Screening for Partial Conjunction Hypotheses

Yoav Benjamini*

Department of Statistics and Operations Research

Tel Aviv University, Tel Aviv 69978, Israel

and

Ruth Heller†

April 30, 2007

SUMMARY. We consider the problem of testing the partial conjunction null, that asks whether less than u out of n null hypotheses are false. It offers an in-between approach to the testing of the global null that all n hypotheses are null, and the full conjunction null that not all of the n hypotheses are false. We address the problem of testing many partial conjunction hypotheses simultaneously, a problem that arises when combining maps of p-values. We suggest powerful test statistics that are valid under dependence between the test statistics as well as under independence. We suggest controlling the false discovery rate (FDR) on the p-values for testing the partial conjunction hypotheses, and we prove that the FDR controlling procedure in (Benjamini and Hochberg (1995)) remains valid under various dependency structures. We apply the method to examples from Microarray analysis and functional

* *email:* ybenja@post.tau.ac.il
† *email:* rheller@post.tau.ac.il

Magnetic Resonance Imaging (fMRI), two application areas where the need for partial conjunction analysis has been idnetified.

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

1. Introduction

In many modern biostatistics applications there is need to combine p-value maps. One 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. Another example from functional magnetic resonance imaging (fMRI) research, is that of looking for the brain regions that participated in most (or at least one) of several cognitive tasks. The maps are independent in the first example, but may be dependent in the second example.

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. While 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 test the ""global null hypothesis", where the alternative is that at least one null hypothesis is false, Fisher's combined p-value being probably the best-known method for this purpose (see e.g. Loughin (2004), Zaykin et al. (2002) and Lazar et al. (2002)).

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 "conjunction null hypothesis" that not all null hypotheses are false, with the alternative conjunction hypothesis that all are false. The need to answer such questions has arisen quite naturally in fMRI analysis (see Friston et al. (1999) and Nichols et al. (2005)), where the difficulty is compounded in the sense that the conjunction null is tested in many locations.

As noted above the global null findings are often too general to be scientifically meaningful. But the conjunction null is often too restrictive, making it practically very difficult to reject when screening a large number of conjunction nulls. A natural compromise is to relax the conjunction null and strengthen the global null by asking instead whether no more that a given number of the null hypotheses hold. In other words, 'can at least u out of my n null hypotheses be rejected?'

Such a test, an in-between approach to the testing of the (full) conjunction null and the global null, is called the partial conjunction test. Formally, consider $n \geq 2$ null hypotheses at each "location" $s \in \{1, \ldots, S\}$, $H_{01}(s), H_{02}(s), \ldots, H_{0n}(s)$, and let $p_1(s), \ldots, p_n(s)$ be their associated p-values in location s. Let $k(s)$ be the (unknown) number of false null hypotheses in location s, then our question 'Can at least u out of n hypotheses be false nulls?' can be formulated as follows:

$$H_0^{u/n}(s) : k(s) < u \quad \text{versus } H_1^{u/n}(s) : k(s) \geq u \tag{1}$$

Friston et al. (2005) have recognized the usefulness of testing $H_0^{u/n}(s)$ 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 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 family-wise 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 6. 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 $H_0^{u/n}(s)$. 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 testing conjunction hypothesis (where $u = n$) and lead to familiar tests for the global null (where $u = 1$.)

We then suggest to screen these valid p-values across locations while controlling for multiplicity. It can be done by controlling the FWE, but we prefer using False Discovery Rate (FDR) control that is commonly used in large multiplicity problems such as microarray and fMRI analyses. In section 3 we prove that the BH procedure (Benjamini and Hochberg (1995)) 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.

This is in contrast to the most immediate procedure of combining maps of p-values, where one first threshold each map separately to control the FDR at level q, and then take their intersection. In the extreme situation where the conjunction of threshold maps is that of the falsely discovered locations, the FDR of such a procedure will be 1.

In sections 4 and 5 we give examples from fMRI and Microarray analysis respectively. In section 6 we discuss the power of the methodology suggested via simulations. In section 7 we give our final remarks.

2. Combining p-values

Many methods for combining p-values, $p_s^{u/n} = f(p_1(s), \ldots, p_n(s))$ can be designed. Under the partial conjunction null $H_0^{u/n}(s)$, let $U_1(s), \ldots, U_{n-u+1}(s)$ be the p-values for which the null hypotheses hold, in the sense that $U_i(s) \succeq_{\text{st}} U(0,1)$ 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 $p_s^{u/n} = f(u_1, \ldots, u_{n-u+1}, p_1, \ldots, p_{u-1})$ be the combined p-value. The following lemma tells us that as long as the combining method makes sense, in that f is nondecreasing in all its components, then the stochastically smallest $p_s^{u/n}$ under $H_0^{u/n}(s)$ will occur when $u-1$ p-values are identically zero.

LEMMA 1. *Under $H_0^{u/n}(s)$, let $h_i(P_i) \leq P_i$ for some function $h_i(\cdot)$ and $i = 1, \ldots, u-1$ and let $P_s^{u/n*} = f(U_1, \ldots, U_{n-u+1}, h_1(P_1), \ldots, h_{u-1}(P_{u-1}))$ and $P_s^{u/n} = f(U_1, \ldots, U_{n-u+1}, P_1, \ldots, P_{u-1})$. Then $P_s^{u/n*} \prec_{st} P_s^{u/n}$.*

Proof. Since f is nondecreasing in all its components

$f(U_1,\ldots,U_{n-u+1},h_1(P_1),\ldots,h_{u-1}(P_{u-1})) \leq f(U_1,\ldots,U_{n-u+1},P_1,\ldots,P_{u-1})$. Therefore, if the event $\{P_s^{u/n} \leq q\}$ occurs then the event $\{P_s^{u/n*} \leq q\}$ occurs and the result follows.

Lemma 1 helps us construct valid pooled p-values. The pooled value $p_s^{u/n}$ will be valid if it depends only on the $n-u+1$ largest p-values using a combining function that satisfies $f(U_1,\ldots,U_{n-u+1},0,\ldots,0) \underset{\text{st}}{\succsim} U(0,1)$. Below we give several valid p-values.

2.1 *Combining p-values under dependence*

Let us recall Simes' test for the intersection of hypotheses $\cap_{i=1}^n H_{0i}(s)$. Given $p_i(s)$ the p-value for testing $H_{0i}(s)$, and the sorted values being $p_{(1)}(s) \leq p_{(2)}(s) \leq \ldots \leq p_{(n)}(s)$, the intersection hypothesis is rejected at level α if there exists an i s.t. $p_{(i)} \leq \frac{i}{n}\alpha$. Equivalently, the Simes test can be conducted by using the adjusted p-value $\min\{np_{(1)}(s), \frac{n}{2}p_{(2)}(s),\ldots,\frac{n}{n-1}p_{(n-1)}(s),p_{(n)}(s)\}$, rejecting the intersection hypothesis if the adjusted p-value is smaller than α.

For testing the partial conjunction null $H_0^{u/n}(s)$, we combine the $n-u+1$ largest p-values similarly, thus creating a restricted and shifted Simes p-value,

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

For example, suppose that the test of 3 conditions end up with p-values $0.5, 0.022, 0.01$. For testing that for all three conditions the alternative hypothesis holds we use $p_s^{3/3} = p_s^{(3)} = 0.5$, for testing that for at least one condition the alternative holds we use $p_s^{1/3} = \min\{3p_{(1)}(s), 1.5p_{(2)}(s), p_{(3)}(s)\} = 0.03$ and for testing that for at least two conditions the alternative holds we

use $p_s^{2/3} = \min\{2p_{(2)}(s), p_{(3)}(s)\} = 0.044$.

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 was the first to show that the Simes test is valid under a specific dependency structure in Sarkar (1998). It is now well established that the Simes test is valid under any of the conditions below:

1. The p-values per location are independent, see theorem in Simes (1986) and in Benjamini and Hochberg (1995).

2. The set of p-values per location satisfy the positive regression dependency on a subset (PRDS) property, as defined in Benjamini and Yekutieli (2001): $P(p_i(s) \in A, i = 1, \ldots, n | p_j(s) = x)$ is non-decreasing in x for any increasing set A and any $p_j(s) \in I_0(s)$, where $I_0(s)$ is the subset of null p-values and $q_i, i = 1, \ldots, n$ are arbitrary constants. Important examples include comparison of various independent treatments with the same control; the set of p-values for testing one-sided hypotheses based on Gaussian test statistics that are positively correlated; and the set of p-values based on t-statistics with a joint estimator of the variability $X_1/S, \ldots, X_n/S$, under the additional assumption that $\{|X_1|, \ldots, |X_n|\}$ satisfy the PRDS property (corollary 3.3 in Benjamini and Yekutieli (2001)). The validity of the Simes test follows from theorem 1.2 in Benjamini and Yekutieli (2001), for the special case that all n p-values come from the null hypotheses.

3. The p-values per location for testing one-sided hypotheses based on t-statistics using positively correlated normals with a joint estimator of

the variability, see case 4 in Benjamini and Yekutieli (2001).

Under these fairly general assumptions, the restricted and shifted Simes p-value can be used:

THEOREM 1. *Let $p_s^{u/n}$ be the pooled p-value using equation (2). If the set of null p-values at location s satisfy either of the conditions 1-3 above, then $p_s^{u/n}$ is a valid p-value for testing $H_0^{u/n}(s)$.*

See appendix A for a proof.

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

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

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

Proof.

$$P(p_s^{u/n} \leq q) = P((n-u+1)p_{(u)}(s) \leq q) \leq P(U_i \leq \frac{q}{n-u+1}, i = 1, \ldots, n-u+1) \leq q.$$

2.2 *Combining independent p-values*

Let $z_{(1)}(s) \leq \ldots \leq z_{(n)}(s)$ 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

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

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

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

These are valid partial conjunction p-values since they are both increasing functions of $p_1(s), \ldots, p_n(s)$ so lemma 1 tells us that the stochastically smallest distribution of $p_s^{u/n}$ occurs when $u-1$ p-values are zero and the remaining $n-u+1$ p-values are $U(0,1)$ random variables.

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 to test the partial conjunction hypothesis.

3. Screening while controlling the FDR

Consider now the situation where we test a large family of partial conjunction hypotheses $H_0^{u/n}(s), s = 1, \ldots, S$. Our approach has two natural components (a) Construct a valid pooled p-value per location using one of equations (2)-(5) as appropriate, and (b)use an FDR controlling procedure on the pooled location p-values. If the p-values within the individual maps are independent and the pooled location p-values are valid, any FDR controlling procedures will obviously control the FDR at the desired level q .

However, the independence assumption is often unrealistic. For example, in fMRI the measured signal of neighboring brain locations are typically positively correlated. Theorem 1.2 of Benjamini and Yekutieli (2001) states that the BH procedure for controlling the FDR is valid when the p-values satisfy the PRDS property. As discussed in section 2.1 this is a fairly general dependency structure that includes special cases that are commonly encountered. 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 non-negatively

correlated. If several p-value maps are combined, and within each map the location p-values satisfy the PRDS property, the following condition guarantee that the combined p-value map also satisfies the PRDS property if the n p-values in each location are independent:

Condition 3.1. For the extreme case under the partial conjunction null $H_0^{u/n}(s)$ that $u-1$ p-values are 0, and the remaining $n-u+1$ p-values are $U(0,1)$ random variables, the combining function $f : \Re^n \to \Re$ can be written as $f(U_1, \ldots, U_{n-u+1}, 0, \ldots, 0) = G(\sum_{i=1}^{n-u+1} g(U_i))$, 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).

THEOREM 3. *Assume the p-values within individual maps satisfy the PRDS property, and that the p-values in each location are independent. Then if furthermore condition 3.1 is satisfied the pooled p-value map also satisfies the PRDS property.*

See appendix B for a proof.

In particular, the combining functions motivated Fisher's and Stouffer's methods for combining p-values satisfy the above conditions. For the Fisher method: $f(U_1, \ldots, U_{n-u+1}, 0, \ldots, 0) = 1 - P(\chi^2_{2(n-u+1)} \geq -2\sum_{i=1}^{n-u+1} \log U_i)$, so $G(x) = 1 - P(\chi^2_{2(n-u+1)} \geq -2x)$ 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: $f(U_1, \ldots, U_{n-u+1}, 0, \ldots, 0) = \Phi(\sum_{i=1}^{n-u+1}(-\Phi^{-1}(1-U_i))/(n-u+1))$, 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.

As a result of Theorem 3 and the above discussion, if within every map the p-values satisfy the PRDS assumption and the n p-values in each location are independent, applying the BH procedure after using equation (4) or (5) to combine the p-values in each location will control the FDR at the desired level q. This follows from the validity of the BH procedure for p-values satisfying the PRDS property, as stated in theorem 1.2 in Benjamini and Yekutieli (2001).

While 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. Still, if the p-values within the individual maps have local dependencies, then the dependencies between the p-values within the combined map remain local. In this case both the BH and other FDR controlling procedures continue to control the FDR at level q asymptotically for the partial conjunction hypotheses tests, , when combined using equations (2)-(5) as appropriate. These methods are valid under the following asymptotic conditions on every map i, $i = 1, \ldots, n$:

$$\lim \frac{S_{0i}}{S} = A_{0i} \text{ Exists and } A_{0i} < 1 \tag{6}$$

$$F_{Si} = \frac{1}{S}\sum_{s=1}^{S} 1[p_i(s) < t | H_{0i}(s)] \overset{a.s.}{\to} A_{0i}F_i(t), \quad F_i(t) \leq t \quad \forall t \in (0,1] \tag{7}$$

$$G_{Si} = \frac{1}{S}\sum_{s=1}^{S} 1[p_i(s) < t | H_{1i}(s)] \overset{a.s.}{\to} (1 - A_{0i})G_i(t) \quad \forall t \in (0,1] \tag{8}$$

where S_{0i} is the number of null locations in map i. These conditions are satisfied, for example, if we have m-dependence between the locations, the data in each map is strictly stationary, and in one of the following two asymptotic settings: (1) increasing domain asymptotics, where the distance between

locations remains fixed but the domain goes to infinity, and (2) infill asymptotics, where the domain remains fixed but the number of points increases to infinity.

The threshold in the BH procedure is

$$t_s^* = \sup\{t : \frac{t}{F_s(t) + G_s(t)} \leq q\},$$

where $F_S = \frac{1}{S}\sum_{s=1}^{S} 1[p_s^{u/n} < t | H_{0s}^{u/n}]$ and $G_S = \frac{1}{S}\sum_{s=1}^{S} 1[p_s^{u/n} < t | H_{1s}^{u/n}]$. It controls the FDR asymptotically at level q (as $S \to \infty$) for any valid pooled p-value (i.e. not only using equation 5 or 4, but also using 2 when valid or 3) if conditions (6)-(8) hold, and $\delta \equiv \sup\{t : t/\lim(F_s(t) + G_s(t)) \leq q\} \in (0, 1]$:

$$\begin{aligned} FDR &= E(\frac{F_s(t^* s)}{F_s(t_s^*) + G_s(t_s^*)}) = E(\frac{t_s^*}{F_s(t_s^*) + G_s(t_s^*)} + \frac{(F_s(t_s^*) - t_s^*)}{F_s(t_s^*) + G_s(t_s^*)}) \\ &\leq q + \sup_{t \geq \delta}\{\frac{(F_s(t) - t)}{F_s(t) + G_s(t)}\} + I\{t_s^* < \delta\} \end{aligned}$$

From equations (7)-(8) the second term is asymptotically negative (because these conditions guarantee that the variance of $F_s(t)$ is asymptotically zero, so $\lim F_s(t) \leq t$) , and from the definition of δ the third term is asymptotically zero. It follows that the asymptotic upper bound for the FDR is q. This result is due to Storey et al. (2004), where more powerful procedures for FDR control are also suggested. The asymptotic validity of these procedures carries over to the pooled p-value map.

4. Application to fMRI

In fMRI, the signal is recorded over time for a series of brain slices while the subject performs various cognitive tasks. Consider several visual stimuli: faces, houses, common man-made objects, and geometric patterns. The researcher is interested in finding the regions that were more active during

the first three visual stimuli than during the viewing of geometric patterns. Since the three contrasts (i.e. faces minus patterns, houses minus patterns and objects minus patterns) are positively correlated, the combining method in equation 2 is used.

Figure B shows the superimposed maps that passed the FDR cut-off of 0.05 for testing that at least one, at least two or all three contrasts were activated. From this figure, the regions that were found to react to all three contrasts at an FDR significance 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 distribution associated with a single contrast - which includes areas whose selectivity is unique to a single object category, such as the FFA (e.g. Kanwisher et al. (1997)), the PPA (e.g. Epstein and Kanwisher (1998)) and other object-related regions (for review see e.g. Hasson et al. (003b), Malach and Levy (2002)). However, when a conjunction of at least two categories are considered (the union of yellow and blue regions), or of all three categories (the blue region) - 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)).

[Figure 1 about here.]

Remark. On single p-value maps from nueroimaging data, Genovese et al. (2002) argue that the FDR procedure controls the FDR at level q since 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.

5. 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 genome-wide TF binding datasets (see details in Pyne et al. (2006)). Pyne et al. (2006) combined these datasets by first applying a cut-off 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 cut-off 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 cut-offs 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 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 full conjunction discoveries, but significantly less partial conjunction discoveries since it does not gain power from pooling together information from several sources. Of course, since 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 about here.]

The finding that the gene was active 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, so in this case the genes that were found to be active in at least 2 datasets may be the most interesting to look at.

6. 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 addition in 100 locations an added signal of size μ ($\mu = 1, \ldots, 6$) was added in k out of the 10 repetitions ($k = 3, 7, 9$) per location.

We pooled the p-values using 5, 4 or 2, then computed the resulting map threshold using the suggested BH procedure.

The simulations results in Figure 2 show that their is not one pooling method that dominates all others. When the partial conjunction hypothesis is false, if most p-values come from the alternative (e.g. $k = 7$ or $k = 9$) than pooling the p-values using equations 5 or 4 is usually more powerful than 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. A more 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. Note the sharp decrease in power when u increases, supporting our motivation for using the partial conjunction test rather than the full conjunction test when screening for many hypotheses.

[Figure 2 about here.]

7. Discussion

The need to combine p-value maps arises in many modern applications, such as genomics and fMRI. In this article we have suggested powerful new methods to combine both independent and dependent p-value maps. A couple of possible extension are discussed below.

In the cases considered, the identity of the null hypotheses rejected in every location is less important, and only the proportion of null hypotheses rejected is of interest. However, in some cases 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 additional identify these u hypotheses, but the level of testing needs to be adjusted so that the FDR on locations is properly defined and controlled.

If n p-values are combined, n combined p-value maps can be created depending on which of the n different partial conjunction null hypotheses $H_0^{u/n}(s),\ u = 1, \ldots, n$ are tested. In some cases, as in the fMRI example, it is interesting to create and examine all n maps. Define an overall location discovery as a discovery if the minimum u of interest tested is rejected (e.g. rejection of $H_0^{1/3}(s)$ in the fMRI example), a false overall location discovery as a discovery where the true unknown number of false null hypotheses $k(s)$ is greater than u (i.e. reject $H_0^{u/n}(s)$ even though $k(s) < u$ for at least one value of u), and the overall FDR as the expected proportion of the overall false discoveries out of the overall discoveries. Then the overall FDR may be larger than q. It is straightforward to show that an upper bound on the overall FDR is nq, so in order to control the overall FDR at level q we can test

each partial conjunction at an FDR level of q/n, but this is not a powerful procedure. How to design more powerful procedures is a point for further research.

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).

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.* **57 (1)**, 289–300.

Benjamini, Y. and Yekutieli, Y. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29 (4)**, 1165–1188.

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

Epstein, R. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature* **392**, 598–601.

Friston, K., Holmes, A., Price, C., Buckel, C. and Worsley, K. (1999). Multisubject fmri studies and conjunction analyses. *NeuroImage* **10 (4)**, 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.

Hasson, U., Harel, M., Levy, I. and Malach, R. (2003b). Large-scale mirror-symmetry organization of human occipito- temporal object areas. *Neuron* **37**, 1027–1041.

Kanwisher, N., McDermott, J. and Chun, M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for the perception of faces. *Neuroscience* **17**, 4302–4311.

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(4)**, 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. *Proc Natl Acad Sci U S A* **92**, 8135–8139.

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

Nichols, T., Brett, M., J., A., Wager, T. and J., P. (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.

Sarkar, S. (1998). Some probability inequalities for ordered mtp_2 random variables: A proof of the simes conjecture. *The Annals of Statistics* **26 (2)**, 494 – 504.

Simes, R. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika* **73 (3)**, 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.

Appendix A

Proof of theorem 1

We have to show that $P_0(P_s^{u/n} \leq q) \leq q$ where the subscript 0 indicates that the probability is calculated under the partial conjunction null hypothesis that at most $u-1$ come from the alternative.

Since $p_s^{u/n}$ in equation (2) is an increasing function of the p-values, lemma 1 tells us that the stochastically smallest distribution of $p_s^{u/n}$ occurs when

$u-1$ p-values are zero (i.e. $h_i(P_i) = 0$ for $i = 1, \ldots, u-1$) and the remaining $n-u+1$ p-values are $U(0,1)$ random variables. So it is enough to show for this case that $P_s^{u/n} \underset{\text{st}}{\succ} U(0,1)$.

Let $U_{(1)} \leq \ldots \leq U_{(n-u+1)}$ be the order statistics of $n-u+1$ $U(0,1)$ random variables.

$$P_0(P_s^{u/n} \leq q) \leq P(\min_{i=1,\ldots,n-u+1} \{\frac{(n-u+1)}{i} U_{(i)}\} \leq q) \leq q$$

where the last inequality follows since the Simes test based on the $n-u+1$ null p-values is valid under any of the conditions 1-3 on these p-values.

Appendix B

Proof of theorem 3

The individual p-value maps are PRDS, so for every map $j = 1, \ldots, n$ with vector of p-values $\tilde{p}^j = (p_1^j, \ldots, p_S^j)$ and a fixed arbitrary vector of constants $\tilde{q}^j$ we can say the following: $P(\tilde{p}^j \in A | p_{sj} = q)$ for any increasing set A is non-decreasing in q for all $s \in I_0^j$, where I_0^j is the subset of null locations.

Let the combined p-value in location s be $p_s^{u/n} = f(p_1(s), \ldots, p_n(s))$ and $\tilde{p}^{u/n} = (p_1^{u/n}, \ldots, p_S^{u/n})$ be the vector of all combined p-values.

We want to show that $P(\tilde{p}^{u/n} \in A | p_s^{u/n} = q)$ for any increasing set A is non-decreasing in q for $p_s^{u/n} \in H_0^{u/n}(s)$. Note that the pooled p-value under the null is stochastically smallest when the unconstrained $u-1$ parameters go to infinity (lemma 1). Therefore the FDR will be largest under this null configuration, and it is enough to address this extreme configuration. Note that for this extreme configuration, the $u-1$ unconstrained p-values are zero and the pooled null p-value $p_s^{u/n} = f(U_1, \ldots, U_{n-u+1}, 0, \ldots, 0)$ is

an increasing function of $n-u+1$ independent $U(0,1)$ random variables $U_i, i = 1, \ldots, n-u+1$.

Let $I_0^{u/n}$ be the subset of locations where the partial conjunction null is true, $I_0^{u/n} = \{s : H_0^{u/n}(s) \text{ is true}\}$. Without loss of generality assume the p-value maps with zero p-values in location s are indexed as the last $u-1$ maps. Let

$$\begin{aligned} h(u_1, \ldots, u_{n-u+1}) &= P(\widetilde{p}^{u/n} \in A | (p_1(s), \ldots, p_n(s)) = (u_1, \ldots, u_{n-u+1}, 0, \ldots, 0)) \\ &= p(\widetilde{p}^{u/n} \in A | U_1 = u_1, \ldots, U_{n-u+1} = u_{n-u+1}) \end{aligned}$$

Since the individual maps are PRDS and since the p-values within every location are independent it follows that $h(u_1, \ldots, u_{n-u+1})$ is a non-decreasing function of u_i for $i \in \{1, \ldots, n-u+1\}$.

Therefore the problem reduces to that of showing that the following probability increases in q:

$$P(\widetilde{p}^{u/n} \in A | f(U_1, \ldots, U_{n-u+1}, 0, \ldots, 0) = q)$$

To prove this, we will use the following theorem due to Efron (1965):

THEOREM 4. *Let $X_1, \ldots, X_n$ be n independent random variables with PF_2 densities $r_1(x), \ldots, r_n(x)$ respectively, let $S = \sum_{i=1}^n X_i$ be their sum, and let $\Phi(x_1, \ldots, x_n)$ be a real measurable function on Euclidean n-space which is non-decreasing in each of its arguments. Then $E(\Phi(x_1, \ldots, x_n)|S = s)$ is a non-decreasing function of s.*

Since $f(U_1, \ldots, U_{n-u+1}, 0, \ldots, 0) = G(\sum_{i=1}^{n-u+1} g(U_i))$ and both $G(\cdot)$ and $g(\cdot)$ are increasing, for every q there exists a constant c such that $\{p_s^{u/n} =$

$q\} = \{\sum_{i=1}^{n-u+1} g(U_i) = c\}$, and $c(q)$ is increasing in q.

$$P(\tilde{p}^{u/n} \in A | \{p_s^{u/n} = q\}) = P(\tilde{p}^{u/n} \in A | \sum_{i=1}^{n-u+1} g(U_i) = c)$$

$$= Eh(U_1, \ldots, U_{n-u+1}) | \sum_{i=1}^{n-u+1} g(U_i) = c)$$

We can apply theorem 4 to conclude that $P(\tilde{p}^{u/n} > \tilde{q} \ | \{p_s^{u/n} = q\})$ increases in q.

The proof now continues as in the proof of theorem 1.2 in Benjamini and Yekutieli (2001), since the required relationship $P(\tilde{p}^{u/n} > \tilde{q} \ | p_s^{u/n} = q)$ (equation (12) in the proof of theorem 1.2) is satisfied.

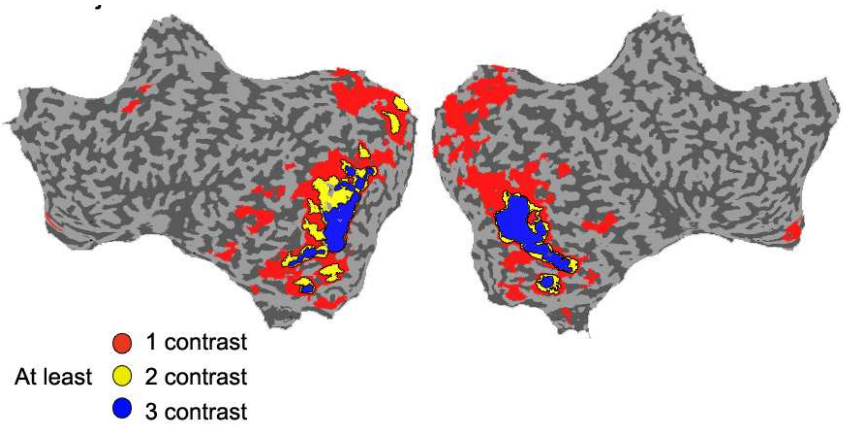

Figure 1. Activation maps for a single subject presented on unfolded cortical hemispheres: blue regions activated in all three contrasts with $FDR < 0.05$; yellow or blue regions activated in at least two contrasts with $FDR < 0.05$; red, yellow or blue regions activated in at least one contrast with $FDR < 0.05$.

[ht!]

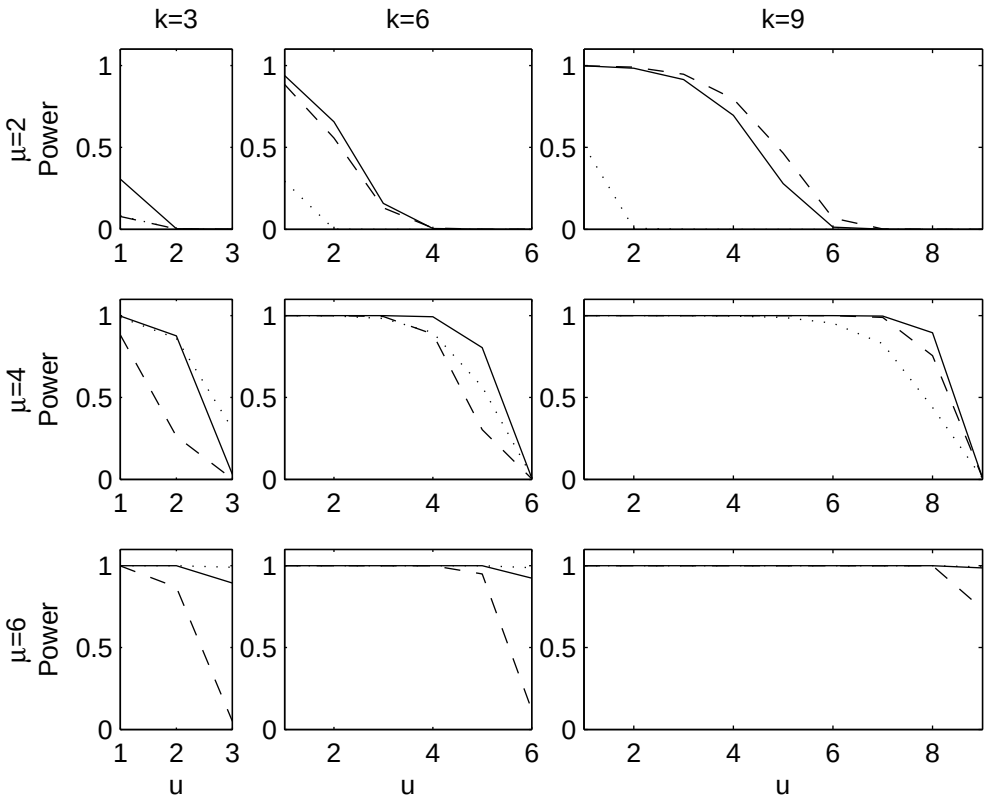

Figure 2. Power as a function of u when the FDR level is 0.05 and the simulated setting is that in which the number of p-values per location that come from the alternative is either null or 3 (first column), 7 (second column) or 9 (third column). The combining method is based on (1) equation 5 (solid line) (2) equation 4 (dashed line) and (3) equation 2 (dotted line). Each row is for a different signal size μ: $\mu = 2$ (top), $\mu = 4$ (middle) and $\mu = 6$ (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.

[!hbp]

Table 1

Number of significant genes for protein Swi4 (that forms part of the TF SBF) with an FDR level of 0.05

	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