T distribution is 2.0.

Rates include:

- PCER = $E(V)/p$ : per-comparison error rate,
- PFER = $E(V)$ : per-family error rate,
- FWER = $Pr(V \geq k)$: family-wise error rate,

- FDR = \[
\begin{align*}
E(V/R) & \quad R \geq 0 \quad : \text{false discovery rate.} \\
0 & \quad R = 0
\end{align*}
\]

In general, the per-family error rate is most conservative and the per-comparison error rate (ignoring the multiplicity problem) is the least conservative (Dudoit et al. (2002)). In the gene expression context, a less conservative error rate is often preferred since researchers view gene expression experiments as exploratory methods and are usually interested in obtaining a fairly large list of candidate genes, even if some proportion of these are likely to be false positives. For this reason, the false discovery rate (Benjamini and Hochberg (2002)) is becoming a popular choice of error rate.

### 3.2.2 Strong Control

Error rates are defined under the true data generating distribution $P$, so that they depend on which hypotheses are in fact true. In practice, we do not know which hypotheses are true, so we have to choose a way to compute the expectations and/or probabilities in the error rate. Weak control means that the error rate is controlled under a null distribution $Q_0$ satisfying the complete null hypothesis $H_0^C = \bigcap_{j=1}^p H_{0,j}$, i.e.: only if all hypotheses are true. In gene expression studies, however, it is usually expected that some of the hypotheses are false, so that weak control is not sufficient. Strong control means that the error rate is controlled under any combination of true and false hypotheses, so that in particular it is controlled under the true distribution $P$. Since control of the error rate under $P$ is our goal, we would like a testing procedure to have strong control (at least asymptotically). While none of the
multiple testing procedures in the literature in fact guarantees strong control for finite samples (and large $p$) without extreme parametric assumptions, we recommend using procedures which at least aim to have strong control as $n \to \infty$. We discuss asymptotic strong control further in the following Section 3.2.3.

The FDR method of Benjamini and Hochberg (2002) takes a different approach. This method does not have level $\alpha$ when the complete null $H_0^C$ is true, but it does control $E(V/R)$ under the truth given certain assumptions about the data generating distribution. The methods proposed here can not handle strong control of the FDR. We note, however, that weak control of the FDR under $P_0 \in \mathcal{M}_0$ (as defined in Section 3.3.1) is equivalent to weak control of the FWER.

### 3.2.3 Asymptotic Strong Control

Let $\alpha_n$ denote the error rate for a sample of size $n$ and consider a target error rate $\alpha$. Asymptotic strong control can then be defined as $\alpha_n \to \alpha$ as $n \to \infty$. In order to guarantee asymptotic strong control, we need convergence in distribution of the test statistics. There are a few situations in which this is the case. First, if the dimension of the data $p$ were finite, then the usual central limit theorem would apply and the asymptotic distribution of the standardized test statistics (as $n \to \infty$) would be multivariate normal. In gene expression studies, however, the number of genes grows with the number of samples. Consider now the situation where $p > n$ so that $p = p(n) \to \infty$ faster than $n$. Under a parametric model for the observed data, one might be able to prove that the distribution of the test statistics is arbitrarily well
approximated by a multivariate normal distribution for $n \to \infty$. Note that when $p \gg n$ there is no multivariate central limit theorem, so that proving an approximation by a multivariate normal will only be possible with restrictive parametric assumptions on the observed data. In the gene expression context, however, we rarely believe such a parametric model. In practice, for any $n$ there will typically be some genes whose marginal distribution is not yet normal. Thus, without the existence of a multivariate normal approximation or limit distribution it is very doubtful that multiple testing procedures will have asymptotic strong control, but this is an area of future research.

We present a few preliminary ideas on this topic. First, it is clear that some error rates should be harder to control than others because they depend on the most extreme gene(s) (e.g.: family-wise error). Second, parameters whose estimators have second order terms (e.g.: regression coefficients) will make error control harder than with sample means. Third, what we can say about the asymptotic distribution of the test statistics depends on the rate at which $p \to \infty$ relative to $n$. Consider the example studied by van der Laan and Bryan (2001), in which $\frac{n}{\log p} \to \infty$. Suppose the parameters of interest are the sample means and the test statistics are $T_j = \sqrt{n}(\hat{\mu}_j - \mu_0)$, $j = 1, \ldots, p$. Let $\Sigma$ denote the covariance matrix of $X$ and let $\Sigma_n$ be the empirical covariance matrix. Then, if the minimum eigen value of $\Sigma$ is bounded away from zero, we have shown (van der Laan and Bryan (2001))

1. $\max_{i,j} |\Sigma_{n,i,j} - \Sigma_{i,j}| \to 0$,
2. $\max_{i,j} |\Sigma^{-1}_{n,i,j} - \Sigma^{-1}_{i,j}| \to 0$.

This does not, however, guarantee that $T \xrightarrow{n \to \infty} N(0, \Sigma)$. If, in addition,
one assumes $X \sim N(\mu, \Sigma)$, then $T \sim N(0, \Sigma)$, which suggests that in this parametric family one should control the error rate under $N(0, \Sigma_n)$. Using the results of van der Laan and Bryan (2001), one can now establish asymptotic strong control of such a procedure. This argument can be generalized to asymptotically linear test statistics $T = \{T_j : j = 1, \ldots, p\}$ with influence curves $IC(X) = \{IC_j(X) : j = 1, \ldots, p\}$ with negligible second order terms (Pollard and van der Laan (2001)).

Given these remarks, it seems unlikely that any multiple testing procedure can achieve asymptotic strong control in the gene expression context. Assuming a parametric model gives asymptotic strong control if the model is correct. But, in practice the model may be far from the truth (at least for some genes), so that one would do better using another choice of null distribution, such as an empirical null distribution. We propose multiple testing procedures which aim to have asymptotic strong control (of PCER, PFER, or PWER), that is they would achieve asymptotic strong control for finite $p$ or $p \to \infty$ slow enough, and we compare these methods based on their finite sample performance.

### 3.3 Null Distribution

In order to decide if any of the observed test statistics are sufficiently unusual to reject the corresponding null hypotheses, we compare them to their joint distribution under the null (Equation 1).
3.3.1 Null Distribution as a Projection Parameter of $P$

In order to control the type I error, we need the null distribution of the data to resemble the true data generating distribution $P$ except with the additional property that Equation 1 holds for all $j$. To understand what this means, first suppose that the complete null $H_0^C$ is true. Then the true distribution $P$ lies in the space $\mathcal{M}_0 = \{ P \in \mathcal{M} : \mu_j(P) = \mu_j^0, j = 1, \ldots, p \} \subset \mathcal{M}$ of all distributions in the model $\mathcal{M}$ for which Equation 1 holds. For example, let $P$ be parametrized by $\mu$ and variation independent nuisance parameters $\eta$. Since the nuisance parameters $\eta$ are not constrained by the hypotheses $H_{0,j}$, then $\mathcal{M}_0 = \{ P_{\mu,\eta} : \mu = \mu^0, \eta \}$. Now, in practice $P$ probably does not lie in $\mathcal{M}_0$, but we can not know if this is the case or not. Hence, we propose to use as null distribution a projection $P_0 = \Pi(P|\mathcal{M}_0)$ of $P$ onto $\mathcal{M}_0$ using a particular distance metric $d(\cdot, \cdot)$.

This data null distribution $P_0$ will be as close to $P$ as possible, and if $H_0^C$ holds then $d(P, \mathcal{M}_0) \equiv \inf\{d(P, P_0) : P_0 \in \mathcal{M}_0\} = 0$. Suppose we use the Kullback-Leibler (K-L) distance. Then we have:

$$
P_0 = P_0(P) = \arg \max_{P_0 \in \mathcal{M}_0, P_0 \ll \mu} \int \log \left( \frac{\partial P_0'(x)}{\partial \mu(x)} \right) dP(x). \tag{3}
$$

Note this defines the optimal null distribution $P_0$ as a function of the true distribution $P$.

The data null distribution $P_0$ directly implies a joint null distribution $Q_0$ for the test statistics. For example, in a shift experiment where the parameter of interest is a location parameter, then for a non-parametric model and test statistics given by Equation 2, $Q_0 = Q(\cdot - \mu^0)$, where $Q$ is the underlying distribution of the test statistics.
3.3.2 Estimation of Null Distribution

In practice, we do not know the true distribution $P$ so that we need to estimate $P_0$ in order to estimate $Q_0$. We can use the substitution estimator

$$P_0(P_n) = \arg \max_{P_0 \in M_0, P_0 \ll \mu_n} \int \log \left( \frac{\partial P_0'(x)}{\partial \mu_n(x)} \right) dP_n(x), \quad (4)$$

which can be defined as a maximum likelihood estimate $\hat{P}_{0,MLE}$ of $P_0$ under the model $M_0$, where $\mu_n$ is a user supplied dominant measure. Then, different choices of the model $M$ for $P$ define different estimators. For example, if $M$ is non-parametric and $\mu_j$ is a location parameter of $x_j$, then $\hat{P}_{0,MLE} = (P_n - \mu^0)$. A disadvantage of this estimator is that for small $n$ the marginal distributions are very discrete with many ties. If $M = N(\mu, \Sigma)$, then $\hat{P}_{0,MLE} = N(\mu^0, \Sigma_n)$. This estimator requires the parametric assumption that the data generating distribution is multivariate normal. For both approaches, the null distribution of the test statistics $Q_0$ is estimated by generating a large number $B$ of bootstrap data sets from the chosen distribution $\hat{P}_{0,MLE}$ and computing the test statistics $T^b_j$ for $b = 1, \ldots, B$. Then, the distribution of the test statistics $T^b_j$ over the $B$ bootstrap data sets estimates $Q_0$.

One advantage of using an estimate $\hat{Q}_{0,MLE}$ of $Q_0$ (derived from the estimated projection $\hat{P}_{0,MLE}$) as the null distribution is that the true $Q_0$ would give strong control of the error rate. Hence, if $p$ were finite then the proposed $\hat{Q}_{0,MLE}$ would give asymptotic strong control as $n \to \infty$. And in general (i.e.: $p \gg n$), $\hat{Q}_{0,MLE}$ at least aims to have strong control in the limit, and we can compare different estimates of $Q_0$ based on their finite sample properties. A directly related benefit of this approach is that any nuisance parameters $\eta$
are not allowed to asymptotically vary from their true values so that under asymptotics $\hat{\eta} \rightarrow \eta$ any rejection can be directly attributed to the parameter of interest (e.g.: $\mu_j \neq \mu_j^0$ for some $j$) and not to $\eta$.

### 3.4 Testing Method

Given an estimated joint null distribution $\hat{Q}_0$, we reject $H_{0,j}$ whenever $T_j$ lies in the tails of the marginal distribution $\hat{Q}_{0,j}, j = 1, \ldots, p$. In traditional testing settings, a common threshold $c$ is used to make the testing decision for every variable, i.e.: reject if $|T_j| > c(\alpha)$ for a specified level $\alpha$. A generalization of the common threshold approach is to use a common quantile for each gene, producing gene-specific thresholds $c_j(\alpha)$. This corresponds with a common threshold only if the marginal distributions have identical tail probabilities, which is not usually the case. Note that single-step max $T$ and min $p$ procedures (Westfall and Young (1993)) produce a common threshold for all genes even though they do use the joint null distribution.

Suppose, for example, that we use a resampling-based null distribution (as defined in Section 3.3.2). If the estimated null distribution is sufficiently smooth (i.e.: $B$ large enough), gene-specific thresholds can be found which achieve a sharp bound for any choice of error rate. Consider the vector of thresholds $\{c_j : j\}$ defined by

$$
\frac{1}{B} \sum_{b=1}^{B} I \left\{ \sum_{j=1}^{p} I\{|T_j^b| > c_j\} > k \right\} < \alpha,
$$

where $B$ is the number of independent bootstrap data sets resampled and $k$ is a pre-specified number of false positives. When $k = 1$ this is the usual family-wise error rate, and when $k > 1$ this controls the expected number
of false positives $E(V) \leq k$ with probability $1 - \alpha$. The common quantile method can be further refined by using a step-down (or step-up) method to adjust the quantiles (See Pollard and van der Laan (2002)).

An alternative approach to multiple testing is to compute a p-value (i.e.: the probability of observing a statistic as or more extreme than $T_j$), adjust it for multiple tests, and simply report this adjusted p-value. Adjusted p-values are defined as the level of the entire testing procedure at which $H_{0,j}$ would just be rejected, given all of the statistics $T_j$. Rejection decisions can later be made by comparing the adjusted p-values to the level $\alpha$ of the test for different choices of $\alpha$. Westfall and Young (1993), Yekutieli and Benjamini (1999), and Dudoit et al. (2002) discuss methods for computing adjusted p-values. These methods usually require more computation than choosing thresholds $c_j$ to control a multiple testing type I error rate directly. For a more detailed comparison of p-value and threshold methods, we refer the reader to Pollard and van der Laan (2002).

4 Example: Two Sample Problem

As a specific example, consider the two sample multiple testing problem, where we observe $p$ variables (e.g.: gene expression measurements) on each subject. The conclusions drawn here can be extended to other testing problems.
4.1 Data and Null Hypotheses

Suppose we observe $n_1$ observations from population 1 and $n_2$ from population 2. We can think of the data as $(X_i, L_i)$, where $X_i$ is the multivariate expression vector $X_{ij}, j = 1, \ldots, p$ for subject $i$ and $L_i \in \{1, 2\}$ is a label indicating subject $i$'s group membership. The two populations can have different gene expression distributions, so that $P(X_i | L_i)$ is one of two different distributions $P_1, P_2$ depending on the value of $L_i$. Let $\mu_{1,j}$ and $\mu_{2,j}$ denote the means of gene $j$ in populations 1 and 2, respectively. Suppose we are interested in testing

$$H_{0,j} : \mu_j \equiv \mu_{2,j} - \mu_{1,j} = 0, j = 1, \ldots, p.$$  \hfill (6)

We can select test statistics, an error control rate, and a resampling-based null distribution as described in Section 3.

4.2 Permutation Null Distribution

Another approach to the two sample problem, which has been applied frequently in the gene expression literature, is to use permutations of the group labels to compute the null distribution. The assumption behind this approach is that the data are identically distributed in both populations, which is often a stronger assumption than we wish to make. In order to understand the permutation distribution within the framework presented in this paper, we first consider the two true null distributions corresponding with permutations and bootstrap resampling and then discuss implications for the corresponding estimated distributions.
Consider a model $\mathcal{M}$ for this two sample problem (e.g.: non-parametric $P_1$ for population 1 and $P_2$ for population 2). First, suppose we are testing the null hypothesis $H_0: P_1 = P_2$. In this case, we have $\mathcal{M}_0^1 = \{P_1 = P_2\}$ and $P_0^1(P) = \Pi(P|\mathcal{M}_0^1)$ (e.g.: as defined in Equation 3). So, $P_0^1(P)$ can be estimated with permutation resampling. Specifically, the distribution of the test statistics over many (possibly all, depending on the sample sizes $n_1, n_2$ and computing power) permutations of the group labels $L_i$ provides a null distribution estimate $P_0^1(P_n)$. Second, suppose we are testing the null hypotheses $H_{0,j}$ given by Equation 6. Then, we have $\mathcal{M}_0^2 = \{\mu_1 = \mu_2\}$ and $P_0^2(P) = \Pi(P|\mathcal{M}_0^2)$. It is clear that permutation resampling is not appropriate for the estimation of $P_0^2$, since $\mu_{1,j} = \mu_{2,j}, j = 1, \ldots, p$ does not imply $P_1 = P_2$. Instead, the bootstrap estimates defined in Section 3.3.2 are recommended.

Thus, the permutation null distribution can be quite problematic for testing the hypotheses $H_{0,j}$ given by Equation 6. Another way to view this problem is to notice that the permutation null distribution lies in $\mathcal{M}_0^2$ but may be quite far from the projection of $P$ onto $\mathcal{M}_0^2$. Consequently, a null hypothesis $H_{0,j}$ may be rejected because the distribution of $T_j$ under permutations depends on some nuisance parameter which is quite different from its permutation null value (an “average” over the two groups), rather than because $\mu_{1,j} \neq \mu_{2,j}$. The following algebraic comparison of null distributions makes this distinction clear. It is important to remark, however, that there are some situations in which the permutation distribution is a reasonable choice. In randomized clinical trials, for example, one may believe in equality of the underlying distributions of two or more groups. In addition, we show that
when $n_1 = n_2$ the permutation distribution can be a good estimate of $Q_0$.

### 4.3 Algebraic Comparison of Permutation and Bootstrap Null Distributions

For simplicity, we suppose that $p = 2$, but note that conclusions about the covariance of two genes can be applied to any pairwise covariance when $p$ is much larger. For gene $j$, denote the variance of $X_i$ by $\sigma_{1j}^2$ in population 1 and by $\sigma_{2j}^2$ in population 2. Let $\phi_1$ be the covariance between the two genes in population 1 and $\phi_2$ be the covariance between the two genes in population 2.

Bootstrap resampling is equivalent to generating $n_1$ new subjects from an estimate of $P_1$ and $n_2$ new subjects from an estimate of $P_2$. Let $(X_i^b, L_i^b)$ denote the bootstrap data. Note that $P(X_i^b | L_i^b = 1)$ is $\hat{P}_1$ and similarly $P(X_i^b | L_i^b = 2)$ is $\hat{P}_2$. Permutation resampling corresponds with randomly reassigning the labels $\{L_i : i = 1, \ldots, n\}$ to the subjects. Let $(X_i^*, L_i^*)$ denote the permuted data. Now $X_i^* \perp L_i^*$ so that the permutation gene expression distribution $g(X_i^* | L_i^*)$ does not depend on $L_i^*$. We can approximate the permutation distribution $g(X_i^*) = g(X_i^* | L_i^*)$ by a mixture of the distributions $\hat{P}_1, \hat{P}_2$ with mixing proportions $p_1 = \frac{n_1}{n}$ and $p_2 = \frac{n_2}{n} = 1 - p_1$, where $n_1, n_2$ are fixed.

Suppose we use test statistics defined by Equation 2. Now, we can write $T_j$ as $\sum_{i=1}^{n} \frac{I(L_i = 2)X_i}{n_2} - \frac{I(L_i = 1)X_i}{n_1}$, where $I(L_i = k)$ is the indicator function that equals one when $L_i = k$. The test statistics are $T_j^b = \sum_{i=1}^{n} \frac{I(L_i = 2)X_i^b}{n_2} - \frac{I(L_i = 1)X_i^b}{n_1}$ and $T_j^* = \sum_{i=1}^{n} \frac{I(L_i^* = 2)X_i^*}{n_2} - \frac{I(L_i^* = 1)X_i^*}{n_1}$, under the bootstrap and permutation null distributions, respectively. In the appendix, we derive ex-
pressions for the variances of $T^b_j$ and $T^*_j$ ($j = 1, 2$) and the covariances of $(T^b_1, T^b_2)$ and $(T^*_1, T^*_2)$. Here we present the results:

$$
\begin{align*}
\text{Var}(T^b_j) &= \frac{\sigma^2_{1,j}}{n_1} + \frac{\sigma^2_{2,j}}{n_2} \\
\text{Var}(T^*_j) &= \frac{\sigma^2_{1,j}}{n_1} + \frac{\sigma^2_{2,j}}{n_2} \\
\text{Cov}(T^b_1, T^b_2) &= \frac{\phi_1}{n_1} + \frac{\phi_2}{n_2} \\
\text{Cov}(T^*_1, T^*_2) &= \frac{\phi_1}{n_2} + \frac{\phi_2}{n_1}
\end{align*}
$$

It is interesting to note that the roles of $n_1$ and $n_2$ are reversed under permutations. These expressions show us that under most values of the underlying parameters, the bootstrap and permutation distributions of $T_j$ are not equivalent. But, when (i) $n_1 = n_2$ or (ii) $\sigma^2_{1,j} = \sigma^2_{2,j} \equiv \sigma^2_j$ ($j = 1, 2$) and $\phi_1 = \phi_2 \equiv \phi$, then they are the same. Thus, unless one of these conditions holds we recommend using a bootstrap distribution since it preserves the correlation structure of the original data. When a study is “balanced” ($n_1 = n_2$), however, these results suggest that one should use the equivalent permutation distribution, because $Z^* \perp L^*$ implies that the variances and covariances are the same for both populations and estimates of these “pooled” values (which make use of all $n$ subjects) are more efficient.

Suppose we were to use the usual standardized t-statistics. By dividing the difference in means by an estimate of its variance, we expect that $\text{Var}(T^b_j) = \text{Var}(T^*_j) = 1$. So, we solve the problem of different variance terms under permutations. The covariances $\text{Cov}(T^b_1, T^b_2)$ and $\text{Cov}(T^*_1, T^*_2)$, however, are still not equivalent unless $n_1 = n_2$ or the correlation structures
are the same in the two populations.

The results of this section can be immediately generalized to other two sample multiple testing problems.

### 4.4 Bias of Permutation Null Distribution

We have found that the permutation null distribution of standardized t-statistics does not have mean zero whenever $n_1 \neq n_2$. The bias is largest when the sample sizes are very different or at least one of the sample sizes is very small. As an illustration, consider the following simple example. Let $n_1 = 2, n_2 = 50$ and suppose that the observations for gene $j$ in population 1 are $(1, 3)$ while the observations in population 2 are a vector of zeros. It is easy to enumerate all of the possible permutations for this data set and compute the expected value of any test statistic under this null distribution exactly. The results for the difference in means and the t-statistic are (rounded to two decimal places for presentation):

$$
E(\mu_1 - \mu_2) = \left(\binom{2}{2} + \binom{50}{1}\right) \times 0.44 + \left(\binom{50}{1}\right) \times 1.48 - \left(\binom{50}{2}\right) \times 0.08 = 0
$$

$$
E\left(\frac{\mu_1 - \mu_2}{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right) = \left(\binom{2}{2} + \binom{50}{1}\right) \times 0.87 + \left(\binom{50}{1}\right) \times 0.99 - \left(\binom{50}{2}\right) \times 1.27 = -1.104
$$

We have confirmed these formulas by simulation.

The consequences of this finding are serious. Whenever there are genes whose means are not equal in a data set with $n_1 \neq n_2$, then the permutation null distribution of the t-statistic will not have mean zero and should not be used to assess whether two population means are equal. Furthermore, the
genes need not be differentially expressed in truth, since the bias follows from
the means being unequal in the observed data. In the simulations below, we
see that the size of this bias depends on the magnitude of the difference in
means and that there is also a bias in the estimation of the variance of both
the difference in means and the t-statistic in unbalanced designs when the two
groups have unequal means. A simple improvement upon the permutation
approach would be to use the mixture distribution described in the previous
Section 4.3, which is a good approximation of the permutation distribution
and will not have this bias.

4.5 Simulations

We have conducted simulations to understand the performance of different
multiple testing procedures. We focus on estimation of the null distribution,
e.g.: mean, variance and covariance of the test statistic under different choices
of \( Q_{0,MLE} \). We also report estimates of the error control rates in Section 4.5.4.
We report the main findings here and refer the reader to Pollard and van der
Laan (2002) for more extensive results.

4.5.1 Data and Null Distributions

First, we simulate \( n_1 \) observations from a \( p \)-variate normal distribution with
equal means \( \mu_1 = 0 \), equal variances \( \sigma_1^2 = 0.1 \), and all pairwise correlations
\( \rho_1 = 0 \). Second, we simulate \( n_2 \) observations from a \( p \)-variate normal dis-
tribution with equal means \( \mu_2 = 0 \), equal variances \( \sigma_2^2 = 5 \) and all pairwise
correlations \( \rho_2 = 0.9 \). We focus on results for \( p = 100 \) genes. The values
of all parameters are chosen in light of the results from Section 4.3 as an
extreme case of unbalanced groups in terms of sample size, variance, and correlation. For each simulated data set, we compute two test statistics: the difference in means and the standardized t-statistic. In this section, we let $D_j$ denote the difference in means and $T_j$ the t-statistic for gene $j$ so that we can distinguish the two test statistics. The null distributions of these statistics are estimated by (i) permutations, (ii) the non-parametric bootstrap (centered empirical distribution), and (iii) the parametric bootstrap (multivariate normal distribution with equal means).

In each case, $B = 1000$ independent resampled data sets are used. This number was chosen because we ran such a large number of simulations. We acknowledge that our results could have been more accurate with larger $B$ and suggest that in practice $B$ should be chosen as large as possible since one usually estimates the resampled distribution only once.

### 4.5.2 Comparison of Test Statistics

We compare $D_j$ and $T_j$ based on the ease with which their null distributions can be estimated for reasonable sample sizes. Since all genes have the same marginal distribution in this simulation, we report the results for one gene and note that they are representative for all genes.

Both test statistics are unbiased with observed means close to zero, though the means for $D_j$ tend to be smaller in absolute value and slightly less variable than those for $T_j$. Table 1 shows results for the variance of the same null distributions. We see that it is very difficult to estimate the variance of $T_j$’s null distribution with the non-parametric bootstrap. This is another argument in favor of using $D_j$. Note that the permutation distribution of
Table 1: Variance of the permutation, non-parametric bootstrap, and parametric bootstrap null distributions of the difference in means $D_j$ and the t-statistic $T_j$ for one gene. The true values are from formulas (approximate for the t-statistics, Moore and McCabe (2002)) and have been confirmed by simulation.

$VAR(D_j)$ is far from the truth, as predicted by the formulas in Section 4.3. Both bootstrap null distributions estimate $COV(D_j, D_{j'})(j \neq j')$ accurately, whereas the permutation distribution does not. The parametric bootstrap is the only distribution that accurately estimates $COV(T_j, T_{j'})(j \neq j')$; the nonparametric bootstrap estimate is larger than the true value, whereas the permutation estimate is much smaller than the true value.
4.5.3 Comparison of Null Distributions

The results in the previous section allow us to compare the three choices of null distribution in terms of how easy they are to estimate and how close they are to the true null distribution.

- **PERMUTATIONS:** As expected, the permutation distribution estimates of $VAR(D_j)$, $COV(D_j, D_{j'})$ and $COV(T_j, T_{j'})$ are far from the true values and close to those given by the formulas in Section 4.3. The permutation distribution estimates for $VAR(T_j)$, however, are much closer to the true values, as we predict. When $n_1 = n_2$, permutations perform well.

- **NON-PARAMETRIC:** The centered empirical distribution is fairly close to the true $Q_0$ for $D_j$, but it’s estimate of the variance of $T_j$ is quite variable and occasionally very far from the truth. This problem is a consequence of there being many ties in the resampling when one or both of the populations has a small sample size. These ties lead to very small variance estimates in denominators of the t-statistics, which produce unrealistically large bootstrap $T^b_j$. Smoothing the empirical distribution reduces this problem.

- **PARAMETRIC:** The parameters of the multivariate normal bootstrap null distribution are the closest to the true values and are quite accurate even in very unbalanced designs. Thus, we see that you gain immensely if you guess the model correctly and use a parametric bootstrap null distribution. In Section 5.2, however, we see that the multivariate normal is not necessarily a good choice of model with real data.
Table 2: Estimates $\hat{\alpha}$ of the error rate $P(V > 10)$ over $I = 200$ independent data sets for the permutation, non-parametric bootstrap, and parametric bootstrap null distributions of $D_j$ and $T_j$. We can expect the error in the estimates to be on the order of 0.05. The target error rate is $\alpha = 0.05$.

### 4.5.4 Error Control

We report results from using Equation 5 to control $P(V > 10) \leq \alpha = 0.05$, where $V$ is the number of false positives. Results for other error rates follow similar patterns. Table 2 shows the estimates of $\alpha$ over $I = 200$ independent data sets. A few interesting points emerge. First, conservative error control is associated with overestimating $\text{VAR}(T_j)$ (causing the upper quantiles $c_j$ to be too large) and conversely, failure to control the error rate is due to underestimation. Second, the direction of the bias in $\text{VAR}(T_j)$ has consequences
in terms of the size of the bias of $\hat{\alpha}$. In particular, the skewedness of type I error means that bias due to an underestimate of the variance is much larger in magnitude than the bias due to a similarly sized overestimate of the variance. Finally, as expected, the permutation null distribution does the worst job of controlling both error rates. Both the non-parametric and the parametric bootstrap methods perform fairly well, though they tend to be conservative for $T_j$ and anti-conservative for $D_j$. When the simulation is repeated with more similar covariance structures in the two populations, both methods control the error rate perfectly.

### 4.5.5 Differentially Expressed Genes

We repeat the simulation giving ten of the genes in population 2 non-zero means (0.5, 1.0, 1.5, ..., 5.0). Most of the results are similar to the first simulation, but we note a few exceptions. First, we confirm the result of Section 4.4 that genes with non-zero differences in means do not have mean zero in the permutation null distribution of $T_j$ when $n_1 \neq n_2$ (See Figure 2). We see that this bias increases with the value of the mean in population 2. Second, we also find a similar bias in the estimated variance of both $T_j$ and $D_j$ for genes with unequal means under permutations. Third, estimated error rates tend to be slightly larger when there are some false null hypotheses. Finally, the methods with the largest error rates have the most power (as high as 0.9). In practice, one might want to use a cost function that accounts for both type I and type II errors in order to optimize both the error rate and power.
Mean of the Permutation Distribution for 100 Genes
(genes 91 – 100 are differently expressed)

Figure 2: Mean of the permutation null distributions of the difference in means and the t-statistic for simulated data. Genes 91 to 100 have increasing non-zero means between 0.05 and 5 in population 2. The average value of the mean over $I = 200$ independent data sets is plotted for each gene. The mean of the null distribution should be zero for all genes.

5 Data Analysis

We apply resampling-based multiple testing methods to a publicly available data set (Alizadeh et al. (2000)). Expression levels of 13,412 clones (relative to a pooled control) were measured in the blood samples of 40 diffuse large B-cell lymphoma (DLBCL) patients using cDNA arrays. According to Alizadeh
et al. (2000), the patients belong to two molecularly distinct disease groups, 21 Activated and 19 Germinal Center (GC). We log the data (base 2), replace missing values with the mean for that gene, and truncate any expression ratio greater than 20-fold to \( \log_2(20) \).

### 5.1 Multiple Testing

Our goal is to identify clones with significantly different mean expression levels between the Activated and GC groups. We compute standardized t-statistics for each gene. We choose to control the usual family-wise error and compare the clones identified as having significantly different means between the two groups using several methods (See Table 3). Interestingly, Equation 5 and single-step Bonferroni common quantiles produce the same subset of clones (for both the bootstrap and the permutation null distributions), though this need not be the case since the single-step Bonferroni quantiles are always smaller. We observe that the variances of the t-statistics across the \( B = 1000 \) samples tend to be smaller in the permutation distribution compared to the bootstrap distribution, resulting in the larger number of rejected null hypotheses with permutations. Based on the results of Section 4, we believe that the permutation subset is likely to be larger and the bootstrap subset to be slightly smaller than the true subset. We note that all of the rejections have unadjusted p-values equal to zero. If the tails of the null distributions had more observations (\( B \) larger), some of these likely would be non-zero and may even be non-significant after multiple testing adjustment.
<table>
<thead>
<tr>
<th>Method</th>
<th>Null Distribution</th>
<th>Rejections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equation 5 common quantiles</td>
<td>non-parametric bootstrap</td>
<td>186</td>
</tr>
<tr>
<td>Bonferroni common quantiles</td>
<td>non-parametric bootstrap</td>
<td>186</td>
</tr>
<tr>
<td>Equation 5 common quantiles</td>
<td>permutations</td>
<td>287</td>
</tr>
<tr>
<td>Bonferroni common quantiles</td>
<td>permutations</td>
<td>287</td>
</tr>
<tr>
<td>Bonferroni common threshold</td>
<td>t-distribution</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 3: Number of rejected null hypotheses (out of \( p = 13,412 \)) for five different choices of thresholds and null distribution. All 32 of the genes in the t-distribution subset are in both the permutation and the bootstrap subset, and the bootstrap and permutation subsets have 156 genes in common. Data are from Alizadeh et al. (2000).

5.2 Simulations

We conduct additional simulations using 100 randomly selected genes from the data set of Alizadeh et al. (2000) centered to all have mean zero in the Activated and GC groups as the true data generating distribution. The idea is to make use of a real data set in order to (i) avoid assumptions about the parametric form of the underlying distribution and (ii) have a more realistic covariance structure between the genes. We treat the 21 Activated and 19 GC patients as the population and randomly sample \( n_1 \) Activated and \( n_2 \) GC patients from it to create an “observed” data set \( I = 200 \) times. We estimate the null distributions of the t-statistic and the difference in means, resampling \( B = 1000 \) times. In each case, we use Equation 5 to control the FWE \( P(V > 10) \leq \alpha = 0.05 \). Overall, the permutation distribution does
the worst job and the non-parametric bootstrap the best job of controlling the error rate. Notice that the parametric bootstrap is no longer the best method, since the model is not normal.

<table>
<thead>
<tr>
<th></th>
<th>Permutation</th>
<th>Non-parametric Bootstrap</th>
<th>Parametric Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1 = 5, n_2 = 15$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_j$</td>
<td>0.21</td>
<td>0.025</td>
<td>0.085</td>
</tr>
<tr>
<td>$T_j$</td>
<td>0.020</td>
<td>0.025</td>
<td>0.020</td>
</tr>
<tr>
<td>$n_1 = 9, n_2 = 11$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_j$</td>
<td>0.13</td>
<td>0.050</td>
<td>0.065</td>
</tr>
<tr>
<td>$T_j$</td>
<td>0.015</td>
<td>0.065</td>
<td>0.015</td>
</tr>
<tr>
<td>$n_1 = 10, n_2 = 10$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_j$</td>
<td>0.17</td>
<td>0.060</td>
<td>0.070</td>
</tr>
<tr>
<td>$T_j$</td>
<td>0.020</td>
<td>0.055</td>
<td>0.035</td>
</tr>
</tbody>
</table>

Table 4: Estimates $\hat{\alpha}$ of the error rate $Pr(V > 10)$ over $I = 200$ independent simulated data sets for permutation, non-parametric bootstrap and parametric bootstrap null distributions of $D_j$ and $T_j$. In each case, Equation 5 was used to adjust for multiple tests. The target error rate is $\alpha = 0.05$

6 Acknowledgment

This research has been supported by a grant from the Life Sciences Informatics Program with industrial partner biotech company Chiron Corporation.
References


APPENDIX: Details of proofs in Section 4.3

The derivations of expressions for (i) the variances of $T_j^b$ and $T_j^*$ ($j = 1, 2$) and (ii) the covariances of $(T_1^b, T_2^b)$ and $(T_1^*, T_2^*)$ are similar, and both make use of the double expectation theorem. For simplicity, assume that the null hypotheses hold for both genes, so that the means for the two populations are zero vectors $\mu_1 = \mu_2 = (0, 0)$. Consider gene $j$. The variance of the difference in means test statistic under bootstrap resampling is:

\[
\text{Var}(T_j^b) = E((T_j^b)^2) - E(T_j^b)^2
\]
\[
= E((T_j^b)^2)
\]
\[
= E\left( \sum_{i=1}^{n} \frac{I(L_j^b = 2)(Z_j^b)^2}{n_1^2} + \frac{I(L_j^b = 1)(Z_j^b)^2}{n_2^2} \right)
\]
\[
= n \left( E \left( \frac{I(L_j^b = 2)(Z_j^b)^2}{n_2^2} + \frac{I(L_j^b = 1)(Z_j^b)^2}{n_2^2} \right) \right)
\]
\[
= n \left( \frac{\sigma_j^2}{n_1^2} \frac{n_1}{n} + \frac{\sigma_j^2}{n_2^2} \frac{n_2}{n} \right)
\]
\[
= \frac{\sigma_j^2}{n_1} + \frac{\sigma_j^2}{n_2}
\]

Similarly, the variance of the test statistic under permutations is:

\[
\text{Var}(T_j^*) = E((T_j^*)^2) - E(T_j^*)^2
\]
\[
= E((T_j^*)^2)
\]
\[
= E\left( \sum_{i=1}^{n} \frac{I(L_j^* = 2)(Z_j^*)^2}{n_1^2} + \frac{I(L_j^* = 1)(Z_j^*)^2}{n_1^2} \right)
\]
\[
= n \left( E \left( \frac{I(L_j^* = 2)(Z_j^*)^2}{n_2^2} + \frac{I(L_j^* = 1)(Z_j^*)^2}{n_2^2} \right) \right)
\]
\[
= n \left( \frac{1}{n} (\sigma_j^2/n_1 + \sigma_j^2/n_2) \right) \frac{n_1}{n} + \frac{1}{n} (\sigma_j^2/n_1 + \sigma_j^2/n_2) \frac{n_2}{n}
\]
\[
= \frac{\sigma_j^2}{n_2} + \frac{\sigma_j^2}{n_1}
\]
Note that in the permutation derivation, the variance of of \( Z^* \) is \( 1/n(\sigma_1^2n_1 + \sigma_2^2n_2) \) for both values of \( L^* \), since \( Z^* \perp L^* \). It is interesting to note that the final expression for the variance under permutations resembles that under bootstrap resampling, except with the roles of \( n_1 \) and \( n_2 \) reversed.

Now, consider the covariance between the test statistics for the two genes.

Under bootstrap resampling we have:

\[
\text{Cov}(T_1^b, T_2^b) = E(T_1^b \cdot T_2^b) = \sum_{i=1}^{n} \left[ \frac{(I(L^b_i = 2)Z_1^b_i) - (I(L^b_i = 1)Z_1^b_i)}{n_1^2} \right] \cdot \left[ \frac{(I(L^b_i = 2)Z_2^b_i) - (I(L^b_i = 1)Z_2^b_i)}{n_2^2} \right] = \frac{\phi_1 \cdot n_1}{n_1^2} + \frac{\phi_2 \cdot n_2}{n_2^2}
\]

Under permutations we have:

\[
\text{Cov}(T_1^p, T_2^p) = E(T_1^p \cdot T_2^p) = \sum_{i=1}^{n} \left[ \frac{(I(L^* = 2)Z_1^* i) - (I(L^* = 1)Z_1^* i)}{n_1^2} \right] \cdot \left[ \frac{(I(L^* = 2)Z_2^* i) - (I(L^* = 1)Z_2^* i)}{n_2^2} \right] = \frac{\phi_1 \cdot n_1}{n_1^2} + \frac{\phi_2 \cdot n_2}{n_2^2}
\]
Note that in the permutation derivation, the covariance of $Z_1^*$ and $Z_2^*$ is $1/n(\phi_1 n_1 + \phi_2 n_2)$ for both values of $L^*$, since $Z^* \perp L^*$. Again, it is interesting to note that the final expression for the covariance under permutations resembles that under bootstrap resampling, except with the roles of $n_1$ and $n_2$ reversed.