Genetic Epidemiology 31: 741–747 (2007)

Improving Power in Genome-Wide Association Studies: Weights Tip the Scale

Kathryn Roeder,[1]* B. Devlin[2] and Larry Wasserman[1]

[1]*Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania*
[2]*Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania*

The potential of genome-wide association analysis can only be realized when they have power to detect signals despite the detrimental effect of multiple testing on power. We develop a weighted multiple testing procedure that facilitates the input of prior information in the form of groupings of tests. For each group a weight is estimated from the observed test statistics within the group. Differentially weighting groups improves the power to detect signals in likely groupings. The advantage of the grouped-weighting concept, over fixed weights based on prior information, is that it often leads to an increase in power even if many of the groupings are not correlated with the signal. Being data dependent, the procedure is remarkably robust to poor choices in groupings. Power is typically improved if one (or more) of the groups clusters multiple tests with signals, yet little power is lost when the groupings are totally random. If there is no apparent signal in a group, relative to a group that appears to have several tests with signals, the former group will be down-weighted relative to the latter. If no groups show apparent signals, then the weights will be approximately equal. The only restriction on the procedure is that the number of groups be small, relative to the total number of tests performed. *Genet. Epidemiol.* 31:741–747, 2007.

Key words: Bonferroni correction; genome-wide association analysis; multiple testing; weighted *P*-values

Contract grant sponsor: National Institute of Mental Health Grant; Contract grant number: MH057881.
*Correspondence to: Kathryn Roeder, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213.
E-mail: roeder@stat.cmu.edu
Received 5 December 2006; Revised 15 February 2007; Accepted 4 April 2007
Published online 4 June 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20237

Thorough testing for association between genetic variation and a complex disease typically requires scanning large numbers of genetic polymorphisms. In a multiple testing situation, such as a whole genome association scan, the null hypothesis is rejected for any test that achieves a P-value less than a predetermined threshold. To account for the greater risk of false positives, this threshold is more stringent as the number of tests conducted increases. To bolster power, recent statistical methods have been proposed that up-weight and down-weight hypotheses, based on prior likelihood of association with the phenotype [Genovese et al., 2006; Roeder et al., 2006]. Weighted procedures multiply the threshold by the weight w, for each test, raising the threshold when $w>1$ and lowering it if $w<1$. To control the overall rate of false positives, a budget must be imposed on the weighting scheme. Large weights must be balanced with small weights, so that the average weight is one. These investigations reveal that if the weights are informative, the procedure improves power substantially, but, if the weights are uninformative, the loss in power is usually small. Surprisingly, aside from this budget requirement, any set of non-negative weights is valid [Genovese et al., 2006]. Although desirable in some respects, this flexibility makes it difficult to select weights for a particular analysis.

The type of prior information readily available to investigators is often non-specific. For instance, single nucleotide polymorphisms (SNPs) might naturally be grouped, based on features that make various candidates more promising for this disease under investigation. For a brain-disorder phenotype we might cross-classify SNPs by categorical variables such as those displayed in Table I. The SNPs in $\mathcal{G}_1$ seem most promising, a priori, while those in $\mathcal{G}_4$ seem least promising. Those in $\mathcal{G}_2$ and $\mathcal{G}_3$ are more promising than those in $\mathcal{G}_4$, but somewhat ambiguous. It is easy to imagine additional variables that further partition the SNPs into various classes that help to separate the more promising SNPs from the others. Although this type of information lends itself to grouping SNPs, it does not lead directly to weights for the groups. Indeed it might not even be possible to choose a natural ordering of the groups. What is needed is a way to use the data to determine the weights, once the groups are formed.

Until recently, methods for weighted multiple testing required that prior weights be developed

TABLE I. An illustration for grouping SNPs and tests

	Functional	Function unknown
Brain expressed	$\mathcal{G}_1$	$\mathcal{G}_2$
Non-brain expressed	$\mathcal{G}_3$	$\mathcal{G}_4$

independently of the data under investigation [Genovese et al., 2006; Roeder et al., 2006]. In this article, we ask the following questions: if the weights are to be applied to tests grouped by prior information, what choice of weights will optimize the average power of the genetic association study? How can we estimate these weights from the data to achieve greater power without affecting control of the family-wise error rate?

When analyzing data one often has candidate genes derived from various sources. For instance, consider a genome-wide analysis for association with phenotype A. One might identify 250 genes known to lie in pathways that affect this phenotype. If all SNPs in these genes are grouped together, this constitutes a group including, say, 2,500 SNPs. Next, suppose phenotype B is similar to phenotype A. The investigators are likely to think genes associated with phenotype B might also affect phenotype A. This delineates a natural choice for a second group of SNPs. If 75 genes are natural candidates for phenotypes, but 25 of these were already included in group 1, then one could place the remaining, say, 500 SNPs in group 2. Continuing with this reasoning one could delineate one or more additional groups by identifying genes in other pathways, SNPs in regulatory regions, SNPs in regions associated with copy number variation, and so forth. Finally, we envision one large group of SNPs for which there is little prior evidence of association. This constitutes the final grouping.

Alternatively, the groupings might be based upon specific studies of the same phenotype, choosing the groups based on reported linkage peaks, results of previous genome-wide association studies, or results of gene expression studies. For instance, using the results of a linkage study, one could place promising genes under peaks in group 1, less promising genes under peaks in group 2, genes near peaks in group 3, SNPs near peaks, but not in genes in group 4, genes in candidate genes, but not in peaks, in group 5, and finally, all other SNPs in the remaining group.

Based on technical requirements our approach places a minimum on the number of SNPs in a grouping, but not a maximum. Hence, it is possible to create hundreds or even thousands of groupings. From a mathematical viewpoint this is acceptable, but from a scientific viewpoint it seems implausible. In a typical study, we anticipate investigators identifying somewhere between 3 and 30 promising groups of SNPs, each containing at least 20, but usually many more SNPs.

METHODS

Consider m hypotheses corresponding to standardized test statistics $T = (T_1, \ldots, T_m)$. The P-values associated with the tests are $(P_1, \ldots, P_m)$. We assume the jth statistic T_j is approximately normally distributed with non-centrality parameter $\xi_j \equiv E[T_j]$, or it is χ^2 distributed with non-centrality parameter ξ_j^2. The non-centrality parameters divide the hypotheses into disjoint sets, the nulls $\mathcal{H}_0 = \{j : \xi_j = 0\}$, and the alternatives $\mathcal{H}_1 = \{j : \xi_j^2 > 0\}$, containing m_0 and $m_1 = m - m_0$ elements, respectively.

In practice, the noncentrality parameters are unknown. We use the data to attempt to discover the partition $(\mathcal{H}_0, \mathcal{H}_1)$. This problem is classified as a multiple testing problem because the type I error can be very large if one does not control for the number of tests performed to determine the partition. Most methods that control for multiple testing do so by adjusting the threshold for rejecting the null hypothesis. When using a Bonferroni correction for m tests, the threshold for rejection is achieved if the P-value $P_j \le \alpha/m$. The weighted Bonferroni procedure of Genovese et al. [2006] is as follows. Specify nonnegative weights $w = (w_1, \ldots, w_m)$ and reject hypothesis H_j on the set $j \in \mathcal{R}$ defined as

$$\mathcal{R} = \left\{ j : \frac{P_j}{w_j} \le \frac{\alpha}{m} \right\}. \tag{1}$$

As long as $m^{-1} \sum_j w_j = 1$, this procedure controls family-wise error rate at level α. For a test of $\xi_j = 0$ versus $\xi_j \neq 0$, the power of a single weighted test is

$$\pi(\xi_j, w_j) = \overline{\Phi}\left(\overline{\Phi}^{-1}\left(\frac{\alpha w_j}{2m}\right) - \xi_j\right) + \overline{\Phi}\left(\overline{\Phi}^{-1}\left(\frac{\alpha w_j}{2m}\right) + \xi_j\right),$$

where $\overline{\Phi}(t)$ is the upper tail probability of a standard normal cumulative distribution function. When the alternative hypothesis is true, weighting increases the power when $w_j > 1$ and decreases the power when $w_j < 1$. We call $\pi(\xi_j, w_j)$ the *per-hypothesis power*. For signals $(\xi_1, \ldots, \xi_m)$ and weights $(w_1, \ldots, w_m)$ the *average power* is

$$\overline{\pi}(\theta, w) = \frac{1}{m_1} \sum_{j \in \mathcal{H}_1} \pi(\xi_j, w_j).$$

The optimal weight vector $w = (w_1, \ldots, w_m)$ that maximizes the average power subject to $w_j \ge 0$ and $m^{-1} \sum_{j=1}^m w_j = 1$ is

$$w(\xi_j) = \frac{m}{\alpha} \overline{\Phi}\left(\frac{|\xi_j|}{2} + \frac{c}{|\xi_j|}\right) \tag{2}$$

[Wasserman and Roeder, 2006], where c is the constant that satisfies the budget criterion on weights

$$\frac{1}{m}\sum_{j=1}^{m} w(\xi_j) = 1. \quad (3)$$

Although we do not know ξ_j it is useful to study the form of the optimal weights. The optimal weights vary with the signal strength in a non-monotonic manner (Fig. 1). For any particular sample, c adjusts the weights to satisfy the budget constraint on weights. In doing so, it shifts the mode of the weight function from left to right depending on the number of small, versus large, signals observed.

The weights act directly on the P-values via (1), which in turn affects the critical values. If chosen optimally, the threshold for rejection varies smoothly with the signal strength. Figure 2 plots the rejection threshold $-\log_{10}(\alpha w_j/m)$, calculated for the data displayed in Figure 1, as a function of the signal strength and contrasts it with the rejection threshold of a Bonferroni corrected test $-\log_{10}(\alpha/m)$. From Figures 1 and 2 it is evident why an optimally weighted test has greater power than a non-weighted test. The weighted threshold is less stringent for signals in the midrange, and more stringent for both large and small signals. Consequently, if the signal is likely to be very strong or very weak, the test is down-weighted (weight less than one). In practice, little power is lost by this tradeoff. For small signals the chance of rejecting the hypothesis is minimal with or without weights. For large signals the P-value is likely to cross the threshold regardless of the weight. Larger weights are focused in the midrange to help to reveal signals that are marginal.

In practice ξ_j is not known, so it must be estimated to utilize this weight function. A natural choice is to build on the two-stage experimental design [Satagopan and Elston, 2003; Wang et al., 2006] and split the data into subsets, using one subset to estimate ξ_i, and hence $w(\xi_i)$, and the second to conduct a weighted test of the hypothesis [Rubin et al., 2006]. This approach would arise naturally in an association test conducted in stages. It does lead to a gain in power relative to unweighted testing of stage 2 data; however, it is not better than simply using the full data set without weights for the analysis [Rubin et al., 2006; Wasserman and Roeder, 2006]. These results are corroborated by Skol et al. [2006] in a related context. They showed that it is better to use stages 1 and 2 jointly, rather than using stage 2 as an independent replication of stage 1.

To gain a strong advantage with data-based weights, prior information is needed. One option is to order the tests [Rubin et al., 2006], but with a large number of tests this can be challenging. Another option is to group tests that are likely to have a signal, based on prior knowledge, as follows:

1. Partition the tests into subsets $\mathcal{G}_1, \ldots, \mathcal{G}_K$, with the kth group containing r_k elements, ensuring that r_k is at least 20–30.
2. Calculate the sample mean Y_k and variance S_k^2 for the test statistics in each group.
3. Label the ith test in group k, T_{ik}. At best only a fraction of the elements in each group will have a signal, hence we assume that for $i = 1, \ldots, r_k$ the

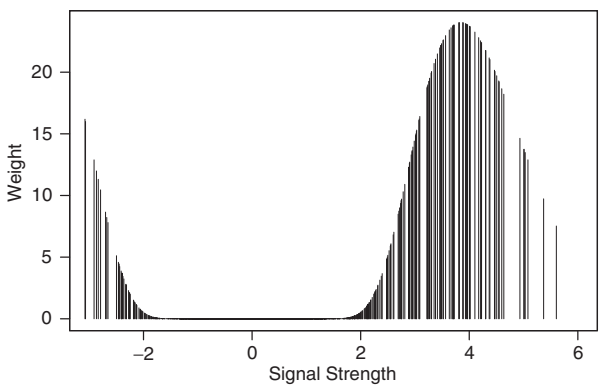

Fig. 1. Distribution of optimal weights for $m = 100{,}000$ simulated tests (a random selection of 5,000 are displayed). The signal strength ξ is the non-centrality parameter for a standard normal test statistic; if the test statistic is χ^2 distributed, the signal strength is the square root of the non-centrality parameter. The weight is w(ξ).

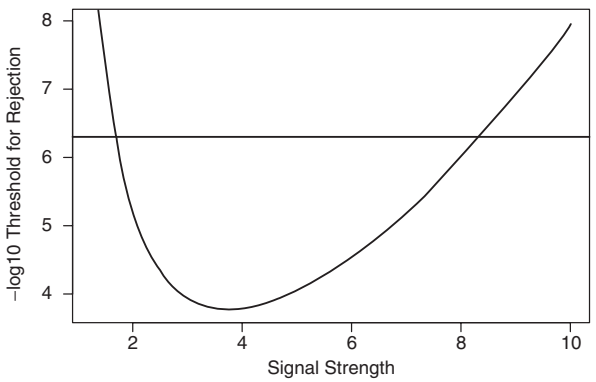

Fig. 2. Threshold for rejecting P-values versus signal strength. The $\log_{10}$ P-value is rejected if it is larger than the threshold. For this illustration $m = 100{,}000$ and $\alpha = 0.05$. The unweighted Bonferroni has a constant threshold value (horizontal line). The weighted threshold varies as a function of the weight (curved line). The optimal weight is calculated as a function of the (estimated) signal strength.

distribution of the test statistics is approximated by a mixture model

$$T_{ik} \sim (1-\pi_k)N(0,1) + \pi_k N(\xi_k, 1)$$

or

$$T_{ik} \sim (1-\pi_k)\chi^2_1(0) + \pi_k \chi^2_1(\xi_k^2),$$

where ξ_k is the signal size for those tests with a signal in the kth group. This is an approximation because the signal is likely to vary across tests. The mixture of normals is only appropriate when the tests are one-sided. For two-sided alternatives, the χ^2 is the natural approach. This test squares the non-centrality parameter, effectively removing any ambiguity about the direction of the associations.

4. Estimate (π_k, ξ_k) using the method of moments estimator (for details see appendix). Because ξ_k has no meaning when $\pi_k = 0$, the $\widehat{\xi}_k$ is set to 0 when $\widehat{\pi}_k$ is close to zero. For the normal model the estimators are

$$\widehat{\pi}_k = Y_k^2/(Y_k^2 + S_k^2 - 1), \qquad \widehat{\xi}_k = Y_k/\pi_k, \qquad (4)$$

provided $\widehat{\pi}_k > 1/r_k$; otherwise $\widehat{\xi}_k = 0$.
For the χ^2 model they are

$$\widehat{\xi}^2 = \frac{(S_k^2 + Y_k^2 + 3)}{Y_k - 1}, \qquad \widehat{\pi}_k = \frac{Y_k - 1}{\widehat{\xi}_k^2} \qquad (5)$$

provided $Y_k > 1$ and $1/r_k < \widehat{\pi}_k < (r_k - 1)/r_k$; otherwise $\widehat{\xi}_k = 0$.

5. For each of the k groups, construct weights $w(\widehat{\xi}_k)$. It is apparent in Figure 1 that if $|\widehat{\xi}_k| < \delta$, for δ near 0, then $w(\widehat{\xi}_k) \approx 0$ and it is unlikely that any tests in the kth group will be significant, regardless of the P-value. The stochastic quantity δ depends upon the relative values of $(\widehat{\xi}_1, \ldots, \widehat{\xi}_K)$, and the number of elements in each group. For this reason we have found that smoothing the weights generally improves power of the procedure. We suggest using a linear combination such as

$$\widehat{w}_k = (1-\gamma)\, w(\widehat{\xi}_k) + \gamma\, K^{-1} \sum_k w(\widehat{\xi}_k)$$

with $\gamma = 0.01$ or 0.05. The larger the choice of γ, the more evenly distributed the weights across groups. Alternatively, one could smooth the weights by using a Stein shrinkage estimator or bagging procedure to obtain a more robust estimator of $(\xi_1, \ldots, \xi_K)$ [Hastie et al., 2001]. Regardless of how the weights are smoothed, one should renorm them to ensure the weights sum to m. Each test in group k receives the weight $\widehat{w}_k$. Another effect of the smoothing is to ensure that each group gets a weight greater than zero.

This weighting scheme relies on data-based estimators of the optimal weights, but with a partition of the data sufficiently crude to preserve the control of family-wise error rate. The approach is an example of the "sieve principle" [Bickel et al., 1993]. The sieve principle works because the number of parameters estimated is far less than the number of observations. Thus, many observations are used to estimate each parameter. Consequently parameters are estimated with substantially less variability than if they were estimated using only the test statistics from the particular gene under investigation. Because the weights are determined by the size of the tests in the entire cluster the probability of upweighting simply because a single test is large, due to chance, is small. More formally this result is stated in the following theorem.

Theorem. Let $b_m = \frac{1}{m}\sum_k \sqrt{r_k}$. If $\sum_{j=1}^m \widehat{w}_j = m$, then the rejection rule defined in (1) controls family-wise error at level $\alpha + O(b_m)$. Proof is in the appendix.

This result establishes control of family-wise error at level α, asymptotically, provided

$$b_m = \frac{\sum_k \sqrt{r_k}}{\sum_k r_k} \to 0, \qquad \text{as } m \to \infty.$$

The inflation term in the error rate is near zero under most plausible circumstances. The basic requirement is that each group contains a sufficient number of elements to permit valid estimation of $(\xi_1, \ldots, \xi_K)$. For instance, suppose we are testing 100,000 SNPs and we have chosen 10 groups, each of size 10,000. In this case the error term is proportional to 0.01. More realistically, suppose we have five groups, four with 1,000 elements each and one with 96,000 elements. Then the error is proportional to 0.002.

Figure 3 illustrates how $w(\widehat{\xi}_k)$ varies with $\widehat{\xi}_k$ and the sample variances (weight is proportional to the diameter of the circle). Notice that weight increases as a function of the signal until it becomes fairly large and then declines.

RESULTS

To simulate a large scale study of association, we generate test statistics from $m = 10,000$ tests to mimic the second stage of a two-stage genome-wide association study, with about 1/3–1% of the initial SNPs tested at stage 2. Each test statistic is simulated as $T_j = \xi_j + \varepsilon_j$ with $\varepsilon_j \sim N(0,1)$. We simulate linkage disequilibrium between test statistics by generating them in correlated "haplotype blocks", including

five tests per block, with squared correlation between $(\varepsilon_j, \varepsilon_{j'})$ equal $r^2 = .8$ when (j, j') are elements in a common block. Tests not in a common block are simulated independently.

The majority of the tests, m_0, are generated under the null hypothesis ($\xi_j = 0$) and the remaining $m_1 = 100$ are generated with a signal ($\xi > 0$). In actual data, clusters of tests in close proximity tend to exhibit a signal. To simulate this feature we choose 20 blocks with signal strengths within each block equal to $(\xi_1, \ldots, \xi_5) = \xi_0 \times (1, 1.5, 2, 2.5, 3)$. This mimics a situation with 20 actual causal variants present in the genome, leading to approximately 100 tests associated with the phenotype at varying levels of intensity. The five levels of signal are meant to represent the causal SNP, with signal 3ξ, and the associated SNPs in the block with slightly weaker signals, between ξ and 2.5ξ.

To illustrate the effect of weights, we group the 10,000 tests into six categories, $\mathcal{G}_1, \ldots, \mathcal{G}_6$, using "prior" information of varying quality. We chose six groups because it seems like a reasonable number to reflect actual prior knowledge. Groups 1–5 are reserved for tests that we expect to yield a signal and group 6 contains what we believe are less promising tests. The number of elements in group 6 is far larger than the number in groups 1–5.

In reality the signal strength of any particular test would be unknown. Nevertheless, in our simulation we use our knowledge of $(\xi_1, \ldots, \xi_m)$ to generate the groupings. We initially create perfect groupings and then permute some tests to degrade the information in the prior groupings. The more tests permuted, the greater the degradation. When we permute tests, we always place all tests within a block in a common grouping.

We start the process by placing all m_0 tests with $\xi = 0$ in $\mathcal{G}_6$. Then the m_1 remaining tests are distributed evenly in the remaining five categories, $\mathcal{G}_1, \ldots, \mathcal{G}_5$. Next, to create more realistic groupings, we move some tests from $\mathcal{G}_1, \ldots, \mathcal{G}_5$ into $\mathcal{G}_6$ and vice versa. Specifically, we move a fraction p_0 of the m_0 null tests to $\mathcal{G}_1, \ldots, \mathcal{G}_5$, and distribute them evenly. Likewise, we move a fraction p_1 of the m_1 tests from $\mathcal{G}_1, \ldots, \mathcal{G}_5$ to $\mathcal{G}_6$. By varying (p_0, p_1), we obtain various levels of informativeness of the groupings, reflecting priors of various value.

After forming the groupings, we measure the information content based upon the clustering of the signals. Let ξ_{ik} be the signal of the ith element in $\mathcal{G}_k$, $\bar{\xi}_{.k}$ be the mean signal in $\mathcal{G}_k$, and $\bar{\xi}_{..}$ be the mean signal of the whole set, respectively. The information in a prior grouping is summarized by the R^2

$$R^2 = 1 - \frac{\sum_k \sum_i (\xi_{ik} - \bar{\xi}_{.k})^2}{\sum_k \sum_i (\xi_{ik} - \bar{\xi}_{..})^2}.$$

For instance, if $p_0 = p_1 = 0$ then R^2 is near 1, but if $p_0 = p_1 = .5$ then R^2 is near 0.

Figure 4 displays the difference in power as a function of R^2, in which the proportion of null tests in $\mathcal{G}_1, \ldots, \mathcal{G}_5$, and alternative tests in $\mathcal{G}_6$ varies as follows: $p_0 \in [0.01 - 0.5]$ and $p_1 \in [0.01 - 0.95]$. These results show that, provided $p_0 < 0.4$ and $p_1 < 0.4$, the weighted method is more powerful than the unweighted method (plot symbol "+"). For R^2 near 0 the loss in power from poorly selected groupings varies from near zero to −5% (plot symbol "∗").

To see the effect of including null loci in the same grouping as the tests with true effects, we fix $(\xi_0 = 2, p_1 = 0, m_1 = 100)$ and vary p_0. Setting $p_0 = 0.5$ (0.1) increases the number of elements in $\mathcal{G}_1, \ldots, \mathcal{G}_5$ to 1,010 (218), but only 20 are true alternatives. For $p_0 = 0.01$, 0.1, 0.25, and 0.5 we find a difference in power (weighted minus the unweighted procedure) of 12, 8, 5, and 1 percent, respectively. Next we explore the effect of failing to place all of the true effects in $\mathcal{G}_1, \ldots, \mathcal{G}_5$. To do so, we fix $(\xi_0 = 2,$

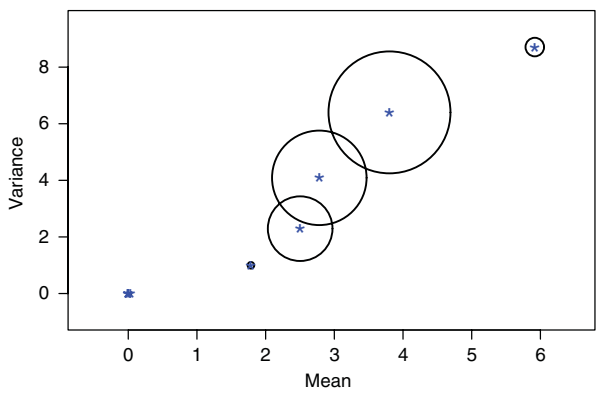

Fig. 3. Weight as a function of $\hat{\xi}_k$ and variance. The diameter of the circle indicates relative weight.

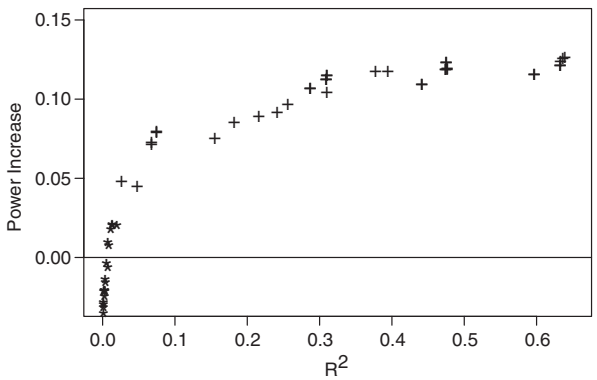

Fig. 4. Net power difference between weighted Bonferroni and unweighted, as a function of R^2. The best cases are $p_0 < 0.4$ and $p_1 < 0.4$ (plot symbol +). All other conditions are depicted with plot symbol ∗.

$p_0 = .1, m_1 = 100$) and vary p_1. For $p_1 = 0.05$. 0.1, 0.5, and 0.9, we find a difference in power of 8, 7, 2, and −2 percent, respectively. Even when 90% of the true alternatives are grouped with large numbers of nulls in group 6, the loss in power is relatively small.

DISCUSSION

Whole genome analysis has generated much discussion about power, the effect of multiple testing on power, and various multistage experimental designs [e.g., Wang et al., 2006]. We investigate the performance of a weighting scheme that allows for the input of weak prior information, in the form of groupings of tests, to improve power in large scale investigations of association. The method can be applied at any stage of an experiment. The beauty of the grouped-weighting concept is that it is likely to lead to an increase in power, provided multiple tests with signals are clustered together in one (or more) of the groups. Little power is lost when many groups contain no true signal. This remarkable robustness is achieved because the procedure uses the observed test statistics in the grouping to determine the weight. If there is no apparent signal, the group will be down-weighted. The only restriction on the procedure is that the number of groups be small, relative to the total number of tests performed, and the number of elements in each group be sufficiently large (say 20 or more).

As discussed in Genovese et al. [2006], under a broad set of conditions, weighted procedures perform in a desirable asymmetrical fashion: good choices of weights lead to substantial gains in power, while poor choices lead to modest losses. The requirement is that the number of up-weighted tests is relatively small, compared with down-weighted tests, so that the majority of the tests are not strongly down-weighted. For the grouped weighting scheme, the weights are determined by a data-based estimate of the effect size in a grouping. Because most groupings will not contain a detectable signal, the sparce weighting criterion is met naturally. The non-monotonic form of the optimal weight function limits the size of weight attributed to each group. Finally, the smoothing of weights and the constraint that limits the weights to sum to m ensures that each test receives at least a baseline weight. These features in total ensure that the procedure typically either gains power and losses are minimal. In our simulations we did not discover any conditions for which the power loss was dramatic; however, we did not explore weighting schemes with large numbers of small groups. It is possible that under these conditions a poor choice of groupings could lead to a more serious loss in power.

Using groupings and weights to interpret the many tests conducted in a large scale association study has potential, regardless of power lost when weights are poorly chosen. Typically some SNPs are favored due to knowledge gleaned from the literature and prior investigations, yet SNPs with no relationship to phenotype can often produce smaller *P*-values than the favored candidates. Moreover, it often happens that promising candidate SNPs do produce small *P*-values, but these *P*-values are not small enough to cross the significance threshold when a Bonferroni correction is applied. After the huge investment of a whole genome scan it would be foolhardy not to pursue both (i) SNPs that produce tiny *P*-values and (ii) SNPs that produce more modest *P*-values that would have been significant had a formal weighting scheme been utilized to incorporate prior information. We suggest using the weighting method of analysis described here as a way to formalize the incorporation of prior information.

Weights can be incorporated into various multiple testing procedures, including false discovery methods. This paper considers controlling family-wise error rate, but similar results hold for false discovery control [Benjamini and Hochberg, 1995] and will be pursued elsewhere.

ELECTRONIC DATABASE INFORMATION

Software for computing weighted *P*-values is available from: http://wpicr.wpic.pitt.edu/WPICCompGen/

REFERENCES

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300.

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA. 1993. Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Series in the Mathematical Statistics. Baltimore, Maryland.

Genovese CR, Roeder K, Wasserman L. 2006. False discovery control with P-value weighting. Biometrika 93:509–524.

Hastie T, Tibshirani R, Friedman J. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Verlag.

Roeder K, Bacanu SA, Wasserman L, Devlin B. 2006. Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet 78:243–252.

Rubin D, van der Laan M, Dudoit S. 2006. Multiple testing procedures which are optimal at a simple alternative. Collection of Biostatistics Research Archive, http://www.bepress.com/ucbbiostat/paper171/

Satagopan JM, Elston RC. 2003. Optimal two-stage genotyping in population-based association studies. Genet Epidemiol 25: 149–157.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213.

Wang H, Thomas DC, Pe'er I, Stram DO. 2006. Optimal two-stage genotyping designs for genome-wide association scans. Genet Epidemiol 30:356–368.

Wasserman L, Roeder K. 2006. Weighted hypothesis testing. http://arxiv.org/abs/math.ST/0604172

APPENDIX

Derivation of (4): $E[T] = \pi\xi$ and $\text{var}[T] = 1 + \xi^2\pi(1-\pi)$. Set $E[T] = Y$ and $\text{var}[T] = S^2$ and solve for π and ξ.

Derivation of (5): $E[T] = \pi\xi^2 + 1$ and $E[T^2] = \pi\xi^4 + 6\pi\xi^2 + 3$. Set $E[T] = Y$ and $E[T^2] = S^2 - Y^2$ and solve for π and ξ^2.

Proof of Theorem 1. Recall that $\mathcal{H}_0$ and $\mathcal{R}$ define the set of tests for which the null is true and null is rejected, respectively. With fixed weights, the family-wise error is

$$P(\#(\mathcal{R} \cap \mathcal{H}_0) > 0) = P\left(P_j \le \frac{\alpha w_j}{m} \text{ for some } j \in \mathcal{H}_0\right)$$
$$\le \sum_{j \in \mathcal{H}_0} P\left(P_j \le \frac{\alpha w_j}{m}\right) = \frac{\alpha}{m} \sum_{j \in \mathcal{H}_0} w_j \le \alpha \overline{w} = \alpha.$$

The estimated signal in the group occupied by the jth test, $\widehat{\xi}_k$, is estimated from a sample of r_k test statistics, consequently $\widehat{\xi}_k = \xi_k + O(r_k^{-1/2})$. Thus with random weights

$$P(\#(\mathcal{R} \cap \mathcal{H}_0) > 0) \le \sum_{j \in \mathcal{H}_0} P\left(P_j \le \frac{\alpha w_j(\widehat{\xi}_k)}{m}\right)$$
$$\approx \frac{\alpha}{m} \sum_{j \in \mathcal{H}_0} \left\{ w_j(\xi_k) + \left(w_j(\widehat{\xi}_k) - w_j(\xi_k) \right) \right\}$$
$$\le \alpha(1 + O(b_m)).$$