JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 4, 2009

Pp. 625–638
DOI: 10.1089/cmb.2008.0144

Properties of Balanced Permutations

LUCINDA K. SOUTHWORTH,[1] STUART K. KIM,[2] and ART B. OWEN[3]

ABSTRACT

This paper takes a close look at balanced permutations, a recently developed sample reuse method with applications in bioinformatics. It turns out that balanced permutation reference distributions do not have the correct null behavior, which can be traced to their lack of a group structure. We find that they can give *p*-values that are too permissive to varying degrees. In particular the observed test statistic can be larger than that of all *B* balanced permutations of a data set with a probability much higher than $1/(B+1)$, even under the null hypothesis.

Key words: functional genomics, gene expression, genomics, statistics.

1. INTRODUCTION

SAMPLE REUSE METHODS SUCH AS PERMUTATION TESTING and the bootstrap are invaluable tools in high-throughput genomic settings, such as microarray analyses. They are used to test hypotheses and compute *p*-values without making strong parametric assumptions about the data, and they adapt readily to complicated test statistics.

Permutation tests, described in more detail below, work by permuting the treatment labels of the data and comparing the resulting values of a test statistic to the original one. Suppose for example, that there are n observations $X_1, \ldots, X_n$ in the treatment group and also n observations $Y_1, \ldots, Y_n$ in the control group. We might measure the treatment effect via $\hat{\Delta} = \bar{X} - \bar{Y}$ where $\bar{X}$ and $\bar{Y}$ are averages of the treatment and control observations, respectively. There are $P = \binom{2n}{n} = (2n)!/(n!)^2$ ways to redo the assignment of treatment versus control labels, and they each give a value for the treatment effect. If the actual treatment effect is larger than that from all of the other permutations, then we may claim a p-value of $1/P$. More generally, if the observed effect beats (is larger than) exactly b of these values we can claim $p = 1 - b/P$. The smallest available p value is $1/P$ because the actual treatment allocation is always included in the reference set.

One reason why permutation tests work and are intuitively reasonable is their symmetry. If $\hat{\Delta}$ is the actual difference and $\hat{\Delta}^*$ is the difference for any reassignment of labels, chosen without looking at the X and Y values, then

$$\Pr(\hat{\Delta} \geq D) = \Pr(\hat{\Delta}^* \geq D) \tag{1}$$

holds for all D, under the null hypothesis $\mathbb{H}_0$ that X_i and Y_i are all independent and identically distributed. That is, $\hat{\Delta}$ and $\hat{\Delta}^*$ have the same distribution.

Recently, a special form of permutation analysis, called balanced permutation, has been employed. In a balanced permutation, we make sure that after relabeling, the new treatment group has exactly $n/2$ members that came from the original treatment group and $n/2$ from the original control group. The number of balanced

[1]Biomedical Informatics, [2]Developmental Biology, and [3]Statistics, Stanford University, Stanford, California.

permutations is $B = \binom{n}{n/2}^2$ which can be much smaller than the total number P of permutations. Balanced permutations require n to be even, but nearly balanced permutations are possible for odd n.

These balanced permutations are mentioned in remark E on page 1159 of Efron et al. (2001). This is the earliest mention we have found. Since then, they have been applied numerous times in the literature on statistical analysis of microarrays. As of 2008, the National Cancer Institute describes balanced permutations in their page on statistical tests at *http://discover.nci.nih.gov/microarrayAnalysis/Statistical.Tests.jsp.* They say that more extreme p-values typically result, but they remark on a granularity problem. For fixed n, $1/B$ is bigger than $1/P$, and so balanced permutations require larger sample sizes n than ordinary ones do, if one is to attain very small p-values.

Balanced permutations are a subset of all possible permutations, and so equation (1) holds for them too. Intuitively, a balanced permutation analysis should be even better than a permutation analysis. If $\mathbb{E}(X_i) = \mu_x > \mu_y = \mathbb{E}(Y_i)$ then $\mathbb{E}(\hat{\boldsymbol{\Delta}}) > 0$ while $\mathbb{E}(\hat{\boldsymbol{\Delta}}^*)$ remains at 0 due to cancellation. By contrast, for ordinary permutations some of the $\mathbb{E}(\hat{\boldsymbol{\Delta}}^*)$ are positive and some are negative.

Equation (1) is not enough to make balanced permutations give properly calibrated p-values. The theory of permutation tests also requires a group structure for the set of permutations used, and balanced permutations do not satisfy this condition.

Though they fail a sufficient condition for giving exact p values, that does not in itself mean that they give bad p values. The goal of this paper is to investigate the p-values produced via balanced permutations. What we find is that those p values can be much smaller than they should be. This is most extreme in the case where the original test statistic beats all of the balanced permutations.

The outline of this paper is as follows. Section 2 describes permutation based inferences including balanced permutations and it distinguishes balanced from stratified permutations. Section 3 shows through theory and simulations that the reference distribution provided by balanced permutations can lead to permissive p-values, sometimes by astonishingly large factors.

The major use of balanced permutations is in estimating a pooled p-value distribution for N test statistics assumed to have the same distribution, and for estimating the false discovery rate (FDR) of a procedure. Section 4 looks at balanced permutations for pooled reference distributions. It is well known that including non-null cases makes a permutation analysis conservative. We find that including non-null cases in the balanced permutation scheme has the opposite effect, making it overly permissive. Even purely null cases can give rise to a permissive test, though this problem diminishes in large scale problems, when the null cases are independent of the one being tested. Section 5 summarizes our conclusions on balanced permutations, recommending against their use.

2. BACKGROUND

To be concrete, suppose that we are comparing gene expression for two groups of subjects, treatment and control. Gene expression is a proxy measurement for the amount of product, either protein or RNA, made by a gene. In the past decade, various high-throughput technologies, generally termed microarrays, have made it possible to measure the expression activity of all genes in a genome with a single assay. We will later look at the issues that arise when there are many gene expression measurements to consider. For simplicity, we begin with the expression of a single gene.

We have $n \geq 2$ subjects in each group. For the treatment subjects we measure $X_i \in \mathbb{R}$ for $i = 1, \ldots, n$ and for the control group we measure $Y_i \in \mathbb{R}$ for $i = 1, \ldots, n$. We assume an unpaired framework, so there is no connection between X_i and Y_i. The X_i are modeled as n independent samples from a distribution F and the Y_i are similarly sampled from G. We will always assume that F and G both have finite means. This is reasonable for the applications we have in mind and it rules out uninteresting corner cases.

We are usually interested in $H_0 : \mathbb{E}(X) = \mathbb{E}(Y)$. We also consider the stricter hypothesis $\mathbb{H}_0 : F = G$. There is no exact nonparametric test for a one-sample mean (Bahadur and Savage 1956) and so we cannot expect one for H_0 either. One often makes do with exact tests for $\mathbb{H}_0$.

The customary test for H_0 is Student's t test. Let $\bar{X} = (1/n)\sum_{i=1}^n X_i$ and $\bar{Y} = (1/n)\sum_{i=1}^n Y_i$. Then under H_0, the statistic

$$t = \frac{\bar{X} - \bar{Y}}{s/\sqrt{2/n}} \tag{2}$$

where

$$s^2 = \frac{1}{2n-2}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{n}(Y_i-\bar{Y})^2\right),$$

has the $t_{(2n-2)}$ distribution. The t-test rejects H_0 if $|t| > t_{(2n-2)}^{1-\alpha/2}$. There are also one tailed versions.

The t test is designed for the case where X_i and Y_i are normally distributed with common variance and possibly different means. The test holds up fairly well for non-normal data. In this particular setting with equal sample sizes of n, it is also reasonably accurate when X_i and Y_i have unequal variances. For a discussion of the properties of t tests when assumptions may be violated, see Miller (1986).

2.1. Permutation tests

Permutation tests are very simple and intuitive. The original treatment assignment is one of $P = \binom{2n}{n}$ possible assignments. Under the null hypothesis, they all have an equal chance to be the most extreme. The explanation is not based on the P test statistics being independent and identically distributed. In fact, they are not independent. The explanation comes from a symmetry argument which in turn stems from the group structure of the set of permutations.

Here we make a formal presentation of permutation tests of $\mathbb{H}_0$, so that we may later highlight their differences with balanced permutations. Balanced permutations also seem intuitively reasonable and even seem intuitively better than ordinary permutations. Readers who primarily want to apply permutation methods may prefer to skip this subsection.

We suppose as before that $X_i \sim F$ and $Y_i \sim G$ are all independent, for $i = 1, \ldots, n$. The sample mean difference is $\hat{\Delta} = \bar{X} - \bar{Y} = \sum_{i=1}^{2n} Z_i\omega_i$, where

$$Z_i = \begin{cases} X_i & 1 \le \mathrm{i} \le \mathrm{n} \\ Y_{i-n} & n+1 \le \mathrm{i} \le 2\mathrm{n}, \end{cases} \quad \text{and} \quad \omega_i = \begin{cases} 1/n & 1 \le i \le n \\ -1/n & n+1 \le i \le 2n. \end{cases}$$

We use $\mathcal{Z} = (Z_1, \ldots, Z_{2n})$ to represent both samples combined.

Suppose that we relabel the data, changing which subjects are treatment and which are control. Any relabeling of the data can be obtained via a permutation of indices as follows. Let $\pi(1), \ldots, \pi(2n)$ be a permutation of the integers $1, \ldots, 2n$. We shuffle the components of $\mathcal{Z}$ obtaining the vector $\pi(\mathcal{Z}) = (Z_{\pi(1)}, Z_{\pi(2)}, \ldots, Z_{\pi(2n)})$. We use the symbol π for two meanings, a permutation of indices and the resulting rearrangement of a vector, but this should cause no confusion.

Now let

$$\hat{\Delta}(\pi(\mathcal{Z})) = \sum_{i=1}^{2n} Z_{\pi(i)}\omega_i. \tag{3}$$

Under $\mathbb{H}_0$ the distribution of $\hat{\Delta}(\pi(\mathcal{Z}))$ is the same as that of $\hat{\Delta}(\mathcal{Z})$, for all permutations π.

The set of all permutations of $\{1, \ldots, 2n\}$ comprises a group, called $\mathcal{S}_{2n}$. The group operation is composition. For two permutations $\pi, \widetilde{\pi} \in \mathcal{S}_{2n}$ their composition, denoted by $\pi \circ \widetilde{\pi}$, is the permutation one gets by applying $\widetilde{\pi}$ first and then applying π to the result.

Although there are $M = (2n)!$ permutations of $\{1, \ldots, 2n\}$ there are only $P = \binom{2n}{n} = (2n)!/(n!)^2$ distinct cases for $\hat{\Delta}(\pi(\mathcal{Z}))$ determined by which n observations get $\pi(i) \le n$ and which get $\pi(i) > n$. The ratio $M/P = (n!)^2$ reflects the fact that there are $n!$ ways to permute the observations with $\pi(i) \le n$, (and similarly the ones with $\pi(i) > n$) without affecting $\hat{\Delta}(\pi(\mathcal{Z}))$).

A permutation test uses the uniform distribution on all M possible sample values of $\hat{\Delta}(\pi(\mathcal{Z}))$ as a reference distribution for $\hat{\Delta}$, to test $\mathbb{H}_0$. In practice we accomplish the permutation test by the shortcut of computing only the P distinct cases. Those cases all have multiplicity M/P and so we can still use a uniform distribution on them. When even P is too large for enumeration to be feasible, then one commonly substitutes a very large random sample of permutations for the complete enumeration.

Here we construct permutation tests of $\mathbb{H}_0$ at level $\alpha \in (0, 1)$ based on Chapter 15.2 of Lehmann and Romano (2005). We will work with all M relabelings, instead of reducing to the P distinct ones, in order to use the simple group structure of $\mathcal{S}_{2n}$.

To test $\mathbb{H}_0$, let T be a real valued function such that $T(\hat{\Delta})$ takes larger values in response to stronger evidence against $\mathbb{H}_0$. For example $T(\Delta) = |\Delta|$ for a usual two tailed test and $T(\Delta) = \Delta$ or $-\Delta$ for one tailed tests. To get a test with level α we reject $\mathbb{H}_0$ if the observed $T(\hat{\Delta})$ is larger than $M\alpha$ of the permuted values. We sort the M values of the permuted test statistics $T(\hat{\Delta}(\pi(\mathcal{Z})))$, as π ranges through $\mathcal{S}_{2n}$, getting

$$T^{(1)}(\mathcal{Z}) \le T^{(2)}(\mathcal{Z}) \le \cdots \le T^{(M)}(\mathcal{Z}).$$

Let $k = M - \lfloor M\alpha \rfloor$. Roughly speaking, the test will reject $\mathbb{H}_0$ if $T(\hat{\Delta}(\mathcal{Z}))$ is larger than $T^{(k)}(\mathcal{Z})$ but the recipe below takes care of ties, the number of which can in principle depend on the data $\mathcal{Z}$. We let $M_+(\mathcal{Z})$ be the number of $T^{(r)}(\mathcal{Z})$ values strictly larger than $T^{(k)}(\mathcal{Z})$ and $M_0(\mathcal{Z})$ be the number of $T^{(r)}(\mathcal{Z})$ values exactly equal to $T^{(k)}$. Then define the tie breaker quantity $a(\mathcal{Z}) = (M\alpha - M_+(\mathcal{Z}))/M_0(\mathcal{Z})$ and let

$$\psi(\mathcal{Z}) = \begin{cases} 1 & T(\hat{\Delta}(\mathcal{Z})) > T^{(k)}(\mathcal{Z}), \\ a(\mathcal{Z}) & T(\hat{\Delta}(\mathcal{Z})) = T^{(k)}(\mathcal{Z}), \\ 0 & T(\hat{\Delta}(\mathcal{Z})) < T^{(k)}(\mathcal{Z}). \end{cases}$$

By construction

$$\sum_{\pi \in \mathcal{S}_{2n}} \psi(\pi(\mathcal{Z})) = M\alpha \tag{4}$$

holds for any $\mathcal{Z} \in \mathbb{R}^{2n}$.

The interpretation of ψ is as follows. If $\psi = 1$ then we reject $\mathbb{H}_0$. If $\psi = 0$ then we don't reject it. If $0 < \psi < 1$ then we reject $\mathbb{H}_0$ with probability ψ. It is best to have all the ψ values equal to 0 or 1 to avoid randomized decisions. We can usually do this for special choices of α. For continuously distributed data all P possible sample differences are distinct with probability one. Then taking $\alpha = A/M$ for a positive integer $A < M$ yields $a(\mathcal{Z}) = 1$ and so eliminates randomized decisions.

Theorem 1. *Suppose that $X_i \sim F$ and $Y_i \sim G$ for $i = 1, \ldots, n$ are all independent and that $\mathbb{H}_0: F = G$ holds. For a test statistic T (z) and the randomization test quantity ψ constructed above from the permutation group $\mathcal{S}_{2n}$, $\mathbb{E}(\psi(\mathcal{Z})) = \alpha$.*

Remark. Theorem 1 follows from Theorem 15.2.1 of Lehmann and Romano (2005) of which it is a special case. We recap their proof, for later use.

Proof. Using $\mathbb{H}_0$ and then equation (4), we get

$$\mathbb{E}(\psi(\mathcal{Z})) = \frac{1}{M} \sum_{\pi \in \mathcal{S}_{2n}} \mathbb{E}(\psi(\pi(\mathcal{Z}))) = \frac{1}{M} \mathbb{E}\left(\sum_{\pi \in \mathcal{S}_{2n}} \psi(\pi(\mathcal{Z})) \right) = \frac{M\alpha}{M} = \alpha. \quad \blacksquare$$

When we have arranged for ψ to be either 0 or 1, then $\mathbb{E}(\psi(\mathcal{Z})) = \Pr(\psi(\mathcal{Z}) = 1)$. Therefore, the test rejects $\mathbb{H}_0$ with probability α. The p value is the smallest α rejected (among the non-randomized choices).

2.2. *Stratified permutations and balanced permutations*

Two different ideas have been called "balanced permutation" in the literature. In one kind of balanced permutation, the subjects can be split into two groups in two different ways. The split of primary interest is treatment versus control. The second split has levels that are variously called strata or blocks. It may already be known to be quite important, but it is not of primary interest. For example, test and control might correspond to young and old mice while the strata might correspond to two different strains. We might want a test for age that adjusts for known differences between strains.

Suppose that we have n test subjects, where n is even for simplicity, with $n/2$ from stratum A and $n/2$ from stratum B. Similarly, there are $n/2$ control subjects from each stratum. This study design has balanced the treatment versus control dichotomy with respect to the strata, and it is reasonable to enforce such a balance on all the permutations. To get a permutation that is balanced with respect to the stratification, we may select $n/2$ of the subjects from stratum A and $n/2$ of the subjects from stratum B to be relabeled as test subjects. The remaining n subjects are relabeled as controls, and of course are also balanced. Stratified permutation tests of this type can be generalized to stratification factors with more than 2 levels and also to settings with more than 1 factor.

Tusher et al. (2001) employed this form of stratification to compare effects of radiation on genes, adjusting for cell lines. Their subjects had a third dichotomy, corresponding to aliquots, that need not concern us here.

The other method called balanced permutations is the one described in the introduction. These balanced permutations are not the same as stratified permutations. The difference is that the former stratifies on a covariate dichotomy, while the latter stratifies on the dichotomy of interest. Stratification on a covariate is sometimes studied as a restricted randomization and has a long history in nonparametric methods; for example, see Good (2000) and Edgington and Onghena (2007).

We will use the term "balanced permutation" to describe only the new method of stratifying on the variable of interest. As described in the introduction, there is reason to be optimistic about the performance of balanced permutations. On the other hand, an ordinary unbalanced permutation test applied to $\bar{X} - \bar{Y}$, is asymptotically equivalent to Student's t-test. That test is the appropriate one for Gaussian data, and is asymptotically well behaved otherwise, under mild moment conditions. We might therefore expect that there is little room for improvement from balanced permutations, at least for $\bar{X} - \bar{Y}$ and moderately large n.

To justify using balanced permutations to construct a reference distribution, we need to apply a result like Theorem 1. The key ingredient is the group structure. Stratification has a group structure. In the setting above we may employ permutations π_A and π_B within strata A and B respectively. The permutation of the whole data set may be represented by the pair (π_A, π_B). These pairs have a group structure. Formally, it is the direct sum of the groups to which π_A and π_B belong. Practically, the permutations for strata A and B are just handled separately and independently. The group structure can extend to three or more strata.

The set $\mathcal{B}_{2n}$ of balanced permutations is not a group. To begin with, the original data $\mathcal{Z}$ are the identity permutation of the data, but the identity permutation is not a balanced permutation. We might simply include the identity permutation, with the same multiplicity as each balanced permutation into $\mathcal{B}_{2n}$, getting the set $\mathcal{B}_{2n+}$. Then our putative p value is the proportion of permutations $\pi \in \mathcal{B}_{2n+}$ for which $T(\mathcal{Z}) \geq T(\pi(\mathcal{Z}))$ holds.

The problem with this approach is that the set $\mathcal{B}_{2n+}$ is not a group either, under composition. A balanced permutation of a balanced permutation is not necessarily another balanced permutation. Therefore, $\mathcal{B}_{2n}$ and $\mathcal{B}_{2n+}$ both fail the closure requirement of a group. An example is illustrated in Figure 1.

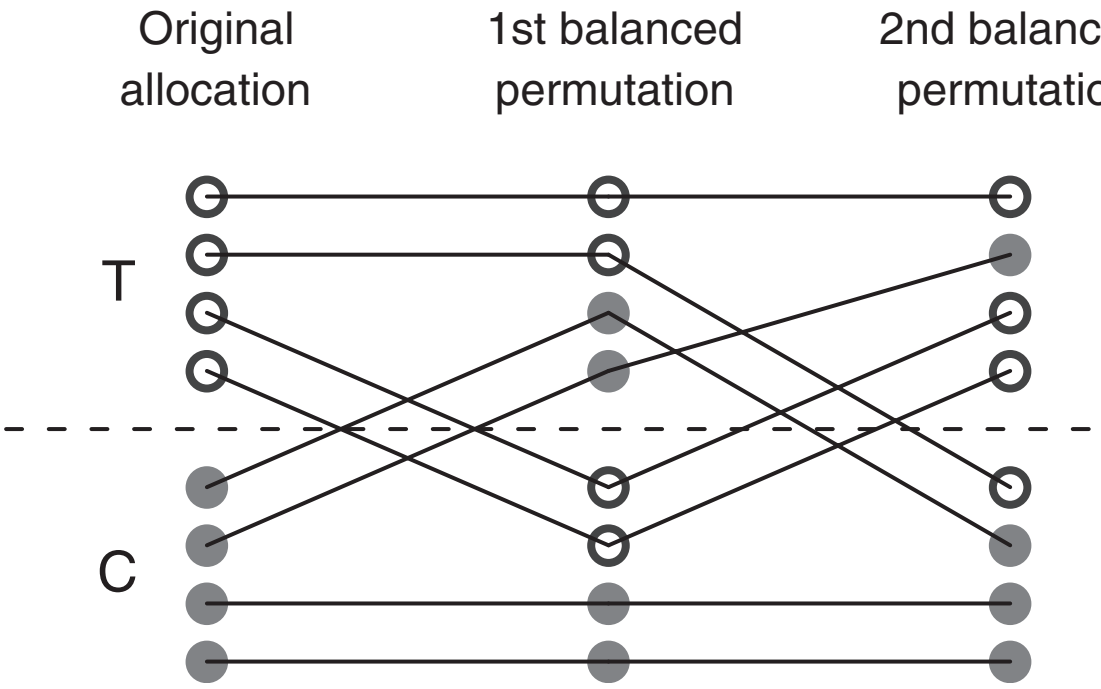

FIG. 1. This figure shows that the result of applying two balanced permutations in succession need not be another balanced permutation. The first column shows treatment and control groups, T and C, of four observations each. The middle column shows the results of a balanced permutation that swapped two of the treatment subjects for two of the controls as indicated by line segments. The third column shows the result of a second such swap. The final treatment group has three members of the original treatment group and one member of the original control group, and so it is unbalanced.

It is not obvious at first sight where the proof of Theorem 1 makes use of the group property of $\mathcal{S}_{2n}$. But if we replace the group $\mathcal{S}_{2n}$ by $\mathcal{B}_{2n}$ (or $\mathcal{B}_{2n+}$), and follow the steps, the proof fails at the third equality sign. We can still pick k so that $T^{(k)}(\mathcal{Z})$ is beaten by a fraction α of $T^{(r)}(\pi(\mathcal{Z}))$ taken over balanced permutations π. For a group of permutations, that means $T^{(k)}$ will be beaten by a fraction α of all $T^{(k)}(\widetilde{\pi}(\pi(\mathcal{Z})))$ as $\widetilde{\pi}$ ranges over the group and π is any fixed group member. This happens for the group $\mathcal{S}_{2n}$, because $\{\widetilde{\pi} \circ \pi \mid \widetilde{\pi} \in \mathcal{S}_{2n}\} = \mathcal{S}_{2n}$. But no such result holds for $\mathcal{B}_{2n}$ or $\mathcal{B}_{2n+}$. For balanced permutations, the set of ordered values of $T^{(r)}(\widetilde{\pi}(\pi(\mathcal{Z})))$ as $\widetilde{\pi}$ varies for fixed π is not constant. That set depends on π and so its $1-\alpha$ quantile will in general depend on π.

3. COVERAGE PROPERTIES OF BALANCED PERMUTATION

Because balanced permutations do not satisfy Theorem 1, we cannot be sure that they provide a suitable p value for $\mathbb{H}_0$. We look at some theoretical and empirical results to see how they do.

In microarray experiments, n is often small, because arrays are costly. For example, Tusher et al. (2001) compare two groups of $n=4$ arrays. Our main interest is therefore in values of n between 4 and 20. We also look briefly at $n=2$ because a mathematical answer is possible there and because it illustrates the issues.

3.1. Theory for $n=2$

For $n=2$ there are only $B=4$ distinct balanced permutations, and so we can easily look at all of them. Suppose that $\hat{\Delta}(\mathcal{Z}) = \bar{X} - \bar{Y}$. Then we can write out results from all of the balanced permutations as a vector

$$\begin{pmatrix} \hat{\Delta}_1^* \\ \hat{\Delta}_2^* \\ \hat{\Delta}_3^* \\ \hat{\Delta}_4^* \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ Y_1 \\ Y_2 \end{pmatrix}.$$

Theorem 2. *Suppose that X_1, X_2, Y_1, and Y_2 are independent random variables with the same continuous distribution F. Let $b = \sum_{j=1}^{4} 1\{\hat{\Delta} > \hat{\Delta}_j^*\}$ be the number of balanced permutations that the original permutation beats. Then* $\Pr(b=k) = 1/6$ *for* $k \in \{0,1,3,4\}$ *and* $\Pr(b=2) = 1/3$.

Proof. We easily find that $\hat{\Delta} > \hat{\Delta}_1^*$ if and only if $X_2 > Y_1$. From the other three cases, we find that $\hat{\Delta}$ beats all 4 balanced permutations if and only if both X_i are larger than both Y_i. This has probability 1/6 by symmetry. A similar argument shows that $\hat{\Delta}$ has probability 1/6 of beating none of the $\hat{\Delta}_j^*$. For $\hat{\Delta}$ to beat exactly 3 of the $\hat{\Delta}_j^*$, the X's must rank first and third among the four data points. Thus $\Pr(b=3) = 1/6$ as well. Similarly $\Pr(b=1) = 1/6$, and by subtracting we find $\Pr(b=2) = 1/3$. ■

Thus, for $n=2$, the balanced permutation histogram for comparing means is not uniform. It is slightly conservative. When the original permutation beats all 4 balanced permutations, we would naively claim a p value of 1/5. But this event only has probability 1/6 of happening. Balanced permutations are not likely to be applied when $n=2$, except perhaps when results from multiple genes are to be pooled as described in Section 4. Unfortunately, when n gets larger, the non-uniformity does not disappear, and the method switches from conservative to permissive.

It is interesting to note that the number of permutations beaten has a double high spike in the middle at 2. If we insert a pseudo value $\hat{\Delta}_5^* = 0$, then by symmetry we split that spike into 2 equal pieces and obtain a uniform histogram for $\sum_{j=1}^{5} 1\{\hat{\Delta} > \hat{\Delta}_j^*\}$. This spike at the median appears in all of the balanced permutation histograms for treatment effects that we have investigated. It is not a simple consequence of the histogram being symmetric. For instance, the histogram for the full permutation set is symmetric but it does not have a double high spike in the middle.

3.2. Numerical results for small $n>2$

The method used to prove Theorem 2 does not give us an answer for $n>2$. When $n=4$ the argument there shows that the observed difference will beat all 36 balanced permutations if and only if the smallest $n/2=2$ of the X_i have a higher sum than the largest 2 of the Y_i. The probability of this event depends on the underlying distribution of X_i and Y_i. We proceed numerically, using the $N(0,1)$ distribution for illustration.

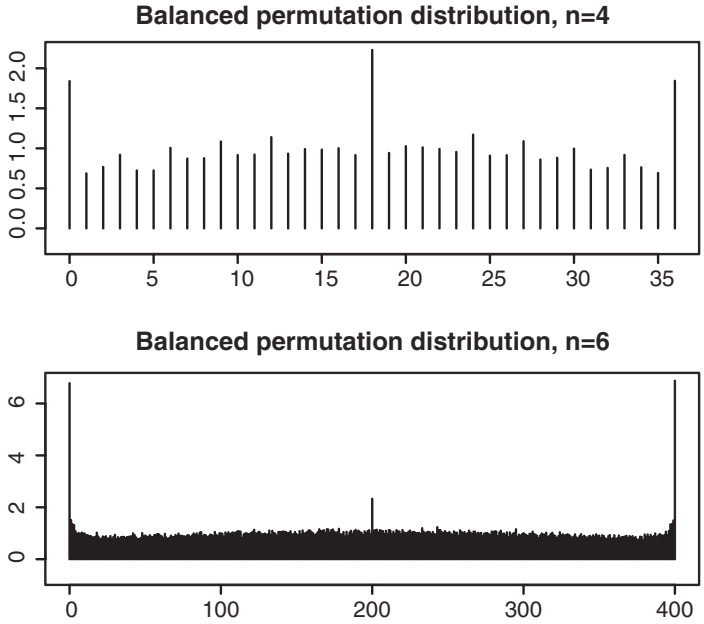

FIG. 2. These figures plot the probability mass function of the number of balanced permutations of $X - Y$ beaten by the original mean difference. The scale is chosen so that the heights of the bars average to 1 in each figure. The top figure is for $n = 4$ $N(0,1)$ random variables contributing to each of $\bar{X}$ and $\bar{Y}$. With $n = 4$ there are 36 balanced permutations. The bottom figure is for $n = 6$ where there are 400 balanced permutations. Both figures differ from the ideal in which all bars have height 1.

For $n = 4$ observations per group there are now $B = \binom{4}{2}^2 = 36$ balanced permutations. The observed $\bar{X} - \bar{Y}$ will beat some number in $\{0, 1, \ldots, 36\}$ of them. Ideally, the number of balanced permutations giving $\pi(\bar{X} - \bar{Y}) < \bar{X} - \bar{Y}$ should have the uniform distribution on those 37 levels. The top plot in Figure 2 shows the actual distribution, from 10,000 simulations of the case where $X_1, \ldots, X_4, Y_1, \ldots, Y_4$ are independent $N(0, 1)$ random variables. The histogram is nearly uniform except that the bars at 0, 18, and 36 are about twice as high as desired. There is also a slight tendency for small and large multiples of 3 to have higher bars than their neighbors.

The bottom plot in Figure 2 considers the case of $n = 6$. Now there are $B = 400$ balanced permutations. The bar at 200 is about twice the desired height, while the ones at 0 and 400 are about 6.8 times as high as they should be.

Histograms for $n = 8$ and 10 (not shown) are also nearly flat except for extra height in the middle and much greater height at the ends. The heights at the extreme ends are described in Section 3.3, for values of n from 2 to 20.

The p values quoted above are one-tailed. For two-tailed p-values one might simply double both the nominal and observed quantities. The true two-tailed p value for a mean that beats all 400 balanced permutations would be about $2 \times 6.8/401 \doteq 0.034$ instead of $2/401 \doteq 0.005$. In principle we might be better

TABLE 1. RESULTS OF 100,000 SIMULATIONS OF A BALANCED PERMUTATION TEST FOR THE DIFFERENCE OF MEANS WITH $N = 4$ PER GROUP

	Exp. < 36	*Exp.* $= 36$
Gaussian < 36	94217	801
Gaussian $= 36$	370	4612

The simulation was run with Gaussian values (all observations in both groups were independent N (0, 1)). Then it was rerun with values from the exponential distribution of mean 1.The simulations were coupled. The table shows counts of the number of times that the original permutation beat all 36 balanced permutations.

off putting both $\bar{X}-\bar{Y}$ and $\bar{Y}-\bar{X}$ into the reference group. Then the two-tailed p-value would be about $2\times 6.8/402$, but changing the denominators from 401 to 402 is a much smaller effect than the others we're studying, so we don't consider it further.

For $n=2$ there is enough symmetry to determine the histogram of p values. For larger n and Gaussian data it may be possible to apply computational geometry. For example, when the treatment difference beats all balanced permutations the combined vector Z lies in an intersection of $\binom{n}{n/2}^2$ half spaces in $\mathbb{R}^{2n}$. But as remarked above, the answer is not distribution free. To confirm this distribution effect, we compared results for exponential data to those for normal data for $n=4$.

Results of a simulation of 100,000 cases for $n=4$ are reported in Table 1. The simulations were run with independent N (0, 1) data for all X_i and Y_i. We counted how many times the observed $\bar{X}-\bar{Y}$ beat all 36 balanced permutation values. Then the simulation was rerun changing the $N(0, 1)$ values to exponential (mean 1) values. The second simulations were coupled to the first by the probability integral transformation, $-\log(1-\Phi^{-1}(X))$ which turns a $N(0, 1)$ random variable X into an exponential one. There were 370 cases where the Gaussian simulation beat all 36 of its balanced permutations while the exponential did not. There were 801 reverse cases. This matched pair difference is statistically significant by McNemar's test. It is equivalent to more than 12 standard deviations using equation 10.4 of Agresti (2002). The size of the effect however is not extremely large. For Gaussian data, the observed treatment effect beats all 36 balanced permutations about 1.843 times as often as it should, while for exponential data the ratio is about 2.003.

3.3. Results for larger n

The most interesting feature of the balanced permutation histogram is the probability that $\hat{\Delta}$ is larger than all $B_n=\binom{n}{n/2}^2$ balanced permutations. It is fairly easy to investigate this probability in the Gaussian setting because of the following lemma.

Lemma 1. *Let $n\geq 2$ be an even integer, and suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are independent $N(\mu, \sigma^2)$ random variables for some $\sigma>0$. Then the observed treatment effect $\hat{\Delta}=\bar{X}-\bar{Y}$ is statistically independent of the whole collection of balanced permutation treatment effects $\hat{\Delta}_1^*, \ldots, \hat{\Delta}_{B_n}^*$.*

Proof. Under the hypothesis $\mathcal{Z}=(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ has a multivariate normal distribution. Therefore, so does $(\hat{\Delta}, \hat{\Delta}_1^*, \ldots, \hat{\Delta}_{B_n}^*)$. Because of the balance, the correlation between $\hat{\Delta}$ and $\hat{\Delta}_j^*$ is zero. Therefore, $\hat{\Delta}$ is independent of $(\hat{\Delta}_1^*, \ldots, \hat{\Delta}_{B_n}^*)$. ■

Lemma 1 lets us estimate the probability that $\hat{\Delta}$ beats all B_n balanced permutations by sampling X_i and Y_i for $i=1, \ldots, n$ many times, computing $\Delta_{(B_n)}^*=\max_{1\leq j\leq B_n}\hat{\Delta}_j^*$ each time, and then averaging the values of

Table 2. Rounded Numerical Estimates of the Probability that an Estimated Treatment Effect from n Observations per Group Beats All $B_n\binom{n}{2}^2$ of its Balanced Permutation Treatment Effects

n	$\hat{p}_n$	$\hat{\theta}_n$	$2.57\mathrm{CV}_n$
2	0.166000	0.83	0.006
4	0.051000	1.89	0.008
6	0.016700	6.71	0.011
8	0.005700	27.90	0.014
10	0.001980	125.00	0.017
12	0.000699	596.00	0.021
14	0.000248	2,920.00	0.026
16	0.000089	14,800.00	0.030
18	0.000032	75,500.00	0.037
20	0.000012	401,000.00	0.046

The estimate $\hat{p}_n$ is in the second column. The normalized estimate $\hat{\theta}_n(B_n+1)\hat{p}_n$ is in the third column. Ideally this should be near 1. The fourth column has $\Phi^{-1}(0.995)=2.57$ times the sampling coefficient of variation of $\hat{\theta}_n$. For example, the approximate 99% confidence interval for θ_8 is 1.4% higher or lower than the estimate 27.9.

$$\eta = \Pr(\hat{\Delta} > \Delta^*_{(B_n)} \mid \Delta^*_{(B_n)}) = \Phi\left(-\sqrt{\frac{n}{2}}\frac{\Delta^*_{(B_n)}}{\sigma}\right).$$

The results of this computation are shown in Table 2. For each even n from 2 to 20 inclusive, 100,000 data sets were investigated. The probability p_n of beating all balanced permutations grows much faster than $1/(B_n+1)$. By $n=20$ the p values are off by roughly 400,000. A least squares fit of $\log(\hat{p}_n(B_n+1))$ versus n has intercept -2.335, slope 0.745 and an R^2 of 0.995. While the fit is good over the range 2 to 20, there is also positive curvature suggesting that extrapolating this model will underestimate the ratio.

3.4. Balanced permutation tests based on t statistics

In high-throughput settings, permutation tests are often run on the t-statistic of equation (2), instead of the raw treatment effect $\hat{\Delta}$. Let t^* be the value of t computed after a reassignment of the treatment versus control labels.

For a single test statistic this may seem redundant, because the permutation test on $\hat{\Delta}$ already behaves similarly to the t statistic. But in high throughput settings we often have many test statistics, for instance one per probe, whose underlying data have extremely different variances. The $\hat{\Delta}^*$ from different probes are not comparable. The t^* values are more comparable because they normalize $\hat{\Delta}^*$ by a standard error. Permutations of t^* from one probe may thus be useful in forming a test for another probe.

The theorem below shows that balanced permutations of t statistics have the same calibration problem that balanced permutations of treatment differences do. As a result, the values in Table 2 apply to t tests too.

Theorem 3. *Let $n \geq 2$ be an even number. Let $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$ be distinct numbers and let $X_1^*, \ldots, X_n^*, Y_1^*, \ldots, Y_n^*$ be those same values after a possible relabeling. Let $\hat{\Delta} = \bar{X} - \bar{Y}$, let t be the t statistic in* (2) *and take $\hat{\Delta}^*$ and t^* to be the analogous quantities on the relabeled values. Then $\hat{\Delta} \geq \hat{\Delta}^*$ holds if and only if $t \geq t^*$*

Proof. The value $\hat{\Delta}^*$ is a strictly increasing function of $\bar{X}^*$, as we vary the labelling. The proof follows by showing that t^* is also strictly increasing in $\bar{X}^*$. For notational simplicity we show that t is strictly increasing in $\bar{X}$.

Let $m = (\bar{X}+\bar{Y})/2 = (\bar{X}^*+\bar{Y}^*)/2$. Then $\hat{\Delta} = 2(\bar{X}-m)$, where m does not depend on the labeling. Let D be the positive square root of $\sum_{i=1}^n (X_i-\bar{X})^2 + \sum_{i=1}^n (Y_i-\bar{Y})^2$, so that $t = 2\sqrt{n(n-1)}(\bar{X}-m)/D$. We need to show that $(\bar{X}-m)/D$ is strictly increasing in $\bar{X}$, where D depends on the labeling.

Introduce $S = \sum_{i=1}^n (X_i-m)^2 + \sum_{i=1}^n (Y_i-m)^2$ which does not depend on labeling. Replacing $X_i - \bar{X}$ by $X_i - m + m - \bar{X}$ in D^2, and similarly for the Y values, we get $D^2 = S - 2n(m-\bar{X})^2$ after simplification. Now

$$\frac{t}{2\sqrt{n(n-1)}} = \frac{\bar{X}-m}{(S-2n(\bar{X}-m)^2)^{1/2}}. \tag{5}$$

The derivative of (5) with respect to $\bar{X}$ is $1/D + 2n(\bar{X}-m)^2/D^3$, which is strictly positive, and so t is increasing in $\bar{X}$. ■

Theorem 3 shows that we get the same calibration problem whether we use t or $\hat{\Delta}$ when using a one tailed criterion. The same equivalence holds for the two tailed criteria $|t|$ and $|\hat{\Delta}|$. Both $|t|$ and $|\hat{\Delta}|$ are increasing in $\bar{X}$ over the range $\bar{X} > \bar{Y}$ both are decreasing in $\bar{X}$ when $\bar{X} < \bar{Y}$, and both are 0 when $\bar{X} = \bar{Y}$.

The assumption that all $2n$ values are distinct can be weakened. It ensures that D is never 0, but that can be enforced with a weaker condition.

3.5. Practical consequences

What we see in the figures is that the reference histogram is roughly uniform but with a spike in the middle and two more at the ends that grow quickly with n.

The spikes at the ends are the most serious because they strongly affect p-values. Suppose for example that $n=10$. Then when the treatment effect beats all balanced permutations we might naively quote a one-tailed p value of $1/(1+\binom{10}{5}^2) \doteq 1.57\times10^{-5}$. However, the proper p-value in that setting is about 1.98×10^{-3}, roughly 125 times larger.

In studies that screen for interesting genes, using a small number n of arrays per group, the case where the observed effect beats all permutations is the most important one. When many genes are investigated, one has to make adjustments for multiple comparisons. Step down methods start by looking at the most significant genes first before choosing a cutoff and quoting a false discovery rate. For an account of multiple testing procedures, see Dudoit and van der Laan (2008). When the very smallest p-values are permissive, the whole process will be adversely affected.

The problem is not confined to the very smallest p-value. If the original treatment effect beats all but one of the balanced permutations, then the claimed p-value is about $(125 + 1)/2 = 63$ times as small as it should be, assuming that the second bar from the right in the reference histogram has nearly its proper height. Similarly, outcomes that are near the ends of the histogram will get permissive p-values.

4. POOLED BALANCED PERMUTATIONS

As mentioned previously, microarray technologies can simultaneously measure the expression levels of thousands of genes at once. To this end, most microarrays use multiple measuring points (probes) for every gene. The main use of balanced permutations in high throughput problems is to calibrate a p value for one probe, by pooling data from a large number N of similar probes. It is also used to compute summaries of N linked tests, such as estimated FDRs. Sometimes there is a near one to one match relating probes to genes, but in other settings there can be numerous probes per gene. Typically, N is in the thousands, while n is much smaller.

Fan et al. (2004) use balanced permutations to get a pooled reference distribution to set N p-values. Jones-Rhoades et al. (2007) use them to estimate a rate for false discovery of genes, in a procedure that merges nearby differentially expressed probes into up and downregulated genomic intervals and then searches for new genes in those intervals.

Suppose that we are making N tests of type H_0, one for each probe. For the i'th probe we get a t statistic t_i. The null distribution of t_i will not be exactly the $t_{(2n-2)}$ distribution, if the underlying data are not normally distributed. Furthermore, one might replace the t test by some quantity that is more robust to very noisy estimates of probe specific standard deviations.

Either a permutation or a balanced permutation analysis generates B permutations of the data and forms a reference distribution of BN test statistics t_{ij} for probes $i = 1, \ldots, N$ and permutations π_j for $j = 1, \ldots, B$. With such a reference distribution, the p value for $t_{i'}$ is

$$\frac{1}{BN}\sum_{i=1}^{N}\sum_{j=1}^{B} 1\{T(t_{i'}) \geq T(t_{ij})\} \tag{6}$$

where the usual choice for T is $T(t) = |t|$. In using equation (6) the original permutation of the data should be one of the B included.

Table 3. This Table Presents the Probability that a Null t Statistic Beats the Largest t Statistic Among Balanced Permutations of Independent Non-null Data

	$n = 4$			$n = 6$	
$\mu = 0$	$\mu = 1$	$\mu = 2$	$\mu = 0$	$\mu = 1$	$\mu = 2$
0.971	1.936	4.234	1.813	6.144	20.64
0.030	0.025	0.017	0.043	0.037	0.021

The values shown in the top row are $\left(1 + \binom{n}{n/2}^2\right) \Pr(t \geq \max_j t_j^*)$, which should be 1 for a properly calibrated reference set. Here t is a t statistic computed with independent X_i, $Y_i \sim N(0, 1)$ for $i = 1, \ldots, n$ and t_j^* are all of the balanced permutations of a t statistic, on data where $X_i \sim N(\mu, 1)$ and $Y_i \sim N(0, 1)$ independently of each other and of t. The second row gives 99% relative error values. There were 10,000 simulations of each case.

4.1. NON-NULL GENES

There are known problems with pooling permutation-based statistics from N problems, whether the permutations are balanced or not. The main one is that the presence of non-null genes among the N cases tested distorts the reference distribution. In a full permutation analysis, many of the sampled permutations are very unbalanced for the treatment effect and they generate a heavy tailed reference distribution. The result is a conservative method that over-estimates FDR. Fan et al. (2005) and Xie et al. (2005) both raise this issue and seek to apply the permutations only to those genes that are not differentially-expressed. Identifying null genes is at least as hard as identifying non-null genes. Yang and Churchill (2007) compare methods that attempt to remove non-null genes.

The opposite occurs in balanced permutations. The balance has the effect of inflating the denominators of t statistics while the treatment effect cancels out of the numerators. The result is a resampled test statistic distribution with tails that are too short. Then, to the extent that these genes are included in the permutation, the FDR will be underestimated.

Table 3 shows some results of this type. Let the non-null data have $X_i \sim N(\mu, 1)$ and $Y_i \sim N(0, 1)$ all independent, and suppose that the null data have a common $N(\tau, \sigma^2)$ for all of X_i and Y_i. The table presents the probability that t computed from the null data beats all the balanced permutation t statistics from the non-null data. The effect grows with μ and with n. The case with $\mu = 0$ and $n = 4$ is not permissive, but $\mu = 0$ is a null case, described in more detail below.

There is partial remedy for the non-null case, described by Pan (2003), which should be helpful when n is large. Notice that each observation belongs to either the test or control group in the original data and to either the test or control group in the permuted data. This yields four different sets of observations in any balanced permutation. The idea of Pan (2003) is to form the variance estimate in the denominator of the t statistic by pooling sample variances from these four subsamples instead of just the permuted treatment and control groups. This will remove mean differences from variance estimate in the denominator of t. Unfortunately, the degrees of freedom for the resulting reference distribution are $2n - 4$ instead of $2n - 2$, which will be an important difference when n is small.

Scheid and Spang (2006, 2007) also note the problem from including non-null genes in the reference distribution. They have proposed a method of filtered permutations. Instead of filtering the list of genes to be mostly null, they filter the set of permutations to be the ones π for which a subsequent permutation $\widetilde{\pi} \circ \pi$ gives nearly $U(0, 1)$ distributed p values for a comparison of the two permuted groups. That is, they employ

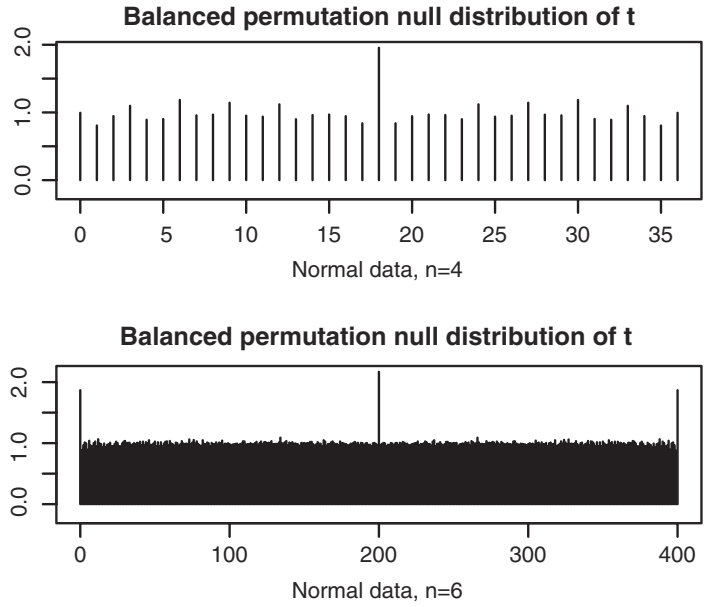

FIG. 3. These plots show histograms of the number of balanced permutations of the t statistic from one data set that are beaten by a t test from an independently sampled data set. All data values were independent $N(0,1)$ in both data sets. The top figure is for $n = 4$, and the bottom one is for $n = 6$.

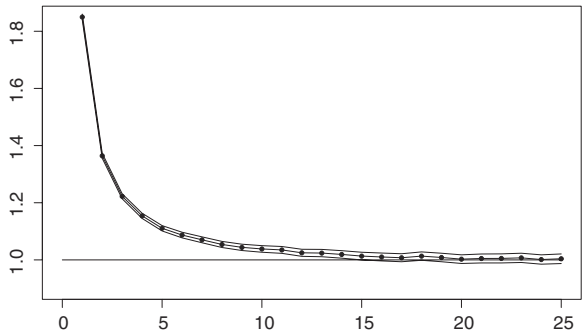

FIG. 4. This figure shows that a reference distribution based on balanced permutation t tests of I > 1 independent null probes is not as permissive as one using a single independent null probe. The horizontal axis shows the number I of independent probes used in a simulation of balanced permutation t tests for two groups of n = 6 observations each. The vertical axis shows 200 I times the probability that the original data has a value of |t| larger than that of all 200 I balanced permutation |t| values in the reference set. The dots show estimated probability ratios, the curves show upper and lower 99% pointwise confidence limits, and there is a horizontal reference line at 1.

an iterated permutation analysis. By considering permutations of permutations their method might be accomplishing numerically what the group structure provides mathematically. But there is as yet no theoretical analysis of filtered permutations.

4.2. Null genes

In the case of balanced permutations, there are also difficulties in the distribution of p values from the null genes. For any probe i, we know that the uniform distribution on $T(t_{ij})$ for $j \in \mathcal{B}_{2n}$ or $j \in \mathcal{B}_{2n+}$ fails to be a good reference distribution. However, when we combine all N of these test distributions into a single reference distribution, then for whichever probe i' we are testing, our reference distribution is dominated by data from probes $i \neq i'$. Hence we need to look at how a reference distribution from one probe works for data from another probe.

A precise comparison of $T(t_{i'})$ with $\pi(T(t_i))$ depends on the dependence between data from probes i and i'. There is usually some dependence, because different genes on the same array tend to be correlated. The case where probes i and i' are highly correlated should be similar to that in which a probe is compared to its own balanced permutations. The case of independent probes is easier to study than probes with arbitrary dependence, and this case should be favorable to the use of balanced permutations. So we consider $X_{i1}, \ldots, X_{ik}, Y_{i1}, \ldots, Y_{ik}$ that are independent of $X_{i'1}, \ldots, X_{i'k}, Y_{i'1}, \ldots, Y_{i'k}$.

For the test statistic $\hat{\Delta} = \bar{X} - \bar{Y}$, on normal data, we find from Lemma 1 that it makes no difference whether probe i is compared to balanced permutations of itself or of another null probe i'.

As mentioned above, when multiple genes are used to calibrate each other, it is more reasonable to permute t statistics. In Table 3, the values for $\mu = 0$ correspond to independent null probes. Figure 3 shows histograms of the number of balanced permuted t statistics from one sample that are beaten by a t statistic from another independently generated sample. Because the second sample is independent of the first we can use the known t distribution to advantage. In this simulation, we generated the list of balanced t values 1000 times. For each of these times, we sorted the 36 possible values, splitting the real line into 37 bins, and evaluated the probability that an independent t random variable would land in each of those bins. The bars shown are the averages of these probabilities over the 1000 simulations.

When $n = 4$ the spikes at the ends have vanished, but the one in the middle remains at about 2.0. So for this case we would get very nearly the desired flat histogram from the device of inserting 0 as a pseudo value. When we repeat the simulation for $n = 6$, there are spikes at the end. The t statistic computed on one data set with $n = 6$ has about twice the chance it should have of beating all the balanced permutations from another data set.

4.3. Multiple independent null genes

Now suppose that probe i' under test is compared to a reference distribution made up of numerous probes indexed by $i = 1, \ldots, I$ and that all $I + 1$ of the probes are statistically independent. Probe i' may tend to beat

all balanced permutations of any probe i more often than it should. But when the I probes are independent of each other it becomes less likely that i' beats all balanced permutations of all I other probes.

We generated 1,000,000 normally distributed data sets of two groups with $n = 6$ observations each. From each of these, we found $t^* = \max_{\pi \in \mathcal{B}_{2n}} |t(\pi(\mathcal{Z}))|$, the maximum of 200 two-tailed t statistics on 10 degrees of freedom each. These million values were then arranged into roughly $10^6/I$ lists of I values each. The probability that an independently generated value of $|t_{(10)}|$ exceeds the largest of these I values should average to $1/(200I)$ for proper calibration. Figure 4 shows corresponding sample averages presented as a multiple of $1/(200I)$, as I varies from 1 to 25. We see that pooling independent null data from balanced permutations is somewhat permissive but that the effect disappears by around $I = 15$ to 25.

The same pattern holds for $n = 8$ observations per group. There the ratio starts nearer to 4.0 but is again close to 1 for $I = 20$.

If we can successfully eliminate the non-null cases and then pool a large number of independent probes, we will get a suitable reference distribution. Then again we might also do that with ordinary permutations, and should we miss some non-null cases, the consequence would be a conservative test instead of a permissive one.

5. DISCUSSION

We find no reason to prefer balanced permutations to ordinary permutations, for genomic applications. The intuitive argument that leads one to expect greater power for them, is countered by them producing permissive p-values when there is no effect. At a minimum, more work is required to justify their use or properly calibrate them.

When used to judge the sampling distribution of a test statistic from N different genes, balanced permutation reference sets are adversely affected by non-null cases among those permuted. So are ordinary permutations. But where ordinary permutations yield a conservative reference distribution, balanced permutations provide a permissive one, when based on the t statistic or on a difference in means. Such permissive methods are more likely to lead to false discoveries.

In practice, it is important to attempt to remove the non-null cases before constructing a permutation-based reference distribution. When the balanced permutations are applied to null cases, we find that balanced permutations can still be mildly permissive. Fortunately, the effect seems to disappear when a large number of null cases have been pooled. However, that disappearance was for null cases that were independent of the case being tested, a condition not required for ordinary permutations. Correlations among the null cases and between the null cases and the gene being tested could result in a distortion that does not decrease for large numbers of pooled reference genes.

For the problems that we have analyzed, it is clear that balanced permutations cannot simply be used as a substitute for full permutation methods. Our analysis does not cover very complicated methods such as the one Jones-Rhoades et al. (2007) used or the empirical Bayes analysis proposed by Efron et al. (2001). We would hesitate to use balanced permutations as an ingredient in more complicated analyses, because they work poorly in simple situations.

We started looking at balanced permutations as an ingredient in a similarly complicated analysis, involving data from the AGEMAP study of Zahn et al. (2007). We considered near balanced permutations with respect to age (old versus very old mice) within strata defined by male and female mice. It could only be near balance because there were 5 mice of each sex at each age. For each mouse, there were 14 tissues. The effects of interest were related to differences in the gene-gene correlation matrices at the two ages. Simulations of even a simple before versus after correlation for just one pair of genes whose correlation does not change revealed permissive p-values like the ones we report here. As a consequence that work continued with ordinary permutations instead of balanced ones.

A final disadvantage of balanced permutations is the one mentioned at the NCI website. There are fewer balanced permutations, so that granularity properties are exacerbated.

ACKNOWLEDGMENTS

We thank Sarah Emerson for helpful discussions. This work was supported by the U.S. National Science Foundation (grant DMS-0604939), the U.S. National Institutes of Health, and a training fellowship from the

National Library of Medicine to the Biomedical Informatics Training Program at Stanford (NLM grant no. 07033).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Agresti, A. 2002. *Categorical Data Analysis*, 2nd ed. Wiley, New York.

Bahadur, R.R., and Savage, L.J. 1956. The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Stat.* 27, 1115–1122.

Dudoit, S., and van der Laan, M.J. 2008. *Multiple Testing Procedures with Applications to Genetics*. Springer-Verlag, New York.

Edgington, E.S., and Onghena, P. 2007. *Randomization Tests*, 4th ed. Chapman and Hall/CRC, Boca Raton, FL.

Efron, B., Tibshirani, R., Storey, J.D., et al. 2001. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc*. 96, 1151–1160.

Fan, J., Tam, P., Woude, G.V., et al. 2004. Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc. Natl. Acad. Sci. USA* 101, 1135–1140.

Fan, J., Chen, Y., Chan, H.M., et al. 2005. Removing intensity effects and indentifying significant genes for Affymetrix arrays in macrophage migration inhibitory factor-suppresed neuroblastoma cells. *Proc. Natl. Acad. Sci. USA* 102, 17751–17756.

Good, P. 2000. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd ed. Springer-Verlag, New York.

Jones-Rhoades, M.W., Borevitz, J.O., and Presus, D. 2007. Genome-wide expression profiling of the *Arabidopsis* female gametophyte identifies families of small, secreted proteins. *PLOS Genet*. 3, 1848–1861.

Lehmann, E.L., and Romano, J.P. 2005. *Testing Statistical Hypotheses,* 3rd ed. Springer, New York.

Miller, R.G. 1986. *Beyond ANOVA, Basics of Applied Statistics*. Wiley, New York.

Pan, W. 2003. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics* 19, 1333–1340.

Scheid, S., and Spang, R. 2006. Permutation filtering: a novel concept for significance analysis of large-scale genomic data. *Lect. Notes Comput. Sci.* 3909, 338–347.

Scheid, S., and Spang, R. 2007. Compensating for unknown confounders in microarray data analysis using filtered permutations. *J. Comput. Biol*. 14, 669–681.

Tusher, V.G., Tibshirani, R., and Chu, G. 2001. Significance analysis of microarrays applied to the ioninzing radiation response. *Proc. Natl. Acad. Sci. USA* 98, 5116–5121.

Xie, Y., Pan, W., and Khodursky, A.B. 2005. A note on using permutation-based false discovery estimates to compare different analysis methods for microarray data. *Bioinformatics* 21, 4280–4288.

Yang, H., and Churchill, G. 2007. Estimating *p*-values in small microarray experiments. *Bioinformatics* 23, 38–43.

Zahn, J., Poosala, S., Owen, A.B., et al. 2007. AGEMAP: a gene expression database for aging in mice. *PLOS Genet.* 3, 2326–2337.

Address reprint requests to:
Dr. Art B. Owen
Department of Statistics
Stanford University
Sequoia Hall
Stanford, CA 94305

E-mail: owen@stanford.edu