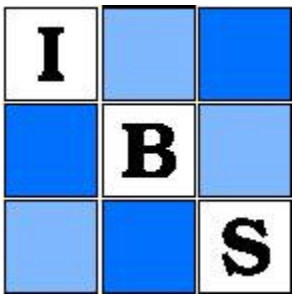

BIOMETRICS 64, 1215–1222
December 2008

DOI: 10.1111/j.1541-0420.2007.00984.x

Screening for Partial Conjunction Hypotheses

Yoav Benjamini

Department of Statistics and Operations Research, Tel Aviv University,
Tel Aviv 69978, Israel
email: ybenja@post.tau.ac.il

and

Ruth Heller

Department of Statistics, University of Pennsylvania, Philadelphia,
Pennsylvania 19104-6340, U.S.A.
email: ruheller@wharton.upenn.edu

SUMMARY. We consider the problem of testing for partial conjunction of hypothesis, which argues that at least u out of n tested hypotheses are false. It offers an in-between approach to the testing of the conjunction of null hypotheses against the alternative that at least one is not, and the testing of the disjunction of null hypotheses against the alternative that all hypotheses are not null. We suggest powerful test statistics for testing such a partial conjunction hypothesis that are valid under dependence between the test statistics as well as under independence. We then address the problem of testing many partial conjunction hypotheses simultaneously using the false discovery rate (FDR) approach. We prove that if the FDR controlling procedure in Benjamini and Hochberg (1995, *Journal of the Royal Statistical Society, Series B* **57,** 289–300) is used for this purpose the FDR is controlled under various dependency structures. Moreover, we can screen at all levels simultaneously in order to display the findings on a superimposed map and still control an appropriate FDR measure. We apply the method to examples from microarray analysis and functional magnetic resonance imaging (fMRI), two application areas where the need for partial conjunction analysis has been identified.

KEY WORDS: False discovery rate; Functional MRI; Global null; Meta-analysis; Microarray; Multiple comparisons.

1. Introduction

In many modern biostatistics applications there is a need to combine p-value maps. In functional magnetic resonance imaging (fMRI) the signal in the brain indicating activity is recorded over time while the subject is involved in a cognitive task. From the map of p-values, regions in the brain that participated in the task are identified. When several cognitive tasks are studied the researcher is interested in the brain regions that participated in most (or at least one) of several cognitive tasks. The map is indexed by the brain location, and the p-values across tasks in the same location may be dependent. Another example from genomics research is that of meta-analysis of microarray experiments to help identify genes that were consistently differentially expressed in most experiments that examine the same problem. The index in the map is the gene, and the p-values for the same gene across experiments are independent.

Pooling together inferences made under different yet related conditions enables the researcher to (1) gain statistical power, or (2) make a stronger scientific statement. The first goal is the more familiar one, as it is in frequent use in meta-analysis. Although there may be only a weak evidence against the null hypothesis at each study, pooling the evidence across studies may yield very convincing results. Methods are abundant for producing a single combined p-value to show that at least one hypothesis is false by testing the *conjunction of null hypotheses*, also known as the *global null hypothesis*, the *intersection null hypothesis*, or the *omnibus null hypothesis*. Fisher's combined p-value is probably the best known method for this purpose (see, e.g., Lazar et al. [2002], Zaykin et al. [2002], and Loughin [2004]).

Even when the above goal is achieved, the scientific conclusion arrived at is quite weak, in the sense that the evidence may stem from a very strong result in a single study and none in the others. Thus the second goal for combining p-values addresses this weakness: we would like to show that the results across studies are consistent in the sense that the null hypothesis at each and every study can be rejected. To show such a result, the disjunction of null hypotheses is tested against the alternative that all hypotheses are false (i.e., against the conjunction of alternative hypotheses). The need to answer such questions has arisen quite naturally in fMRI analysis

(see Friston, Holmes, and Worsley [1999] and Nichols et al. [2005]), where the disjunction of null hypotheses is known as the conjunction null. The analysis is challenging because this null is tested in many brain locations.

As noted above the findings from the rejection of the conjunction of null hypotheses are often too general to be scientifically meaningful. Yet rejecting the disjunction of null hypotheses is often too restrictive, making it practically very difficult to reject anywhere when screening a large number of such hypotheses. A natural compromise is to test instead the partial conjunction null that at least a prespecified number of the null hypotheses hold, against the alternative that at least u out of the n null hypotheses are false.

Such a test is called the *partial conjunction test.* Formally, consider $n \geq 2$ null hypotheses at each "location" $s \in \{1, \ldots, S\}, \mathrm{H}_{01}(s)$, $\mathrm{H}_{02}(s), \ldots, \mathrm{H}_{0n}(s)$, and let $p_1(s), \ldots, p_n(s)$ be their associated p-values. Let $k(s)$ be the (unknown) number of false null hypotheses in location s, then our question "Are at least u out of n null hypotheses false?" can be formulated as follows:

$$\mathrm{H}_0^{u/n}(s) : k(s) < u \quad \text{versus } H_1^{u/n}(s) : k(s) \geq u. \tag{1}$$

Friston, Penny, and Glaser (2005) have recognized the usefulness of testing $\mathrm{H}_0^{u/n}(s)$ in fMRI research, when searching for regions in the brain that participate in u different cognitive tasks out of n tasks of similar nature. They suggested using the maximum p-value at each location as the test statistic, adjusting its distribution to take care of both the u-out-of-n and of the multiple locations simultaneously by controlling the familywise error rate. However, this method has two drawbacks. First, it has very low power at a location even if the location responds to all but one condition, as noted by McNamee and Lazar (2004) and demonstrated in Section 7. Second, unless the conjunction hypothesis where $u = n$ is tested, the method is only valid for independent test statistics within every brain location.

The approach we suggest here is different. First, in Section 2 we present a simple general principle for combining the p-values at each location s to derive a valid p-value for testing $\mathrm{H}_0^{u/n}(s)$. The actual choice should further rely on the dependency structure between the p-values at each location, as discussed in Sections 2.1 and 2.2. All choices lead to the use of the maximum p-value when the tested null is the disjunction of null hypotheses (where $u = n$) and lead to familiar tests when the tested null is the conjunction of null hypotheses (where $u = 1$).

We then suggest to screen the valid partial conjunction p-value map across locations while controlling for the false discovery rate (FDR). The (perhaps more) intuitive procedure in such settings, to apply an FDR controlling procedure on each p-value map separately and then take the intersection of the discovered locations, does not control the FDR of the combined discoveries. In the extreme situation where the conjunction of threshold maps is that of the falsely discovered locations, the FDR will be 1. In Section 3 we prove that the procedure in Benjamini and Hochberg (1995), hereafter BH, on the pooled p-values for partial conjunctions, controls the FDR when the original maps are independent even when the p-values within every map are dependent and discuss the validity of this procedure in other realistic settings. Because it may be of interest to look at all levels of partial conjunction $\mathrm{H}_0^{u/n}$, for $u = 1, \ldots, n$, in Section 4 we define an appropriate error measure to control in this case and prove that it is controlled by the procedure in BH.

In Sections 5 and 6 we give examples from fMRI and microarray analysis, respectively. In Section 7 we discuss the power of the methodology suggested via simulations. In Section 8 we give our final remarks.

2. Combining p-Values

Many methods for combining p-values can be designed. Under the partial conjunction null $\mathrm{H}_0^{u/n}(s)$, let $U_1, \ldots, U_{n-u+1}$ be the p-values for which the null hypotheses hold, in the sense that $U_i \overset{\geq}{\text{st}} U(0,1)$ for $i = 1, \ldots, n-u+1$, and let $P_1, \ldots, P_{u-1}$ be the other p-values. Without loss of generality, for a vector of p-values from the partial conjunction null, let the first $n-u+1$ entries correspond to the p-values where the null hypothesis holds and let $P^{u/n}(s) = f(U_1, \ldots, U_{n-u+1}, P_1, \ldots, P_{u-1})$ be the combined p-value. As long as the combining method makes sense, in that f is nondecreasing in all its components, $f(U_1, \ldots, U_{n-u+1}, h_1(P_1), \ldots, h_{u-1}(P_{u-1})) \leq f(U_1, \ldots, U_{n-u+1}, P_1, \ldots, P_{u-1})$ for functions $h_i(x) \leq x$, $i = 1, \ldots, u-1$. Therefore, if the event $\{P^{u/n}(s) \leq q\}$ occurs then the event $\{f(U_1, \ldots, U_{n-u+1}, h_1(P_1), \ldots, h_{u-1}(P_{u-1})) \leq q\}$ occurs and we just proved the following lemma.

LEMMA 1. *Under* $\mathrm{H}_0^{u/n}(s)$, *let* $h_i(P_i) \leq P_i$ *for some function* $h_i(\cdot), i = 1, \ldots, u-1$, *and let* $P_*^{u/n}(s) = f(U_1, \ldots, U_{n-u+1}, h_1(P_1), \ldots, h_{u-1}(P_{u-1}))$ *and* $P^{u/n}(s) = f(U_1, \ldots, U_{n-u+1}, P_1, \ldots, P_{u-1})$. *Then* $P_*^{u/n}(s) \overset{\leq}{\text{st}} P^{u/n}(s)$.

Lemma 1 helps us construct valid pooled p-values. Because the stochastically smallest $P^{u/n}(s)$ under $\mathrm{H}_0^{u/n}(s)$ will occur when $u-1$ p-values are identically zero, the pooled value $P^{u/n}(s)$ will be valid if it depends only on the $n-u+1$ largest p-values using a combining function that satisfies $f(U_1, \ldots, U_{n-u+1}, 0, \ldots, 0) \overset{\geq}{\text{st}} U(0,1)$. Below we give several valid p-values.

2.1 *Combining p-Values under Dependence*

Given $p_i(s)$ the p-value for testing $\mathrm{H}_{0i}(s)$, and the sorted values being $p_{(1)}(s) \leq p_{(2)}(s) \leq \cdots \leq p_{(n)}(s)$, the intersection of hypotheses $\cap_{i=1}^n \mathrm{H}_{0i}(s)$ is rejected at level α by Simes' test if there exists an i such that $p_{(i)} \leq \frac{i}{n}\alpha$. Equivalently, use the adjusted p-value $\min_{i=1,\ldots,n}\{\frac{n}{i}p_{(i)}(s)\}$, rejecting the intersection hypothesis if the adjusted p-value is smaller than α.

For testing the partial conjunction null $\mathrm{H}_0^{u/n}(s)$, we combine the $n-u+1$ largest p-values similarly, thus creating a restricted and shifted Simes p-value,

$$p^{u/n}(s) = \min_{i=1,\ldots n-u+1} \left\{ \frac{(n-u+1)}{i} p_{(u-1+i)}(s) \right\}. \tag{2}$$

For example, suppose that the test of three conditions ends up with p-values 0.5, 0.022, and 0.015. For testing that the alternative hypothesis holds for all three conditions we use $p^{3/3}(s) = p_{(3)}(s) = 0.5$, for at least two conditions we use $p^{2/3}(s) = \min\{2p_{(2)}(s), p_{(3)}(s)\} = 0.044$, and for at least one condition we use $p^{1/3}(s) = \min\{3p_{(1)}(s), 1.5p_{(2)}(s), p_{(3)}(s)\} = 0.033$.

The Simes test was originally developed for independent test statistics, where it is an exact test. Efforts over the last

years have extended its applicability. Sarkar (1998) was the first to show that the Simes test is valid under a specific dependency structure. It is now well established that the Simes test is valid under any of the following conditions. (D1) The p-values per location are independent (Simes, 1986). (D2) The p-values per location satisfy the positive regression dependency on a subset (PRDS) property, as defined in Benjamini and Yekutieli (2001): $P(\{P_i(s),\ i = 1,\ldots,n\} \in A \mid P_j(s) = x)$ is nondecreasing in x for any j in the subset of null hypotheses and any increasing set A, where set A is increasing if $x \in A$ and $y \geq x$ implies that $y \in A$. Important examples include comparison of various independent treatments with the same control and the set of p-values for testing one-sided hypotheses based on Gaussian test statistics that are positively correlated. (D3) The p-values per location for testing one-sided hypotheses are based on t-statistics from positively correlated normals with a joint estimator of the variability (Benjamini and Yekutieli, 2001; case 4).

THEOREM 1. *Let $p^{u/n}(s)$ be the pooled p-value using equation* (2). *If the set of p-values corresponding to null hypotheses at location s satisfy either of the conditions D1–D3 above, then $p^{u/n}(s)$ is a valid p-value for testing $H_0^{u/n}(s)$.*

See Web Appendix A for the proof.

For general dependence we may always revert to Bonferroni, leading to

$$p^{u/n}(s) = (n - u + 1)p_{(u)}(s). \tag{3}$$

Because $P(P^{u/n}(s) \leq q) = P(P_{(u)}(s) \leq \frac{q}{(n-u+1)}) \leq P(\cup_{i=1}^{n-u+1}\{U_i(s) \leq \frac{q}{(n-u+1)}\}) \leq q$, we have the following theorem:

THEOREM 2. *Let $p^{u/n}(s)$ be the pooled p-value using equation* (3). *Then $p^{u/n}(s)$ is a valid p-value for testing $H_0^{u/n}(s)$.*

2.2 *Combining Independent p-Values*

Let $z_{(1)}(s) \leq \cdots \leq z_{(n)}(s)$ be the sorted z-scores corresponding to the n p-values ($z_i(s) = \Phi^{-1}(1 - p_i(s))$). For the partial conjunction null $H_0^{u/n}(s)$, the p-value motivated by the Stouffer method for combining p-values is

$$p^{u/n}(s) = 1 - \Phi\left(\frac{\sum_{i=1}^{n-u+1} z_{(i)}(s)}{\sqrt{n-u+1}}\right) \tag{4}$$

and the p-value motivated by the Fisher method for combining p-values is

$$p^{u/n}(s) = P\left(\chi^2_{2(n-u+1)} \geq -2\sum_{i=u}^{n} \log p_{(i)}(s)\right). \tag{5}$$

These are valid partial conjunction p-values because they are both increasing functions of $p_1(s),\ldots,p_n(s)$, so Lemma 1 can be invoked, and the combining function f in each case satisfies $f(U_1,\ldots,U_{n-u+1},\ 0,\ldots,0) \sim U(0,\ 1)$.

Many other valid combining p-values can be generated. For a systematic comparison of combining methods for testing the global null and for further references see Loughin (2004). A similar modification of these combining methods can be used for more partial conjunction tests.

3. Screening While Controlling the FDR

Consider now the situation where we test a large family of partial conjunction hypotheses $H_0^{u/n}(s)$, $s = 1,\ldots,S$. Once we have a valid combined p-value per location utilizing one of equations (2)–(5) as appropriate, we can use an FDR controlling procedure on the combined location p-values. The question arises whether conditions on the individual p-value maps that permit the use of an FDR controlling procedure per map separately, endow the combined partial conjunction p-values with the condition that allows the use of the same FDR controlling procedure.

If the p-values within the individual maps are independent the answer is simple, any FDR controlling procedure for independent test statistics will obviously control the FDR at the desired level q. However, the independence assumption is often not met. For example, in fMRI a single null hypothesis tested is often one sided (did the stimulus increase the activity in the brain location?) and the p-values are based on (approximately) Gaussian test statistics that are nonnegatively correlated across neighboring brain locations. Such PRDS structure allows the use of the procedure in BH. Now, if several p-value maps are combined, within each map the location p-values satisfy the PRDS property and the n p-values in each location are independent, the following condition guarantees that the combined p-value map also satisfies the PRDS property:

Condition 1. If $f : \Re^n \to \Re$ is the combining function and $U_1,\ldots,U_{n-u+1}$ are $U(0,\ 1)$ random variables, then $f(U_1,\ldots,U_{n-u+1},0,\ldots,0) = G(\sum_{i=1}^{n-u+1} g(U_i))$, where $G(\cdot)$ and $g(\cdot)$ are increasing functions and the probability density of $g(U_i)$ is a Polya frequency function of order 2 (PF_2) (see Efron [1965] for details on these functions).

In particular, the combining functions motivated by Fisher's and Stouffer's methods for combining p-values satisfy the above conditions but the combining functions motivated by the Simes or Bonferroni methods do not. For the Fisher method: $f(U_1,\ldots,U_{n-u+1},0,\ldots,0) = P(\chi^2_{2(n-u+1)} \geq -2(\sum_{i=1}^{n-u+1} \log U_i))$, so $G(x) = P(\chi^2_{2(n-u+1)} \geq -2x)$ is increasing in x for $x \leq 0$ and $g(u) = \log u$ is increasing in u; $g(U_i) = \log U_i$ has an exponential distribution and therefore a PF_2 density. For the Stouffer method: $f(U_1,\ldots,U_{n-u+1},0,\ldots,0) = \Phi(\sum_{i=1}^{n-u+1}(-\Phi^{-1}(1-U_i))/(n-u+1)))^{\frac{1}{2}}$, so $G(x) = \Phi(x/(n-u+1))$ is increasing in x and $g(u) = -\Phi^{-1}(1-u)$ is increasing in u; $g(U_i) = -\Phi^{-1}(1-U_i)$ has a standard normal distribution and therefore a PF_2 density.

THEOREM 3. *Assume that the p-values within individual maps satisfy the PRDS property, and that the p-values in each location are independent. Furthermore if condition* 1 *is satisfied, the BH procedure on the partial conjunction p-value map controls the FDR at the desired level q.*

See Web Appendix B for the proof.

It follows that applying the procedure in BH after using equation (4) or (5) to combine the p-values in each location will control the FDR at the desired level q if within every map the p-values satisfy the PRDS assumption and the n p-values in each location are independent. Although it is quite

likely that BH screening after using equation (2) to combine the p-values at each location also controls the FDR, we do not have such a result. Simulations with PRDS dependency across locations in the same map and across maps in same locations, detailed in Web Appendix C, suggest that the BH procedure on the pooled map using equation (2) controls the FDR.

Previous works show that the BH procedure controls the FDR for p-value maps with many dependency structures other than PRDS. Reiner (2007) shows via a combination of simulations and analytic results that applying the BH procedure on p-values from two-sided tests of correlated normal test statistics with any correlation structure controls the FDR. In an asymptotic framework, Storey, Taylor, and Siegmund (2004) gave convergence conditions on the distribution of the p-values for the inference on a single map of dependent p-values using BH to be valid. We show in Web Appendix D the asymptotic validity of partial conjunction screening when these conditions are satisfied for every map. Storey et al. (2004) also suggested more powerful procedures than the BH for FDR control. The asymptotic validity of these procedures carries over to the partial conjunction p-value map. To summarize, we believe that in most practical situations where the BH procedure is appropriate when screening individual maps, it is also so for screening partial conjunction hypotheses. (If in doubt it can always be applied at a more conservative level of $q/(\sum_{j=1}^{S} \frac{1}{j})$ to guarantee FDR control at level q, see Benjamini and Yekutieli [2001]).

4. Screening at All Levels u

As partial conjunction $H_0^{u/n}$ is a flexible in-between approach for any $1 \leq u \leq n$, it may be tempting to look at all n such maps. In Figure 1, for example, we display three such maps superimposed, where at each location s the largest u for which $H_0^{u/n}(s)$ can be rejected is presented. In applications where it is interesting to create and examine all n maps (e.g., screening for conjunctions in a group of people, see Heller et al. [2007]), it is necessary to define an overall error measure to control for multiplicity. We will define the FDR for screening at all levels u and prove that it is controlled when the BH procedure is applied at every level u to the combined p-values based on Simes (2).

Define an overall location discovery at s if for some u $H_0^{u/n}(s)$ is rejected. Let $\tilde{k}(s) = \max\{u \mid H_0^{u/n}(s) \text{ is rejected}\}$ be the strongest overall result we can claim about s, and define the discovery at location s false if the claim is too strong, in that $\tilde{k}(s) > k(s)$. Then, the overall FDR is the expected proportion of the overall false discoveries out of the overall discoveries.

THEOREM 4. *If each partial conjunction map based on Simes* (2) *is PRDS, and is tested by the procedure in BH at level q, the overall FDR of the superimposed analysis is also less or equal to q.*

See the Appendix for the proof.

Actually the proof only makes use of the fact that the combining function satisfies the following monotonicity property $p^{(u-1)/n}(s) \leq p^{u/n}(s)$. To achieve control of the overall FDR for the combining methods (3)–(5), we can introduce the monotonicity requirement by defining $p_*^{1/n}(s) = p^{1/n}(s)$ and $p_*^{u/n}(s) = \max\{p_*^{(u-1)/n}(s), p^{u/n}(s)\}$ for $u = 2, \ldots, n$. Then, if the resulting maps are PRDS, using the procedure in BH on $\{p_*^{u/n}(s)\colon s = 1, \ldots, S\}$ for $u = 1, \ldots, n$ will also assure overall FDR control.

5. Application to fMRI

In this example, the subject viewed at different time points different pictures belonging to four categories: faces, houses, common man-made objects, and geometric patterns. The researcher is interested in finding the regions in the brain that were more active during the presentation of the first three picture categories than during the viewing of geometric patterns. For each of the first three categories, a t-test statistic was computed for testing the contrast that the brain activity during the presentation of pictures from that category was larger than during the presentation from the fourth category. Because the resulting three test statistics in each brain location are positively correlated, the combining method in equation (2) is used.

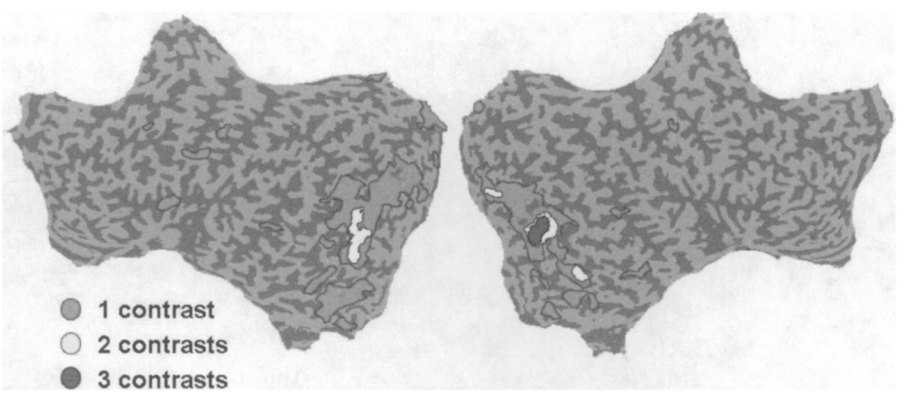

Figure 1. Activation maps for a single subject presented on unfolded cortical hemispheres: blue regions activated in all three contrasts with $FDR < 0.05$ (blue regions in color version at website); white or black (yellow, or blue) regions activated in at least two contrasts with $FDR < 0.05$; dark gray encircled with black, white, or black (red, yellow, or blue) regions activated in at least one contrast with $FDR < 0.05$.

Figure 1 shows the superimposed maps that passed the FDR cutoff of 0.05 for testing that at least one, at least two, or all three contrasts were greater than zero. From this figure, the regions that were found to react to all three contrasts at an FDR level of 0.05 are colored in blue; the regions that were found to react to at least two contrasts at an FDR level of 0.05, are colored in blue or yellow; and the regions that reacted to at least one contrast at an FDR level of 0.05 are colored in red, yellow, or blue. The partial conjunction analysis reveals the much wider region associated with a single contrast. However, when a conjunction of at least two categories are considered (the union of yellow and blue regions), then the delineated regions shrink and become confined to a well-studied cortical region, the object-related lateral occipital complex (LOC), whose most robust functional signature is a preferential activation to images of objects compared to texture patterns (Malach et al., 1995; Malach and Levy, 2002). See Heller et al. (2007) for more details on this example as well as for more examples.

Remark. On single p-value maps from neuroimaging data, Genovese, Lazar, and Nichols (2002) argue that the FDR procedure controls the FDR at level q because the correlations are local and tend to be positive. This reasoning carries over to the pooled p-value map, so the BH procedure is justified by the asymptotic argument in Section 3. Moreover, simulations in Web Appendix C show the FDR control for finite samples.

6. Application to Microarray Meta-Analysis

Microarray technology is used to measure simultaneously the expression of thousands of genes under various experimental conditions. Rapidly growing collections of large datasets are becoming available for subsequent analysis. Given the differences in characteristics of the raw datasets, combining the results can help identify the consistently true signals as well as give indications about possibly inconsistent findings.

Chromatin immunoprecipitation (ChIP) is a well-established procedure used to investigate interactions between proteins and DNA. Coupled with whole-genome DNA microarrays, ChIPs allow one to determine the entire spectrum of in vivo DNA binding sites for any given protein. Proteins called transcription factors (TFs) regulate transcription by binding to DNA motifs upstream of their target genes. The availability of the genome sequence for budding yeast allowed ChIP to be coupled to high throughput analysis on microarrays ("chips") to monitor and measure the binding of a given set of TFs to the upstream regulatory regions of thousands of genes. We applied our combining methods to three well-known ChIP–chip genomewide TF binding datasets (see details in Pyne, Futcher, and Skiena [2006]). Pyne et al. (2006) combined these datasets by first applying a cutoff value for each p-value map with a conservative FDR threshold so that only p-values that were below their FDR threshold are combined using the truncated Fisher method (adjusted as suggested by Zaykin et al. [2002]), then the combined map cutoff is chosen with an FDR controlling procedure. Pyne et al. (2006) added a calculation for finding the genes where at least two or all three datasets cleared their cutoffs under the global null hypothesis. So in fact their definition of a discovery in at least two or all three datasets is different from ours. Moreover, the p-values in the combined map are calculated under the assumption that the map thresholds are fixed even though the thresholds are data dependent, so the control of the FDR is not guaranteed. We apply the Fisher- and Stouffer-based methods for combining the p-values, and then threshold the combined p-value maps with an FDR level of 0.05. We adjusted for missing values conservatively by marking their p-values as 1. In Table 1 we compared our method with the naive method of cutting off every dataset with its own nominal FDR level of 0.05, and with the results in Pyne et al. (2006). We discover more than Pyne et al. (2006), suggesting our procedure is more powerful. The naive method makes more discoveries for $u = 3$, but significantly less discoveries with $u < 3$ because it does not gain power from pooling together information from several sources. Of course, because it does not guarantee control of FDR, the naive method is not recommended. Note that for global testing, the naive method FDR is bounded above by $3q$, so a simple solution is to threshold each map at the $q/3$ level. However, if every map is threshold at 0.05/3, only 118 rejections of the global null are made.

Table 1
Number of significant genes for protein Swi4 (that forms part of TF SBF)

	All 3	At least 2	At least 1
Pyne et al. (2006)	64	103	162
Stouffer method	73	195	321
Fisher method	73	176	305
Naive method	78	121	161

The finding that the gene was differentially expressed in at least one dataset may be too weak scientifically, and the requirement that the gene should be significant in all three datasets may be too severe, so it ignores interesting gene discoveries. Therefore, the genes that were found to be differentially expressed in at least two datasets may be the most interesting to look at.

7. A Simulation Example

We considered different settings in order to compare the power of the suggested methods of pooling p-values, as well as examine how the choice of u affects the power. In each of 1000 locations 10 independent unit variance Gaussian noise measurements were simulated, and in 100 locations a signal of size μ was added in k out of the 10 repetitions ($k = 3, 7, 9$) per location. The signal size μ was independently sampled for each location and map from a $N(\mu_0, \sigma_0^2)$ distribution, where we varied $\mu_0 = 2, \ldots, 6$ and $\sigma_0 = 0, \ldots, \max(2, \mu_0/2)$.

We pooled the p-values using equations (2), (4), or (5), as well as using the maximum p-value method (see Friston et al. [2005] for details) because this is the only method used up to now for $1 < u < n$. Next, we computed the resulting map threshold using the suggested BH procedure.

The simulation results show that none of the pooling methods dominate. The power of each method depends both on the configuration of signal μ and on the proportion k/n of false hypotheses. A careful examination of the identifiable factors that affect the choice between the combining methods in terms of power are outside the scope of this manuscript (see Loughin [2004] for some insight when $u = 1$). Our key

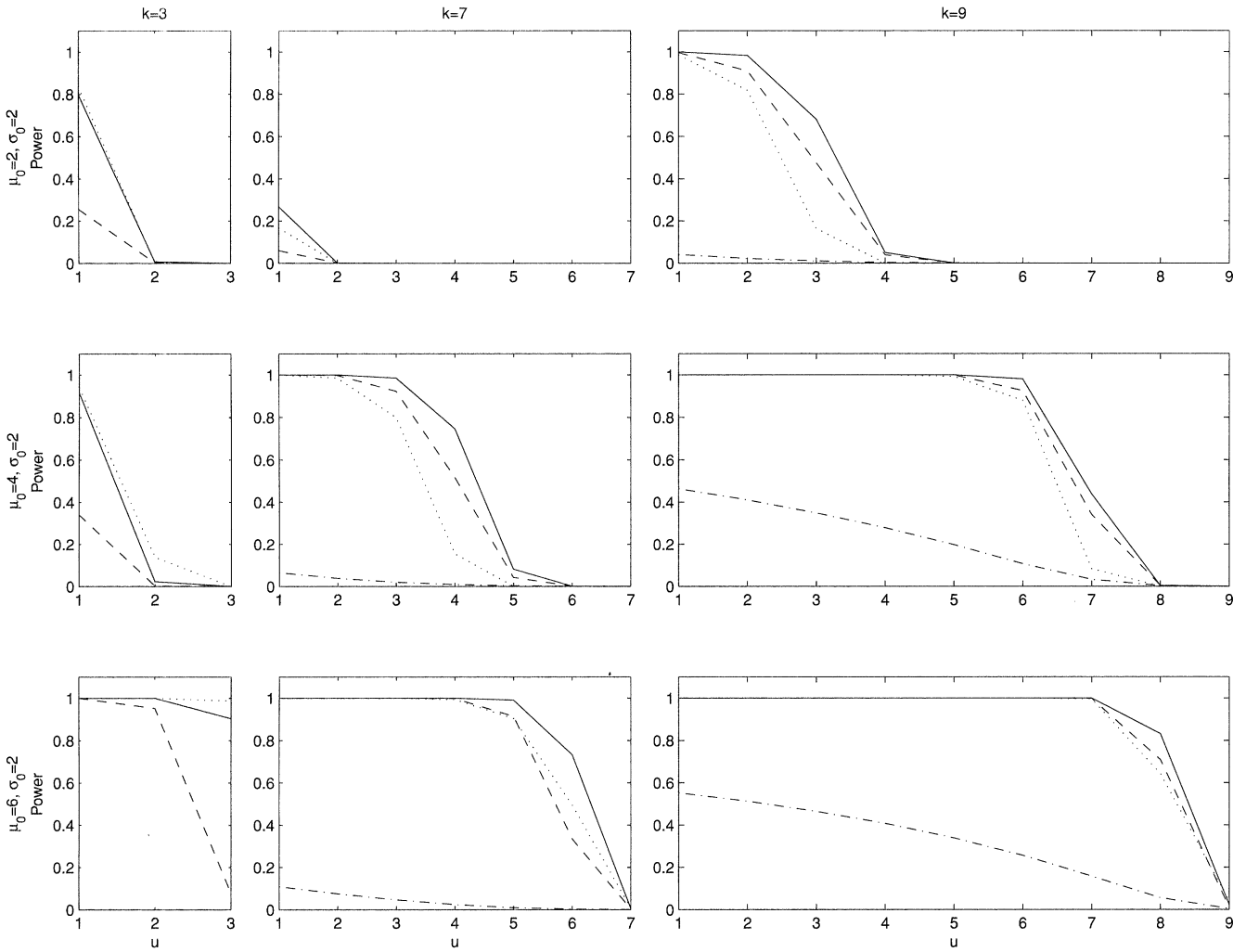

Figure 2. Power as a function of u when the FDR level is 0.05 and the simulated setting is that in which the number of p-values per location that come from the alternative is either null or 3 (first column), 7 (second column), or 9 (third column). The combining method is based on (a) equation (5) (solid line) (b) equation (4) (dashed line) (c) equation (2) (dotted line), and (d) the maximum p-value method (dash–dot line). Each row corresponds to a different $\mu_0(\sigma_0 = 2)$: $\mu_0 = 2$ (top), $\mu_0 = 4$ (middle), and $\mu_0 = 6$ (bottom). There is not one pooling method that is more powerful than all others in all simulation settings; there is a sharp decrease in power when u increases.

observations are that (1) the maximum p-value method has little power to reject the partial conjunction null whenever at least one p-value comes from a null hypothesis, regardless of how small the other p-values may be, and (2) the power decreases sharply when u increases, supporting our motivation for the partial conjunction test with $u < n$ rather than testing of the disjunction of nulls ($u = n$) when screening for many hypotheses. In the representative Figure 2 we see that when the partial conjunction hypothesis is false, if most p-values come from the alternative (e.g., $k = 7$ or $k = 9$) then pooling the p-values using equations (4) or (5) is usually more powerful than using equation (2), but when the number of p-values that come from the alternative is small (e.g., $k = 3$) pooling the p-values using equation (2) may be more powerful even under independence between p-values within each location.

8. Discussion

In this article we have suggested powerful new methods to combine both independent and dependent p-value maps for testing partial conjunctions. We showed that the power decreases as a function of the conjunction parameter u, and discussed the advantages of choosing a u to be larger than 1 but smaller than n. If screening at all levels of u is of interest, we suggested superimposing the maps to discern the results and we showed that if the procedure in BH was used for each partial conjunction at level q then for the superimposed map the expected proportion of false overall discoveries is bounded at the same level q. The result is restricted to the above procedure, but we suspect the procedure can be generalized while maintaining control of the overall FDR and advances on this front are desirable. Several other extension are discussed below.

In the cases considered, the identity of the null hypotheses rejected in every location is not important, and only the proportion of null hypotheses rejected is of interest. If the identity of the rejected null hypotheses is of interest as well, stepwise procedures can be applied in every location (e.g., in Tamhane and Dunnett, 1999) to discover whether at least u out of n null hypotheses are rejected and in addition identify these u hypotheses, but the level of testing needs to be adjusted so that the FDR on locations is properly defined and controlled.

For combining a large number of maps n sampled from a population, for example, when each map corresponds to a subject, it may be interesting to estimate rather than test the proportion of nonnull hypotheses per location, say by getting a lower confidence bound on this proportion (Friston, Holmes, and Worsley [1999] and Friston, Holmes, Price, Buchel, and Worsley [1999] address this issue in fMRI). This is an interesting point for further research, in particular using the approach of confidence intervals after selection suggested in Benjamini and Yekutieli (2005).

9. Supplementary Materials

Web Appendices and Figures referenced in Sections 2.1–7 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

We wish to thank Yulia Golland and Rafael Malach for supplying the fMRI data and for valuable comments on the fMRI example, and Yosef Rinott for referring us to Efron (1965). We also thank the referees for their suggestions for improving the manuscript.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate—A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Metallurgy)* **57,** 289–300.

Benjamini, Y. and Yekutieli, Y. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29,** 1165–1188.

Benjamini, Y. and Yekutieli, Y. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association* **100,** 71–93.

Efron, B. (1965). Increasing properties of polya frequency functions. *Annals of Mathematical Statistics* **36,** 272–279.

Friston, K., Holmes, A., and Worsley, K. (1999). Comments and controversies: How many subjects constitute a study? *NeuroImage* **10,** 1–5.

Friston, K., Holmes, A., Price, C., Buchel, C., and Worsley, K. (1999). Multisubject fMRI studies and conjunction analyses. *NeuroImage* **10,** 385–396.

Friston, K., Penny, W., and Glaser, D. (2005). Conjunction revisited. *NeuroImage* **25,** 661–667.

Genovese, C., Lazar, N., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* **15,** 870–878.

Heller, R., Golland, Y., Malach, R., and Benjamini, Y. (2007). Conjunction group analysis: An alternative to mixed/random effect analysis. *NeuroImage* **37,** 1178–1185.

Lazar, N., Luna, B., Sweeney, J., and Eddy, W. (2002). Combining brains: A survey of methods for statistical pooling of information. *NeuroImage* **16,** 538–550.

Loughin, T. (2004). A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics and Data Analysis* **47,** 467–485.

Malach, R. and Levy, I. (2002). The topography of high-order human object areas. *Trends in Cognitive Sciences* **6,** 176–184.

Malach, R., Reppas, J., Benson, R., Kwong, K., Jiang, H., Kennedy, W., Ledden, P., Brady, T., Rosen, B., and Tootell, R. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences of the United States of America* **92,** 8135–8139.

McNamee, R. and Lazar, N. (2004). Assessing the sensitivity of fmri group maps. *NeuroImage* **22,** 920–931.

Nichols, T., Brett, M., Anderson, J., Wager, T., and Poline, J. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage* **25,** 653–660.

Pyne, S., Futcher, B., and Skiena, S. (2006). Meta-analysis based on control of false discovery rate: Combining yeast ChIP-chip datasets. *Bioinformatics* **22,** 2516–2522.

Reiner, A. (2007). FDR control by the bh procedure for two-sided correlated tests with implications to gene expression data analysis. *Biometrical Journal* **49,** 107–126.

Sarkar, S. (1998). Some probability inequalities for ordered mtp_2 random variables: A proof of the Simes conjecture. *The Annals of Statistics* **26,** 494–504.

Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73,** 751–754.

Storey, J., Taylor, J., and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B* **66,** 187–205.

Tamhane, A. and Dunnett, C. (1999). Stepwise multiple test procedures with biometric applications. *Statistical Planning and Inference* **82,** 55–68.

Zaykin, D., Zhivotovsky, L., Westfall, P., and Weir, B. (2002). Truncated product method for combining p-values. *Genetic Epidemiology* **22,** 170–185.

Received April 2007. *Revised November* 2007. *Accepted November* 2007.

Appendix

Proof of Theorem 4. Let m_i be the number of locations s for which exactly i hypotheses are false, namely $k(s) = i$, and let I_i be their corresponding location indices for $i = 0, \ldots, n$. Then $\sum_{i=0}^{n} m_i = S$ is the total number of locations and $\cup_{i=1}^{n} I_i = \{1, \ldots, S\}$.

Let R_u be the number of hypotheses rejected when testing $\text{H}_0^{u/n}(s)$ on all S locations using the procedure in BH for FDR control at level q. Note that $R_1 \geq R_2 \geq \cdots \geq R_n$ if $p^{n/n}(s) \geq \cdots \geq p^{u/n}(s) \geq \cdots \geq p^{1/n}(s)$. To see this, note that because $\sum_{i=1}^{m} 1[p_i^{u/n} \leq R_u q/m] = R_u$ we have $\sum_{i=1}^{m} 1[p_i^{(u-1)/n} \leq R_u q/m] \geq R_u$. But

because $R_{u-1} = \max\{k : \sum_{i=1}^{m} 1[p_i^{(u-1)/n} \leq kq/m] \geq k\}$ then $R_{u-1} \geq R_u$. Moreover, the number of false discoveries is $\sum_{j=1}^{n} \sum_{s \in I_{j-1}} 1(P^{j/n}(s) \leq R_j \frac{q}{m})$. For example, if $s \in I_1$ then if s is rejected when testing $u = 3$ it is also rejected when testing $u = 2$ because $p^{2/n}(s) \leq p^{3/n}(s)$ and $R_3 \leq R_2$ so we should count it only once as an error (at the level $u = 2$). Moreover, because $\mathrm{H}_0^{1/n}(s)$ was also rejected the number of overall discoveries is R_1. Therefore,

$$
\begin{aligned}
&E\left(\frac{\sum_{j=1}^{n}\sum_{s\in I_j} 1\left(P^{j/n}(s) \leq R_j \frac{q}{m}\right)}{R_1}\right) \\
&= \sum_{j=1}^{n}\sum_{s\in I_{j-1}} E\left(\frac{1\left(P^{j/n}(s) \leq R_j \frac{q}{m}\right)}{R_1}\right) \\
&\leq \sum_{j=1}^{n}\sum_{s\in I_{j-1}} E\left(\frac{1\left(P^{j/n}(s) \leq R_j \frac{q}{m}\right)}{R_j}\right) \\
&\leq \sum_{j=1}^{n}\sum_{s\in I_{j-1}}\sum_{k=1}^{S} \frac{1}{k} P\left(P^{j/n}(s) \leq kq/m \cap C_{j,k}^{(s)}\right) \\
&\leq \sum_{j=1}^{n}\sum_{s\in I_{j-1}} \frac{q}{m} \leq \sum_{j=1}^{n} \frac{m_{j-1}}{m} q = \left(1 - \frac{m_n}{m}\right) q \leq q,
\end{aligned}
\tag{A.1}
$$

where $C_{j,k}^{(s)}$ is the set that when screening for partial conjunctions with parameter $u = j$ exactly k p-values are rejected including the s from the partial conjunction null, so the third inequality follows if the partial conjunction p-values are PRDS (or independent) according to the proof in Benjamini and Yekutieli (2001).