Statistics and Its Interface Volume 6 (2013) 79–89

On permutation procedures for strong control in multiple testing with gene expression data

Grzegorz A. Rempala* and Yuhong Yang†,‡

We consider two popular, permutation-based, step-down procedures of p-values adjustment in multiple testing problems known as $\min P$ and $\max T$ and intended for strong control of the family-wise error rate, under the so-called subset pivotality property (SPP). We examine key but subtle issues involved in ascertaining validity of these methods, and also introduce a new, slightly narrower notion of strong control which ensures proper bounds on the family-wise error rate in $\min P$ and $\max T$ without SPP.

AMS 2000 subject classifications: Primary 62H15, 62J15; secondary 92D10.
Keywords and phrases: Multiple testing, Strong control of family-wise error rate, Step down p-value adjustment methods, Subset pivotality property, Permutation test.

1. INTRODUCTION

Over the last decade, permutation-based methods for strong control of the family-wise error rate have received considerable attention in genomic applications (e.g., microarray data analysis under treatment and control) in the context of multiple testing. Naturally, one may test the difference between control and treatment separately for each gene. However, since typically a large number of genes (thousands or more) are investigated in a modern microarray-based or other genomic experiment, the simple-minded approach of conducting the tests one at a time is bound to have enormous probability of type I error due to the familiar problem of multiple testing. This is why the multiple-testing-adjusted statistical procedures have received much attention in genomic data analysis. For a general introduction to the topic, see, e.g. the recent monograph by Dudoit and van der Laan (2008).

The familiar single-step Bonferroni procedure and the likes provide strong control of the error rate, but they are typically too conservative. Westfall and Young (1993) proposed less conservative step-down $\min P$ and $\max T$ procedures (see Appendix) intended to properly take the dependence between tests into account in a non-specific way. Westfall and Young (1993) also pointed out that since under most circumstances the joint and marginal null distributions of test statistics are unknown, a practical way of implementing proper family-wise error control is via the usual permutation (or randomization) approximation to these null distributions.

One of the widely debated issues (see e.g., Westfall and Troendle 2009 and references therein) is the precise nature of the family-wise error rate control provided by these permutation-based, multiple-testing adjustment procedures and the appropriate assumptions which are required for such control. As pointed out in Chapter 2 of Dudoit and van der Laan, in order to properly control the rate of type I errors, whether in a multiple testing problem or not, one needs to specify the joint distribution of the test statistics. Hence, the key idea seems to lie in the specification of the appropriate *test statistic null distribution* (as opposed to the data generating distribution) which ensures the control of the error rates under the true null distribution (often partially unknown). Whereas in their monograph Dudoit and van de Laan (2008 Chapter 2) considered the asymptotic control properties of the $\min P$ and $\max T$ procedures under the *null dominance* assumption, Westfall and Young (1993) argued that $\min P$ and $\max T$ provide exact (non-asymptotic) strong control under the null distributions of the test statistics having the *subset pivotality property* or SPP (see, Westfall and Young 1993, Chapter 2; Dudoit and van der Laan Chapter 2, as well as Definition 3 below). Seemingly, there has been some confusion/misunderstanding/debate surrounding the SPP concept itself as well as its validity in the specific context of the microarray data. For example, it was stated in the literature that if each individual test on a gene depends only on the observations on that gene, then SPP holds. It seems, however, that the issue is quite subtle, and the above statement may or may not be correct, depending on the setup of the hypotheses. In particular, as we demonstrate herein, a strange paradox about this requirement seems to be at work when considering permutation-based methods in pursuit of strong control. On one hand, we can argue that SPP does not hold generally in the gene expression data context; but on the other hand, we can also argue that it does not really matter if subset pivotality holds or not. This casts some doubt on the validity of the permutation-based methods, which is further supported by our examples. However,

*Grzegorz Rempala's research was partially sponsored by NSF grants DMS-0840695 and DMS-1106485 as well as NIH grant R01-DE019243.
†Yuhong Yang's research was partially supported by NSF grant DMS-1106576.
‡Corresponding author.

as we show in the current paper, despite these difficulties, it turns out that for adjusting p values based on marginal tests, the $\min P$ and $\max T$ always provide strong control in a weaker sense *exactly* (not just *approximately*), whether or not SPP holds.

In the current paper, for the sake of clarity and simplicity, we focus on one specific setting, that is, the microarray experiment comparing a treatment condition with a control one. We first illustrate some subtle issues/difficulties surrounding the problem of multiple testing in this setting and subsequently propose potential remedies and offer several clarifications. In the next section (Section 2) we introduce our notation and briefly review the main concepts. Section 3 of the paper discusses the meaning of the multiple testing null hypotheses and points to some difficulties with applying the notion of partial null hypothesis in permutation-based step-down procedures. In Section 4, we address SPP and the issue of its validity in microarray data. In particular, we give a general result on broad existence of families with SPP (Theorem 1). In Section 5 we show that the permutation $\min P$ and $\max T$ procedures do provide strong control in a less strict sense. To make this concept rigorous we introduce a notion of *partial* strong control and give a formal result as Theorem 2. Finally, in Section 6, some concluding remarks are given. The relevant proof and auxiliary results are deferred to the appendix.

2. FAMILY-WISE ERROR RATE

Throughout the paper we shall use the following notation. Let $\underline{X}_j = (X_{1j}, \ldots, X_{Nj})$, $1 \leq j \leq n_1$ be iid observations of gene expression levels (possibly after a suitable transformation of the raw data) of N genes under the experimental (treatment) condition and $\underline{Y}_j = (Y_{1j}, \ldots, Y_{Nj})$, $1 \leq j \leq n_2$ be iid observations of the expression levels of the same genes, in the corresponding order, but under the control condition. Herein $\underline{X}$ and $\underline{Y}$ are always assumed to be independent. Note that N is typically much larger than n_1 and n_2. We are interested in whether the treatment affects the gene expression relative to the control condition.

Let X_i and Y_i denote the random expression levels for gene i under the treatment and under the control, respectively. A gene i ($1 \leq i \leq N$) is said to be *differentially expressed* if the distribution of X_i is different from that of Y_i. Note that a more restrictive definition is also often used: gene i is differentially expressed if the mean of X_i is different from that of Y_i. Clearly, the latter definition addresses only the difference in mean. In this work, we will focus on the former definition.

Let $\mathcal{H}_i$ denote the hypothesis that gene i is not differentially expressed and let $H_i \in \{0, 1\}$ be the corresponding indicator function, i.e., $H_i = 0$ when the null hypothesis $\mathcal{H}_i$ is true and $H_i = 1$ otherwise. Following Ge, Dudoit and Speed (2003) in the sequel we shall use $\mathcal{H}_i$ and H_i interchangeably. For testing $\mathcal{H}_i$ (or H_i) a test statistic T_i is proposed, and large values of $|T_i|$ or large (small) values of T_i provide evidence against $\mathcal{H}_i$, depending on the specification of the alternative hypothesis as two-sided or one-sided. Herein we assume that T_i depends only on the observations on gene i, that is, T_i is a function of $(X_{i1}, \ldots, X_{in_1})$ and $(Y_{i1}, \ldots, Y_{in_2})$. We call such a test statistic a marginal one. Examples of T_i include two-sample t-statistic, F-statistic and many other test statistics (e.g. Wilcoxon and Mann-Whitney statistic, see, for instance, Bain and Engelhardt 1992). The main statistical issue in analyzing microarray data stems from the fact that, since many tests are performed, the size of the critical set for an individual test may no longer be a meaningful quantity for characterizing the confidence level associated with the set of genes declared to be differentially expressed based on the individual tests.

Let $M_0 = \{i : H_i = 0\}$ be the collection of indices corresponding to true null hypotheses and $M_1 = \{i : H_i = 1\}$ be its complement (i.e., the false null hypotheses). We let $M = \{i : 1 \leq i \leq N\}$ and note that $M = M_1 \cup M_0$ and $M_1 \cap M_0 = \emptyset$. For a given multiple testing procedure, if any hypothesis in M_0 is rejected, a type I error occurs. The associated probability is called the family-wise error rate (FWER).

Following Ge et al. (2003) (with some abuse of notation, see below), let $H_{M_0} = \cap_{i \in M_0}\{H_i = 0\}$ denote the state that all the null hypotheses in M_0 are true but all the hypotheses in M_1 are false. Let $H_M = \cap_{i \in M}\{H_i = 0\}$ denote the *complete null hypothesis* (i.e., the state of nature when all the gene-specific null hypotheses are true). We note that H_{M_0} with $M \backslash M_0 \neq \emptyset$ is also referred to in the sequel as the *partial null hypothesis*.

Let δ be a multiple testing procedure with $\delta(i)$ indicating the decision on gene i: $\delta(i) = 0$ if H_i is accepted (or not rejected) and $\delta(i) = 1$ if H_i is rejected. Thus the family-wise error rate of δ is

$$FWER(\delta) = Pr\left(\delta(i) = 1 \text{ for at least one } i \in M_0 | H_{M_0}\right).$$

Deferring the discussion of the meaning of the above conditional probability to the next section, we note also the following definitions (see Ge et al. 2003, Westfall and Young 1993, p. 10, Hochberg and Tamhane 1987, p. 3, Dudoit and van der Laan 2008, p. 95). Throughout the paper we assume $0 < \alpha < 1$.

Definition 1 (Weak control)**.** A multiple testing procedure δ is said to weakly control the FWER at level α if

$$Pr\left(\delta(i) = 1 \text{ for at least one } i \in M \,|H_M\right) \leq \alpha.$$

Definition 2 (Strong control)**.** A multiple testing procedure δ is said to strongly control the FWER at level α if for every possible choice of $M_0 \subset M$, we have

$$Pr\left(\delta(i) = 1 \text{ for at least one } i \in M_0 \,|H_{M_0}\right) \leq \alpha.$$

It is obvious that the strong control implies the weak control. The concept of strong control, at first glance, seems to be well-defined in the context of microarrays and elsewhere. However, in a rigorous sense, the definitions are not completely clear since, as discussed e.g. in Dudoit et al. (2004) and Pollard and van der Laan (2004), each subset M_0 of null hypotheses corresponds to a family of possible null distributions of the test statistics. As we argue below, the issue seems fundamental for interpreting correctly the outcomes of any permutation-based analysis intended for strong control.

3. ISSUES WITH H_{M_0}

3.1 Joint or marginal distributions?

Peter Westfall in his discussion of the paper by Ge, Dudoit and Speed (Ge, Dudoit and Speed 2003, p. 63), brings up the issue of how to interpret the joint null hypotheses H_M and H_{M_0} (see also Westfall and Troendle 2008). He points out that there are two interpretations. One is that H_{M_0} and H_M do not address the joint distributions of the expressions of the genes. That is, H_{M_0} (or similarly H_M) means that the marginal distributions of X_i and Y_i are the same for each $i \in M_0$ (or $i \in M$) but nothing else can be said about the joint distributions of $\{X_i, i \in M_0\}$ and $\{Y_i, i \in M_0\}$ (or $\{X_i, i \in M\}$ and $\{Y_i, i \in M\}$). This interpretation matches the interest of comparing the treatment and control marginally over the genes. As expected, there can be infinitely many different joint distributions of $\{X_i, i \in M_0\}$ and $\{Y_i, i \in M_0\}$ that yield the same marginal distribution for X_i and Y_i for $i \in M_0$.

An alternative interpretation of H_{M_0} is that the joint distribution of $\{X_i, i \in M_0\}$ is the same as that of $\{Y_i, i \in M_0\}$. Westfall views that in order to ensure the validity of the procedures $\min P$ and $\max T$ this latter interpretation of H_{M_0} should be adopted, and even though not explicitly stated, Ge, Dudoit and Speed (2003) seem to share this view. While not disagreeing with this, our main point of this subsection is that each interpretation above has some undesirable consequences, and adopting the joint interpretation does not necessarily solve the problem. In fact, under joint interpretation a permutation-based procedure may wrongly find "differentially expressed" genes with high probability due to the difference of the joint distributions of $\{X_i, i \in M_0\}$ and $\{Y_i, i \in M_0\}$. Thus, with the second interpretation of H_{M_0}, the nature of the problem is no longer that of the usual multiple testing (i.e., finding the marginally differentially expressed genes in our context of microarray analysis). This point may be illustrated in the following computer simulation example.

Example 1. Consider $N = 10$ genes and two sets of microarray replicates under treatment and control, with $n_1 = 2$ and $n_2 = 3$. We shall compare the $\max T$ permutation procedure for the nominal control of FWER $= \alpha = 0.1$ under two scenarios in which the marginal distributions of gene expressions in both conditions are the same, but their joined distributions differ. Under the first scenario, we take $\underline{X}_i$, $i = 1, 2$ as two independent vectors of replications of two generated standard normal variables X_1, X_2 respectively, and take $\underline{Y}_i$, $i = 1, 2, 3$ as three independent vectors of independent standard normal variables. Under the second scenario, all five vectors are iid with iid standard normal components. In both cases the one-sided t-statistic is used for the $\max T$ procedure. In this setting, the number of possible permutations is 10 and hence the permutation procedure can be performed exactly for $\alpha = 0.1$. By repeating each of the scenarios a large number of times, we compare the nominal and empirical FWER in the strong control problem in a one-sided test for both scenarios. The results of the analysis performed with the help of R software (http://cran.r-project.org/) and the associated Bioconductor library *"multtest"* are presented in the last section of the Appendix. As we can see from the computer output, under the 100,000 replicates of our first scenario, the empirical rate is seen to be about 20% above the nominal $\alpha = 0.1$ rate, with the difference exceeding the size of the simulation error. This is in contrast to our second scenario where the empirical error rate is seen to agree well with the nominal one, based on the same number of replicates. It seems, therefore, that in the first scenario the $\max T$ procedure implemented in the "multest" library does not really strongly control FWER, even in the approximate sense.

As seen in the next example, the difference between the nominal and true FWER for step-down permutation procedures may be even more pronounced in some specific circumstances.

Example 2. Suppose that X_i and Y_i all have the same continuous distribution with mean $\mu_{X,i}$ and $\mu_{Y,i}$ respectively and unit variance. We assume that the common distribution has an unbounded support on $(-\infty, \infty)$. We are interested in testing $H_{0i} : \mu_{X,i} = \mu_{Y,i} \leq \mu_0$ versus $H_{1i} : \mu_{Y,i} > \max(\mu_0, \mu_{X,i})$ for the genes for a given constant μ_0. For illustration purposes, suppose that there is only one observation for each of the treatment and control. Consider the test statistic $T_i = Y_i - \max(\mu_0, X_i)$ which provides evidence against H_{0i} when $T_i \geq c$ for some constant c. As we briefly outline below, in this setup and with large N, the permutation-based methods reject at least one gene with very high probability, even when none of the genes are differentially expressed.

Suppose that, similarly to Example 1, the true distributions of the observations are given by $X_1 = X_2 = \cdots = X_N$ with X_1 normally distributed with mean μ_1 and unit variance and $Y_1, \ldots, Y_N$ independent and identifiably distributed with mean μ_2 and unit variance. Consider $\max T$ step-down procedure with $\alpha = 0.5$. For the permutation distribution of the two observations, there are only two possibilities: the original data or the switch of $\underline{X}$ and $\underline{Y}$ each

occurring with equal probability of 1/2. Let $T_{\max}$ denote the maximum of the test statistics over all the genes. Then it has the value $\max Y_i - \max(\mu_0, X_1)$ under the original observations, or the value $X_1 - \min\{Y_i : Y_i \geq \mu_0\}$ otherwise (define $\min\{Y_i : Y_i \geq \mu_0\}$ to be μ_0 when $Y_i < \mu_0$ for all i). Clearly $X_1 - \min\{Y_i : Y_i \geq \mu_0\} \leq X_1 - \mu_0$ and $\max Y_i$ is large with high probability when N is large. Therefore, under the null hypotheses $\mu_1 = \mu_2$, with large enough N, for any given $\epsilon > 0$ the value of $T_{\max}$ under the original observation is greater than that under the switch of $\underline{X}$ and $\underline{Y}$ with probability $\geq 1 - \epsilon$. Thus for $\alpha \geq 0.5$, with the permutation approach, we will make type I error with probability close to one when N is large. In other words, the permutation-based approximations to the adjusted p-values for the $\max T$ procedure are not trustworthy in this case, and a similar argument can be also made for the $\min P$ procedure.

The above examples illustrate the following point. In general, with the dependence structure unaccounted for, the permutation approach to the multiple testing problem can perform very poorly and wrongly declare genes to be differentially expressed, due to the changes of the genes dependence structure across the experimental conditions, rather than due to the changes of the marginal distributions. As seen in Example 1, even in the case when there are multiple observations for both the treatment and the control, the problem still exists to some degree. The difficulty is compounded by the nature of the gene expression data, which makes it unclear if the standard asymptotic analysis that assumes a large sample size ($n_1, n_2 \to \infty$ relative to N) would be practically useful/relevant.

The problem described in Examples 1 and 2 also indicates that if one is interested in marginal testing, the permutation approach does not serve that purpose correctly. If one conducts a permutation procedure, its conclusion seems to be about the joint distributions under the control and treatment. Thus the permutation approaches intrinsically are not quite in line with multiple marginal testing.

From the literature, one may get the impression that the permutation procedures $\min P$ and $\max T$ are less conservative compared to the Bonferroni method (or the like). Based on the discussion above, this may not be correct. In fact, Bonferroni method does control the FWER in the multiple marginal testing sense (note that our use of the term "multiple marginal testing" is to emphasize that each test concerns a "marginal" distribution in the sense that there is no real interest in the relationship between the tests, even though a "marginal" distribution can be multi-dimensional), but the permutation procedures do not necessarily control FWER in that sense, as shown in the examples.

Clearly, the definitions of weak and strong controls still make sense when we follow the first interpretation of H_{M_0}. However, in that direction, to our knowledge, there is no multiple marginal testing method with strong control of FWER that goes much beyond the Bonferroni-type procedures (like, e.g., the method introduced in Holm 1979). This considerably weakens the usefulness of the concept of strong control of FWER. Intuitively, this is not too surprising, because with large N and small n_1 and n_2 it seems unlikely that one can get very far without additional restrictions on the joint distribution of $\underline{X}$ and $\underline{Y}$.

In defense of the permutation procedures, one can argue that the multiple marginal testing may not be the correct objective in the first place. Ultimately, if possible, one wants to know the difference between the joint distributions under the control and treatment. However, since genes are often related to one another in nontrivial manners, the statement that the treatment and control are different, by itself may not be very useful, and searching for marginally differentially expressed genes is a constructive way to proceed. It seems that even though challenging, understanding how treatment affects genes both marginally and jointly is an important direction in microarray data analysis.

3.2 What is the state of nature?

Another difficulty with the definition of strong control under the joint interpretation of H_{M_0} is that the state of nature may not be uniquely defined or even cannot be defined at all. This obviously makes the joint interpretation somewhat problematic. Let's consider two examples.

Example 3. Suppose $N = 2$ and that X_1 and X_2 are iid with $N(0,1)$ distribution and $Y_1 = Y_2$ also with $N(0,1)$ distribution. Then what is the state of nature? When the marginal distributions are concerned, clearly, we have $H_1 = H_2 = 0$. Thus, following the first interpretation of H_{M_0}, $M_0 = \{1, 2\}$. With the second interpretation, however, what is H_{M_0}? Actually, we see clearly that H_{M_0} cannot be $\{H_1 = 0, H_2 = 0\}$. Further, it cannot be $\{H_1 = 0\}$ because otherwise for gene no. 2, the distributions of X_2 and Y_2 would have to be different (recall that M_1 is the collection of the false null hypotheses), and the same argument applies to $\{H_2 = 0\}$ as well. Thus M_0 is not well-defined.

Example 4. Let X_1, X_2, X_3, Y_1, Y_2 be iid with $Unif[0,1]$ distribution, and $Y_3 = Y_1 + Y_2 \bmod 1$ (i.e., $Y_3 = Y_1 + Y_2$ if $Y_1 + Y_2 \leq 1$ and $Y_3 = Y_1 + Y_2 - 1$ otherwise). Then, like in Example 3, marginally we have $H_1 = H_2 = H_3 = 0$. Again, under the second interpretation of H_{M_0}, the state of nature is unclear. Obviously H_{M_0} cannot be $\{H_1 = 0, H_2 = 0, H_3 = 0\}$, but how about $\{H_1 = 0, H_2 = 0\}$, or $\{H_1 = 0, H_3 = 0\}$, or $\{H_2 = 0, H_3 = 0\}$? Apparently, H_{M_0} cannot be any of them because otherwise there is only one gene left in M_1 yet the distributions under the two conditions are not different for that gene. Similarly H_{M_0} is none of $\{H_1 = 0\}$, $\{H_2 = 0\}, \{H_3 = 0\}$.

From the above two examples, we see that M_0 (for the concept of strong control) is not properly defined when H_{M_0} (the state of nature) is interpreted in terms of the

joint distribution instead of marginally. For Example 3, $\{H_1 = 0\}$, $\{H_2 = 0\}$ hold separately, but their intersection $\{H_1 = 0, H_2 = 0\}$ does not hold in terms of the joint interpretation. Thus the approach of adding a joint distribution requirement on top of marginal assumptions for conducting a permutation test has an essential difficulty, related to the very validity of the permutation approach.

One may consider two ways in an attempt to overcome the aforementioned problem. One is to define M_0 to be the largest collection of the unaffected genes in the sense that the genes in the set have the same joint distribution of expressions under both the treatment and control conditions but adding any additional gene in the set would make the joint distributions different. This ensures that H_{M_0} does always exist, but it is not hard to see that M_0 is not necessarily unique. Indeed, with this definition, in Example 3, M_0 can be both $\{1\}$ and $\{2\}$. Another thought may be to reinterpret H_{M_0} and H_{M_1} to mean that the genes in M_0 have the same joint distribution under the two conditions and that the genes in M_1 have non-identical joint distributions under the two conditions. Then, in both examples, $M_0 = \emptyset$ is the only choice for the state of nature (which does not seem to agree well with intuition). However, in general, if a choice of M_0 is not empty, then moving any member to M_1 still satisfies the requirement. Even if one puts a maximal requirement on M_0, one still cannot overcome the non-uniqueness of the state of nature.

In any event, due to undefiniteness or non-uniqueness of H_{M_0}, when one performs the permutation procedures, it is unclear how one should interpret the outcome, which is, obviously, undesirable and challenges the usefulness of permutation-based methods for strong control. One might consider putting some restrictive conditions on the joint distributions, so that the state of nature is well defined. This, however, moves away from one's desire of not making strong assumptions on the joint distributions of the test statistics in the multiple test problems.

4. SUBSET PIVOTALITY

Strong control of the FWER, in theory, can be obtained by using the so called closed testing method of Marcus, Peritz and Gabriel (1976) (see also Hochberg and Tamhane 1987, p. 54 and Hsu 1996, p. 137). To implement this method, however, one must have a size α test for every possible intersection of the individual hypotheses. In the context of gene expression data, it seems difficult (to say the least) to construct a meaningful size α test for an intersection hypothesis in which many genes are involved, without restrictive assumptions on the dependence among the genes.

Westfall and Young (1993) proposed two permutation based procedures, $\min P$ and $\max T$, designed to control the family-wise error rate without modeling the dependence among the individual tests. For readers convenience, the details of these procedures are presented in the appendix. For understanding $\min P$ and $\max T$ properties, it is critically important to make a clear distinction between the theoretical probability distributions of the minimal p-value statistics and the permutation distributions. It seems that failing to do so when arguing for strong control property in the literature contributed much to the confusion on the validity of the permutation-based methods. When the theoretical adjustments of the p-values are done by $\min P$ and $\max T$ procedures (by assuming that the distributions of the minimal p-value statistics are known), it is quite clear that the methods are closed without any additional assumptions, like e.g. SPP (see below). Nevertheless, when the implementations of these methods are done via permutations, the matter becomes complicated and subtle. It is not hard to see that the methods are closed with respect to the permutation distributions, this however, may not be sufficient for the strong control in $\min P$ and $\max T$. Indeed, the permutation distribution depends on the data and when the state of nature of the data is not the complete null (that $\underline{X}$ and $\underline{Y}$ have the same distribution), the conditional permutation distribution may have little in common with the true state of nature. Thus the $\min P$ and $\max T$ procedures when practically implemented via permutations, are not really necessarily closed methods in the sense of Marcus, Peritz and Gabriel (1976) since the sizes for all the tests are not necessarily properly controlled under the true data distributions.

The strong control properties of the permutation-based $\min P$ and $\max T$ procedures are stated under a critical assumption, namely, the *subset pivotality* property (SPP) defined below. In a sense, SPP seems to be an attempt to obtain the closed testing method in a practical way. Let us now consider SPP assumption and examine whether it is likely to be satisfied or not, in the context of gene expression data analysis.

Let $P_1, \ldots, P_N$ be the p-value statistics of the test statistics $T_1, \ldots, T_N$. Let $\underline{P} = (P_1, \ldots, P_N)$. The following definition is given in Westfall and Young (1993, p. 42). (We note that the concept is revisited on p. 115 of the book in a less formal fashion.)

Definition 3 (Subset pivotality)**.** The distribution of $\underline{P}$ has the subset pivotality property if the joint distribution of the sub-vector $\{P_i : i \in M_0'\}$ is identical under the restrictions $H_{M_0'}$ and H_M, for all subsets $M_0' = \{i_1, \ldots, i_j\}$ of true null hypotheses.

The definition may seem to be quite clear, but there are subtleties in its statement. Actually, we are aware of two understandings of the definition and the part in question is "*for all subsets* $M_0' = \{i_1, \ldots, i_j\}$ *of true null hypotheses*". One interpretation of the requirement in the definition is that the joint distribution of the sub-vector $\{P_i : i \in M_0'\}$ is identical under the restrictions $H_{M_0'}$ and H_M for all subsets $M_0' = \{i_1, \ldots, i_j\}$ (see Ge, Dudoit, Speed 2003, p. 14). Another interpretation is that there is a fixed true H_{M_0} and the definition requires that the joint distribution of the sub-vector $\{P_i : i \in M_0'\}$ is identical under the restriction $H_{M_0'}$

and H_M for all subsets M_0' of M_0. Obviously the first interpretation is more stringent. We will focus on the first interpretation in the following discussion (however we note that some of the difficulties described below are also encountered under the second interpretation).

From the definition, the property is pertaining to the distribution of $\underline{P}$. Of course, the joint distribution of $\underline{X}$ and $\underline{Y}$ and the test statistics are also in the picture through their effects on $\underline{P}$. Given the choice of the test statistics, is it a property for the single joint distribution of $\underline{X}$ and $\underline{Y}$ or for a family of distributions? Apparently the answer should be the latter, because in the statement, different choices of M_0' are allowed (and obviously they correspond to different distributions of $\underline{X}$ and $\underline{Y}$). But then the meaning of the definition is not quite clear. Does it mean that we start with a collection of the joint distributions of $(\underline{X}, \underline{Y})$ and the condition is required on the corresponding set of distributions of the test statistics? Or can we start with all possible joint distributions of $(\underline{X}, \underline{Y})$ and just consider the subset of the distributions that satisfy the requirement?

The first interpretation seems to be legitimate. The second one may look legitimate at first sight, but it is actually misleading since the restriction to a set of distributions that satisfy SPP cannot be verified in any meaningful way in practical settings. With the understanding that SPP really is a property of a given (verifiable) collection of joint distributions of $(\underline{X}, \underline{Y})$, let Ω denote such a collection. An important question then is: Under what conditions on Ω and $T = (T_1, \ldots, T_N)$, can we expect SPP to hold?

With $\mathcal{H}_i$ clearly defined, but nothing specifically said otherwise, one may think about interpreting the setup in Ge, Dudoit and Speed (2003) as one with Ω including all possible joint distributions of $(\underline{X}, \underline{Y})$ (of course, we still assume that $\underline{X}$ and $\underline{Y}$ are independent). In their argument on why SPP is usually satisfied for the case of gene expression data analysis, they appeal to the fact that T_i depends only on gene i. However, without further assumptions on the joint distributions of genes, this generally seems unlikely be true, as will be seen.

4.1 Is SPP typically satisfied for gene expression data?

As before, let M_0 denote the set of genes that have the same marginal distributions under the treatment and control. According to Definition 3, SPP requires that the joint distribution of $\{P_i : i \in M_0\}$ stay unchanged under H_{M_0} and H_M. However, when the treatment and control conditions have complicated effects on the dependence structure between the genes expressions, even though the tests are done with one gene at a time, one cannot expect SPP to hold in general. To see this, let us consider a simple setting with only three genes.

Example 5. Suppose that Y_1, Y_2, Y_3 are iid standard-normally distributed (thus, under the control condition, the expression levels of the genes are independent). The treatment may or may not change this distribution. Under the complete null hypothesis, X_1, X_2, X_3 are also iid with standard normal distribution. Now suppose that one possible effect of the treatment is that X_1 equals in distribution to $\beta_{11}Y_1 + \beta_{12}Y_2 + \beta_{13}Y_3 + \mu_1$, X_2 equals in distribution to $\beta_{21}Y_1 + \beta_{22}Y_2 + \beta_{23}Y_3 + \mu_2$, and X_3 equals in distribution to $\beta_{31}Y_1 + \beta_{32}Y_2 + \beta_{33}Y_3 + \mu_3$, where μ_1, μ_2, μ_3, and the β parameters are real numbers. Then one can easily arrange the constants so that X_1, X_2, X_3 all have mean zero and variance one, marginally. However, their joint distribution is not necessarily the same as that of Y_1, Y_2, Y_3.

Note again that in the above example true M_0 may not be well-defined when the joint interpretation of H_{M_0} is used and it is then not meaningful to consider SPP under the joint interpretation. On the other hand, under the marginal interpretation of M_0 it is clear that in this family of distributions, SPP fails. Conceptually, this is quite possible to happen in gene expression: the treatment can make some genes co-expressed. This simple example indicates that for a general family of distributions SPP in the gene expression context may not hold without imposing restrictions that are hard to verify/justify (e.g., the assumption that the genes in M_0 are independent among themselves, regardless of the treatment conditions, or that the treatment and control distributions differ only by location parameter shifts).

4.2 A sufficient condition for SPP

If the control and treatment differ in the gene expressions only in terms of location-shift, then SPP holds for the location family. More precisely, consider the following. Let $q(x_1, \ldots, x_N)$ be a given joint probability density function. Assume that the joint distribution of $(X_1, \ldots, X_N)$ is $q_\mu(x_1, \ldots, x_N) = q(x_1 - \mu_1, \ldots, x_N - \mu_N)$ for $\mu = (\mu_1, \ldots, \mu_N) \in R^N$ and the joint distribution of $(Y_1, \ldots, Y_N)$ is of the form of $q_{\tilde{\mu}}$ for some $\tilde{\mu} \in R^N$. Then we say that the treatment and control differ in location-shifts. In this case, SPP obviously holds. For related discussion and more examples in other multiple testing scenarios, see Chapter 3 of Westfall and Young (1993).

4.3 A paradox?

In the discussion of the previous sub-section we pointed out that for a given family of marginal distributions of the observations, with no additional assumptions made on their joint dependence structure across the experimental conditions, SPP does not hold generally. We emphasize that here we have a fixed family to begin with (which is required for considering SPP). Very surprisingly and interestingly, however, when we somewhat change the angle of looking at the problem, it seems that we may claim that in a sense SPP always holds for permutation-based methods via the following result which is argued in the appendix.

Theorem 1. *Let F_M be a true joint distribution of all the genes under both treatment and control, i.e., the true distribution of the vector $(\underline{X}, \underline{Y})$ where $\underline{X} = \{X_i, i \in M\}$ and $\underline{Y} = \{Y_i, i \in M\}$ as well as $M = M_0 \cup M_1$ with $M_0 \neq \emptyset$. Under joint interpretation of H_{M_0} there exists a collection Ω of joint distributions of $(\underline{X}, \underline{Y})$ which contains F_M and satisfies SPP.*

The result of Theorem 1 leads to a paradox in the justification of the use of the permutation-based procedures when pursuing strong control via the step-down adjusting methods. If we start with a given family of the joint distributions of $\underline{X}$ and $\underline{Y}$, we are told that we need SPP for the permutation-based procedures to work. As already mentioned, when the treatment and control conditions give rise to different dependence structure among the genes, SPP does not hold generally. This suggests that the permutation procedure may not be valid (if SPP is really relevant). On the other hand, if one starts from whatever the true joint distribution of $(\underline{X}, \underline{Y})$ is, one can construct a "friendly" family in which SPP always holds. The key point here is that the permutation-based step-down procedures ($\min P$ and $\max T$) do not in any way depend on the specification of the family of the joint distributions and thus it doesn't matter whether the "friendly" family is known explicitly or not. Consequently, any properties of the permutation procedure for approximating type I error probabilities that hold under the true distribution in the "friendly" family have to hold in the original family (whatever it might be) as well. Therefore, in our context, when using the permutation procedures, it seems that one does not need SPP after all under whatever state of nature!

The issue of general utility of SPP seems somewhat unclear and we leave its resolution to more studies. For instance, as is apparent from the proof of Theorem 1, when the test statistics involve multiple genes the construction of the family with SPP fails and, in general, some aspects of the multiple testing procedures in such circumstances differ from our present setting.

In the next section, we argue that for the step-down permutation-based methods the control issue becomes much easier with a slight modification of the original definition of strong control.

5. PARTIAL STRONG CONTROL

The resampling-based methods for p-value adjustment were thought to provide approximations to the joint distributions of the p-value statistics. Consequently, the strong control property was not expected to hold exactly. This was clearly stated, for instance, in Ge, Dudoit and Speed (2003, Section 4). Westfall and Young (1993, Chapter 2) made general statements that when the adjustments of the p-values cannot be done without error, the resampling methods only approximately control the FWER in the strong sense. However, as seen in our Examples 1 and 2 in Section 3, even the approximate control may be questionable, and therefore the understanding of the accuracy of the permutation-based approximations to $\min_{i \in M_0} P_i$ (or $\max_{i \in M_0} T_i$) appears to be key for the understanding of the strong control properties of the permutation-based methods. Unfortunately, it seems that too-often the accuracy issue is brushed aside in discussing the practical aspects of implementing the strong control algorithms. In fact, at first glance, one might not be very optimistic about the accuracy of the approximations. Imagine, for example, that M_0 is of size 150 and $n_1 = n_2 = 30$ (which may not be atypical in the gene expression context for the time being). Together with the complexity of the dependence between the genes, it seems perhaps unrealistic to expect the distribution of $\min_{i \in M_0} P_i$ to be well approximated.

In Section 4 we have made an attempt to illustrate the potential problems regarding the strong control property (or lack of strong control) of the permutation-based methods and the nature of SPP. All these seem to cast some doubt on the permutation-based methods for strong control. Several other researchers (see, for instance, Storey 2003) had expressed some concerns about the satisfaction of SPP requirement for gene expression data.

It is then perhaps surprising to find out that the issues surrounding SPP are actually largely irrelevant and the Westfall and Young procedures with permutation do strongly control the FWER in a certain sense.

The essence of an idea behind a permutation test is in exploiting the symmetry between the observations from treatment and control conditions under the null hypothesis (i.e. identification of the appropriate permutation group). In the setting of gene expression data, this means that under the null hypothesis of no difference between treatment and control, the observations under the two conditions are exchangeable in distribution. More precisely, under this null hypothesis, conditionally on the observed values of $\underline{X}, \underline{Y}$, all subjects have exactly the same probability to be associated with any given vector of expression levels from both experimental conditions. Consequently, the control of the type I error probability conditional on the observed values of $\underline{X}, \underline{Y}$ ensures the control of the unconditional error probability.

The key for obtaining strong control for the permutation-based adjustment is to understand the distribution of $\min_{i \in M_0} \tilde{P}_i$, where $\tilde{P}_i$ is the adjusted p-value. There is no need to require the accuracy of the permutation approximation for the genes which are not in M_0. Therefore, at the heart of the matter is really the issue of whether or not, under the partial null H_{M_0}, there is still the desired symmetry on the set of null genes M_0 that guarantees the validity of the permutation approach. It turns out to be the case under a bit more restrictive definition of strong control given in the following

Definition 4 (Partial strong control)**.** A multiple testing procedure δ is said to partially strongly control the FWER

at level α if for every possible choice of $M_0 \subset M$, when the joint distribution of X_i, $i \in M_0$ is the same as that of Y_i, $i \in M_0$, we have

$$Pr\left(\delta(i) = 1 \text{ for at least one } i \in M_0 \,|H_{M_0}\right) \leq \alpha.$$

The following result then holds. The proof is presented in the appendix.

Theorem 2. *The Westfall and Young's* $\min P$ *and* $\max T$ *permutation procedures partially strongly control the family-wise error rate at the exact nominal level* α.

Theorem 2 formally justifies the use of $\min P$ and $\max T$ procedures for strong control in a bit more restrictive sense (partial strong control instead of strong control). This result is noteworthy for two reasons: (i) despite the prevailing belief that $\min P$ and $\max T$ procedures provide strong control only approximately (and with little ability to assess the accuracy of the approximation), they actually provide (partial) strong control exactly; (ii) there is no need for the subset pivotality assumption (nor for considering a distribution family to which the true distribution is assumed to belong).

As mentioned already, with the definition of M_0 as the collection of all the genes i such that X_i and Y_i have the same distribution marginally, in general the condition that the joint distribution of X_i, $i \in M_0$ is the same as that of Y_i, $i \in M_0$ may not hold. Then there is no guarantee that the probability of type I error is under the desired level for the permutation-based $\min P$ and $\max T$ procedures (see Example 1). To have a better understanding of Theorem 2, we introduce the following definition.

Definition 5 (Maximal joint null set)**.** For a subset of $\{1, \ldots, N\}$, say S, if the joint distribution of X_i, $i \in S$ is the same as that of Y_i, $i \in S$, we call S a joint null set. If a joint null set S is such that when any additional gene is added to the set, the enlarged set is no longer a joint null set, then we call S a maximal joint null set.

Note that as shown already in Section 3, there may be multiple maximal joint null sets. Obviously S has to be a subset of M_0. In any case, consider a maximal joint null set S^* and let $S^{'}$ be the complement of S^* in M_0 (i.e., $S^{'} = M_0 \setminus S^*$). We call $S^{'}$ the set of individual null genes.

Theorem 2 means that for the $\min P$ and $\max T$ procedures, the probability of making any false discovery in a maximal joint null set S^* is always under the intended control. When there are multiple maximal joint null sets, the probability control is for each of them separately (but not necessarily jointly). In general, the type I error in $S^{'}$ may not be well controlled. In the extreme case that S^* has size 1 (i.e., all the null genes are actually individual null genes), Theorem 2 is not useful at all.

In applications, it may be proper to envision that sometimes a treatment leaves most genes completely unaffected or practically unaffected in terms of their joint distribution. Then a maximal joint null set S can be chosen to be this set. Among the rest of genes, even though the treatment has changed their joint distribution, there may still be some whose marginal distributions under treatment happen to be the same as under the control (or almost the same in a practical sense). If the number of individual null genes in $S^{'}$ is much smaller compared to the size of S^*, i.e., $|S^{'}|/|S^*|$ is small, say upper bounded by β, then the number of false discoveries is properly bounded for the permutation based methods. The corollary below follows immediately from the above considerations.

Corollary 1. *Assume that ratio of the number of individual null genes and the size of the maximal joint null set is upper bounded by* β. *Let* α *be the chosen test size when applying the Westfall and Young's* $\min P$ *or* $\max T$ *procedure. Then with probability at least* $1-\alpha$, *the proportion of the genes in* M_0 *declared to be significant is at most* $\frac{\beta}{1+\beta}$.

Remarks:

1. The proof of Theorem 2 follows the spirit of Westfall and Young's (1993) idea of FWER control in step-down procedures, even though the original proof there does not lead to Theorem 2. There are subtleties that are easily confusing in the derivation. It is crucially important to do the conditioning right. In the complete null case, the conditioning is straightforward and one does it right away and presents the whole argument in terms of the conditional probability. However, under the partial null, working with the conditional/unconditional probabilities is trickier, and seems to be a major cause of the confusion on the validity of the permutation-based methods for strong control. If one starts by conditioning on the expressions of all the genes, then the argument cannot go through. The clever adjustment of the Westfall and Young procedures allows one to formally drop the irrelevant genes before making a conditional argument.
2. As mentioned already, the permutation implementation in the $\min P$ and $\max T$ procedures were viewed as approximations to the theoretical distributions of the minimal p-value statistics. From the proof, it is clear that whether and how well the permutation distributions approximate the unconditional distributions of the minimal statistics are not directly relevant for bounding the type I error probability.
3. As pointed out in Ge, Dudoit and Speed (2003), the marginal distributions of the test statistics for the genes do not need to be the same, and additionally the test statistics can be completely different across the genes (if desirable).

6. CONCLUSIONS

It seems that the notion of strong control of FWER in gene expressions analysis, even though appealing, has some

challenging difficulties to overcome. Beyond the Bonferroni and Holm's methods, the only known methods intended for (partial) strong control without assuming additional conditions on the distributions of the p-value statistics (e.g., independence) rely on permutation in their implementation. This approach, however, moves away from the starting point of the usual multiple testing problems because it concerns the joint distributions of the test statistics rather then the marginal ones. Consequently, it could happen with high probability that some genes might be declared differentially expressed simply because the joint distributions of the expression levels of the genes under the treatment and control are different even though the respective marginal distributions remain identical (or practically identical). In addition, under the joint interpretation of the null hypotheses of no difference between treatment and control, the state of the nature may be no longer properly defined.

When one considers a family of joint distribution for the gene expression levels, since the treatment often has effects on the relationship between the genes (e.g., changing independently expressed genes into co-expressed genes), the SPP requirement may be restrictive. Interestingly enough, the step-down methodology utilizing permutations to adjust the p-values, i.e., the Westfall and Young's $\min P$ and $\max T$ procedures actually do ensure strong control in a partial sense without SPP. The difference between partial strong control and strong control is that whereas for the latter we require the appropriate probability to be bounded for all subsets of M_0, for the former we only require this for subsets for which the joint distributions of $\{X_i\}$ and $\{Y_i\}$ are the same. It seems that the partial strong control, although weaker than strong control, can still be practically very useful. If a treatment leaves a large collection of genes totally unaffected (and thus their joint distribution is the same as that under the control), then the $\min P$ and $\max T$ procedures will not falsely pick up any of those genes with a desirably high probability.

7. BIBLIOGRAPHICAL NOTE

The idea of $\min P$ and $\max T$ was introduced in Chapter 2 of Westfall and Young (1993) monograph along with the concepts of a subset pivotality and a partial null set. The permutation based approximations were discussed by Westfall and Young in quite a general setting and not necessarily with gene expressions analysis in mind. The paper of Ge Dudoit and Speed (2003) gave a more computationally feasible implementation of the $\min P$ procedure for gene expression data. The recent discussion of the issues related to applying permutation tests in step-down procedures was given in Kropf et al. (2004). Some of the points raised herein were also discussed in Westfall and Wolfinger (1997) and Westfall and Troendle (2008).

Finally, we also note that some results intended to control strongly FWER in an asymptotic sense under conditions weaker than SPP were presented by Pollard and van der Laan (2004) in the context of single parameter hypothesis.

APPENDIX A

A.1 Step-down $\max T$ and $\min P$ methods

For completeness and readers' convenience, we present here, in our particular context, the Westfall and Young's $\min P$ and $\max T$ procedures for p-values adjustments in multiple testing problems (see Westfall and Young 1993, Chapter 2).

Let $p_1, \ldots, p_N$ be the raw p-values of the genes based on test statistics T_i which depend only on gene i respectively. Let $t_1, t_2, \ldots, t_N$ be the realized values of the test statistics from the data. Let the ordered p-values be $p_{r_1} \le p_{r_2} \le \cdots \le p_{r_N}$ and the ordered values of the statistics be $t_{s_1} \ge t_{s_2} \ge \cdots \ge t_{s_N}$. Note that the p-values may or may not be based on resampling. Note also that there may be ties in the p-values (which can happen with a positive probability, e.g., when the p-values are discrete or when some genes are perfectly correlated with each other). In such cases we could choose any reasonable tie breaking method known in the literature.

Let J be the total number of $(n_1 + n_2)!$ permutations of the subjects. For each permutation $j = 1, \ldots, J$ compute the corresponding p-values $p_1^{(j)}, \ldots, p_N^{(j)}$ and the test statistic values $t_1^{(j)}, t_2^{(j)}, \ldots, t_N^{(j)}$. Then let $\widetilde{p}_{r_1}^{(j)} = \min_{1 \le i \le N} p_i^{(j)}$, $\widetilde{p}_{r_2}^{(j)} = \min_{i \ne r_1} p_i^{(j)}, \ldots, \widetilde{p}_{r_l}^{(j)} = \min_{i \notin \{r_1, \ldots, r_{l-1}\}} p_i^{(j)}, \ldots, \widetilde{p}_{r_N}^{(j)} = p_{r_N}^{(j)}$; Similarly, let $\widetilde{t}_{s_1}^{(j)} = \max_{1 \le i \le N} t_i^{(j)}$, $\widetilde{t}_{s_2}^{(j)} = \max_{i \ne s_1} t_i^{(j)}, \ldots, \widetilde{t}_{s_l}^{(j)} = \max_{i \notin \{s_1, \ldots, s_{l-1}\}} t_i^{(j)}, \ldots, \widetilde{t}_{s_N}^{(j)} = t_{s_N}^{(j)}$.

Now, denote $l_1 = \#\{p_{r_1} \ge \widetilde{p}_{r_1}^{(j)}, 1 \le j \le J\}$, $l_2 = \#\{p_{r_2} \ge \widetilde{p}_{r_2}^{(j)}, 1 \le j \le J\}, \ldots, l_N = \#\{p_{r_N} \ge \widetilde{p}_{r_N}^{(j)}, 1 \le j \le J\}$ and $h_1 = \#\{t_{s_1} \le \widetilde{t}_{s_1}^{(j)}, 1 \le j \le J\}$, $h_2 = \#\{t_{s_2} \le \widetilde{t}_{s_2}^{(j)}, 1 \le j \le J\}, \ldots, h_N = \#\{t_{s_N} \le \widetilde{t}_{s_N}^{(j)}, 1 \le j \le J\}$.

Finally, for the $\min P$ procedure, let the adjusted p-values be $\widetilde{\widetilde{p}}_{r_1} = l_1/J, \widetilde{\widetilde{p}}_{r_2} = \max(l_2/J, \widetilde{\widetilde{p}}_{r_1}), \ldots, \widetilde{\widetilde{p}}_{r_N} = \max(l_N/J, \widetilde{\widetilde{p}}_{r_{N-1}})$; and for the $\max T$ procedure, let the adjusted p-values be $\overline{p}_{s_1} = h_1/J$, $\overline{p}_{s_2} = \max(h_2/J, \overline{p}_{s_1}), \ldots, \overline{p}_{s_N} = \max(h_N/J, \overline{p}_{s_{N-1}})$. In order to control the test FWER at level α, each adjusted p-value needs to be now compared with α.

A.2 Proof of Theorem 1

We first note that under the joint interpretation of H_{M_0}, the joint distribution of $\{X_i, i \in M_0\}$ is the same as that of $\{Y_i, i \in M_0\}$. In order to construct our family of distributions which satisfies SPP, let $M' \subset M$ be any subset of M and consider a joint distribution, say $F_{M'}$, of the sub-vector of $(\underline{X}, \underline{Y})$ consisting only of the components $\{X_i, i \in M'\}$ and $\{Y_i, i \in M'\}$. We first modify these sub-vectors into $\{\widetilde{X}_i, i \in M'\}$ and $\{\widetilde{Y}_i, i \in M'\}$ by replacing for $i \in M' \backslash M_0$ the corresponding components with the mutually independent, standard Gaussian variables independent of each other and of both original sub-vectors.

We then augment the sub-vector $\{\widetilde{X}_i, i \in M'\}$ by an independent vector of mutually independent standard Gaussian components $\{Z_i^x, i \in M \backslash M'\}$. Similarly, we augment the sub-vector $\{\widetilde{Y}_i, i \in M'\}$ by an independent sub-vector $\{Z_i^y, i \in M \backslash M'\}$ of mutually independent, unit-variance Gaussian components with mean one which are additionally independent of $\{Z_i^x, i \in M \backslash M'\}$ as well as $\{\widetilde{X}_i, i \in M'\}$ and $\{\widetilde{Y}_i, i \in M'\}$. Let $\widetilde{F}_{M'}$ be the joint distribution of a $2N$-vector obtained from the sub-vectors $\{ X_i, i \in M'\}$ and $\{ Y_i, i \in M'\}$ by the described above replacement and augmentation procedure. With all subsets $M' \subset M$, we consider a family of distributions given by $\{\widetilde{F}_{M'} : M' \subset M\}$. Finally, when $M_0 \neq M$, we also add to this family an additional distribution, namely that of the sub-vectors $\{X_i, i \in M_0\}$ and $\{Y_i, i \in M_0\}$ augmented to $2N$ vector by adding to each one of them a vector of mutually independent standard Gaussian components. As above, these added Gaussian vectors are taken to be independent of each other as well as of the original sub-vectors $\{X_i, i \in M_0\}$ and $\{Y_i, i \in M_0\}$. Note that this latest distribution corresponds to a complete null hypothesis.

With our construction, we now have a family of distributions of $(\underline{X}, \underline{Y})$. Note that for each $M' \subset M$ there is a member of the family which has M' as exactly the set of genes which are not differentially expressed. Furthermore, it is not difficult to see that SPP holds for this new family. This completes the proof.

A.3 Proof of Theorem 2

We focus on the $\min P$ procedure. The $\max T$ can be handled similarly.

Suppose that $S = \{k_1, k_2, \ldots, k_m\}$ is a joint null set, $m \geq 1$ (obviously, when $m = 0$, there cannot be any type I error). We want to show that the probability of at least one of the null genes in S being declared to be significant is no greater than α, i.e., that

$$\Pr\left(\widetilde{\widetilde{p}}_i \leq \alpha \text{ for at least one } i \in S\right) \leq \alpha.$$

Let p_{k^*} be the smallest p-value in $p_{k_1}, \ldots, p_{k_m}$. When there are ties, we follow the same order as in $p_{r_1} \leq p_{r_2} \leq \cdots \leq p_{r_N}$ to break the ties. Then the above requirement is $\Pr(\widetilde{\widetilde{p}}_{k^*} \leq \alpha) \leq \alpha$. Now by the definition of $\widetilde{\widetilde{p}}_i$,

$$\begin{aligned}\Pr\left(\widetilde{\widetilde{p}}_{k^*} \leq \alpha\right) &\leq \Pr\left(\frac{\#\{p_{k^*} \leq \widetilde{p}_{k^*}^{(j)}\}}{J} \leq \alpha\right) \\ &\leq \Pr\left(\frac{\#\{p_{k^*} \leq \min_{k \in S} p_k^{(j)}\}}{J} \leq \alpha\right),\end{aligned}$$

where the second inequality holds because by construction the values $p_k^{(j)}$ for $k \in S$ are included in the minimization used for obtaining $\widetilde{p}_{k^*}^{(j)}$. Notice that for the last probability above, only the genes in S are involved, and since each test statistic T_i involves only single gene i, the last expression depends on the set S only. Now, because the distribution of X_i, $i \in S$ is the same as that of Y_i, $i \in S$, conditional on the values of expression of the genes in S for all the subjects, due to symmetry, each permutation of the subjects to be associated with the given values of the expression has exactly the same probability. Then it follows directly that the conditional probability of the event "$p_{k^*} \leq \min_{k \in S} p_k^{(j)}$ for no more than a fraction of $\alpha \times 100\%$ times" is no greater than α. Since the upper bound α does not depend on the expression values, obviously, the unconditional probability of the type I error is also upper bounded by α. This completes the proof of the theorem.

A.4 R code

The following code was used to conduct the numerical simulation for Example 1. The output below was obtained from R software version 2.13.1 (with 'multtest' library version 2.8.0) running on the Mac Pro with dual quad-core Intel Xeon processor. The Monte-Carlo error of the simulation ("err bounds" in the output below) was estimated by computing the corresponding lower and upper bound in the law of the iterated logarithm for $B = 100,000$ replicates. In this particular setting, the difference between the empirical and the nominal FWER was considered to be within the simulation margin of error if the lower value in "err bounds" fell below the nominal FWER $= \alpha = 0.1$.

```
> require('multtest')
> B=100000;
> n=10;
> cl<-c(0,0,1,1,1);
>
> test=function(k=1,classlab=cl) {
+ sink('teka'); #re-direct irrelevant output
+ cnt=0;
+ for (i in 1:B){
+ data=cbind(rnorm(k),rnorm(k),rnorm(n),rnorm(n),
rnorm(n));
+ mt.maxT(data,classlab,test='t.equalvar',
side='lower')->res;
+ cnt=cnt+ifelse(sum(res[,4]<.11)>0,1,0);
+ } #compare to nominal level of 1/10
+ sink(); #return output to console
+ cat('emp.FWER=',cnt/B,'\n');
+ a=sqrt(2*(1-cnt/B)*cnt/B)*log(log(B))/sqrt(B);
+ cat('err bounds=',c(cnt/B-a,cnt/B+a),'\n');
+ }
>#Dataset One: Marginal t-Stat Equidistribution
> test(k=1);
emp.FWER= 0.12025
err bounds= 0.1166958 0.1238042
>
>#Dataset Two: Joint t-Stat Equidistribution
> test(k=n);
```

```
emp.FWER= 0.09926
err bounds= 0.09599255 0.1025274
```

ACKNOWLEDGMENTS

This work was initiated during both authors year-long visit to the Institute for Mathematics and Its Applications (IMA) at the University of Minnesota as participants in the annual program "Probability and Statistics in Complex Systems: Genomics, Networks, and Financial Engineering" in the academic year 2003–2004. The authors would like to thank the IMA for its support. They would also like to express their gratitude towards the annual program organizers as well as the invited guests and workshops speakers for introducing them to the statistical problems of genomics and thus creating an opportunity for their collaboration. Finally, the authors full-heartedly thank the referees and the associate editor for constructive comments which helped them improve the paper.

Received August 18, 2011

REFERENCES

Bain, L. J. and Engelhardt, M. (1991). *Introduction to Probability and Mathematical Statistics*, Duxbury Press, North Scituate, MA.

Dudoit, S., Shaffer, J. P. and Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**(1) 71–103. MR1997066

Dudoit, S. and van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics. MR2373771

Ge, Y., Dudoit, S. and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *TEST* **12**(1) 1–44 (discussion pp. 44–77). MR1993286

Ge, Y., Sealfon, S. and Speed, T. P. (2009). Multiple testing and its applications to microarrays. *Statistical Methods in Medical Research* **18**(6) 543–563. MR2750559

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley & Sons, New York. MR0914493

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6** 65–70. MR0538597

Hsu, J. C. (1996). *Multiple Comparisons. Theory and Methods*. Chapman & Hall Ltd, New York. MR1629127

Korn, E. L., Troendle, J. F., McShane, L. M. and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* **124** 379–398. MR2080371

Marcus, R., Peritz, E. and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056

Kropf, S., Läuter, J., Eszlinger, M., Krohn, K. and Paschke, R. (2004). Nonparametric multiple test procedures with data-driven order of hypotheses and with weighted hypotheses. *Journal of Statistical Planning and Inference* **125**(1) 31–47. MR2086887

Petrondas, D. A. and Gabriel, K. R. (1983). Multiple comparisons by rerandomisation tests. *Journal of the American Statistical Association* **78** 949–957.

Pollard, K. and van der Laan, M. (2004). Choice of a null distribution in resampling-based multiple testing *Journal of Statistical Planning and Inference* **125**(1–2) 85–100. MR2086890

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* **31**(6) 2013–2035. MR2036398

Westfall, P. H. and Young, S. S. (1993). *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. John Wiley & Sons, New York.

Westfall, P. H. and Wolfinger, R. D.(1997). Multiple tests with discrete distributions. *The American Statistician* **51** 3–8.

Westfall, P. H. (2003). Discussion of the paper by Ge, Dudoit and Speed. *TEST* **12**(1) 60–65.

Westfall, P. H. and Troendle, J. F. (2008). Multiple testing with minimal assumptions, *Biometrical Journal* **50** 745–755. MR2542340

Grzegorz A. Rempala
Department of Biostatistics and Cancer Research Center
Georgia Health Sciences University
USA
E-mail address: grempala@georgiahealth.edu

Yuhong Yang
School of Statistics
University of Minnesota
USA
E-mail address: yyang@stat.umn.edu