8

Multiple Tests for Genetic Effects in Association Studies

Peter H. Westfall, Dmitri V. Zaykin, and S. Stanley Young

1. Introduction

Many common human diseases have a genetic component as measured by familial studies. Metabolic disorders such as diabetes, cardiovascular diseases such as high blood pressure, psychiatric disorders such as schizophrenia, and neurodegenerative diseases such as Alzheimer's disease all are thought to have a hereditary component. In some diseases the genetic control is through a single gene, while in others, multiple genes interact in complex ways with environmental factors to produce the disease ***(1–5)***.

Data are and will become increasingly available to attempt to link genes to disease phenotype(s). Linkage studies, although powerful for screening relatively large chromosomal regions, lack needed precision because of the constraints imposed by the number of recombination events during generations contained in the pedigree ***(6)***. Recently, researchers have attempted to develop techniques that exploit possibilities of fine mapping due to linkage disequilibrium between genetic markers and disease genes. Typing single nucleotide polymorphism markers (SNPs) inside of candidate regions provides a potential means for such analysis ***(7)***; however, the problem remains in that the complex diseases are very likely to have multiple etiologies. Consider control of essential hypertension. It has a measured heritability of 3 45%, yet the identification of specific genes remains unclear. Many candidate genes for essential hypertension have been identified and, in a particular individual, a combination of some few of these genes might lead to disease.

There is a need for a statistical strategy to analyze these complex experiments, given the multiple testing implied by multiple candidate genes and the risk of false associations. In this chapter we discuss primarily methods for controlling

From: *Methods in Molecular Biology, vol. 184: Biostatistical Methods*
Edited by: S. W. Looney © Humana Press Inc., Totowa, NJ

familywise error rate (FWE) with multiple genetic tests, identifying single and epistatic effects, and discuss readily available software (PROC MULTTEST of SAS/STAT®) for this purpose. The benefit of the method is sound inference in the evaluation of case-control genotype–phenotype association studies.

2. Multiple Testing Principles for Disease–Genotype Association

Our focus is primarily on multiple contingency table-type tests described in, for example, Sasieni ***(8)***, and extensions thereof. In the simplest analysis, subjects are cross-classified in a 2 × 2 table, according to disease status (case or control) and presence or absence of a particular allele at a given locus.

As an initial screening procedure, one may perform a test for each genetic locus in a genome scan or dense SNP map. Such tests are associational only, and further study is needed to establish causation; however, they can be very useful to identify candidate genes. Follow-up analyses can proceed using, for example, linkage analysis or haplotype-level tests ***(9)***.

When these tests are performed separately over multiple loci, there can be hundreds, even thousands, of tests, and false-positives are expected, as discussed throughout the statistics and genetics literature (e.g., ***10–12***). Various methods have been proposed to control this risk in genetic studies, such as FWE-controlling methods ***(12)***; informal, global-based testing methods ***(13)***; and false discovery rate (FDR) controlling methods ***(14)***.

We suggest controlling FWE and justify it in two ways. First, control of FWE has a simple operational interpretation: If the FWE is set at 10% (say), then we expect that in only one out of every 10 studies will one or more false significant results be claimed. Therefore, the analyst may gamble upon the occurrence that the given study was not one of those 10%, and claim that all identified associations are real and repeatable. The FDR controlling procedure of Benjamini and Hochberg ***(15)***, described in Weller et al. ***(14)*** for genetic QTL analysis, while more powerful than FWE for gene finding, does not allow such a clear operational definition. In a given study, the number of erroneous significances is a random variable, and therefore somewhat unpredictable. Furthermore, while FDR-controlling methods allow only an average of 100α% of the claimed significances to be in error, the false discovery rate can be substantially larger than that in studies where one or more genes have been declared significant ***(16)***. Thus, although FDR-controlling methods are indeed more powerful, their operational interpretation is not as useful as that of FWE-controlling methods.

Second, advances in modern computing have made the powerful FWE-controlling "closed testing" methods accessible for the analysis of genetic tests. In particular, these methods can accommodate discreteness and genetic correlation structures (including linkage) to improve power. In our examples we will incorporate such features through exact testing methods.

For readers unfamiliar with multiple testing methods, closed testing, and/or PROC MULTTEST, it may be helpful to read Westfall and Wolfinger's (2000) article "Closed Multiple Testing Procedures and PROC MULTTEST," available on the SAS® website (http://www.sas.com/service/library/periodicals/obs/obswww23/). The remainder of this chapter is a condensed summary of material therein, with special emphasis on genetics applications.

2.1. The Closure Principle

FWE-controlling methods can be made less conservative and more powerful by using the closure principle of Marcus et al. ***(17)***. The procedure is as follows: one considers all possible combination hypotheses obtained via intersection of the set of base hypotheses of interest. If the base hypothesis, and all intersections that contain it as a subcomponent, are all rejected by an appropriate α-level test (we will use exact tests here), then the closure principle allows that the given hypothesis can be rejected, at FWE level α. Thus, if there are k base tests, there are $2^k–1$ tests to consider. For small studies, this procedure is ideal; however, for typical genotype/phenotype association studies where thousands of genotypes are considered, the number of intersection subsets to evaluate seems astronomical, and uncomputable even by current standards. However, there are simplifications that make this methodology computationally feasible, as we now discuss.

2.2. Application of Closure to the Min P Statistic

Given the typical genome scan, with each test yielding a p-value for genetic association, the first impulse is to locate the minimum value (min P). The question then becomes, "How unusual is the min P, given the number of genetic features scanned?" This question can be answered using an hypothesis testing approach, where one tests the global null hypothesis of no feature effect by evaluating the probability that the min P can be as low as the observed value, under the global null. This is similar to the approaches described in ***(18,19)***, except that they do not apply the closure principle to isolate particular loci. Their analysis at the first step is essentially equivalent, but by applying the closure principle to their test procedure, one can obtain multiple candidate (single-level) associations between quantitative trait loci (QTLs) and the trait, all with familywise error protection, even under the case where there are some null and some non-null QTL locations.

Fortunately, one need not consider all $2^k–1$ subsets for the closed procedure. If each subset is tested using min P from that subset, then one need only evaluate the k subsets that correspond to the ordered p-values, and not the entire set of $2^k–1$. Formally, let the observed p-values be $p_1,\cdots,p_k$ ordered as $p_{(1)}\leq\cdots\leq p_{(k)}$ with corresponding hypotheses $H_{(1)},\cdots,H_{(k)}$ with $p_{(j)}=p_{i_j}$. Let the random p-values prior to observation be denoted by P_j.

The closed min *P*-based method collapses ***(20)*** to the following sequential procedure:

Algorithm: Closed min *P* Testing
reject $H_{(1)}$ if $P\,(\min_{j\in\{i_1,\dots,i_k\}} P_j \le p_{(1)}) \le \alpha$
reject $H_{(2)}$ if $H_{(1)}$ was rejected and $P\,(\min_{\mathrm{j}\in\{\mathrm{i}_2,\dots,\mathrm{i}_k\}} P_j \le p_{(2)}) \le \alpha$
.
.
.
reject $H_{(k)}$ if $H_{(k\text{-}1)}$ was rejected and $P\,(\min_{\mathrm{j}\in\{\mathrm{i}_k\}} P_j \le p_{(k)}) \le \alpha$

In accordance with the closure principle, all probabilities are calculated under the assumption of no genetic effect in the respective subsets of hypotheses.

Using the Bonferroni inequality

$$P\,(\min_{l\in\{i_j,\dots,i_k\}} P_l \le p_{(j)}) \le (k - j + 1)\,p_{(j)}\,,$$

the closed min *P*-based procedure becomes the Holm method ***(21)***. However, this method is needlessly conservative: the upper bound $(k - j + 1)p_{(j)}$ is too large, implying that 5%-level significance might not be attained. This conservativeness arises because (1) there are correlations, sometimes large, among the genes due to linkage, and (2) the distributions of the tests are discrete ***(22)***.

The correlation structure and discreteness of distributions can be taken into account by calculating the probabilities

$$P\,(\min_{l\in\{i_j,\dots,i_k\}} P_l \le p_{(j)})$$

directly and exactly using permutation tests. To do this, one randomly permutes the vectors of genetic indicators over the set of all subjects, so that in a given resampled data set, the first n_1 vectors are assumed to have phenotype 1, and the remaining n_2 are assumed to have phenotype 2. Thus, in this permutation model, the null hypothesis of no genetic effect holds for all subsets of hypotheses, as required by both the closure principle and the "subset pivotality" criterion of Westfall and Young ***(23)***, p. 42. The probabilities

$$P\,(\min_{l\in\{i_j,\dots,i_k\}} P_l \le p_{(j)})$$

are then exactly computed as the proportion of possible permutations for which the value of $_{l\in\{i_j,\dots,i_k\}} P_l^{\,*}$, as calculated from the permuted data set, is less than or equal to the value $p_{(j)}$, as calculated from the original data set. Because resampled (or permuted) data sets preserve the correlation structure and discreteness characteristics, the resulting probabilities are typically less than the conservative Bonferroni approximations $(k - j + 1)p_{(j)}$.

As the number of possible permutations can be exceedingly large, a simple and accurate approximation can be obtained by permutation resampling, that is, by sampling with replacement from the finite population of possible permutations. The resulting method is a statistically permutationally exact method under the case of infinitely many Monte Carlo samples. Monte Carlo error bounds and detailed algorithms are described by Westfall and Young ***(23)***.

Fortunately, software to perform this exact, closed min *P*-based analysis is readily available in PROC MULTTEST of SAS/STAT® ***(24)***. This software requires a binary or ordinal phenotype such as (diseased)/(not diseased), or (severely diseased)/(moderately diseased)/(not diseased). The software runs more quickly when the phenotype is coded as binary. To take full advantage of the discreteness, it also requires binary genotype representations, although it can analyze ordinal genotype representations in exact fashion as well.

2.3. Application of Closure to the Simes–Hommel Test for Genetic Association

As an alternative to the use of the min *P* statistic for testing each subset homogeneity hypothesis, one may use Simes test ***(25)***, which considers the entire distribution of *p*-values, rather than just the minimum. For a given set of k genetic association tests with *p*-values $p_1,...,p_k$, the hypothesis of no genetic effect is rejected if $\min\{kp_{(j)}/j\} \leq \alpha$, where the $p_{(j)}$ are the ordered *p*-values. Like the case with closed testing and the min *P* test, closed testing with Simes' test allows shortcuts so that all $2^k - 1$ subsets need not be evaluated. The simplification occurs because, for each subset size (say, s), one need only consider the combined test that contains the gene of interest, and the $s - 1$ remaining largest *p*-values, rather than all $\binom{k}{s}$ subsets of size s. Hommel ***(26)***, Wright ***(27)***, and Grechanovsky and Hochberg ***(28)*** describe such shortcut methods. PROC MULTTEST of SAS/STAT® (as of Version 8.1) can perform these tests with $O(k^2)$ operations, rather than $O(2^k)$, which makes the method feasible for genetics screening tests.

The Simes test is valid (has type I error rate $\leq \alpha$) when the tests are positively dependent, as shown by Sarkar ***(29)***. In negatively dependent cases, the error rate may exceed α, but the excess is typically slight and not troubling ***(30)***.

While it would be preferred to use the discreteness of the distributions for the Simes test ***(31)***, as shown in **Subheading 2.2.** for the min *P* test, such an analysis would greatly increase the computational complexity. Studies have shown that the Simes-based approach tends to be more powerful than the min *P*-based approach when there are greater numbers of affected hypotheses ***(32,33)***. In genetics experiments where multiple gene effects are expected, or with tight linkage, this might indeed be the case. In such a case, the Simes-based approach might have superior power to the min *P*-based approach. Further research is needed to develop computa–tionally convenient Simes-based closed testing algorithms that incorporate distributional characteristics.

2.4. Application of Closure to the Fisher Test for Genetic Association

Yet another possibility is to apply the Fisher combination test ***(34)*** for each subset homogeneity hypothesis. For a given subset, the combination test statistic is $T = -2\Sigma \ln p_i$, which is distributed as χ^2_{2k} when (1) the subset homogeneity hypothesis is true, (2) the *p*-values are uniformly distributed, and (3) the tests are independent. Assumptions (2) and (3) are rather crucial here, but may be reasonable for the analysis of candidate genes that are expected to be only weakly linked, and when sample sizes are large. In gene expression tests, there is no linkage and the independence assumption might be more reasonable than in the case of gene–disease association tests.

Like the Simes-based tests, the Fisher combination-based test often allows several small *p*-values to reinforce one another to produce a more powerful test (than the min *P*-based method). The same $O(k^2)$ computational simplification seen for the closed Simes-based method described above holds for the closed Fisher combination method, making it also feasible for genetic association tests, and the method is available in PROC MULTTEST of SAS/STAT® (Version 8.1).

Pesarin ***(35)*** avoids the independence and uniformity assumptions, developing algorithms for exact Fisher combination tests that incorporate relevant distributional characteristics, including correlations. Further research is needed to develop computationally convenient closed testing algorithms that incorporate such tests.

3. Applications to Gene–Disease Associations

While one typically views the phenotype as a response (or penetrance) resulting from genetic predisposition, it is often reasonable (e.g., in case-control studies) to turn the problem on its head, and view genotype frequency as a function of the phenotype. In this section we apply the general closed testing methods described in **Subheading 2.** to specific genetic association tests, with the point of view of multiple comparisons of gene frequencies between cases and controls.

3.1. Multiple "Serological" Tests with Binary Phenotype

Consider the following 2×2 contingency table, cross-classifying disease status with presence of a particular allele at a given locus. The sample size is deliberately small to illustrate the main ideas.

Group	Allele A present	Allele A absent	Total
Case	5(100%)	0(0%)	5(100%)
Control	2(40%)	3(60%)	5(100%)

Sasieni ***(8)*** calls this a "serological" test because it "was common when HLA typing was done by serology, so that it was not possible to distinguish between [homozygous and heterozygous states]." He also notes that the resulting contingency table test (chi-square, χ^2) is completely efficient when allele A is dominant. Our analyses will consider the Fisher exact test instead of the χ^2 ***(36)***.

Now consider the following arrangement of the contingency table in a "flat file" representation amenable to computer input.

Subject	Group	*D1*
01	Case	1
02	Case	1
03	Case	1
04	Case	1
05	Case	1
06	Control	1
07	Control	0
08	Control	1
09	Control	0
10	Control	0

Here, *D1* stands for "dominance coding at locus 1," and the 0s and 1s denote presence or absence of allele *A* at that locus. Now, in genomic scans, for example, using SNPs ***(37)***, we will have multiple such indicators for a large collection of loci, resulting in a data set like that in **Table 1**, shown with just three loci for convenience.

For this data set, the Fisher exact (two-sided) *p*-values for testing associations between case-control status and locus are 0.1667, 1.0000, and 0.5238, respectively, for loci 1, 2, and 3. Nothing is significant, as expected with the small sample sizes; these values are used for illustration purposes only.

The closure principle described in **Subheading 2.1.** requires that additional *p*-values be computed for intersection hypotheses H_{12}: *D1* and *D2* are unaffected; H_{13}: *D1* and *D3* are unaffected; H_{23}: *D2* and *D3* are unaffected; and H_{123}: *D1, D2,* and *D3* are unaffected. By "unaffected" we mean that the distributions of the binary vectors are identical between Cases and Controls.

To calculate the exact closed min *P*-based multiple test procedure described in **Subheading 2.2.**, there are simplifications, and we require *p*-values only for the intersection hypotheses corresponding to the ordered *p*-values, H_{123}, H_{23}, and H_2. The *p*-value for H_{123} using the min *P* statistic is then $p_{123} = P(\min(P_1,P_2,P_3) \le 0.1667 \,|H_{123})$. To calculate this quantity exactly, one can enumerate all 10! permutations of the three-dimensional vectors, calculate $\min(P_1,P_2,P_3)$ for each permutation and note whether it is smaller than 0.1667, and take p_{123} to be the proportion of the 10! permutations (actually, only

Table 1
Input Form for Multiple Dominance Tests

Subject	Group	*D1*	*D2*	*D3*
01	Case	1	0	1
02	Case	1	0	1
03	Case	1	1	1
04	Case	1	0	0
05	Case	1	1	1
06	Control	1	0	1
07	Control	0	0	0
08	Control	1	1	1
09	Control	0	1	0
10	Control	0	0	0

10!/[5!5!] are required) yielding a min *P* smaller than 0.1667. Alternately, one can sample from the permutation distribution. The following table shows one random sample from the multivariate permutation distribution:

Subject	Group	*D1*	*D2*	*D3*
07	Case	0	0	0
02	Case	1	0	1
10	Case	0	0	0
01	Case	1	0	0
08	Case	1	1	1
03	Control	1	1	1
09	Control	0	1	0
05	Control	1	1	1
04	Control	1	0	0
06	Control	1	0	1

For this sample, the *p*-values are, respectively, 1.0000, 0.5238, and 1.0000, with min P=0.5238. Thus, this is one of the 10! permutations for which min *P* is not smaller than 0.1667. Sampling all permutations, 21.43% of the permutations yield min *P* smaller than 0.1667, so $p_{123}=0.2143$ is the exact *p*-value for the composite H_{123} when the min *P* test is used. According to the closed min *P* testing algorithm in **Subheading 2.2.**, the hypothesis H_1 (which happens to correspond to the smallest *p*-value) would not be rejected, and no further inference could be made. However, if H_1 were rejected, then we could proceed to test H_3 using the *p*-value $p_{23}=P(\min(P_2,P_3)\leq 0.5238|H_{23})$; H_3 would have been rejected if this probability were less than 0.05 (or whatever FWE level is chosen).

This analysis is automated in PROC MULTTEST of SAS/STAT®. The invoking code and testing portion of the output are as follows:

```
proc multtest data=table1 stepperm n=1000000 seed=121211;
  class group;
  test fisher(D1 D2 D3);
  contrast “compare” –1 1; run;
```

p-Values

Variable	Contrast	Raw	Stepdown Permutation
D1	Compare	0.1667	0.2149
D2	Compare	1.0000	1.0000
D3	Compare	0.5238	0.7855

The results of the closed testing algorithm are conveniently reported as adjusted *p*-values in the “Stepdown Permutation” column: if the adjusted *p*-value is <0.05, then the corresponding genetic association is significant at the FWE = 0.05 level using the closed min *P*-based testing algorithm of **Subheading 2.2.** Note also that the reported *p*-value 0.2149 differs slightly from the *p*-value 0.2143 obtained via direct enumeration of all 10! permutations; this difference reflects Monte Carlo error. As MULTTEST sampled 1,000,000 times, with replacement, from the population of permutations, the Monte Carlo standard error is just $\{0.2149(1-0.2149)/1000000\}^{1/2} = 0.00041$; thus the Monte Carlo estimate is 1.46 standard errors from the exact value, or acceptably close.

We have chosen 1,000,000 samples from the permutation distribution in this case, and the analysis takes less than a minute on a typical (as of the present date) PC workstation. In larger problems with more loci, it will take longer. We suggest at least 1000 samples to estimate the *p*-values with reasonable precision, although as large a number of samples as is convenient should ordinarily be chosen.

3.2. Testing Both Dominant and Recessive Modes of Inheritance

We may allow for recessive effects by considering 2 × 2 tables where genetic effect is coded as either (1) the gene is homozygous for the allele in question or (2) the gene is not homozygous for the allele in question. There is a high degree of dependence among such tests; this will be accommodated exactly in the closed multiple testing procedure. Following from **Table 1**, **Table 2** represents the input form suggested for such an analysis. Each gene has been coded two ways, with dominance coding *D* as shown in **Table 1**, and recessive coding *R*.

Note that there is positive correlation between the two codings, as a person who is “recessive” with respect to one allele is also “dominant” with respect to the other. There can also be strong positive correlation between closely linked genes owing to the linkage disequilibrium; nevertheless, these correlations are properly modeled via vector resampling as described previously.

Table 2
Dominant and Recessive Codings

Subject	Group	*G1*	*D1*	*R1*	*G2*	*D2*	*R2*	*G3*	*D3*	*R3*
01	Case	*AA*	1	1	*aa*	0	0	*AA*	1	1
02	Case	*AA*	1	1	*aa*	0	0	*AA*	1	1
03	Case	*AA*	1	1	*AA*	1	1	*AA*	1	1
04	Case	*AA*	1	1	*aa*	0	0	*aa*	0	0
05	Case	*AA*	1	1	*Aa*	1	0	*AA*	1	1
06	Control	*Aa*	1	0	*aa*	0	0	*AA*	1	1
07	Control	*aa*	0	0	*aa*	0	0	*aa*	0	0
08	Control	*Aa*	1	0	*AA*	1	1	*Aa*	1	0
09	Control	*aa*	0	0	*AA*	1	1	*aa*	0	0
10	Control	*aa*	0	0	*aa*	0	0	*aa*	0	0

Using the binary coding shown in **Table 2**, the specific hypotheses tested are $H_{0ij} : \pi_{1ij} = \pi_{2ij}$, where π_{1ij} denotes prevalence of coding i ($i = R,D$) for gene j ($j = 1,2,3$) among controls; and where π_{2ij} denotes the corresponding quantity among cases. These hypotheses again are testable using the two-sided Fisher exact test, and the exact closed testing method is applicable as well. Code and output follow.

```
proc multtest data=table2 stepperm n=1000000 seed=121211;
  class group;
  test fisher(D1 R1 D2 R2 D3 R3);
  contrast "compare" –1 1; run;
```

```
                              p-Values

                                                    Stepdown
  Variable       Contrast              Raw         Permutation

     D1           Compare             0.1667         0.3258
     R1           Compare             0.0079         0.0161
     D2           Compare             1.0000         1.0000
     R2           Compare             1.0000         1.0000
     D3           Compare             0.5238         0.7855
     R3           Compare             0.2063         0.3490
```

There are several points to make about the results. First, the recessive genotype at locus 1 is considered statistically significant at the FWE = 0.05 level using the exact min *P*-based closed testing procedure, as the Stepdown Permutation *p*-value is < 0.05. Second, the adjustment of the unadjusted *p*-value 0.0079 to the adjusted 0.0161 is substantially less than one might expect with Bonferroni correction (6 × 0.0079 = 0.0474); this savings comes as a result of using exact closed testing methods that incorporate discreteness as well as correlations. Third, it is somewhat unusual to find a more significant result when the family size is expanded, as we see here comparing the "dominance" analysis with three tests to the "dominance + recessive" analysis with six tests. However, when the expanded family contains tests that are more powerful, then it is certainly possible that there will be more significance in the expanded family, despite the larger multiple testing penalty. This example is suggestive of a situation where locus 1 has a purely recessive and fully penetrant effect.

This method can be extended to multiallelic genes as well. With multiple alleles the number of tests expands considerably: for $L > 2$ alleles, there will be $2L$ tests. (However, when $L = 2$ there are only two tests as, e.g., the dominant and recessive tests for allele *a* are completely determined by the corresponding tests for allele *A*.) Caution is recommended here, as large numbers of multiallelic genes can increase the family size substantially, thereby reducing power (in most cases).

3.3. Multiple Tests for Epistatic Effects

When two or more genes are necessary for the expression of the phenotype, we have an *epistasis*. It is thought that many complex traits and diseases are the result of the interaction of several rather common genotypes.

One possible method for screening gene combinations is to compare frequencies of the combinations occurring in either the case or the control populations. Let us revert to the "dominance" coding shown in **Table 1**, and consider whether combined effects of genes might signal differences in cases vs controls. The resulting data look like this:

Subject	Group	*D1*	*D2*	*D3*	*D1D2*	*D1D3*	*D2D3*
01	Case	1	0	1	0	1	0
02	Case	1	0	1	0	1	0
03	Case	1	1	1	1	1	1
04	Case	1	0	0	0	0	0
05	Case	1	1	1	1	1	1
06	Control	1	0	1	0	1	0
07	Control	0	0	0	0	0	0
08	Control	1	1	1	1	1	1
09	Control	0	1	0	0	0	0
10	Control	0	0	0	0	0	0

There are obviously correlations between the columns; in fact, in these data the *D1D3* column is identical to the *D3* column. The exact closed testing procedure automatically accounts for such dependencies, in effect reducing the multiplicative adjustment by one for each perfect dependency. The SAS code and output are as follows:

```
proc multtest data=table3 stepperm n=1000000 seed=121211;
```

```
class group;
test fisher(D1 D2 D3 d1d2 d1d3 d2d3);
contrast “compare” –1 1; run;
```

p-Values

Variable	Contrast	Raw	Stepdown Permutation
D1	Compar	0.1667	0.2864
D2	Compar	1.0000	1.0000
D3	Compar	0.5238	0.7855
D1D2	Compar	0.1667	0.2864
D1D3	Compar	0.5238	0.7855
D2D3	Compar	1.0000	1.0000

In this analysis, nothing would be considered significant, as none of the Stepdown Permutation (or closed exact Min *P* adjusted) *p*-values are less than the FWE 0.05 level. However, had there been a synergistic effect of two of these genes in dominant form, we might have seen some significant results.

One should be very cautious about using the multiplicative factors as shown here to discover epistatic effects; indiscriminant selection can greatly increase family size and thereby reduce power. For example, if there are 1000 genes, one might consider 1000(999)/2 = 499,500 possible combinations. It is preferred to keep the family size smaller; thus this method is suggested when the number of genes is small, say 100 or less (assuming reasonable sample sizes in the case and control groups).

3.4. Stratification

One can conceive of several situations in which gene–disease associations should be analyzed using stratification. Two cases of major importance are as follows:

1. Epistasis involving a known gene. A known gene, say *G1*, contributes to disease. However, there are questions of epistasis concerning other genes. In this case, the epistatic effects should not be modeled as illustrated in **Subheading 3.3.** as the prevalence of *G1Gi* will surely differ between cases and controls, owing simply to the main effect of *G1*. In such a case it will be appropriate to compare the prevalence of genotype *Gi* among patients who share a common value of *G1*.
2. Environmental factors. An environmental factor, smoking, for example, might be a known contributor to disease. In such a case, it would be better to assess genetic contributions by partialling out the smoking variable, both to improve sensitivity of the tests and to remove a possible source of confounding.

Stratified analyses can be handled in an exact fashion, using essentially the same methods as described in **Subheadings 3.1.–3.3.**, but using exact stratified (Mantel–Haenszel) tests instead of Fisher exact tests. Exact *p*-values for these tests (analogous to the Fisher exact *p*-values) are easily obtained using existing software. The hypotheses differ depending on whether one is using stratified or unstratified analysis; with stratified analysis, the composite null hypothesis states that the distributions of the binary vectors are identical for both groups *within each stratum* (although the distributions are allowed to differ between strata). To test these hypotheses, we permute the observation vectors as before, but independently within strata.

The following invocation of PROC MULTTEST uses the data in **Table 1**, and treats *D3* as if it were a known gene contributing to disease (in dominant form), and performs exact, closed, stratified multiple testing.

```
proc multtest data=table1 stepperm n=1000000 seed=121211;
  class group;
  strata d3;
  test ca(D1 D2/permutation=20);
  contrast "compare" –1 1; run;
```

p-Values

Variable	Contrast	Raw	Stepdown Permutation
D1	Compare	0.2500	0.3993
D2	Compare	1.0000	1.0000

In this example we find no significant difference of *D1* frequency between cases and controls when analyzed within groups defined by *D3* status. Of course, this is a very small data set; in practice, we might apply this to hundreds of candidate genes, after stratifying on one known gene.

The syntax "ca" in the preceding SAS code stands for "Cochran–Armitage Trend test" ***(38,39)***, which is equivalent to the Fisher exact test in the unstratified case, and which gives an exact stratified Fisher exact test in the stratified case. The syntax "permutation = 20" specifies exact permutation tests when the total number of observed *Gi* genotypes is < 20, or in this case, always. With large numbers of cases and controls, it is reasonable to specify "permutation = 100" or so to calculate exact permutation tests when the totals are < 100, but otherwise to use the normal approximation.

With sufficient sample size, this could be used as a forward stepwise procedure: select the most significant gene at step 1 (if significant by adjusted *p*-value); select the second major contributing gene (if significant by adjusted *p*-value), while partialling out the first as a "stratum" variable; select the third major contributing gene (if significant by adjusted *p*-value) while partialling the first and second selected variables as a combined "stratum" variable. This procedure has the attractive property that FWE is controlled at each stage, under the assumption of fixed ordering of variables. However, because the variables selected at earlier stages are random, it is possible that FWE is uncontrolled; *see* **ref. *40*** for further details in a related application.

Table 3
Ordinal Phenotype

Group	*D1*	*D2*	*D3*
Severe	1	0	1
Severe	1	0	1
Severe	1	1	1
Mild	1	0	0
Mild	1	1	1
None	1	0	1
None	0	0	0
None	1	1	1
None	0	1	0
None	0	0	0

3.5. Ordinal Phenotypes

Some phenotypic traits, for example, mental diseases, are best expressed ordinally, i.e., not diseased, mildly diseased, diseased, and badly diseased. One can perform a logistic regression of the binary genotype on the phenotypical outcome and test for significance of phenotype. The resulting logistic regression score test is equivalent to the Cochran–Armitage linear trend test that compares proportions of genotypes among the ordinal categories ***(8)***.

In our paradigm of conditioning on the phenotype and examining the distribution of the genotypes, such an analysis can easily be accommodated as shown in **Subheadings 3.1.–3.4.**, with the exception that the tests are based on the exact permutation distribution of the (possibly stratified) Cochran–Armitage test instead of the Fisher exact two-sample tests. The null hypothesis for any given set of genotypes is that the multivariate binary genotype distribution is identical across all phenotype categories; or equivalently, that all permutations of the *n* multivariate binary vectors are equally likely. The trend test statistics are most sensitive to linear (or at least monotonic) departures from the null.

The exact closed min *P*-based analysis shown in the previous subheadings can be performed just as easily in this case. Consider the data in **Table 1**, but with the “Case/Control” variable recoded as “Severe,” “Mild,” and “None” **(Table 3)**. The following SAS code and output show how to perform the exact, closed multiple testing procedure with these data.

```
proc multtest data=table4 order=data stepperm n=1000000 seed=121211;
   class group;
   test ca(D1 D2 D3/permutation=20);
   contrast “trend” –1 0 1; run;
```

p-Values

Variable	Contrast	Raw	Stepdown Permutation
D1	Trend	0.1417	0.2349
D2	Trend	1.0000	1.0000
D3	Trend	0.1667	0.3108

Here the "raw" *p*-values are exact Cochran–Armitage permutation *p*-values, and the Stepdown Permutation *p*-values are obtained by evaluating the distribution of min *P* over subsets corresponding to the ordered raw *p*-values, as described in **Subheading 3.1.**

Trend tests can be more powerful than Fisher exact tests that collapse the phenotype into two categories (e.g., here we might classify both "Severe" and "Mild" into "Case," and leave "None" as "Control.") However, one should limit the number of categories (say, to five or fewer); otherwise, the tables can become sparse and the tests can lose power. (Note: the MULTTEST procedure also computes exact stratified Cochran–Armitage trend tests for ordinal phenotype, as shown in **Subheading 3.3.** for the case of binary phenotype.)

3.6. Ordinal Genotypes: Cumulative Polygenic Effects

Suppose the disease is associated with allele *A* on a biallelic gene. In some disease models, the effect or penetrance of the gene is higher for heterozygotes than for homozygotes *aa*, while the effect or penetrance of the gene is still higher for homozygotes *AA* than for heterozygotes. This suggests a linear model relating phenotype *Y* to the ordinal genotype *X*, where $X = 0$, 1, or 2 for genotypes *aa*, *Aa*, and *AA*, respectively. Assuming *Y* is binary or ordinal as we have done, and again turning the problem on its head, we can compare the distribution of *X* for the different categories of *Y*, and test for trend, much as with the Cochran–Armitage test. Exact nonparametric tests for trend are available in, for example, StatXact ***(41)***, so, in principle, the problem of exact closed multiple tests is solved for this case as well. However, there is no ready-made software for this purpose. A reasonable solution is available in PROC MULTTEST where one uses the parametric test for genotype *i* to obtain *p*-values p_i, then finds the multiplicity-adjusted *p*-values $P(\min P_j \leq p_i)$. The final result is obtained by permuting vectors as before, so the final analysis is exact (modulo Monte Carlo error, which can be reduced to an arbitrarily low level). Using the parametric unadjusted *p*-values in the exact min *P* test can cause problems of imbalance ***(42,43)***; however, this problem is often not a major issue ***(44)***.

Another source of ordinal variables that can be handled similarly is the cumulative effect of several genes. For example, it may be thought that the

Table 4
Ordinal Genotypes

Subject	Group	*G1*	*X1*	*G2*	*X2*	*G3*	*X3*	*X4*
01	Case	*AA*	2	*aa*	0	*AA*	2	4
02	Case	*AA*	2	*aa*	0	*AA*	2	4
03	Case	*AA*	2	*AA*	2	*AA*	2	6
04	Case	*AA*	2	*aa*	0	*aa*	0	2
05	Case	*AA*	2	*Aa*	1	*AA*	2	5
06	Control	*Aa*	1	*aa*	0	*AA*	2	3
07	Control	*aa*	0	*aa*	0	*aa*	0	0
08	Control	*Aa*	1	*AA*	2	*Aa*	1	4
09	Control	*aa*	0	*AA*	2	*aa*	0	2
10	Control	*aa*	0	*aa*	0	*aa*	0	0

alleles A_1, A_2, and A_3 (in genes 1, 2, and 3, respectively) contribute cumulatively to the phenotype Y. In this case the variable $X_4 = X_1 + X_2 + X_3$ is suggested. Other codings are possible, such as $X_4 = I(X_1 > 0) + I(X_2 > 0) + I(X_3 > 0)$, where $I(\bullet)$ denotes the indicator function, as would be suggested if the disease is cumulatively related to dominant expressions of the A_i only, with no extra "bump" for recessivity. Note also that such a combination presupposes that the directions of allelic associations are known for all genotypes, which might be rare. Nevertheless, the method is shown below to illustrate the possibility, and to note that perfect dependencies of the type induced by the X_4 variable cause no problems with the exact min P-based testing method.

The data in **Table 4** relate directly to **Table 2**, with $X_1 – X_4$ as just described. Exact closed multiple testing is accomplished via the following code, and the results are shown as follows:

```
proc multtest data=table5 stepperm n=1000000 seed=121211;
  class group;
  test mean(X1–X4);
  contrast "compare" –1 1; run;
```

p-Values

Variable	Contrast	Raw	Stepdown Permutation
X1	Compare	0.0002	0.0161
X2	Compare	0.7599	1.0000
X3	Compare	0.1151	0.3490
X4	Compare	0.0497	0.1272

The result here is that the gene *G1* ordinal variable has a different mean for the two phenotypes, as its adjusted *p*-value is < 0.05.

In this analysis the “Raw” *p*-values are inexact, being based on the normal-theory *t*-test, but the Stepdown Permutation *p*-values are exact modulo Monte Carlo error. The inexactness of the Raw *p* is suggested by the fact that it is so small, relative to the exact multiplicity-adjusted *p*-value.

4. Application to a Large Simulated Data Set

The data for this study (ftp://statgen.ncsu.edu/pub/zaykin/cand/) were simulated according to the following model. We simulated a genetic map of 20 candidate regions. Each 100-kb candidate region contained 10 uniformly, randomly spaced SNPs. Candidate regions themselves were assumed unlinked; however, the recombination process for SNPs inside candidate regions was modeled directly, assuming Haldane’s mapping function (no interference) and Poisson-distributed number of recombination events with mean equal to the genetic length in Morgans. Three of 10 candidate regions contained disease genes. Four SNPs in each of first two regions and three SNPs in the third region were assumed to be contributing to the disease.

We used an additive model with weak interaction to model penetrances. According to this model, one allele for each of 11 SNPs was assigned a uniform random genetic effect, additively contributing to the total probability of developing disease (genetic penetrance), but the final penetrance for each genotype class was given a 0–5% uniform random deviation. Finally, $3^{11} = 177{,}147$ individual penetrances for individual multilocus genotypes were scaled between 0 and 1, so that the “typical” penetrance value of a multilocus genotype was about 50%.

We allowed for separate sexes, with no selfing, and no allowance of sib matings. Generations were assumed to be discrete. We simulated five originally homogeneous equilibrium populations of 500 individuals each, and allowed for 100–200 generations of genetic drift with population growth rate of 1.2, and a migration rate of 0.2 from each of the four populations into the fifth during the first 35 generations. The maximum population size was set to 15,000 individuals. We kept only populations with the final disease prevalence in the range 5–15%. We sampled 500 of affected and 500 of nonaffected individuals from the admixed population at the final generation.

To illustrate the method we simulated sets of data using two different models. The first model (model 1) is as described in the preceding. The second model (model 2) differs in that the chromosome regions are themselves closely linked, so that the association may extend over all 20 regions.

Table 5 contains part of the resulting analysis of a typical data set (20 smallest *p*-values) simulated under the first model, and **Table 6** contains part of the analysis for data that were simulated under the second model (200 generations). The

Table 5
***p*-Values for Simulated Data, Model 1**

Genotype	Unadjusted *p*-value	Closed Bonferroni	Closed Sidak	Closed Permutation
D7	0.0000000	0.00000	0.00000	0.000
D9	0.0000000	0.00000	0.00000	0.000
R9	0.0000000	0.00000	0.00000	0.000
D10	0.0000000	0.00000	0.00000	0.000
R10	0.0000000	0.00000	0.00000	0.000
R7	0.0000006	0.00019	0.00019	0.000
R5	0.0000198	0.00695	0.00692	0.005
D5	0.0000723	0.02548	0.02515	0.015
D8	0.0002896	0.10276	0.09767	0.056
R132	0.0004779	0.16953	0.15597	0.091
R17	0.0008843	0.31228	0.26832	0.201
D133	0.0019127	0.67437	0.49084	0.384
R3	0.0023851	0.83925	0.56838	0.452
D127	0.0024807	0.86887	0.58101	0.462
D70	0.0026634	0.92573	0.60423	0.482
D113	0.0034234	1.00000	0.69534	0.572
D136	0.0034387	1.00000	0.69570	0.572
R8	0.0039253	1.00000	0.74653	0.635
R172	0.0046352	1.00000	0.80447	0.693
D129	0.0095221	1.00000	0.96444	0.915

analysis for both tables was performed using PROC MULTTEST, Version 8.1 (an example of the invoking program code is given in the Appendix).

Actual regions contributing to the probability of developing the disease were typed with markers labeled 1–15, so the algorithm correctly identifies SNPs typed in all three regions. Originally small *p*-values corresponding to the false regions become nonsignificant after proper multiplicity adjustment over the set of 400 tests. Note that closed permutation *p*-values are smaller than the closed Bonferroni and Sidak (independence-assuming) corrections. The effect is more pronounced when long regions of densely mapped SNPs are considered **(Table 6)**. For example, the closed Bonferroni-adjusted *p*-value for R134 is 0.0089834, but corresponding exact (modulo Monte Carlo error) min P permutation adjustment is 0.002.

5. Application to Gene Expression Data

Gene expression data may be analyzed using similar techniques. Data given in Golub et al. ***(45)*** are available at http://waldo.wi.mit.edu/MPR/data_set_ALL_AML.html) for relating gene expression from 7129 genes to

Table 6
***P*-Values for Simulated Data, Model 2**

Genotype	Unadjusted *p*-value	Closed Bonferroni	Closed Sidak	Closed Permutation
R138	0.000014178	0.0043713	0.0043618	0.001
D42	0.000014739	0.0045309	0.0045207	0.001
R83	0.000017126	0.0052764	0.0052625	0.002
R115	0.000018588	0.0057627	0.0057462	0.002
R178	0.000022408	0.0068857	0.0068621	0.002
D194	0.000026979	0.0082854	0.0082512	0.002
D100	0.000028380	0.0087390	0.0087011	0.002
R134	0.000029131	0.0089834	0.0089432	0.002
D6	0.00003150	0.009618	0.009572	0.003
R191	0.00003233	0.009875	0.009827	0.004
R149	0.00003882	0.011811	0.011741	0.005
R73	0.00003975	0.012020	0.011949	0.006
D123	0.00004579	0.013792	0.013697	0.007
R100	0.00006178	0.018604	0.018433	0.009
D130	0.00006465	0.019422	0.019236	0.009
D21	0.00006529	0.019568	0.019378	0.009
R22	0.00007886	0.023409	0.023138	0.010
R104	0.00008572	0.025422	0.025102	0.011
D177	0.00008980	0.026545	0.026196	0.013
R173	0.00010426	0.030858	0.030388	0.015

disease status. (Golub et al. ***[45]*** apparently consider only 6817 of the 7129 available on the data set.) There are 11 patients with acute myeloid leukemia (AML) and 27 with acute lymphoblastic leukemia (ALL).

As discussed in **ref. *45***, one goal is to discriminate between the known AML and ALL populations on the basis of the observable gene expressions. Discriminant analysis (DA) is commonly used for this purpose, and a first step in DA is often to test for differences between the groups using Hotelling's T^2 test ***(46)***. However, the T^2 test requires a nonsingular covariance matrix, and in this case the 7129×7129 covariance is quite singular, having rank somewhere near $11+27=38$, and the test cannot be applied. Nevertheless, the min P test can be carried out exactly to test for global differences; in addition, the exact min P-based closed testing procedure allows one to specify particular genes where simple associations exist, with full FWE protection.

In the gene expression data the response variable is continuous, and exact, distribution-free closed testing methods are available, as described in **Subheading 3.5.** One may test a global hypothesis using the max T statistic, where T is calculated as

$$T = \frac{\bar{X}_{ALL} - \bar{X}_{AML}}{s_p\sqrt{1/11 + 1/17}}$$

where $\bar{X}_{ALL}$ and $\bar{X}_{AML}$ refer to average expression in the ALL and AMR groups, and s_p is the pooled standard deviation. However, because max T is monotonically related to min P, where the p-values are calculated using the t -distribution with df = 11 + 27 – 2, this method is exactly equivalent to the min P testing method described in **Subheading 3**. PROC MULTTEST accomplishes this by resampling the 7129-dimensional vectors of gene expressions without replacement into like data sets having 27 ALL and 11 AML patients, then recomputing max T^* for the resampled data set. The p-value for the appropriate intersection hypothesis is then reported as the proportion of resampled data sets yielding max T^* greater than the original observed max T.

Golub et al. ***(45)*** performed a related permutation based-analysis using the statistic $T^{/} = (\bar{X}_{ALL} - \bar{X}_{AML})/(s_1 + s_2)$. It would be equally possible to perform the exact closed testing procedure max $T^{/}$ as the base test, if desired. The benefits of the MULTTEST analysis are that (1) it is easily available and (2) it is known to control FWE via the closure principle.

Note that, as in the case of gene–disease tests, vector correlations are incorporated via vector resampling. However, in the case of gene expression data, there is no linkage, and therefore large correlations are not expected. Nevertheless, there is sample-specific dependence because the number of variables far exceeds the number of observations. This dependence is used to reduce the p-values, legitimately, because the tests are exact. Furthermore, as noted previously, the sample 7129 × 7129 covariance matrix among the gene expressions is massively singular, but this poses no difficulties whatsoever; the exact multiple testing procedure legitimately incorporates such sample-specific dependencies into the multiplicity adjustments via vector permutation resampling.

Such a test is an exact permutation test when all $\binom{38}{27}$ distinct resampled data sets (more than a billion) are enumerated. However, a reasonable approximation can be obtained by sampling randomly and with replacement from that set of permutations, and this is the PROC MULTTEST approach for testing global hypotheses. To make inferences about the specific genes, the closure method is used, and once again, only the subsets corresponding to the 7129 ordered p-values need to be evaluated, not the entire set of 2^{7129} subsets. Thus, once again, the MULTTEST procedure provides a closed testing method that is computationally feasible.

Table 7
***p*-Values for Golub Leukemia Data Set**

	Unadjusted		
Gene	*p*-value	Bonferroni–Holm	Closed Min *P*
GENE3320	1.3824E-10	0.000001	0.0001
GENE4847	2.4355E-10	0.000002	0.0001
GENE2020	6.578E-10	0.000005	0.0001
GENE1745	0.000000010	0.000070	0.0004
GENE5039	0.000000010	0.000072	0.0004
GENE1834	0.000000015	0.000108	0.0005
GENE461	0.000000036	0.000257	0.0005
GENE4196	0.000000062	0.000438	0.0009
GENE3847	0.000000072	0.000510	0.0010
GENE2288	0.000000089	0.000635	0.0011
GENE1249	0.000000174	0.001239	0.0017
GENE6201	0.000000176	0.001250	0.0017
GENE2242	0.000000195	0.001386	0.0020
GENE3258	0.000000211	0.001500	0.0021
GENE1882	0.000000319	0.002267	0.0024
GENE2111	0.000000366	0.002606	0.0027
GENE2121	0.000000578	0.004115	0.0041
GENE6200	0.000000623	0.004428	0.0042
GENE6373	0.000000819	0.005823	0.0058
GENE6539	0.000001120	0.007961	0.0082
GENE2043	0.000001260	0.008954	0.0092
GENE2759	0.000001309	0.009304	0.0092
GENE6803	0.000001429	0.010156	0.0101
GENE1674	0.000001480	0.010519	0.0103
GENE2402	0.000001523	0.010821	0.0107
GENE2186	0.000001657	0.011770	0.0111
GENE6376	0.000002092	0.014856	0.0142
GENE3605	0.000002553	0.018133	0.0157
GENE6806	0.000002584	0.018352	0.0159
GENE1829	0.000002727	0.019364	0.0168
GENE6797	0.000003014	0.021399	0.0180
GENE6677	0.000003439	0.024412	0.0196
GENE4052	0.000003701	0.026268	0.0220
GENE1394	0.000004925	0.034948	0.0282
GENE6405	0.000005353	0.037980	0.0300
GENE248	0.000006381	0.045267	0.0346
GENE2267	0.000006488	0.046019	0.0352
GENE6041	0.000007802	0.055335	0.0421
GENE6005	0.000008019	0.056861	0.0428
GENE5772	0.000008994	0.063771	0.0471
GENE6378	0.000009591	0.067993	0.0500

The results are given in **Table 7** and the invoking MULTTEST code is given in the Appendix. Surprisingly, several results are significant, despite the small sample sizes and large degree of multiplicity. The association of leukemia subtype with the expression phenotype is confirmed; tests with closed permutation-based adjusted *p*-values < 0.05 indicate significant associations at the 0.05 FWE level.

Also note that 10,000 samples are generated from the permutation distribution, and all 7129 ordered tests were processed for each sample. This took only 20 min on a Windows NT workstation.

The effect of incorporating the sample-specific dependencies among the *p*-values is not as great as one might hope with such massive singularity in the covariance matrix. Naively, one might expect (or hope) that the effective Bonferroni multiplier would be on the order of 38 = 27 + 11, the approximate rank of the 7129 × 7129 covariance matrix, when the dependence structure is incorporated correctly. However, this is not so. Dividing the adjusted *p*-values by the unadjusted *p*-values gives the effective multipliers; for example, the effective multiplier for the test involving "GENE248" is 0.0445267/0.000006381 = 7094.0 for the Bonferroni–Holm procedure, but only 0.0346/0.000006381 = 5422.2 for the exact min *P*-based closed procedure. The savings from using the correlation structure is to reduce the multipliers some, but not nearly to the extent suggested by the rank of the covariance matrix.

The Simes–Hommel method described in **Subheading 2.3.** also was applied to these data; the results were almost identical to Bonferroni–Holm, but the run took nearly 24 h because of the large number of tests.

Finally, we note that there are occasionally extreme outliers in the gene expression data. The negative effects of outliers can be diminished through log transformation as in **ref. *45*** or one can use the rank transformation to avoid taking the logarithm of numbers that are less than or equal to zero. Use of the rank transformation in conjunction with permutation resampling in PROC MULTTEST provides an exact permutation-based closed testing procedure as before. However, this procedure is also attractive because the marginal tests are approximately valid rank-based permutation tests as well, being based on the rank transform ***(47)***. When the analysis is performed on the rank-transformed expression data, there are a few changes in **Table 7**, mostly additions of variables where a large outlier masked the difference using the two-sample *t*-test.

Appendix

The following SAS/STAT® code was used to produce the results shown in **Tables 5** and **6**. It is assumed that the input file (in this case "SNP.DAT") has the result of each gene test coded as *AA* = 1 1, *Aa* = 1 0, *aA* = 0 1, and *aa* = 0 0;

and has the binary phenotype in the first column. The macro “%trans” recodes these data into “*AA* vs not *AA*” and “*aa* vs not *aa*” categories.

```
   %macro trans;
      %do i = 1 %to 200;
      %let i1 = %eval(2*&i-1);
      %let i2 = %eval(2*&i);
      d&i = (bin&i1+bin&i2)=0;
      r&i = (bin&i1+bin&i2)=2;
   %end;
%mend;
data snp;
   infile “snp.dat” lrecl=10000;
   input y bin1-bin400;
   %trans;
   keep y d1-d200 r1-r200;
   run;
proc multtest data=snp stepbon stepsid noprint out=pval stepperm n=10000;
   class y;
   test ca(d1-d200 r1-r200/permutation=100);
   contrast “dis v nondis” 0 1;
run;
proc sort data=pval;
   by raw_p;
proc print data=pval;
   var _var_ raw_p stpbon_p stpsid_p stppermp;
   where raw_p<.05;
run;
```

The following SAS/STAT® code was used to analyze the data shown in Table 7. It is assumed that the SAS data set (“gene.express”) has the result of each gene expression test in variable GENEi, and that the treatment indicators (AML or ALL) are contained in the variable called “disease.”

```
proc multtest data = gene.express out=adjp stepperm holm n=10000 noprint;
   class disease;
   test mean(gene1-gene7129);
   contrast “AML vs ALL” –1 1;
run;
proc sort data=adjp(where=(stppermp le .05));
   by raw_p;
proc print data=adjp(where=(stppermp le .05)) noobs label;
   var _var_ raw_p stpbon_p stppermp;
run;
```

References

1. Bevan, S., Popat, S., and Houlston, R. S. (1999) Relative power of linkage and transmission disequilibrium test strategies to detect non-HLA linked coeliac disease susceptibility genes *Gut* **45,** 668–671.
2. Barnes, K. C. (1999) Gene-environment and gene-gene interaction studies in the molecular genetic analysis of asthma and atopy. *Clin. Exp. Allergy* **29** (Suppl 4)**,** 47–51.
3. El-Gabalawy, H. S., Goldbach-Mansky, R., Smith, D., Arayssi, T., Bale, S., Gulko, P., et al. (1999) Association of HLA alleles and clinical features in patients with synovitis of recent onset. *Arthrit. Rheum.* **42,** 1696–1705.
4. Tomer, Y., Barbesino, G., Greenberg, D. A., Concepcion, E., and Davies, T. F. (1999) Mapping the major susceptibility loci for familial Graves' and Hashimoto's diseases: evidence for genetic heterogeneity and gene interactions. *J. Clin. Endocrinol. Metab*. **84,** 4656–4664.
5. Wicker, L. S., Todd, J. A., and Peterson, L. B. (1995) Genetic control of autoimmune diabetes in the nod mouse. *Annu. Rev. Immunol.* **13,** 179–200.
6. Bodmer, W. F. (1986) Human genetics: the molecular challenge. *Cold Spring Harbor Symp. Quant. Biol.* **51,** 1–13.
7. Hagmann, M. (1999) A good SNP may be hard to find. *Science* **285,** 21–22.
8. Sasieni, P. D. (1997) From genotypes to genes: doubling the sample size. *Biometrics* **53,** 1253–1261.
9. Chiano M. N. and Clayton D. G. (1998) Fine genetic mapping using haplotype analysis and the missing data problem. *Ann. Hum. Genet.* **62,** 55–60.
10. Miller, R. C. (1981) *Simultaneous Statistical Inference*, 2nd edit. Springer-Verlag, New York.
11. Smouse, P. E. and Williams, R. C. (1982) Multivariate analysis of HLA–disease association. *Biometrics* **38,** 757–768.
12. Lander, E. and Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet*. **11,** 241–247.
13. Zaykin, D. V., Zhivotovsky, L. A., Weir B. S., and Westfall, P. H. (2000) Truncated product method for combining *p*-values. Unpublished manuscript.
14. Weller, J. I., Song, J. Z., Heyen, D. W., Lewin, H. A., and Ron, M. (1998) A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150,** 1699–1706.
15. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate — a practical and powerful approach to multiple testing. *JRSS-B* **57,** 289–300.
16. Zaykin, D. V., Young, S. S., and Westfall, P. H. (2000) Using the false discovery rate approach to the genetic dissection of complex traits: a response to Weller et al. *Genetics* **154,** 1917–1918.
17. Marcus, R., Peritz, E., and Gabriel, K. R. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63,** 655–660.
18. Churchill, G. A. and Doerge, R. W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138,** 963–971.
19. Doerge, R. W. and Churchill, G. A. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142,** 285–294.

20. Westfall, P. H. and Wolfinger, R. D. (2000) Closed Multiple Testing Procedures and PROC MULTTEST. SAS Observations, http://www.sas.com/service/library/periodicals/obs/observations.html.
21. Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6,** 65–70.
22. Westfall, P. H. and Wolfinger, R. D. (1997) Multiple tests with discrete distributions. *Am. Stat.* **51,** 3–8.
23. Westfall, P. H. and Young, S.S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. John Wiley & Sons, New York.
24. SAS Institute Inc. (1999) SAS OnlineDoc ®, Version 8, Cary, NC: SAS Institute Inc.
25. Simes, R. J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73,** 751–754.
26. Hommel, G. (1988) A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75,** 383–386.
27. Wright, S. P. (1992) Adjusted *p*-values for simultaneous inference. *Biometrics* **48,** 1005–1014.
28. Grechanovsky, E. and Hochberg, Y. (1999) Closed procedures are better and often admit a shortcut. *J. Stat. Plan. Infer.* **76,** 79–91.
29. Sarkar, S. (1998) Some probability inequalities for ordered MTP_2 random variables: a proof of the Simes conjecture. *Ann. Statist.* **26,** 494–504.
30. Sarkar, S. and Chang, C. K. (1997) Simes' method for multiple hypothesis testing with positively dependent test statistics. *JASA* **92,** 1601–1608.
31. Krummenauer, F. and Hommel, G. (1999) The size of Simes' global test for discrete test statistics. *J. Stat. Plan. Infer.* **82,** 151–162.
32. Dunnett, C. W. and Tamhane, A. C. (1993) Power comparisons of some step-up multiple test procedures. *Statist. Prob. Lett.* **16,** 55–58.
33. Dunnett, C. W. and Tamhane, A. C. (1995) Step-up multiple testing of parameters with unequally correlated estimates. *Biometrics* **51,** 217–227.
34. Fisher, R. A. (1932) *Statistical Methods for Research Workers*. Oliver and Boyd, London.
35. Pesarin, F. (1999) *Permutation Testing of Multidimensional Hypotheses by Nonparametric Combination of Dependent Tests*. CLEUP University Publisher, Padova.
36. Weir, B. S. (1996) *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.
37. Martin, E. R., Lai, E. H., Gilbert, J. R., Rogala, A. R., Afshari, A. J., Riley, J., et al. (2000.) SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *Am. J. Hum. Genetics* **67,** 383–394.
38. Cochran, W. (1954) Some methods for strengthening the common χ^2 tests. *Biometrics* **10,** 417–451.
39. Armitage, P. (1955) Tests for linear trend in proportions and frequencies. *Biometrics* **11,** 375–386.
40. Westfall, P. H., Young, S. S., and Lin, D. K. J. (1997) Forward selection error control in the analysis of supersaturated designs. *Statist. Sinica* **8,** 101–117.

41. Mehta, C. and Patel, N. (1998) *StatXact: statistical software for exact non-parametric inference*. CYTEL Software, Cambridge, MA.
42. Beran, R. (1988) Balanced simultaneous confidence sets. *JASA* **83,** 679–686.
43. Beran, R. (1988) Prepivoting test statistics: a bootstrap view of asymptotic refinements. *JASA* **83,** 687–697.
44. Tu, W. and Zhou, X. H. (2000) Pairwise comparisons of the means of skewed data. *J. Stat. Plan. Infer.* **88,** 59–74.
45. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537.
46. Johnson, R. A. and Wichern, D. W. (1998) *Applied Multivariate Statistical Analysis*, 4th edit. Prentice Hall, Englewood Cliffs, NJ.
47. Conover, W. J. and Iman, R. L. (1981) Rank transformation as a bridge between parametric and nonparametric statistics. *Am. Statist.* **35,** 124–129.