Truncated Product Method for Combining p-values

Dmitri V. Zaykin, Lev A. Zhivotovsky, Peter H. Westfall, and Bruce S. Weir

Reference: Genet Epidemiol. 2002 Feb;22(2):170-85.

(D.V.Z.: Statistical Genetics Group, Department of Bioinformatics, GlaxoWellcome Inc., Research Triangle Park, NC 27709 and Programs in Statistical Genetics and Biomathematics, Department of Statistics, North Carolina State University, Raleigh NC 27695-8203

L.A.Z.: Institute of General Genetics, Russian Academy of Sciences, 3 Gubkin St., Moscow 117809, Russia

P.H.W.: Inf. Systems and Quant. Sci., Texas Tech University, Lubbock, TX 79409

B.S.W.: Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh NC 27695-8203)

Running title: Combining p-values

Abstract

We present a new procedure for combining p-values from a set of L hypothesis tests. Our procedure is to take the product of only those p-values less than some specified cut-off value and to evaluate the probability of such a product, or a smaller value, under the overall hypothesis that all L hypotheses are true. We give an explicit formulation for this p-value, and find by simulation that it can provide high power for detecting departures from the overall hypothesis. We extend the procedure to situations when tests are not independent. We present both real and simulated examples where the method is especially useful. These include exploratory analyses when L is large, such as genome-wide scans for marker-trait associations and meta-analytic applications that combine information from published studies, with potential for dealing with the "publication bias" phenomenon. Once the overall hypothesis is rejected, an adjustment procedure with strong family-wise error protection is available for smaller subsets of hypotheses, down to the individual tests.

Key words: meta analysis, multiple tests, genome-wide scans, microarrays, Bonferroni.

1 Introduction

When L hypothesis tests are conducted with the same significance level α, then the probability of finding at least one significant result among the L is greater than α. The resulting problem of interpreting the results from multiple tests has been considered many times, [e.g., Hochberg and Tamhane, 1987]. It is also recognized [e.g., Rosenthal, 1978] that a series of non-significant results may together suggest significance: "Two 0.06 results are much stronger evidence against the null than one 0.05." This situation led Fisher [Fisher, 1932] to his method for combining significance values, although it is the problem of possibly spurious single significant tests that is of more concern to us. We have been faced with this situation in different genetic contexts, including testing for allelic independence at several loci from several samples [Zaykin et al., 1995], and testing for marker-disease associations at several marker loci [Kaplan and Weir, 1995].

We concentrate here on global tests, or procedures that combine p-values from several tests in order to provide a p-value for the overall hypothesis that all single hypotheses are true.

We review the methods of Edgington [1972], Fisher [1932], Stouffer et al. [1949], and Wilkinson [1951] and describe a new procedure, the "truncated product method", hereafter TPM, that appears to have good properties, especially when the total number of tests is large. This appears to be of special interest in the light of recent advances in DNA microarrays and gene chip technologies.

In the language of Hedges and Olkin [1985], the procedures we discuss are termed omnibus or nonparametric because they depend only on the significance values of individual tests and not on the form of the underlying data.

2 Combination methods

We suppose that tests have been conducted for each of L hypotheses $H_i, i = 1, 2, \ldots, L$. For each test the p-value p_i is calculated: if H_i is true, this is the probability of observing a test statistic as extreme as or more extreme than the observed value in the direction of rejection [Shaffer, 1995]. However, $(1 - p_i)$ is also the value of the probability distribution function of the test statistic and therefore it is uniformly distributed on the interval [0, 1]. This holds for any continuous test statistic. Moreover, $-2 \ln p_i$ has a chi-square distribution with two degrees of freedom. This led Fisher [1932] to note that the statistic

$$t \quad = \quad -2 \sum_{i=1}^{L} \ln p_i = -2 \ln(\prod_{i=1}^{L} p_i)$$

has a chi-square distribution with $2L$ degrees of freedom when all L hypotheses are true (denoted by H_T in the sequel). Therefore, the p-value for the hypothesis that all H_i are true is the probability of a χ^2_{2L} variable being greater or equal to the observed value t. The Fisher combination test is shown to be asymptotically Bahadur optimal by Pallini [1994].

Edgington [1972] preferred to work with the sum of the p-values, and he gave the probability of the sum of L uniform [0, 1] variables being less than or equal to S as

$$\sum_{r=0}^{S'} (-1)^r \binom{L}{r} \frac{(S-r)^L}{L!}$$

where S' is the largest integer less than S. This probability also serves as the p-value for the hypothesis H_T. This and related distributions were also derived by Feller [1966]. Hedges and Olkin [1985] pointed out that one large p-value can overwhelm many small p-values with this approach.

For L independent tests, Wilkinson [1951] noted that the number of p-values less than some quantity τ has a binomial distribution $B(L, \tau)$ when H_T is true. The probability of finding at least

k values less than τ is

$$\sum_{i=k}^{L} \binom{L}{i} \tau^i (1-\tau)^{L-i}$$

Wilkinson set τ to the single-test significance level α. He used this binomial expansion to note, for example, that two tests significant at the 5% level in 14 tests does not imply that two events have occurred where each has a 5% probability, but that one event with a 15% probability has occurred. If $k = 1$, the probability becomes $[1-(1-\tau)^L]$ suggesting that $\tau = 1-(1-\alpha)^{1/L}$ for an overall α-level test. This is the basis of Šidák's correction [Šidák, 1967]. Any individual test must be significant at the level τ in order for the overall level to be α.

An alternative procedure [Stouffer et al., 1949] uses normal-transformed p-values. If $\Phi(x)$ denotes the probability distribution function for the standard normal distribution

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

then each p_i-value can be transformed to a standard normal score, when the hypothesis is true, by

$$\begin{aligned} 1-p_i &= \Phi(z_i) \\ z_i &= \Phi^{-1}(1-p_i) \end{aligned}$$

and $z = \sum_i z_i/\sqrt{L}$ is also standard normal. The p-value for hypothesis H_T is, therefore,

$$p = 1-\Phi\left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L} \Phi^{-1}(1-p_i)\right)$$

Finally, a recent competitor is Simes' test [Simes, 1986]. If the L p-values are ordered, then

$$\Pr\left\{\bigcup_{i}^{L} (p_{(i)} \leq i\alpha/L)\right\} \leq \alpha$$

This suggests a test: reject the global null-hypothesis, $H_0 = \{H_1, ..., H_L\}$, if $p_{(i)} \leq i\alpha/L$ for at least one i. Equivalently, the global p-value is given by $\min\{Lp_{(i)}/i\}$.

3 Truncated product method

As a procedure that combines the features of Fisher's product method and Wilkinson's truncation method, we suggest the use of the product W of all those p_i values that do not exceed some fixed value τ:

$$W = \prod_{i=1}^{L} p_i^{I(p_i \leq \tau)}$$

where $I(\cdot)$ is the indicator function.

Useful aspects of the TPM test include the following:

1. Experience shows that the ordinary Fisher product test loses power in cases where there are a few large p-values. This can happen when tests are one-sided, with noncentrality in the "wrong" direction, or when there are a predominance of near-null effects. By truncating, these large components are removed, thereby providing more power, much like a "trimmed mean" gains efficiency in the presence of outliers [Huber, 1977].

2. A natural, although somewhat arbitrary choice for τ is α (commonly 0.05). This allows easy use of the test in cases where only p-values for "statistically significant" results are given; it also allows estimation of "file drawer effects" in meta-analysis as shown below.

3. The truncated combination emphasizes smaller p-values, somewhat like the Simes and Šidák methods. However, Simes and Šidák p-values can never be smaller than $p_{(1)}$, the smallest p-value, whereas the TPM p-value will be smaller than $p_{(1)}$ when there are several small and reinforcing p-values in the set. In genome scans, this case is likely to occur in a local neighborhood of a susceptibility gene.

4. One can incorporate weights into the analysis as in Good [1955], as $W = \prod_{i=1}^{L} p_i^{w_i I(p_i \leq \tau)}$, thereby allowing studies or tests with more precision to play a larger role.

5. Isolation of individual significances is possible (and computationally feasible even for large numbers of tests as we show in the appendix), through the closure principle of Marcus et al, [1976] when using the TPM test.

3.1 Independence case

First we consider the case when all p-values are independent. Under the null hypothesis H_T, the distribution of W for $w < 1$ can be evaluated by conditioning on the number, k, of the p_i's less than τ:

$$\begin{aligned} \Pr(W \leq w) &= \sum_{k=1}^{L} \Pr(W \leq w \mid k) \Pr(k) \\ &= \sum_{k=1}^{L} \binom{L}{k} (1-\tau)^{L-k} \left(w \sum_{s=0}^{k-1} \frac{(k \ln \tau - \ln w)^s}{s!} I(w \leq \tau^k) + \tau^k I(w > \tau^k) \right) \end{aligned} \quad (1)$$

When L is large ($>$ 1,000 tests, assuming double precision calculations), the probability in (1) should be computed through a Monte Carlo algorithm described in the next section. The derivation of equation (1) is given in Appendix 1. At one extreme, setting $\tau = \min p$ results in Šidák's correction. At the other extreme, when $\tau = 1$, equation (1) provides Fisher's combined p-value. Thus, the method we describe here is "intermediate" between combination and individual adjustment techniques. Note that setting $\tau = 1$ provides a way of calculating Fisher's combined p-value directly. Instead of looking up the cumulative probability from the tail of a chi-square distribution, it can be obtained as

$$\Pr(W \leq w) = w \sum_{s=0}^{L-1} \frac{(-\ln w)^s}{s!} \quad (2)$$

C++ code for calculating the TPM p-value is available at

`http://statgen.ncsu.edu/zaykin/tpm/` .

An executable for a specific OS can be requested from Dmitri Zaykin (zaykind@niehs.nih.gov).

3.2 Correlated tests

The Fisher combination test is somewhat motivated by the independence case; however, Pallini [1994] notes that the simple combination can be also consider optimal for correlated tests. The problem remains as to finding the critical value in the presence of correlations. We consider several possibilities for dealing with non-independent tests. The first method exploits the fact that correlation is approximately invariant under monotone transformations:

$$\begin{aligned} \text{Corr}\left\{g(X_i), g(X_j)\right\} &\approx \frac{[g^{'}(\mu)]^2 \text{Cov}(X_i, X_j)}{\sqrt{[g^{'}(\mu)]^4 \text{Var}(X_i)\text{Var}(X_j)}} \\ &= \text{Corr}\{X_i, X_j\} \end{aligned} \tag{3}$$

Let $\mathbf{\Sigma}$ be a non-degenerate correlation matrix for the vector of p-values, $\mathbf{R}$. If $\mathbf{\Sigma}$ is positive definite, then there exists a matrix $\mathbf{C}$ (Cholesky factor), such that $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}^T$. Let $\mathbf{Z} = \Phi^{-1}(\mathbf{1} - \mathbf{R}^*)$ for some random vector $\mathbf{R}^*$ with Uniform(0,1) independent components. Then $\text{Var}(\mathbf{C}\ \mathbf{Z}) = \mathbf{C}\ \text{Var}(\mathbf{Z})\ \mathbf{C}^T = \mathbf{C}\ \mathbf{I}\ \mathbf{C}^T = \mathbf{\Sigma}$ and $\mathbf{C}\ \mathbf{Z}$ has a multivariate normal distribution with the mean zero and the covariance matrix $\mathbf{\Sigma}$. We can relate vectors $\mathbf{R}^*$ and $\mathbf{R}$ as follows:

$$\mathbf{R} = \mathbf{1} - \Phi\{\mathbf{C}\ \Phi^{-1}(\mathbf{1} - \mathbf{R}^*)\} \tag{4}$$

and

$$\mathbf{R}^* = \mathbf{1} - \Phi\{\mathbf{C}^{-1}\Phi^{-1}(\mathbf{1} - \mathbf{R})\} \tag{5}$$

and then apply equation 1 to values in $\mathbf{R}^*$. Thus, the correlated p-values in $\mathbf{R}$ are transformed to uncorrelated values in $\mathbf{R}^*$ and then the method derived for the independent case is applied. An assumption of this method is that the correlation completely describes the dependence between components of $\mathbf{R}$, as in the case when the test statistics jointly follow a multivariate normal distribution. As an example, consider a vector of two p-values with a known correlation, r. Since the Cholesky factor is

$$\mathbf{C} = \begin{pmatrix} 1 & 0 \\ r & \sqrt{1-r^2} \end{pmatrix},$$

the transformation would modify the vector in the following way:

$$\left[p_1,\ 1 - \Phi\left(\frac{\Phi^{-1}(1-p_2) - r\Phi^{-1}(1-p_1)}{\sqrt{1-r^2}} \right) \right]$$

If $r = 0$, the vector is just $[p_1, p_2]$.

A second method is based on the following Monte Carlo algorithm.

1 Decide on the value of the truncation point, τ.

2 Calculate $W_0 = \prod_i^L p_i^{I(p_i \leq \tau)}$ for a sample of p_i's (L observed p-values). Set $A = 0$.

2a Generate L independent uniform random numbers, $u_1^*, ..., u_L^*$, on $(0, 1)$, i.e. form the vector $\mathbf{R}^*$, then transform it to $\mathbf{R}$ with components $u_1, ..., u_L$, e.g. by equation 4.

2b Calculate $W = \prod_i^L u_i^{I(u_i \leq \tau)}$.

2c If $W \leq W_0$, increment A by one.

3 Repeat steps (2a)-(2c) B times.

4 The combined p-value is A/B.

In practice, sums of $-\ln(p_i)$ should be calculated in order to avoid numerical underflows. This method has the advantage that L can be very large. It can also be easily modified to incorporate combining p-values in a weighted fashion, by assigning weights, w_i, as $W_0 = \prod_i^L p_i^{w_i I(p_i \leq \tau)}$ and $W = \prod_i^L u_i^{w_i I(u_i \leq \tau)}$.

Both methods require that the correlation structure is known. In many cases it will be unknown and have to be estimated from the data. We will give such an example of calculating $\hat{\mathbf{\Sigma}}$ from marker-disease association tests with a dense map of genetic markers. The results obtained with the estimated covariance structure have to be viewed with caution. This problem is similar to the one of the generalized least squares technique, when the model is being pre-multiplied with the Cholesky factor obtained from $\hat{\mathbf{\Sigma}}$ [Rawlings, 1988].

A third approach is to estimate the distribution in W through resampling, when appropriate, as described in general terms by Westfall and Young [1993]. Churchill and Doerge [1994], and Doerge and Churchill [1996], describe permutation tests for quantitative trait loci. These test involve permuting the quantitative trait vector and re-evaluating the min (p-value) statistic for every permutation, then taking the combined p-value to be the proportion of permutations yielding a statistic smaller than that which was observed in the original sample. Such an approach can be applied with the truncated product test as well, and easily accommodates weights, for which analytic solutions are less tractable. Weights are assigned the same way as with the second approach.

4 Power considerations

We evaluated procedures described above on the basis of power to detect departures from null hypotheses concerning the mean of a normal distribution, $H_0 : \mu = \mu_0$, when the variance σ^2 is

known. The alternative hypotheses are $H_A : \mu = \mu_A > \mu_0$. We transform the mean $\bar{x}$ of a sample of size n to give the usual z test statistic

$$z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$$

Under H_0, z has the standard normal distribution, and under H_A it is distributed $N(\gamma, 1)$ where $\gamma = \sqrt{n}(\mu_A - \mu_0)/\sigma$.

When H_0 is true, the p-values could be written as p_0:

$$\begin{aligned} p_0 &= \Pr(Z \geq z) = 1 - \Phi(z) \\ z &= \Phi^{-1}(1 - p_0) \end{aligned}$$

where Z is the standard normal variable. The values of p_0 are distributed uniformly on [0, 1]. When H_A is true the p-value is

$$\begin{aligned} p_A &= \Pr(Z \geq z - \gamma) \\ &= 1 - \Phi\left(\Phi^{-1}(1 - p_0) - \gamma\right) \end{aligned}$$

Therefore, if u is drawn randomly from the uniform distribution on [0,1], p-values under the null and alternative hypotheses can be calculated as $(1 - u)$ and $\Phi\left(\Phi^{-1}(1 - u) + \gamma\right)$. This provides us with a means for generating p-values under each hypothesis rather than generating samples, calculating test statistics and then determining p-values.

5 Results

In the first place we simulated the case where all L hypotheses were true, and rejected the overall hypothesis H_T that all are true when the overall p-value was less than 0.05.

In Table I we show the proportion of 100,000 simulations that resulted in rejection, for several values of τ. TPM gave values consistent with the nominal 0.05 value.

For non-independent tests, we assumed that the test statistic has a multivariate normal distribution with the following variance-covariance matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & ... & & \\ ... & 1 & e^{-|i-j|\xi_{ij}} & \\ & & & \\ & e^{-|i-j|\xi_{ij}} & 1 & ... \\ & & ... & 1 \end{pmatrix}$$

$\mathbf{\Sigma}$ approximates the exponential decay in the correlation expected among marker loci in marker-trait association studies [e.g. Lander and Botstein, 1989]. ξ_{ij} were set to Uniform$(1/4, 3/4)$, so that the correlation between adjacent p-values ranged from 47% to 78%. We transformed vectors of correlated p-values as in (5) before calculating combined p-values. The results, averaged over 10,000 simulations are shown in Table II. They confirm that the size of the tests remain consistent with the nominal 0.05 value when the correlation is taken into account.

We then allowed h_A of $L = 25$ hypotheses to be false, but with the same value of γ so that the tests of each of these h_A hypotheses has the same power. We set $\gamma = 1.64$ so that these powers had the nominal value of β =0.50. All other $L - h_A$ hypotheses were set to be true. The p-value for the overall hypothesis H_T was then calculated by the methods of Edgington, Fisher, Stouffer, Wilkinson (with $\tau = 0.05$), Simes, and by the TPM with $\tau = 0.05$. Empirical powers based on 10,000 simulations are shown in Table III. Fisher's procedure and the TPM had substantially the same performance, although the TPM performed better when few hypotheses are false and the

other methods, except Simes, did not perform well for low numbers of false hypotheses. Simes method has the highest power in the situation when only a single hypothesis is false. In this case, Simes' method provides power at least as good as Bonferroni-adjusted minimum p-value (Tippett's procedure). When all hypotheses are false ($L = 25, \beta = 0.25$), procedures with truncation (TPM, Wilkinson's, and Simes') perform not as well as other methods, but TPM has the best power among the three.

The alternative hypotheses h_A are all one sided, but in Table IV we allow some tests to have positive γ and some negative γ. With this combination of different directions of departure from the null, it is clear that the TPM performs the best. For some values of h_A, Wilkinson's and Simes' procedures perform better than Fisher's procedure, but the other methods do not perform well. When all hypotheses are false ($L = 25, \beta = 0.25$), TPM still has the highest power.

Although the power of the TPM is satisfactory in comparison with the other combination procedures, it is not our intention to recommend its usage for this reason. The more important point is that the alternative hypothesis of interest is different (with the exception of Wilkinson's and Simes' methods). When applying the TPM, rejection of the null hypothesis leads one to declare that there is at least one false hypothesis among the ones resulting in p-values $\leq \tau$. Other combination procedures, except for Wilkinson's, state that there is at least one false hypothesis among *all* L hypotheses tested. Rice [1990] discussed further points distinguishing multiplicative and additive combination methods. The distinction becomes important when L is large, e.g. in whole genome scans or mRNA expression studies using microarrays. We recommend that in these cases the value of τ should be no larger than the significance level.

In the following simulations (Table V) we set $L = 25,000$ and $\tau = 0.05$. Power values for Simes'

and Fisher's methods are lower, however we stress that it is not our intention to compare the power values of TPM and Simes' to Fisher's, as they refer to the different subsets of hypotheses.

As the first example, we simulated a 143 cM map with a 2610 bi-allelic markers (SNP's) and a single bi-allelic disease gene with the penetrance values 0.56, 0.23, 0.05 for disease genotypes AA, Aa, aa. A founder population of 20 people expanded to 50,000 and was allowed to drift for 200 generations. The recombination process between the markers and the disease gene was modeled according to Haldane's mapping function. The growth rate was 1.1. The population disease prevalence was 0.06. We sampled 1500 people (750 affected + 750 unaffected) and at each marker position performed 2×2 Fisher's exact tests [Fisher, 1932] for association of marker alleles with the disease. We plotted the $-\ln p$ against marker map position, together with the $-\ln$ of 1% threshold, Bonferroni-adjusted by the total number of tests (0.01/2610). For similar plots some researchers [e.g. Juo et al, 1997] require that flanking markers around significant tests must also show some significance. Goldin et al. [1999] suggested that p-values should be averaged across certain genetic distances. Our approach is to take into account the information from all neighboring markers up to the distance to which the correlation extends. In our simulations the estimated correlation extended, on average, up to 5 markers in each direction, corresponding to a distance about of 0.5 cM. Figure 1 is a plot of raw -ln(p-values). The disease gene was between markers 1823 and 1824. Figure 2 is a graph of -ln(p-values), combined by TPM in sliding windows of 11 markers, with τ=0.05. Each time the window moved a marker ahead, the 11 probabilities were combined by (1) after applying the transformation given in (5). Because of the large number of markers, there is enough information to estimate correlations with reasonable precision. The correlation matrix, as a function of distance, was estimated from the data with `acf` function of

S-Plus (S-PLUS 3.4, MathSoft Inc.), using the information from all 2610 markers. Figure 3 is a similar graph, but with τ=1. The graphs show that the combined p-value reveals the peak around the true location, while keeping false peaks below the Bonferroni-adjusted significance threshold [see Terwilliger et al., [1997] and Knapp, [1998] for discussion of similar phenomena]. The 0.05 value of τ resulted in a higher peak in these simulations. This happened because of the unequal informativeness of the markers, due to variations in allele frequencies and linkage disequilibria with the disease gene.

As a final example we present an application of the method of combining p-values across genetic linkage studies. The web-based database at http://cooke.gsf.de/asthmagen/main.cfm [Wjst and Immervoll, 1998] reports 644 linkage results for chromosome 6 for asthma related phenotypes. While there are varying degrees of information in the studies, and combination of p-values should therefore incorporate weights, we perform the unweighted analysis to illustrate the precise solution (1). 55 of the p-values are smaller than 0.05 and many of remaining p-values are reported as "non-significant", without indicating the actual values. Note that this alone would prevent one from utilizing methods that require values from all L tests. Taking into account the possibility of the "publication bias", when only successful findings are reported, the value of L=644 is a lower estimate of the number of tests actually performed. None of the p-values would remain significant after adjusting by a false discovery rate controlling procedure by Benjamini and Hochberg [1995], or family-wise error rate controlling procedures, such as Bonferroni, or more powerful method of Hochberg [1988]. However, the TPM ($\tau = 0.05$) yields the combined p-value of 0.00005, giving a strong indication of genetic linkage. The nature of the method also allows us to provide the maximum possible value of L for the result to remain significant. It is equal to 904 in this example.

Thus, the number of non-reported studies with negative results can be as large as $904-644 = 260$ for the combined p-value to remain below 5% level. There have been models developed for estimating the number of non-reports [Givens et al., 1997, Gleser and Olkin, 1996].

6 Discussion

Debate over the use of p-values in summarizing the results of hypothesis tests continues [Goodman, 1999] but the values are routinely calculated and reported in many disciplines. Many genetic studies lead to p-values from tests on multiple loci in multiple samples and it is necessary to take this multiplicity into account when drawing inferences. Among the approaches that have been considered in the past it appears that Fisher's method, of taking minus twice the logarithm of the product of a set of L p-values and recognizing that this quantity is distributed as chi-square with $2L$ degrees of freedom when all L hypotheses are true, is a good way of providing an overall p-value.

We have suggested here an alternative procedure, which also takes into account the size of L p-values but uses only the small ones. By "small" we mean a conventional values such as 0.05. Instead of asking whether the set of L tests contains any evidence for departures from all L hypotheses, our procedure can be used to ask whether any of the significant test statistics are indeed significant. By focusing on only the set of small p-values we appear to have increased the overall power, especially in situations where a small subset of the hypotheses are false.

Acknowledgments

This work was supported in part by NIH Grant GM45344. Helpful advice was given by Drs. R. Berger, E. Martin, S. Ghosh, and L. Stefanski.

7 Appendix 1: Distribution of W

When H_T is true and $\tau < 1$, the number of small p-values (k) has a binomial distribution, and p_i's are observations from the uniform $(0,1)$ distribution, truncated at τ (i.e. the distribution of p_i's is uniform on $(0,\tau)$).

Given k, the conditional distribution of the product (W) can be calculated directly. Let $X_1, ..., X_k$ be independent uniform $(0,\tau)$ random variables. Consider the transformation:

$$\begin{aligned} Z_1 &= X_1 \\ Z_2 &= X_1 X_2 \\ &... \\ Z_k &= X_1 X_2 ... X_k \end{aligned}$$

with inverse

$$\begin{aligned} X_1 &= Z_1 \\ X_2 &= Z_2/Z_1 \\ &... \\ X_k &= Z_k/Z_{k-1} \end{aligned}$$

The Jacobian of the transformation ($\mathbf{J}$) has the following structure:

$$\partial x_i/\partial z_j = \begin{cases} 1 & i=j=1 \\ 1/z_{i-1} & i=j; \geq 1 \\ -z_i/z_{i-1}^2 & j=i-1 \\ 0 & \text{otherwise} \end{cases}$$

Therefore

$$\mid \mathbf{J} \mid = \prod_{\mathrm{i}=1}^{\mathrm{k}-1} 1/\mathrm{z_i}$$

and the joint density is

$$f(\mathbf{Z}) = \frac{1}{\tau^{\mathrm{k}} \prod_{\mathrm{i}=1}^{\mathrm{k}-1} \mathrm{z_i}}$$

Integrating out z_1 through z_{k-1} from the joint density gives the conditional probability, $P(W \leq w \mid k)$:

$$\begin{aligned} \Pr(W \leq w \mid k) &= \int_0^w \left[\int_t^{\tau^k} \int_{z_{k-1}}^{\tau^k} \ldots \int_{z_2}^{\tau^k} \frac{\prod_{i=1}^{k-1} dz_i}{\tau^k \prod_{i=1}^{k-1} z_i} \right] dt \\ &= \int_0^w \frac{(\ln \tau^k - \ln t)^{k-1}}{(k-1)!\ \tau^k} dt \end{aligned} \tag{6}$$

Then the unconditional distribution is found as follows:

$$\begin{aligned} \Pr(W \leq w) &= \int_0^w \sum_{k=1}^{L} \frac{(\ln \tau^k - \ln t)^{k-1}}{(k-1)!\ \tau^k} \\ &\times \quad I(\ln \tau^k > \ln t) \binom{L}{k} \tau^k (1-\tau)^{L-k} dt \end{aligned} \tag{7}$$

The probability calculated in (7) corresponds to the combined p-value. After τ^k in (7) is canceled, this probability is

$$\begin{aligned} \Pr(W \leq w) &= \int_0^w \sum_{k=1}^{L} \frac{(k \ln \tau - \ln t)^{k-1}}{(k-1)!} \\ &\times \quad I(\tau^k > t) \binom{L}{k} (1-\tau)^{L-k} dt \end{aligned} \tag{8}$$

or equivalently

$$\begin{aligned} \Pr(W \leq w) &= \sum_{k=1}^{L} \binom{L}{k} \frac{(1-\tau)^{L-k}}{(k-1)!} \left[\int_0^w (k \ln \tau - \ln t)^{k-1} \right. \\ &\times \quad \left. I(\tau^k > t) dt \right] \end{aligned} \tag{9}$$

Provided $\tau^k > t$, the integral in (9), which we denote by I_k is:

$$\begin{aligned} I_k &= \int_0^w (\ln \tau^k - \ln t)^{k-1} dt && (10) \\ &= (\ln \tau^k - \ln t)^{k-1} t|_0^w \\ &- \int_0^w t d[(\ln \tau^k - \ln t)^{k-1}] && (11) \\ &= w(\ln \tau^k - \ln w)^{k-1} \\ &- (k-1) \int_0^w t(-\frac{1}{t})(\ln \tau^k - \ln t)^{k-2} dt && (12) \\ &= (k-1) I_{k-1} + w A(\tau, k, w)^{k-1}. && (13) \end{aligned}$$

where $A(\tau, k, w) = k \ln \tau - \ln w$. Since $I_1 = w$, then

$$\begin{aligned} I_k &= (k-1)! \left[w + w \sum_{s=1}^{k-1} \frac{A(\tau, k, w)^s}{s!} \right] && (14) \\ &= w(k-1)! \sum_{s=0}^{k-1} \frac{A(\tau, k, w)^s}{s!} && (15) \end{aligned}$$

Therefore,

$$\begin{aligned} \Pr(W \leq w) &= w \sum_{k=1}^{L} \binom{L}{k} \frac{(1-\tau)^{L-k}}{(k-1)!} (k-1)! \sum_{s=0}^{k-1} \frac{A(\tau, k, w)^s}{s!} && (16) \\ &= w \sum_{k=1}^{L} \binom{L}{k} (1-\tau)^{L-k} \sum_{s=0}^{k-1} \frac{A(\tau, k, w)^s}{s!} && (17) \\ &= w \sum_{k=1}^{L} \sum_{s=0}^{k-1} \binom{L}{k} (1-\tau)^{L-k} \frac{A(\tau, k, w)^s}{s!} && (18) \end{aligned}$$

8 Appendix 2: Closed testing with Truncated Fisher Test

Adjustments for subsets of hypotheses and individual adjustments are available through the application of the closure principle of Marcus et al. (1976). Generally, the procedure considers all possible combination hypotheses obtained via the intersection of the set of individual hypotheses of

interest. If an individual hypothesis and all intersections that contain it as a component are rejected by an appropriate α-level test, then the closure principle states that the given hypothesis can be also rejected, at the level α. The closure procedure controls the family-wise error rate (FWER) strongly, meaning that FWER $\leq \alpha$ regardless of which subset of null hypotheses happens to be true (Hochberg and Tamhane, 1987). The total number of combination hypotheses (N_h) is

$$N_h = \sum_{i=1}^{L} \binom{L}{i} = 2^L - 1 \tag{19}$$

which grows quickly with L and often limits applicability of the method.

Fortunately, this is not the case for the TPM test. Noting that (1) is an increasing function of L and a decreasing function of W, we see that, among all intersections of a given size s (where $s \leq L$) that include H_i, only the combination that includes H_i and the remaining $s-1$ largest p-values needs to be tested. Thus, significance for any given hypothesis can be determined using L tests; and when all L component tests are considered, the maximum number of evaluations is L^2. However, many of these evaluations are redundant, and in practice the number is less than L^2. In some cases, e.g., with the Šidák combined tests, the number of evaluations is as small as L. To illustrate this argument, consider the case when $\tau = 1$. Then for the ordered set of p_j's ($j = 1, ..., L$) the adjusting procedure for any subset of p-values, $\mathcal{P}_i$, such that $p_{(i)}$ is the largest p-value in the set, is as follows. Compute (2) for most stringent subsets:

$$\{\mathcal{P}_i,\ p_{(L)}\}$$

$$\{\mathcal{P}_i,\ p_{(L)}, p_{(L-1)}\}$$

$$\{\mathcal{P}_i,\ p_{(L)}, p_{(L-1)}, p_{(L-2)}\}$$

...

$$\{\mathcal{P}_i,\ p_{(L)}, p_{(L-1)}, p_{(L-2)}, ..., p_{(i+1)}\}$$

The adjusted p-value for the subset $\mathcal{P}_i$ is given by the maximum of these values. Many subsets, such as, for example, $\{\mathcal{P}_i,\ p_{(L-1)}\}$ do not need to be considered, because they will yield p-values smaller than the one for $\{\mathcal{P}_i,\ p_{(L)}\}$.

9 References

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society series B. 57:289–300.

Churchill GA. Doerge RW. Empirical threshold values for quantitative trait mapping. 1994. Genetics. 138:963-971

Doerge RW. Churchill GA. Permutation tests for multiple loci affecting a quantitative character. 1996. Genetics. 142:285-294.

Edgington ES. 1972. An additive method for combining probability values from independent experiments. J Psychology 80:351–363.

Feller W. 1966. An Introduction to Probability Theory and its Applications. Wiley and Sons, New York.

Fisher RA. 1932. Statistical Methods for Research Workers. Oliver and Boyd, London.

Givens GH, Smith DD, Tweedie RL. 1997. Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. Statistical Science 12: 221–240.

Goldin LR, Chase GA, Wilson AF. 1999. Regional inference with averaged p-values increases the power to detect linkage. Genetic Epidemiology 17:157–164.

Good IJ. 1955. On the weighted combination of significance tests. Journal of the Royal Statistical Society series B 17:264–265.

Goodman SN. 1999. Toward evidence-based medical statistics. I: The P-value fallacy. Annals of Internal Medicine 130:995–1004.

Hedges LV, Olkin I 1985. Statistical Methods for Meta-Analysis. Academic Press, New York.

Hochberg Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75:800–802.

Hochberg Y, Tamhane AC 1987. Multiple Comparison Procedures. Wiley and Sons, NY.

Holm S. 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6:65–70.

Huber P. 1977. Robust statistical procedures. SIAM Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Juo SH, Beaty TH, Duffy DL, Maestri NE, Prenger VL, Zeiger J, Lei HH, Coresh J. 1997. A comprehensive analysis of complex traits in problem 2A. Genetic Epidemiology 14:815–820.

Kaplan, NL, Weir BS. 1995. Are moment bounds on the recombination fraction between a marker and a disease locus too good to be true? Allelic association mapping revisited for simple genetic diseases in the Finnish population. Am J Human Genetics 57:1486–1498.

Knapp M. 1998. Discriminating between true and false-positive peaks in a genomewide linkage

scan, by use of the peak length. Am J Hum Genetics, 62:1561–1562.

Lander ES, Botstein D. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199.

Marcus R, Peritz E, Gabriel KR. 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63:655–660.

Pallini A. 1994. Bahadur exact slopes for a class of combinations of dependent tests. Metron 52: 53–65.

Rawlings JO. 1988. Applied Regression Analysis. Wadsworth Inc., California.

Rice WR. 1990. A consensus combined p-value and the family-wide significance of component tests. Biometrics 46:303–308.

Rosenthal R. 1978. Combining results of independent studies. Psychological Bull 85:185–193.

Shaffer JP. 1995. Multiple hypothesis testing: A review. Ann Rev Psychology 46:561–584.

Simes RJ. 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751–754.

Šidák, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 78:626–633.

Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM, Jr. 1949. The American Soldier,

Vol. 1. Adjustment During Army Life. Princeton Univ. Press, Princeton.

Terwilliger JD, Shannon WD, Lathrop GM, Nolan JP, Goldin LR, Chase GA, Weeks DE. 1997. True and false positive peaks in genomewide scans: applications of length-biased sampling to genome mapping. Am J Hum Genetics 61:430–438.

Westfall PH, Young SS. 1993. Resampling-Based Multiple Testing. Wiley, New York.

Wilkinson B. 1951. A statistical consideration in psychological research. Psychological Bull. 48:156–158.

Wjst M, Immervoll T. 1998. An internet linkage and mutation database for the complex phenotype asthma. Bioinformatics 14:827–828.

Zaykin D, Zhivotovsky L, Weir BS. 1995. Exact tests for association between alleles at arbitrary numbers of loci. Genetica 96:169–178.

Table I Power levels for the TPM when all L hypotheses are true.

	L					
Truncation point	2	3	5	10	25	50
0.05	0.04910	0.05001	0.04981	0.04992	0.04991	0.05054
0.10	0.04978	0.05070	0.04955	0.05008	0.05016	0.04946
0.25	0.05002	0.04983	0.04942	0.05056	0.05070	0.04958
0.50	0.04886	0.05064	0.05045	0.04990	0.04959	0.04964
1.00	0.04996	0.04936	0.04996	0.05010	0.04996	0.04972

Table II Power levels for the TPM when $L = 25$ hypotheses are true and tests are correlated (10,000 simulations).

Truncation point	Ignoring correlation	Accounting for correlation
0.05	0.100	0.046
0.50	0.154	0.051
1.00	0.155	0.048

Table III Power levels when h_A of $L = 25$ hypotheses are false and $L - h_A$ are true (10,000 simulations, $\beta = 50\%$ for $L = 1 - -6$ and $\beta = 20\%$ for $L = 25$).

	h_A						
Method	1	2	3	4	5	6	25
Edgington	0.077	0.124	0.186	0.266	0.361	0.470	0.987
Fisher	0.119	0.229	0.321	0.502	0.678	0.793	0.981
TPM, $\tau = 0.05$	0.158	0.265	0.401	0.538	0.653	0.774	0.809
Stouffer	0.091	0.159	0.256	0.373	0.502	0.628	0.991
Wilkinson, $\tau = 0.05$	0.075	0.137	0.234	0.355	0.482	0.604	0.759
Simes	0.177	0.224	0.349	0.428	0.486	0.538	0.227

Table IV Power levels when h_A of $L = 25$ hypotheses are false in each direction and $L - 2h_A$ are true (10,000 simulations, $\beta = 60\%$ for $L = 1 - -6$ and $\beta = 25\%$ for $L = 25$).

	h_A						
Method	1	3	5	7	9	11	25
Edgington	0.000	0.000	0.001	0.023	0.227	0.709	0.015
Fisher	0.049	0.307	0.709	0.932	0.991	0.999	0.303
TPM, $\tau = 0.05$	0.389	0.725	0.903	0.972	0.993	0.999	0.473
Stouffer	0.000	0.000	0.008	0.099	0.450	0.848	0.032
Wilkinson, $\tau = 0.05$	0.147	0.492	0.787	0.931	0.979	0.994	0.363
Simes	0.456	0.311	0.753	0.841	0.897	0.938	0.317

Table V Power levels when h_A of $L = 25,000$ hypotheses are false ($\beta = 70\%$) and $L - h_A$ are true (10,000 simulations).

	h_A					
Method	50	75	100	125	150	175
TPM, $\tau = 0.05$	0.389	0.647	0.853	0.956	0.989	0.999
Fisher	0.305	0.524	0.741	0.893	0.963	0.991
Simes	0.361	0.480	0.576	0.665	0.737	0.801

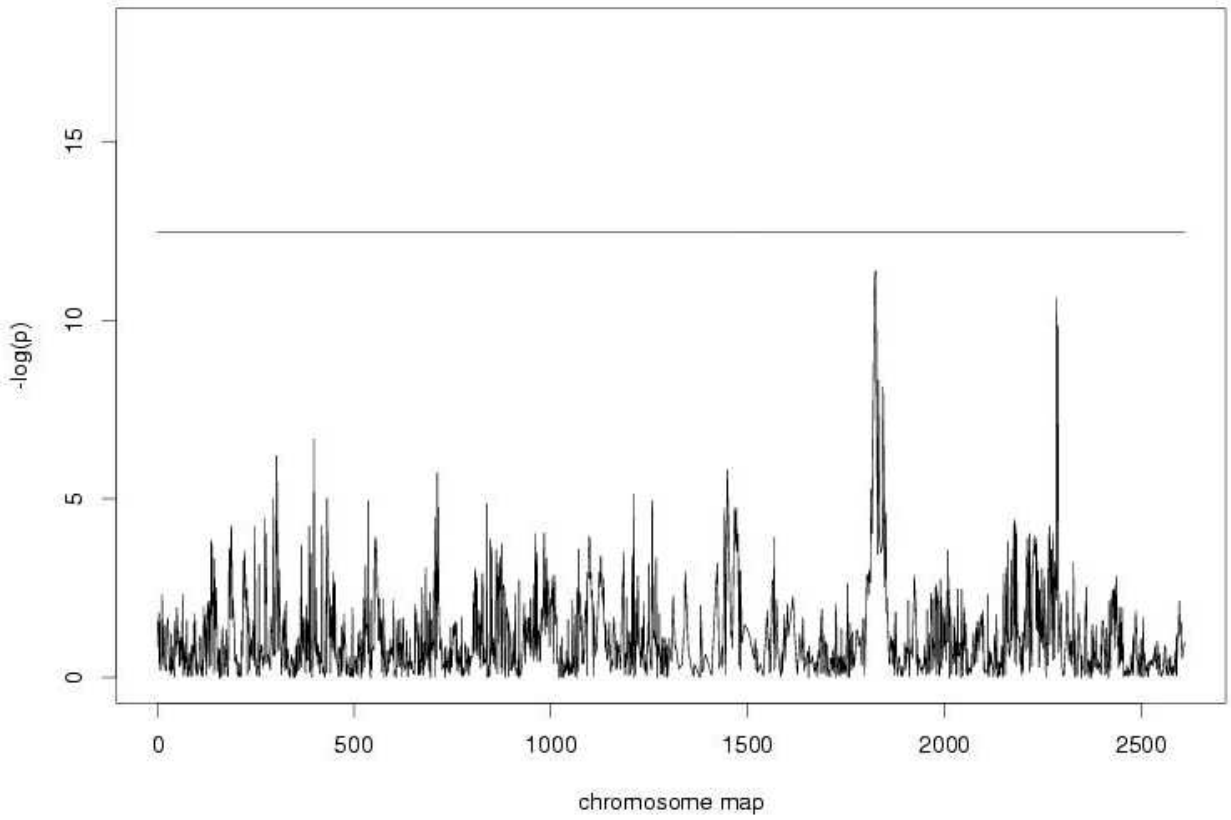

Figure 1: Plot of $-\ln p$

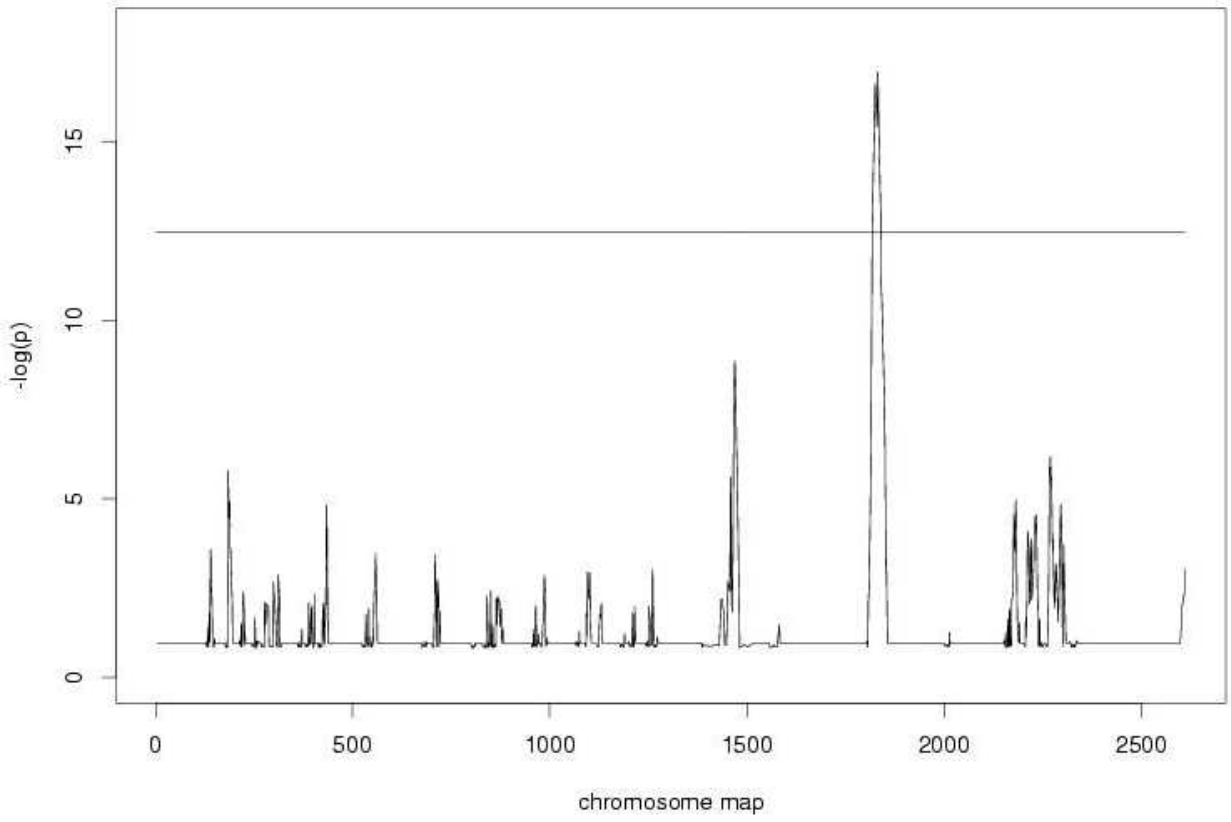

Figure 2: Plot of $-\ln p$ combined in a window, $\tau = 0.05$

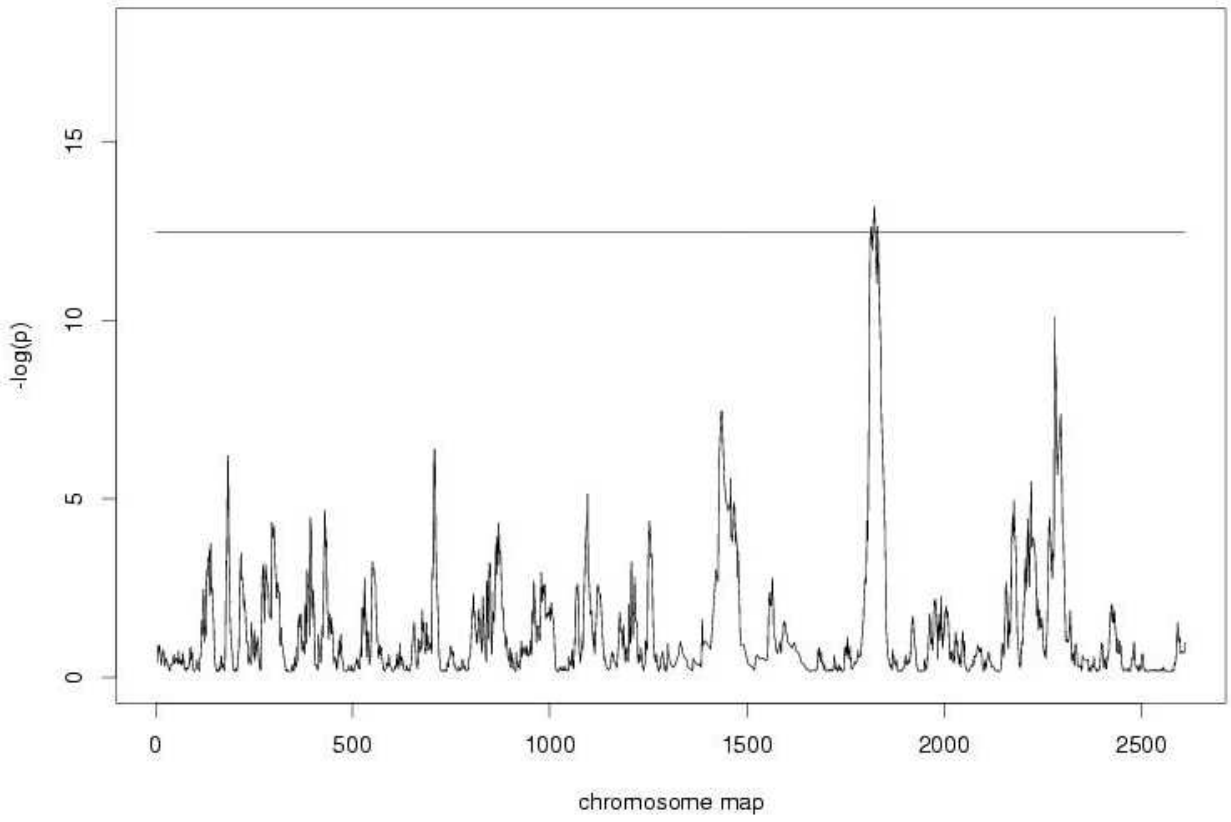

Figure 3: Plot of $-\ln p$ combined in a window, $\tau = 1$