2-20-2007

Power Boosting in Genome-Wide Studies Via Methods for Multivariate Outcomes

Mary J. Emond
University of Washington, emond@u.washington.edu

1. Introduction With increasingly refined genomic knowledge and technology, whole-genome studies are becoming a mainstay of biomedical research. Examples include expression array experiments, comparative genomic hybridization analyses and case-control studies for detecting single nucleotide polymorphism (SNP)/disease associations. Because of the large number of loci studied per subject and the resulting numbers of statistics that are generated, statistical methods for controlling the number of potentially spurious findings are needed. While the idea of controlling the False Discovery Rate (FDR)(Reiner et al, 2003; Storey and Tibshirani, 2003) or Positive False Discovery Rate (pFDR)(Storey, 2002) has become popular, existing implimentations of these methods might be inadequate for some purposes. For example, when used in practice, the true FDR (or pFDR) is unknown and is replaced by an estimated proportion. Methods for estimating the FDR (or pFDR) are subject to varying degrees of bias and variation (Pawitan et al, 2005; Broberg, 2005; Owen, 2005), and their behavior can depend on the dependence stucture of the data. One popular pFDR method has been shown to markedly underestimate the false discovery rate in the presence of a modest number of true positives and modest correlation between test statistics (Yang et al, 2005). The development of confidence intervals for FDR estimates in the presence of correlated data will help alleviate this uncertainty. In the meantime, the Bonferroni correction remains a reliable method for controlling the marginal false positive rate. This method is employed in a multi-stage design to detect SNP-disease associations in the Women's Health Initiative (Prentice and QI, 2006). The well-known shortcoming of the Bonferroni correction is that it is conservative, controlling the family-wise type I error at lower than the nominal level, though the exact level of control is unknown. The Bonferroni correction is least conservative when applied to uncorrelated test statistics. Moderate to high levels of correlation among some groups of test statistics can be expected in most whole-genome studies. Potential statistical power is lost when this correlation is ignored.

In this article, we show that regression methods for misspecified multivariate outcomes can be applied in whole-genome studies in order to gain power over methods that model the outcome independently for each locus. (Note that regression for multivariate outcomes should not be confused with the phrase "multivariate logistic regression" that refers to a univariate outcome modeled as a function of more than one independent variable. The phrase is commonly used in elementary biostatistics textbooks.) In situations such as a case-control study where the outcome usually is not regarded to be multivariate (case versus control status), it is mathematically valid to view the outcome as a set of (identical) repeated measurements, one associated with each genetic locus. When applying multivariate outcome methods, it is common to regard all data collected on one subject as a belonging to a single multivariate observation. Instead, we apply multivariate methods to subgroups of outcome/covariate pairs. This represents an intermediate between determining a joint distribution for all test statistics and making the assumption of independence among all statistics, with the latter being a special case of the proposed method. We provide some guidelines on how to form these subgroups below. In fact, most of the thousands of test statistics generated in genome-wide studies can be expected to have negligible pair-wise correlations, and it is likely of little use to estimate the entire joint distribution even if it were practically possible. For example, in a SNP-disease association study, only test statistics for SNP locus pairs that are close spatially or that are both related to the outcome are expected to be correlated. Hence, effective joint modeling of such SNPs within a relatively small subgroup ("cluster") should provide much of the possible power gain that one could ideally obtain. (For brevity, we use the terms "SNP" and "SNPs" to mean SNP locus and SNP loci, respectively, where appropriate.)

Any multivariate estimation method that remains valid under misspecification of the correlation structure could be applied in the manner we describe, but we focus here on the use of generalized estimating equations (GEE) for analysis of case-control studies. GEE can provide power gains both by taking into account correlation during the estimation procedure to produce smaller standard errors for single SNP test statistics and by producing an estimated joint distribution. However, it is believed that GEE is not statistically appropriate for whole-genome studies because the number of outcomes far exceeds the number of subjects. This belief is not true, in general. GEE and the General Linear Model (GLM) comprise families of models with different correlation structures. Estimates obtained from these techniques will be valid in the setting of many more outcomes than subjects as long as the modeled correlation between most pairs of observations is weak and valid standard error estimates are used (c.f. Lumley and Haegerty, 1999). For example, one particular formulation of GEE (with the identity matrix as the working correlation matrix) is equivalent to performing logistic regression for every SNP in a case-control study. By modeling only subgroups of data jointly, we maintain conditions for valid estimation of test statistics.

In the rest of this paper, we provide a detailed description of how to implement the proposed method using GEE software to analyze data from a case-control study of SNP-disease association. This description is intended to be accessible to non-statisticians. We then report the design and results for a large simulation study of the method's operating characteristics and follow with discussion.

2. Methods In this section we describe how one can apply GEE with assumptions of non-independence to increase statistical power over the application of independence-based logistic regression models to each SNP locus. We show how one may adjust for confounding such as population structure, if desired. The utility of adjusting is addressed in the Discussion. This section also describes the extensive simulation studies performed for the method. For simplicity of presentation and analysis of the method's properties, we assume that the chosen genetic model assigns a score of either 0 or 1 to each locus, and the goal is to find loci for which the odds of disease is elevated when the score is 1 relative to a score of 0. (No loss of generality for the qualitative results of this article should occur by assuming this basic model, since more complex models can be reduced to this case.) We assume that strong control of the family-wise Type I error rate is desired.

2.1 Modeling and Estimation with GEE To apply GEE, one chooses a model for each locus-outcome pair that would be appropriate if that locus-outcome were the only one studied. In practice, the same type of model is usually employed for all locus/outcome pairs. For our case-control example, the outcome, Y_i, for the i^{th} subject is either case or control status ($Y_i = 1$ or 0, respectively). In both the independence method applied by Prentice and QI (2006) and the non-independence method proposed here, and a logistic model of the following form is assumed for each SNP score:

$$\text{Prob}[Y_i = 1|Sj_i] = \frac{\exp(\alpha_j + \beta_j Sj_i)}{1 + \exp(\alpha_j + \beta_j Sj_i)} \tag{1}$$

where Sj_i is the genetic score at the j^{th} SNP locus for the i^{th} subject. Note that a separate intercept, α_j, and a separate association coefficient, β_j, is estimated for each SNP locus. (This has practical implications for specification of the design matrix when fitting the models in GEE (see below)).

Details regarding estimation in multivariate regression models under misspecified correlation structure can be found in White (1982) and Gourieroux et al (1984) with further extensions appearing as GEE in Liang and Zeger (1986). For the purposes of this article, the two essential ingredients for increasing power are to (1) group correlated observations (within subject) into what are denoted as "clusters" in the statistical literature; and (2) to choose a working correlation matrix structure that captures at least part of the true correlation structure in the data. More specifically, for a cluster of n correlated observations on subject i, the working correlation matrix, R, specifies the pair-wise correlations between the centered observations:

$$\text{cor} \begin{bmatrix} Y_{i1} - f(\alpha_1 + \beta_1 S1_i) \\ Y_{i2} - f(\alpha_2 + \beta_2 S1_i) \\ Y_{i3} - f(\alpha_3 + \beta_3 S2_i) \\ \vdots \\ Y_{in} - f(\alpha_n + \beta_n Sn_i) \end{bmatrix} = R = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{nn} \\ r_{21} & 1 & r_{23} & \cdots & r_{2n} \\ r_{31} & r_{32} & 1 & \cdots & r_{3n} \\ \vdots & & & & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \cdots & 1 \end{bmatrix}, \tag{2}$$

where f is the logistic transformation used in model (1), given by $f(x) = exp(x)/(1 + exp(x))$, the SNP loci assigned to this particular cluster have been numbered from S1 to Sn, and R is a symmetric correlation matrix, so that $r_{kl} = r_{lk}$ and $|r_{jk}| \leq 1$. Note that in our case-control example, $Y_{ij} = Y_{ik} = Y_i$ for all j and k since the case-control status is fixed for the i^{th} subject. In the standard GEE modeling situation, observations in different clusters are assumed to be uncorrelated. Because designated clusters for this method need not be independent under the true state of nature for the testing method proposed here to be valid, we replace the term "cluster" with "subgroup" to avoid confusion.

As shown in Gourieroux et al (1984), the correlation structure chosen for the estimation procedure need not be the true correlation structure in order for the estimation to be valid, as long as "robust" standard errors of the coefficient estimates are obtained via a "sandwich estimator." As the chosen working correlation matrix comes closer to the truth, we can expect more efficient estimation of the association parameter in β_j in (1). In the setting of genomic data, non-independence working correlation matrices will generally be closer to the truth for subgroups of data than an independent working correlation matrix. Note that the ordering of the observations within the subgroup is meaningful when the working correlation matrix specifies higher correlation between observations that are placed closer together, as in the case of the AR-1 working correlation matrix. The order of the observations is unimportant when the exchangeable working correlation matrix is used.

2.2 Applying the Method We next provide a step-by-step description of the proposed method in application. K represents the total number of SNP loci that are tested for association with disease, and N is the total number of subjects.

1. For each subject, replicate the case or control status (the value of Y) K times to create the observations $(Y_i, Sj_i), j = 1 \ldots K, i = 1 \ldots N$.

2. Place the K SNPs into a smaller number of subgroups. Each subgroup of SNPs will define a set of observations to be modeled according to (1) and (2) for each subject. The goal is to place SNPj and SNPk in the same subgroup when it is suspected that there may be high correlation between $Y_i - f(\hat{\alpha_j} + \hat{\beta}_j Sj_i)$ and $Y_i - f(\hat{\alpha_k} + \hat{\beta}_k Sk_i)$. In

practice, this means that the regression method will be applied to each subgroup of SNPs as though the SNPs in the subgroup were the only observed SNPs. *This is analogous to the application of a separate logistic regression model for each SNP, except that now there is a separate GEE application for each subgroup of SNPs.* Note that a SNP may be placed in more than one subgroup, as described below, if appropriate. For purposes of algorithm convergence, computational speed and asymptotic validity, subgroups of size 6 to 20 are expected to work well when the total sample size is 400 or more. In a statistical sense, the goal of the subgrouping is to obtain test statistics that have little pairwise correlation in and across subgroups.

(a) Place SNPs in the same subgroup if they are known to be close spatially or if they are known to be or suspected of being close in function. SNPs in the same cluster could be:

   i. SNPs that are linked so as to result in positive correlation between the scores at these loci in a random sample of the population (correlation greater than approximately .10), including SNPs in the same haplotype block

   ii. SNPs in exons for the same gene

   iii. SNPs for genes in the same pathway

   iv. SNPs in genes encoding proteins that are later spatially linked (as in anti-body formation)

   v. SNPs from genes that are empirically related to the outcome, say, via segregation analyses or HLA associations

   vi. Potential effect modifiers such as smoking history; these may be thought of as "auxiliary SNPs" and placed in the same subgroup with SNPs whose action they are thought to modify.

(b) a SNP may appear in more than one analytic subgroup, since a given SNP might be part of more than one functional group.

3. Choose a working correlation structure that is expected to capture at least some of the true correlation structure in the data. Either "auto-regressive-lag-1" (AR-1) or "exchangeable" are recommended choices (see appendix), and both are available in all standard GEE software packages.

4. Choose the "fixed" option for the working correlation matrix. This means that the parameters for the working correlation matrix are supplied by the user and not estimated.

5. Organize the data for input to the GEE software. This includes ordering the observations within a subgroup if not using the exchangeable working correlation matrix. Also, readying the data for input to the algorithm usually also will include the creation of an augmented matrix of independent variables in order to generate separate intercept and regression coefficient estimates (α_j and β_j) for each SNPj in the subgroup. This is necessary if the software is designed only to accomodate the same variables for every observation. In this case, when no adjustment variables are included, all but two entries in each row of the augmented independent variable matrix must be made identically zero, resulting in a matrix with nN rows and $2n$ columns. The "no intercept" option should be specified, since there is no intercept common to all observations. More details on the creation of this augmented matrix are given in the Appendix. Finally, the software will require the creation of an ID vector of length nK that indicates which observations are from the same subject.

6. Apply GEE with the fixed working correlation matrix to each subgroup and obtain the Wald statistic for each SNP's coefficient and its corresponding p-value. (The Wald statistic is the coefficient estimate divided by its standard error, $\hat{\beta}/s.e.(\hat{\beta}) \equiv GW1$, and the p-value is obtained by comparing this statistic to a standard normal distribution. Both are provided by any software package).

7. Apply a Bonferroni correction to the K p-values to obtain a correction for multiple testing.

Comments:

1. SNPs with scores that are negatively correlated *in the sample* should not be grouped together, as power loss can obtain in this situation.

2. We have retained the use of the Bonferroni correction to maintain control over the false positive rate. However, when effective subgrouping is done, the Bonferroni correction applied to the GEE-based statistics will be less conservative than using this correction on the set of test statistics obtained by univariate modeling. This happens because the estimation procedure produces test statistics that are more nearly uncorrelated than does the independence-based estimation procedures.

3. If a SNP appears in more than one subgroup, the user should choose only one test statistic for that SNP to be the primary test statistic. This test statistic should come from a subgroup in which the SNP is considered to have a potential main effect. The SNP may appear in other subgroups as an effect modifier based on known/suspected function or as an entry akin to a "precision variable" if the SNP score is correlated (linked) to the SNP(s) of main interest in these other subgroups. The effect of the "precision variable" is made more clear in the Results.

3. The subgroups may be of different sizes.

4. The choice of working correlation matrix can be different for different subgroups.

5. It is recommended to choose a fixed working correlation matrix in order to speed the estimation and possibly avoid problems with convergence of the estimation algorithm. This recommendation is facilitated by the fact that specifying high correlation between observations still provides a power advantage to the GEE method even when no correlation between SNP scores is present. However, one could choose only the form of the working correlation matrix and not the actual values if the user finds this to be computationally feasible for a given data set.

2.3 Simulation Study of Performance A large simulation study was performed to evaluate the power of this method to detect an association between disease and an hypothetic SNP (denoted as SNP1). In each scenario, a score of SNP1=1 confers a 20% elevation in the risk of disease if no effect modifier is present in the subgroup. We studied the power of the GEE-based method to detect the association between SNP1 and disease when: (A) SNP2 is an effect modifier in the subgroup and does not have an independent effect on disease; (B) SNP2 is an effect modifier in the subgroup and also has an independent effect on disease; (C) SNP2 in the subgroup affects disease risk independenltly of SNP1 but does not modify the effect of SNP1; and (D) none of the foregoing. In these four scenarios, the proposed method is expected to have power greater than or equal to the power of the independence-based method. If negative correlation exists between SNP scores in the subgroup, the GEE method can have less power than the independence-based method. Scenario E was included to illustrate this situation: (E) SNP2 in the subgroup affects disease risk as in scenario C, while SNP1=1 and SNP2=1 cannot occur simultaneously in the same individual. This produces strong negative correlation between SNP1 and SNP2 in the case-control sample. This simulates a situation where the co-occurrence is lethal. Rather than simulating data to fit a convenient mathematical model, we simulated data according to plausible and scientifically interesting biological conditions. The disease probabilities used for each scenario are given in Table 1.

SNP1 and SNP2 were placed in a subgroup of 12 SNPs such that scores for the SNPs placed adjacent to each other in the subgroup had that same pairwise correlation, with decreasing correlation between SNP scores that were farther apart in the subgroup. All other SNPs in the subgroup had no relationship to disease. An AR-1 working correlation matrix was used for all simulations. The parameter for the AR-1 matrix, ρ, was fixed at 0.6, except for the set of simulations that study the effect of varying ρ (see below). Results were the same for a subsample of simulations re-done with 100 SNPs in the subgroup (results not shown).

Power was examined for the five scenarios above when SNP scores were moderately correlated in the population and when SNPs were uncorrelated in the population. This provides ten basic simulation scenarios. Additionally, the effect of the working correlation matrix choice was studied by varying the choice of correlation parameter under a fixed simulation scenario, Scenario B with moderate correlation among SNPs. For this part of the study, ρ was varied among $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, .85\}$. Finally, the performance of the method in adjusting for population structure was studied by adding population structure at a third SNP site (SNP3), again under Scenario B with moderate correlation among SNPs. The simulation data were created in a prospective manner (i.e. as they would appear in the general population), and then an equal number of cases and controls was sampled for each simulation study.

The distribution of the genetic scores was generated as follows: The marginal probability that S1=0, P[S1=0], was set to equal 0.5 so that P[S1=0]=P[S1=1]=0.5. This facilitates the simulation by allowing a subgroup of SNPs with a given mutual pair-wise correlation structure to be generated non-iteratively. This also maximizes power with respect to the distribution of the genetic scores, reduing the sample sizes needed and the sampling/computation time. In preliminary studies, P[S1=1] was set to smaller values (0.1 and 0.2) with no apparent qualitative effect on the relative powers of the methods (not shown).

For simulation studies with "moderate" correlation between SNPs, the probability that S1=1 given S2=1 ($P[S1 = 1|S1 = 2]$) was set to 0.6, slightly higher than the overall probability that S1=1. This corresponds to the situation where having the deleterious polymorphism at site 2 increases the probability of having the deleterious polymorphism at site 1. This results in a population correlation of 0.20 for the SNP scores from adjacent SNP sites.

For the final study on adjustment for confounding, SNP3 was generated so as to be confounded by population structure (Balding, 2006) and placed adjacent to SNP1 in the subgroup. By placing SNP3 adjacent to SNP1 in the subgroup, the confounding effect at SNP3 is allowed to have the largest impact on estimation of association between disease and SNP1 when using the AR-1 working correlation matrix. The case-control sample was simulated such that

1. sub-population 1 has SNP3=1 80% of time, SNP3=0 20% of time

2. sub-population 2 has SNP3=1 20% of time, SNP3=0 80% of time

3. 60% of cases are from sub-population 1, 40% from sub-population 2

4. 32.5% of controls are from sub-population 1, 67.5% from sub-population 2.

This creates an apparent (marginal) odds ratio (OR) of 2.0 for association of SNP3=1 with disease when not adjusted. SNP3 was correlated with SNP1 according to the design above for the scenarios with moderate correlation between scores at adjacent SNP loci. To adjust for this confounding, the following (marginal) model was fitted in lieu of model (1):

$$\text{Prob}[Y_i = 1|Sj_i] = \frac{\exp(\alpha_j + \beta_j Sj_i + \beta_I X_{Ii})}{1 + \exp(\alpha_j + \beta_j Sj_i + \beta_I X_{Ii})}, \tag{3}$$

where X_{Ii} is an indicator variable such that $X_{Ii} = 1$ if the i^{th} subject is from population 1 and $X_{Ii} = 0$ if the i^{th} subject is from population 2.

All GEE results are compared to the results from the independence-based logistic regression method. In addition to the Wald statistic for individual coefficients, the GEE method allows for joint tests of association by forming a joint Wald statistic using the estimated covariance among coefficients. For example, the joint significance of $(\hat{\beta}_1, \hat{\beta}_2)$ may be tested using

$$GW2 = (\hat{\beta}_1 \hat{\beta}_2) \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_{11} \end{bmatrix}^{-1} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \tag{4}$$

where $(\hat{\beta}_1, \hat{\beta}_2)$ are the estimated coefficients for the effects of SNP1 and SNP2, respectively, $\hat{\sigma}_j$ is the robust estimate of the variance for $\hat{\beta}_j$, and $\hat{\sigma}_{jk}$ is the robustly estimated covariance between $\hat{\beta}_j$ and $\hat{\beta}_k$. The $\hat{\sigma}$'s are obtained from the "working covariance" matrix returned by the GEE software. A joint test for any set of elements of a subgroup (including the entire subgroup) can be formed in an analogous manner. We report the power of $GW2$ for all simulations as well as for $GW3$ defined to be

$$GW3 = (\hat{\beta}_j\hat{\beta}_1\hat{\beta}_2)\begin{bmatrix} \hat{\sigma}_{jj} & \hat{\sigma}_{j1} & \hat{\sigma}_{j2} \\ \hat{\sigma}_{j1} & \hat{\sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{j2} & \hat{\sigma}_{12} & \hat{\sigma}_{22} \end{bmatrix}^{-1} \begin{pmatrix} \hat{\beta}_j \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \tag{5}$$

In our simulation studies, "SNP3", corresponding to the regression coefficient β_3 above, is placed adjacent to SNP1 in the subgroup but does not affect disease. The reason to study this particular GW3 statistic is to determine the loss in power that might be inflicted by adding a coefficient for a truly unassociated SNP site (SNP3) to the joint test statistic. In the simulations investigating adjustment for confounding, SNP3 in (5) was spuriously associated with disease through confounding as described above.

The standard practice for investigating the joint effect of two or more variables on the disease outcome is to include both variables as covariates in a univariate outcome logistic regression. Specifically, to investigate the joint effect of SNP1 (S1) and SNP2 (S2) on disease, the following model is usually fitted:

$$\text{prob}[Y_i = 1|S1_i, S2_i] = \frac{\exp(\alpha_{12} + \beta_1 S1_i + \beta_2 S2_i)}{1 + \exp(\alpha_{12} + \beta_1 S1_i + \beta_2 S2_i).} \tag{6}$$

Hence, GEE results and logistic regression based on model (1) were also compared to logistic regression results based on model (6). Model (6) is often termed "multivariate logistic regression" in elementary biostatistics text books, with the "multivariate" aspect referring to the independent variable rather than the outcome.

To make all results comparable across scenarios, we determined the total number of SNP sites tested, K, that would result in 50% power for the independence-based logistic regression method with model (1) after a Bonferroni correction to obtain a family-wise Type I error rate of 0.05. We then applied this K to all results from a given simulation to determine the corresponding power for the remaining test statistics. The reported power for the remaining test statistics is then the proportion of test statistics that are more extreme than the median (absolute value) of the independece-based test statistics. This provides a direct comparison of the distributions of two test statistics (a test statistic with a distribution shifted further from zero than another will provide more power.) Also, for powers near 50%, the percent increase or reduction in power relative to the independence-based result is approximately the percent decrease or increase in sample size, respectively, that one would need to obtain the same power as that attained by the independence-based method. Five hundred data sets were simulated for each reported result; sample sizes were 600 cases and 600 controls for Scenarios A-C, and 2000 cases and 2000 controls for scenarios D-E. Simulations were run using a function written by the author in the statistical computing language R (www.r-project.org) and may be obtained by writing the author at emond-at-u.washington.edu.

	Scenario				
SNP Scores	**A**	**B**	**C**	**D**	**E**
S1=0 & S2=0	0.10	0.10	0.10	0.10	0.10
S1=1 & S2=0	0.10	0 .12	0.12	0.10	0.12
S1=0 & S2=1	0.12	0.12	0.12	0.12	0.12
S1=1 & S2=1	0.24	0.24	0.12	0.12	NA

Table 1: Probabilities of Disease in Simulation Studies The right portion of the table shows the simulated probabilities that Y=1 (disease is present) in the population for the four possible configurations of genetic scores at SNP locus 1 (S1) and at SNP locus 2 (S2)(left portion of the table).

Results Results are summarized in Tables 2 and 3. In Scenarios A and B, the GEE-based test statistic $GW1 = \hat{\beta}_1^G/\text{se}(\hat{\beta}_1^G)$ achieved 35% greater power than the corresponding statistic from independence-based logistic regression, $L1 = \hat{\beta}_1^L/\text{se}(\hat{\beta}_1^L)$, when modest correlation was present between SNP scores in the population. (The superscripts G and L have been added to make clear that the estimates are obtained from the different estimation procedures.) In these two scenarios, SNP2 was an effect modifier for SNP1 as described in Table 1. In scenario C, the effects of the deleterious SNP scores are not synergistic as they are in scenarios A and B, but the GW1 statistic still has 8.4% greater power to detect the assocation between SNP1 and diseas as long as correlation is present between SNP1 and SNP2 in the population. There was no power gain in scenario C when no correlation is present between SNP1 and SNP2. In scenario D, where SNPs is completely unrelated to disease, there was no appreciable power gain for GW1 rleative to L1 in scenario D for either SNP correlation structure.

The statistic ML from independence-based logistic regression using model (6) to account for the affect of SNP2 performed very poorly relative to L1 except in the cases where no positive correlation between SNP scores was induced in the sample (scenarios C and D with no correlation between SNPs in the population).

				Power (%)				
Risk	**pop.**	**sample**	**AR-1**	**Test Statistic**				
Scenario	**cor($S1, S2$)**	**cor($S1, S2$)**	$\boldsymbol{\rho}$	**L1**	**GW1**	**GW2**	**GW3**	**ML**
A	0.20	0.27	0.60	50.0	67.4	81.0	78.2	11.2
B	0.20	0.26	0.60	50.0	67.4	91.0	89.2	8.6
C	0.20	0.18	0.60	50.0	54.2	73.2	75.2	37.2
D	0.20	0.20	0.60	50.0	51.2	48.4	46.4	46.4
E	0.20	-0.42	0.60	50.0	36.0	91.8	93.0	81.8
A	0.00	0.08	0.60	50.0	57.4	85.0	79.8	40.4
B	0.00	0.06	0.60	50.0	54.2	93.8	93.2	37.4
C	0.00	-0.02	0.60	50.0	49.2	78.8	84.2	51.6
D	0.00	0.00	0.60	50.0	50.6	47.6	44.2	50.6
E	0.00	-0.52	0.60	50.0	30.2	96.4	95.6	87.6
B	0.20	0.26	0.00	50.0	50.0	91.4	89.4	8.6
B	0.20	0.26	0.10	50.0	54.6	91.4	89.6	8.6
B	0.20	0.26	0.20	50.0	59.4	91.4	89.8	8.6
B	0.20	0.26	0.30	50.0	62.8	91.4	89.8	8.6
B	0.20	0.26	0.40	50.0	65.0	91.4	89.6	8.6
B	0.20	0.26	0.50	50.0	66.6	91.0	89.4	8.6
B	0.20	0.26	0.60	50.0	67.4	91.0	89.2	8.6
B	0.20	0.26	0.70	50.0	67.6	91.0	89.0	8.6
B	0.20	0.26	0.80	50.0	66.8	91.0	88.8	8.6
B	0.20	0.26	0.85	50.0	66.8	90.6	88.8	8.6

Table 2: Simulation Results The power to detect the association between disease and a genetic score of 1 at SNP locus 1 is given in the right portion of the table for 5 test statistics and the 5 disease probability scenarios described in the Methods. $L1 = \hat{\beta}_1/\text{se}(\hat{\beta}_1)$ from the independence-based logistic regression method using model (1); $GW1 = \hat{\beta}_1/\text{se}(\hat{\beta}_1)$ from GEE using (1) as the (marginal) model; GW2 and GW3 are joint test statistics given by equations (4) and (5); and $ML = \hat{\beta}_1/\text{se}(\hat{\beta}_1)$ from independence-based logistic regression using model (6); **pop. cor($\boldsymbol{S1, S2}$)** is the correlation between scores at loci 1 and 2 in the population; **sample cor($\boldsymbol{S1, S2}$)** is the correlation between socres at loci 1 and 2 in the case-control sample; ρ is the fixed value used in the AR-1 working correlation matrix.

The joint test of assocation between disease and SNPs 1 and 2 (given by GW2, Table 2) performed very well except in scenario D where SNP2 was not associated with disease in any way (neither as an effect modifier alone nor via a main effect). GW3 was always less powerful than GW2, except in scenario C.

When the AR-1 working correlation parameter for the GEE method varied from 0 to 0.85 in scenario B, the greatest power was attained when $\rho = 0.70$. This optimal value of ρ is related to both cor(S1, S2) and the combined effects of SNP1 and SNP2 on disease. The power for $GW2$ and $GW3$ did not change appreciably as ρ varied. When $\rho = 0$, the GEE method reduces to the independence-based method. Note that no power is lost (relative to the independence method) by assuming $\rho = 0.6$, even in the scenario where SNP2 is unrelated to disease and SNPs are uncorrelated (Table 2, row 9).

For the case when confounding through population structure was present to give SNP3 an apparent assocation with disease (Table 3), the power for both L1 and GW1 was increased relative to the results when no confouding was present (Table 2). This is because the score at SNP1 is correlated with the score at SNP3, resulting in some degree of confounding of the association between disease and SNP1. After adjustment for confounding, GW1 still attains a 34% increase in power relative to L1, though the power is reduced relative to the unconfounded, unadjusted case in Table 2.

			Power (%)						
			Test Statistic						
Adjusted	**sample**	**sample**	**SNP1**					**SNP3**	
Fit	**cor($S1, S3$)**	**cor($S1, X_I$)**	**L1**	**GW1**	**GW2**	**GW3**	**ML**	**L1**	**GW1**
no	0.22	0.16	85.6	97.6	99.4	100	37.4	79.2	85.6
yes	0.22	0.16	38.6	55.8	85.8	82.8	7.0	0	0

Table 3: Confounding Simulation Results The right portion of the table shows powers for test statistics when the relationship between the SNP3 score (S3) and disease is confounded by population structure. SNP3 is placed adjacent to SNP1 in the subgroup under disease scenario B with moderate correlation between SNPs, as described in the Methods. The unadjusted results are obtained by fitting (marginal) model (1) for either independence-based logistic regression or GEE to obtain the test statistics L1, GW1, GW2, GW3 and ML as described in Table 2. Adjusted results are obtained analogously by fitting (marginal) model (3) for each SNP in the subgroup. Powers for all five test statistics are given for the tests of association involving SNP1 either singly or jointly; powers are given for testing association between disease and SNP3 singly.

In the case where SNP1 and SNP2 had strong negative correlation in the sample (scenario E), GW1 performed poorly. The relative descrease in power for GW1 is less when the magnitude of the negative correlation is less.

Theoretical results by Gourieroux *et al* (1984) show that test sizes should be at the nominal levels. As a check, we computed test sizes on data simulated under the global null hypothesis. All test sizes were between 0.02 and 0.04.

Discussion The method proposed here provides power gains that have practical potential to provide better discovery of associations between subject status and risk factors measured via high throughput gene-based assays. The method is easy to apply, and GEE-derived tests are known to have good behavior, at least when the subgroup size is small relative to the sample size (1:20 ratio or less). Examination of formulas for the variance of GEE-based parameter estimates in multivariate models indicates that the proposed power boosting method can be expected to have greater power when SNPs in the same subgroup either have positive sample correlation between scores and/or increase the probability of disease. The exception to this rule is the situation where co-occurrence of the deleterious SNP scores is reduced enough among cases that negative correlation is induced between SNP scores as in simulation scenario D. Placing SNPs with negatively correlated scores in the same subgroup can potentially reduce power.

A suggested use for this method is to base the sample size calculations on considerations such as those discussed in Prentice and QI (2006), where it is assumed that an independece-based analysis will be done. The multivariate analysis can then be substituted for the former for potential power gains. It might be tempting to try to use this method to reduce the sample size needed to obtain a specified power. However, we don't currently recommended this as a general strategy. One reason to avoid this strategy is that the nature of this method makes analytic power formulas intractable at this point in time. Hence, one must resort to simulation studies to estimate power. While simulations were used by Prentice and QI (2006) to estimate power, it is much easier to specify the magnitude of a biologically interesting marginal odds ratio in a simulation than it is to specify the underlying sample correlation structure of the chosen subgroups. In cases where the joint distribution of interesting SNP scores is known, power simulations for this multivariate method could be done.

While we believe it is of interest to report our power results for GW_2 and GW_3, the (markedly) superior performance of these statistics is limited to the situation where at least two of the SNP loci in the joint test are associated with disease. Hence, in order for a joint test to be superior to GW1, one would need to be reasonably sure in advance that the majority of SNPs involved in the joint test are truly associated with disease – a rather unlikely scenario. A possible use for these statistics is for joint testing of very tightly linked SNPs, rather than discarding some of these tightly linked SNPs to provide a "tag" SNP set for testing. This suggestion needs further research, however.

A nice feature of this method is that it provides extra power to detect a main effect when effect modifiers are included in the subgroup. In this sense, including effect modifiers in a subgroup is analogous to including "precision variables" in a linear regression model. This feature could also help elucidate the effect modifiers. With the independence-based method, one would need to add many more tests in the form of additional test coefficents from a model that includes interaction terms to identify effect modification (epistasis), with a possibly dramatic reduction in power. Power loss would occur for two reasons: due to a substantial increase in the multiplicity of tests performed and due to lower power attained when using models with multiple correlated covariates. The latter is the well-known problem of "collinearity." In addition, it is infeasible to model all potentially synergistic combinations of SNPs. The method proposed here could help narrow the candidate field in a search for effect modification: one could identify significant genes from a testing phase using model (1), and then perform additional tests for potential effect modification by other SNPs in the same subgroup as the candidate SNP. One could allow for a relatively small number of new tests to be performed in this second phase when planning the multiple testing corrections that should be taken into account in the Bonferroni correction.

We have shown that the suggested subgrouping can result in increased power in biologically interesting scenarios. Admittedly, subgrouping based on considering every the location and/or possible functional implications for every SNP locus in the study would be impractical to apply. However, it is quite worthwhile to apply the subgrouping in a careful manner just for those SNPs for which the investigator has some *a priori* knowledge or suspicion of relationship to each other, either spatially or functionally. This is especially true when these SNPs are considered most likely *a priori* to have an association with disease. Placing SNPs in the same subgroup when they act independently on disease risk (i.e. no interaction/epistasis effects) will also increase power in a case-control design. Functional groups of genes have been published that can make the task easier when the investigator's loci can be matched to genes within these groups (Subramania *et al.*, 2005). For other SNPs, one can subgroup based solely on location. Alternatively, SNPs that cannot be placed into a natural subgroup in via a practical procedure can be grouped randomly. It is worth re-iterating that SNPs with negatively correlated scores should not be placed in the same subgroup, since this can result in power losses. SNPs of known interest could be screened for negative correlation prior to the subgrouping. The risk of including negatively correlated SNP scores in the same subgroup by accident is probably quite low, and the impact would only be meaningful if these SNPs were associated with disease.

The reader may have noted that model (1) does not specify the true relationship between the outcome and the covariates of interest for any of the Scenarios except D. When using GEE it is usually considered important to correctly specify the model for the mean. For a logistic model, the mean is the formula for Prob$[Y = 1]$. Correct mean specification is not needed here where we are interested in the test statistics $\hat{\beta}/s.e.(\hat{\beta})$ and not explicitly interested in correct estimation of $\hat{\beta}$. In a one-stage study, the independence-based logistic regression technique will provide unbiased estimates of the marginal effect of SNP1 (that is the average effect of SNP1, averaging over the effects of all other SNPs), while GEE-based estimates using a non-independence working covariance matrix do not (Pepe and Anderson, 1994; Emond et al, 1997). Hence, the proposed procedure is not intended for estimation, only testing. If the estimates themselves are of interest, these can be obtained via independence-based logistic regression after the testing step is completed. In general, one cannot hope to specify the correct relationship between probability of disease and all SNP sites that might affect risk.

The proposed method performs well relative to the independence-based procedure for removing confounding effects. However, when adjustment variables are correlated with the variables of interest, adding these adjustment variables to the model reduces the power to detect true effects relative to omitting them. Hence, we suggest that potential confounding be investigated after significant associations have been identified using the non-adjusted model (1) in a testing phase. This would be akin to identifying patients who meet inclusion criteria for a study and then assessing for specific exclusion criteria in a second step. This procedure provides more power for finding true associations than simultaneous testing and adjustment using model (3).

Use of multivariate modeling for power boosting is not limited to whole genome SNP-disease association studies. The method can be applied to numerous kinds of multiple outcome data types, including those that result from the multiplicity of genes in an organism. As a second example, consider the case of a two-sample paired RNA microarray expression study, where RNA samples from diseased and non-diseased subjects are competitively hybridized on the same array to obtain expression ratios for thousands of genes. Let Y_{ij} denote the log expression ratio for the i^{th} pair and the j^{th} oligonucleotide. It is common to use a paired t-test to obtain a p-value for the test of no difference in expression for every outcome. This is equivalent to testing $\alpha_j = 0$ in the simple model $Y_{ij} = \alpha_j$, a regression model with only an intercept term. The multivariate power boosting method can be applied here by subgrouping oligonucleotides and applying GEE to the subgroups. A coefficient must be estimated for each oligonucleotide in the subgroup, which usually will require the use of dummy variables in the design matrix as in the main example. Note the that sampling method in this “two-sample” experiment is a case-control design as in our main example, so analogous considerations hold.

A few methods have been proposed for potentially increasing the power over the Bonferroni method in genome-wide studies. These include the truncated product method for combining p-values when the test statistics are uncorrelated (Zaykin, et al, 2002), resampling (Reiner *et al*, 2003; Pollard and van der Laan, 2005), and estimation of analytic distributions for combined evidence (Dudbridge and Koeleman, 2004). In addition, a new approach for defining significance in the multiple testing situation is proposed by Storey (2006). Each of these methods is a “post-processing” method for the test statistics or p-values and any can be applied in addition to multivariate power boosting when their conditions are met. Among the methods listed, resampling is the most studied. Resampling provides potentially more power by preserving the correlation structure within individuals during the resampling process. This is similar in principle to multivariate power boosting, and it is not clear whether much is to be gained by applying both methods. The operating characteristics of combined methods is a topic for further study.

In summary, the proposed method of employing multivariate modeling with robust estimation is a relatively easy method for increasing statistical power in high-throughput studies if effective subgrouping can be done. The method can be implemented using readily available software.

Appendix A.1 Working correlation matrix forms The AR-1 working correlation matrix has the form:

$$R = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & & & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$$

The exchangeable working correlation matrix has the form:

$$R = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & & & & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

When using the "fixed" working correlation matrix option with GEE software, the user will usually need to supply an actual matrix with numerical values. Otherwise, the parameter ρ (and other parameters, if applicable for other forms of working correlation matrix) will be estimated by the software procedure. The user should consult the manual for details specific to the software.

A.2 Creating the matrix of independent variables GEE software generally assumes that the coefficients to be estimated are common to all elements of the subgroup. This is not the case for this approach. Hence, in order to estimate the unique intercept and coefficient term for each SNP locus (α_j and β_j, respectively), dummy variables must be added to the independent variable matrix. For a subgroup with n elements (loci), this matrix will have $2n$ columns corresponding to the n sets of (α_j, β_j)s. Suppose we number the elements of the subgroup from 1 to n. To reformulate the model in terms of covariates that are common to all elements of the subgroup, the linear predictor term in (1) can be re-written as

$$\alpha_1 + \beta_1 S1_i + \alpha_2 + \beta_2 S2_i + \alpha_3 + \beta_3 S3_i + \ldots + \alpha_j + \beta_j Sj_i + \cdots + \alpha_n + \beta_n Sn_i,$$

where all Sjs are identically zero except for Sj_i. The independent variable matrix will have nN rows corresponding to the n observations from each of N subjects. For example, if $n = 3$, the 3 rows of the independent variable matrix for the i^{th} subject will be

$$\begin{array}{cccccc} 1 & S1_i & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & S2_i & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & S3_i \end{array}$$

where the actual observed values are substituted in place of the Sj's. If adjustment variables (potential confounding variables) are fitted, an additional column will be included for each adjustment variable.

References

Balding, D. (2006). A tutorial on statistical methods for populations association studies. *Nature Reviews Genetics* 7 781–791.

Broberg, P. (2005). A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics* 6 1–20.

Gourieroux, C., Monfort, A. & Trognon, A. (1984). Pseudo maximum likelihood methods: theory. *Econometrica* 52 681–700.

Liang, K. & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73 13–22.

Lumley, T. & Heagerty, P. (1999). Weighted empirical adaptive variance estimators for correlated data regression. *Journal of the Royal Statistical Society B* 61 459–477.

MJ Emond, J. R. & Oakes, D. (1997). Bias in gee estimates from misspecified models for longitudinal data. *Communications in Statistics: Theory and Methods* 26 15–32.

Owen, A. (2005). Variance of the number of flase discoveries. *Journal of the Royal Statististical Society, Series B* 67 411–426.

Pepe, M. & Anderson, G. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics: Simulation and Computation* 23 939–951.

Pollard, K. & van der Laan, M. (2005). Choice of null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference* 125 85–100.

Prentice, R. & QI, L. (2006). Aspects of the design and analysis of high-dimensional snp studies for disease risk estimation. *Biostatistics* 7 339–354.

Q Yang, I. C. L. A. C., J Cui & Demissie, S. (2005). Power and type i error rate of false discovery rate approaches in genome-wide association studies. *BMC Genetics* 6(Suppl 1) S134.

Reiner A, B. Y., Yekutieli D (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 12 368–75.

Storey, J. & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceeding of National Academy of Sciences, USA* 100 9440–5.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* 64 479–498.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50 1–25.

Y pawitan, S. M., KR Krishna Murthy & Ploner, A. (2005). Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics* 21 3865–72.