Genetic Epidemiology 33: 442–452 (2009)

Adaptively Weighted Association Statistics

Michael LeBlanc* and Charles Kooperberg

Fred Hutchinson Cancer Research Center, Seattle, Washington

We investigate methods for testing gene-disease outcome associations in situations where the genetic relationship potentially varies among subjects with differing environmental or clinical attributes. We propose a strategy which modestly increases multiple testing by evaluating weighted test statistics which focus (or enrich) association tests within subgroups and use a Monte-Carlo method, based on simulating from the approximate large sample distribution of the statistics, to control type 1 error. We also introduce a stage-wise calculated test statistic which allows more complex weighting on multiple environmental variables. Results from simulation studies confirm improved power of the proposed approaches compared to marginal testing in many situations. *Genet. Epidemiol*. 33:442–452, 2009. © 2009 Wiley-Liss, Inc.

Key words: score tests; association tests; data adaptive; gene-environment interactions

Contract grant sponsor: National Institutes of Health; Contract grant numbers: CA90998; CA125489; CA74781; CA53996.
*Correspondence to: Michael LeBlanc, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M3–C102, Seattle, WA 98109.
E-mail: mleblanc@fhcrc.org
Received 25 April 2008; Revised 26 September 2008; Accepted 14 November 2008
Published online 23 January 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20397

INTRODUCTION

Assessing the association of genomic attributes with disease outcomes is an important and ongoing area of applied research. Commonly, many univariate marginal tests are calculated for a large number of genomic features. For instance, in human genetic association studies with single nucleotide polymorphisms (SNPs) thousands of SNPs are tested for association with disease outcomes. However, given the typical focus on univariate or marginal testing, there are concerns that more complex relationships, such as multiple genes acting in concert or gene-environment (or gene-treatment) interactions, could attenuate the marginal effect size and reduce the power to detect true associations. For instance, the association of the gene with survival may be present only in smokers or prior smokers, or in patients taking a specific treatment.

One strategy is to exhaustively test gene-environment interactions in addition to testing marginal gene associations with outcome. However, interaction testing with many genes is often more difficult due to limited power and can lead to conducting a very large number of tests. Alternatively, if there is reason to suspect that the genetic association is stronger within a subgroup of subjects with specific treatment or set of clinical/environmental attributes, other more powerful test statistics can be constructed. We propose computationally simple test statistics that can exploit such subgroup associations if they exist. The goal is to modestly increase the search by weighting the association tests within subgroups of subjects. We note that constructing tests based on ordered overlapping subgroups of subjects has been previously investigated for linkage analysis [Hauser et al., 2004]. In that case, partial sums of lod scores were calculated for families ranked on a covariate interest. We also introduce a natural extension of the subgroup statistics motivated by statistical "boosting" to define weights.

A subgroup weighted test can be more powerful than a marginal test of association under some gene-environment interaction models. We limit the complexity of the statistical exploration to control the variability of the search. Other authors have studied methods for modeling gene-gene and gene-environment interactions that increase the power of finding marginal associations [Chatterjee et al., 2006; Marchini et al., 2005]. Our proposal directly focuses on weighted score test statistics rather than full modeling. When there are truly gene-environment interactions, power for detecting genetic associations with outcome can be substantially increased by appropriate, yet parsimonious, weighting. The presentation of the proposed statistics is general and we expect the weighted gene association tests to have application in both SNP and gene expression association studies.

METHODS

MOTIVATION FOR EFFICIENCY OF SUBGROUP WEIGHTING

Our strategy is to focus on enriched subgroups of subjects to test for genetic association. Figure 1 suggests that if one could identify the appropriate environmental subgroup, stronger associations between the genetic variable and subject outcome could be seen: as the difference in shading becomes more intense it indicates a greater difference in disease probabilities.

Some motivation for the potential statistical power of considering subgroup tests can be based on testing main effects and subgroups within a simple multiplicative interaction model. Let Z denote an environment or

© 2009 Wiley-Liss, Inc.

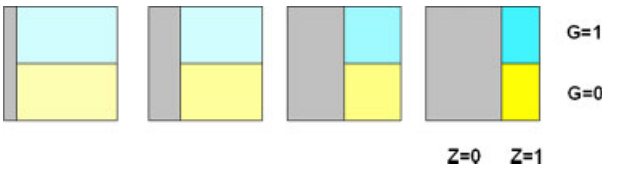

Fig. 1. Enriching the environmentally defined subgroup. Let Z denote an environment or treatment variable and G a genetic factor. Intensity of shading represents the probability of disease.

treatment variable and G a genetic factor and assume that the association may depend on the level of factor Z. Assume a regression model is indexed by a mean function

$$\eta = \beta_0 + \beta_1 G + \beta_2 Z + \beta_3 GZ,$$

where G and Z are independent and binary. Also assume that $P(G = 1) = p_g$ and $P(Z = 1) = p_z$. Assume a continuous phenotype with constant variance; then standard linear model testing leads to easy calculation of the power and relative efficiency of the subgroup test, marginal test and interaction test. The ratios of the square of the mean to variance, the non-centrality parameters, are (1) for the marginal test (using only the univariate model including G) $\delta_m^2 = (\beta_1 + p_z\beta_3)^2 p_g(1 - p_g)$, (2) for the subgroup test for the $Z = 1$ group $\delta_s^2 = (\beta_1 + \beta_3)^2(1 - p_g)p_z p_g$, and (3) for the interaction test $\delta_i^2 = \beta_3^2 p_z p_g(1 - p_z)(1 - p_g)$.

Therefore, in the simple setting in which there is only a genetic association with outcome in the subgroup ($Z = 1$) it is clear that using the subgroup test can yield significantly better power than a marginal test. The ratio of the non-centrality parameters δ_s^2 to δ_m^2 for $\beta_1 = 0$ is

$$\frac{1}{p_z},$$

so the gain can be substantial depending on the fraction of individuals in the subgroup. For instance, relative to the marginal test the impact on non-centrality is largest in the case where p_z is small. While this case induces an interaction, the power to specifically test the interaction will be lower: the ratio of the subgroup to interaction δ_s^2 to δ_i^2 for $\beta_1 = 0$ is

$$\frac{1}{1 - p_z}.$$

Of course, there is reduced power for testing the ; subgroup relative to the marginal test if the interaction effect β_3 is small and the main effect $\beta_1 \neq 0$. We note that the non-centrality parameter formulas above show that even for the case $\beta_1 \neq 0$, the ratios of the subgroup non-centrality parameters of δ_s^2 to δ_m^2 and of δ_s^2 to δ_i^2 do not depend on the gene/allele probability, p_g (of course, the absolute power would depend on this). Our proposal to consider the adaptive test statistic based on the maximum of the marginal and subgroup test statistics can also be motivated in this simple regression example. In large samples one can utilize the analytic forms for mean and variance of the maximum of bivariate normals. Then in this case the expected value of the maximum of the marginal and subgroup tests is

$$\Delta = \delta_m \Phi(-\gamma) + \delta_s \Phi(\gamma) + a\phi(\gamma), \quad (1)$$

where Φ and ϕ are the normal distribution and density functions, respectively [Clark, 1961]. The term $\Phi(\gamma)$ is the probability that the subgroup test statistic is larger than the marginal test statistic. For this case, $a = 2(1 - \rho)$, where ρ is the correlation between the test statistics and $\gamma = (\delta_m - \delta_s)/a$, so the expected value (1) is a linear combination weighted by the power properties of the marginal and the subgroup test. The last term in the expectation, $a\phi(\gamma)$ represents the additional impact of adaptive selection. Therefore, we expect that data adaptive selection will lead to a more powerful test on a wider range of underlying disease models.

We provide plots of the non-centrality parameters under different values for the main effect and interaction parameters for the marginal, subgroup, interaction, maximum of marginal and subgroup, and a global (“full”) 2 degree of freedom test for the case $p_z = p_g = 0.5$. Under the case of no interaction, the marginal test is preferred over all other options (Fig. 2, upper left panel); in the case of both a marginal and subgroup effect in the same direction (Fig. 2, upper right panel) the subgroup test is the most powerful; the full 2 degree of freedom test loses power due to the extra degree of freedom, and the maximum test is the second best. In the case of a negative main effect and positive interaction (Fig. 2, lower left panel), the marginal test can have very poor power, the interaction test performs best for large values of β_3, and the maximum and full 2 degree of freedom tests can provide reasonable compromises over a range of interaction values. However, one would anticipate that as the complexity (and degrees of freedom) the global (full) test increases in real applications (1+ number of effect modifier terms) the power of the test would be significantly and negatively impacted.

These simple calculations indicate that the adaptive selection of maximal select test subgroup statistics or subgroup weighting may provide a simple yet powerful mechanism for testing. In the context of data analysis, one would expect to evaluate two or more subgroup test statistics. These more complex tests of multiple subgroups, and their type I error control and power are explored in other sections of the article.

SUBGROUP WEIGHTED SCORE STATISTIC

Since many useful association tests are score type statistics, we consider weighted versions of such tests. Let Z denote an environment or treatment variable and G_j a genetic factor. A standardized univariate score test of the association of G_j with patient outcome can be expressed as

$$T_j = \frac{U_j}{\sqrt{V_j}},$$

where $U_j = \sum_{i=1}^{n} U_{ji}$ and U_{ji} is the score component for individual i and n is the total sample size. Here, the index j corresponds to gene G_j, and the denominator V_j is the estimated variance of U_j. For example, in the case of binary outcome data (in the setting of a case-control study), with

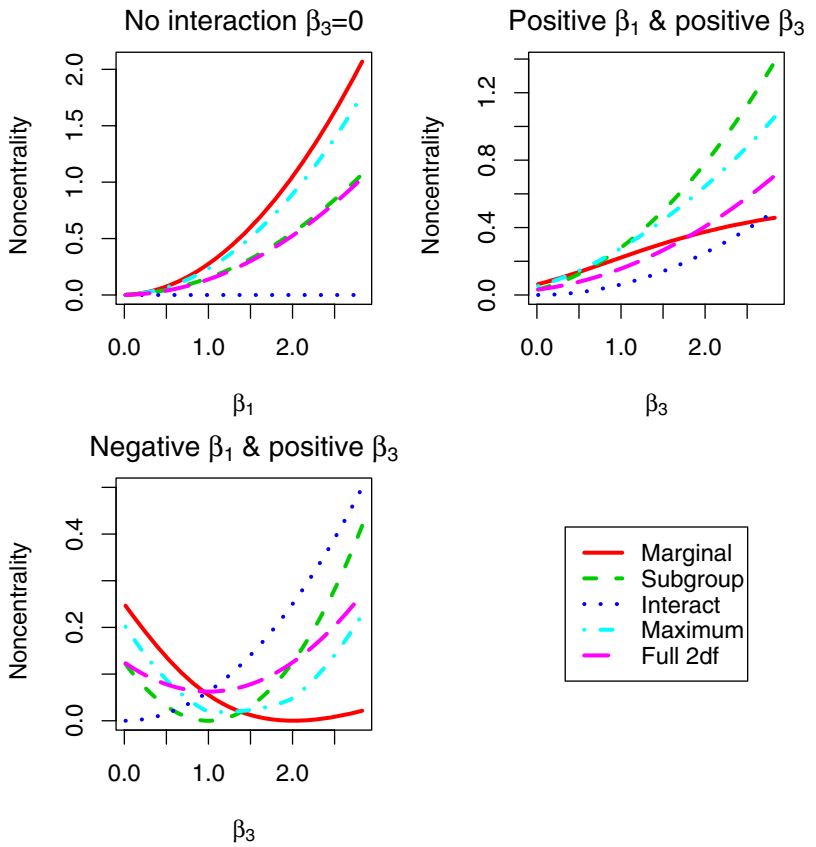

Fig. 2. Non-centrality parameters for various testing strategies under an additive regression model.

$y_i = 1$ or 0 for case or control, G_{ji} the gene j value and Z_{ki} for environmental factor k for individual i, the score component would be

$$U_{ji} = G_{ji}\left(y_i - \frac{e^{\alpha+\beta z_{ki}}}{1+e^{\alpha+\beta z_{ki}}}\right).$$

We note that score components for other models for other outcomes such as time-to-event with the proportional hazards model are easily constructed.

We construct weighted marginal test statistics indexed by a parameter θ,

$$U_W(Z, G_j; \theta) = \sum_{i=1}^{n} h(Z_i, \theta) U_{ji}.$$

The parameter θ can describe a subgroup based on Z, or it could more generally parameterize a subgroup weighting function. As above, U_j represents the score component of G_j. A simple empirical estimate of the variance of $U_W(Z, G_j; \theta)$ is

$$V_W(Z, G_j; \theta) = \sum_{i=1}^{n} h(Z_i, \theta)^2 V_j.$$

The standardized statistic is $T_W(Z, G_j; \theta) = U_W(Z, G_j; \theta)^2 / V_W(Z, G_j; \theta)$. A subgroup weighting function, often called a basis function in adaptive regression methodology [e.g. Hastie et al., 2001], is $h(Z; \theta) = I\{Z > \theta\}$ which can be used if the genetic association was thought to be stronger with larger values of the environmental factor Z. Another popular basis function is the truncated linear spline function $h(Z; \theta) = I\{Z > \theta\}(Z - \theta)$ or the smooth logit function

$$h(Z; \theta) = \exp(a + bZ)/(1 + \exp(a + bZ)).$$

Thus, in our setting, the basis functions are just transformations of the environmental factors. In the case of multiple environmental factors, one could envision smooth additive combinations of the environmental predictors or subgroup rules based on Boolean combinations. However, computationally efficient algorithms would need to be derived for powerful weighting functions. Even in the case of a single environmental factor, it will not be known a priori if the association would be stronger as a marginal effect or within a subgroup defined by environmental or clinical factor Z. For instance, Z could measure body mass index or some ordered measure of smoking history. With respect to subgroup weighting functions, a small number of θ values (say cut-points c, or logit weighting functions)

$$h(Z; \theta_0), \ldots, h(Z; \theta_k)$$

would allow investigation of subgroup effects where one can assume $h(Z; \theta_0)$ represents the overall marginal effect, $h(Z; \theta_0) = 1$.

Therefore, panels of correlated weighted test statistics $T(Z, G_j; \theta_0), \ldots, T(Z, G_j; \theta_k)$ for each gene $j = 1, \ldots, p$ can be constructed and the maximal statistic, $\max\{T(Z, G_j; \theta_0), \ldots, T(Z, G_j; \theta_k)\}$ for each genetic factor G_j can be calculated. It is hoped that this strategy of using the maximum of the

marginal and subgroup tests could lead to improved power over marginal testing.

INFERENCE

The numerators of subgroup weighted test statistics

$$(U_{1j}, \ldots, U_{kj}) = U(Z, G_j; \theta_1), \ldots, U(Z, G_j; \theta_k)$$

tend to be correlated and hence, the standardized versions

$$T(Z, G_j; \theta_1), \ldots, T(Z, G_j; \theta_k)$$

for a specific gene will also be substantially correlated. Therefore, overall multiple testing control using a Bonferroni correction will be very conservative. If all the Kp statistics

$$U_i = (U_{11}, \ldots, U_{1K}, \ldots, U_{p1}, \ldots, U_{pK})_i,$$
$$j = 1, \ldots, p, \quad k = 1, \ldots, K$$

are considered as a vector, in large samples they have an approximately multivariate normal distribution with covariance

$$V = \sum U_i U_i^{\mathrm{T}}. \tag{2}$$

To acknowledge the correlation, we propose resampling or permutation methods to control for multiple testing. Given, the simple additive form of the subgroup weighted test statistics, a Monte Carlo method can be used [e.g. Lin, 2005]. We construct modulated (or simulated) versions of the test statistics

$$\widetilde{U}_{\mathrm{W}}(Z, G_j; \theta) = \sum_1^n h(Z_i, \theta) U_{ji} Q_i,$$

where Q_i are standard normal random variables, and $\widetilde{T}_{\mathrm{W}}(Z, G_j; \theta) = \widetilde{U}_{\mathrm{W}}(Z, G_j; \theta)^2 / V_{\mathrm{W}}(Z, G_j; \theta)$ as the standardized statistic. Therefore, denoting each element by $\widetilde{U}_{jk} = \widetilde{U}_{\mathrm{W}}(Z, G_j; \theta_k)$, conditional on the data, then $\widetilde{U}_i = (\widetilde{U}_{11}, \ldots, \widetilde{U}_{1K}, \ldots, \widetilde{U}_{p1}, \ldots, \widetilde{U}_{pK})_i$, $j = 1, \ldots, p$, $k = 1, \ldots, K$, are multivariate normal with the same covariance given in equation (2).

Then the simulated standardized test statistic is $\widetilde{T}_{\mathrm{W}}(Z, G_j; \theta) = \widetilde{U}_{\mathrm{W}}(Z, G_j; \theta)^2 / V_{\mathrm{W}}(Z, G_j; \theta)$. Based on a large number of random realizations one can calibrate the observed test statistics, for instance, by the family-wise error rate. Note that in instances where the individual genes are not substantially correlated, one could calibrate the type I error for the subgroup tests for each gene, and then use the Bonferroni method to adjust across the multiple genes.

STAGE-WISE WEIGHTED ASSOCIATION TESTS

Just as boosting regression models can be used to improve prediction error [e.g. Friedman et al., 2000], we believe that boosting can motivate a strategy to enhance the strength of the association testing by repeated adjustment of the weighting function. In particular, it can be used to construct a smoothly weighted test statistic or a statistic weighted by several environmental factors. For instance, suppose the current weighting function is $g(Z)$ then one could choose a new cut-point function $h(Z,\theta)$ to maximize the standardized version of

$$U_{\mathrm{W}(\lambda)}(Z, G_j; \theta) = \sum_{i=1}^n (g(Z) + \lambda h(Z, \theta)) U_{ji}.$$

As above, U_{ji} represents the score component on G_j. The goal is to find the new weight that maximizes the correlation with the score vector. A simple empirical estimate of the variance of $U_{\mathrm{W}(\lambda)}(Z, G_j; \theta)$ is

$$V_{\mathrm{W}(\lambda)}(Z, G_j; \theta) = \sum_{i=1}^n (g(Z) + \lambda h(Z, \theta))^2 V_j.$$

Consider small steps of magnitude λ. Each boosting step would select the basis function that maximizes the standardized statistic $T_{\mathrm{W}(\lambda)}(Z, G_j; \theta) = U_{\mathrm{W}(\lambda)}(Z, G_j; \theta)^2 / V_{\mathrm{W}(\lambda)}(Z, G_j; \theta)$. Finally, the resampled version

$$U_{\mathrm{W}(\lambda)}(Z, G_j; \theta) = \sum_{i=1}^n (g(Z) + \lambda h(Z, \theta)) U_{ji} Q_{ji}$$

can be used to construct a null distribution for boosted statistics after any number of boosting steps. Note that there is reasonable computational efficiency since the score components do not need to be updated, yet computation is linear in the number of genes $\times$ number of boosting steps, which makes it prohibitive for large numbers of genes. In addition, for this implementation one would need to determine the total number boosting steps to be used to yield good power characteristics. Too many boosting steps would lead to increased variance and reduced power; this is analogous to overfitting with smoother error functions [e.g. Friedman et al., 2000].

However, rather than using simple boosting, we focus on increased computational speed by utilizing efficient regression algorithms to construct the weighted score statistics. For instance, one can construct a regression basis $X = HG_j$ using a matrix of potential basis weights $H = [h(Z; \theta_0), \ldots, h(Z; \theta_k)]$ multiplied by the genetic variable of interest G_j, then use a computationally efficient stage-wise regression such as least angle regression (LAR) [Efron et al., 2003] with a mean centered outcome. Typically, we propose basis weights or basis functions that are piecewise constant functions $h(Z; \theta) = I\{Z > \theta\}$ of an environmental variable Z.

We briefly outline the components of the LAR algorithm below. For regression, where there are n independent observations $(y_i, x_{i1}, \ldots, x_{ik})$ of the response and k predictor variables it can be viewed as a continuous and fast implementation of stage-wise regression methods. An outline of the algorithm is given below:

LAR

1. Start with $r = y, \widehat{\beta}_1, \widehat{\beta}_2 \cdots \widehat{\beta}_k = 0$. Assume the x_r are standardized to have variance equal to 1.
2. Find the predictor x_m most correlated with r.
3. Increase $\widehat{\beta}_m$ in the direction of sign(cor(r,x_m)) until another predictor x_j has equal correlation to r as x_r. Put r in set of active predictors, S.

4. Move $(\widehat{\beta}_m : m \in S)$ in the joint least squares direction for $(x_m : m \in S)$ until another predictor has equal correlation with the current residual.
5. Repeat step 4 until $\text{cor}(r, x_m) = 0$ for all m.

So in our case, $(x_{i1}, \ldots, x_{ik}) = (h(Z_i; \theta_0), \ldots, h(Z_i; \theta_k))G_j$ and the response is the adjusted score component, $y_i = U_{ji}$. The LAR algorithm generates a vector curve denoting the solution for each step of the algorithm and value of the L_1 norm of the parameter vector. The algorithm is similar to forward step-wise regression, but instead of adding variables at each step, the estimated parameters are increased in a smoothly equiangular fashion to each one's correlations with the residual.

There is a strong connection to the LAR algorithm and LASSO [Tibshirani, 1996] which uses an L^1-penalty on the regression coefficients and leads to both shrinkage and variable selection. The LASSO estimate $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_m)^{\mathrm{T}}$ is defined as the minimizer of

$$g(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \sum_k \beta_k x_{ik} \right)^2 + \lambda_1 \sum |\beta_k|^1,$$

where λ_1 is a non-negative penalty parameter. Often the response and predictors are standardized so that $\sum_i y_i = 0$ and $\sum_i x_{ik} = 0$ and $\sum_i x_{ik}^2 = 1$. One motivation for LAR was an efficient algorithm for LASSO type estimates. LAR gives answers that are often close to LASSO, and are identical if the predictors are orthogonal.

We use the LAR strategy for estimating weights in the simulations given below and call the statistic a stage-wise weighted test statistic. A separate model selection, potentially using the residual error for the LAR fit, for every possible genomic factor would likely not be computationally feasible. Therefore, we propose to use a single tuning parameter set across all genes. We choose a tuning parameter relating to the number of non-zero weights (LAR steps), going between (1 and k+1) where there are k+1 basis functions for weighting the score test statistics; k+1 would correspond to the full unconstrained least squares weights. Non-integer values of the tuning parameters, between k' and k'+1 correspond to the relative L_1 norm between those two steps. Using tuning parameters between 2 and 4 corresponds to 2–4 parameter weights and would add some flexibility over the maximal test statistic but without adding too much additional variance.

To construct the null distribution of the stage-wise statistics, we calculate resampled versions of the score components, $y_i = U_{ji}Q_{ji}$ and use them in the LAR algorithm to construct weights and statistics at the sample tuning parameter as the observed data. Alternatively, one could permute the adjusted score statistics across the basis weight matrix HG_j. We note that the tests must use weights that are recalculated on each random sample, not just those from original data, to appropriately construct the null distribution and control type I error.

RESULTS

MYELOMA EXAMPLE

We first present the weighted testing strategy on data from patients diagnosed with multiple myeloma, a cancer of the plasma cells found in the bone marrow. The data were obtained from three consecutive clinical trials evaluating aggressive chemotherapy regimens in conjunction with autologous transplantation conducted at the Myeloma Institute for Research and Therapy, University of Arkansas for Medical Sciences [Barlogie et al., 2006]. The outcome for patients with myeloma is known to be variable and is associated with clinical and laboratory measures [Greipp et al., 2005]. In this data set, potential predictors include age, gender, several laboratory variables measured at the baseline and 348 SNPs for candidate genes representing functionally relevant polymorphisms playing a role in normal and abnormal cellular functions, inflammation and immunity, as well as for some genes thought to be associated with differential clinical outcome response to chemotherapy. Data on a total of 818 patients were available for analysis. The primary endpoint was early failure of treatment, defined as death, disease progression or fatal toxicity within 18 months of registration. The available SNPs included either individual or small numbers of tag SNPs on genes of interest. We coded each of the SNPs as two binary predictors corresponding to a dominant and a recessive effect rather than investigating additive genetic effects. We imputed missing SNPs based on marginal frequencies. The power of any association analysis with this number of subjects and events is limited, so we consider the analysis here only as illustrative of what can be conducted using weighted association tests.

For weighting, we considered two laboratory variables: serum β2 microglobulin and serum creatinine, which are both associated with extent of disease and renal function. None of the marginal test statistics were significant after either Bonferroni or permutation adjustment of the test statistics to deal with the multiple SNP comparisons. To construct adaptive subgroup weighted statistics we constructed the following basis functions for Z equal to serum β2 microglobulin and Z equal to serum creatinine: $\{Z < c_{.25}\}, \{Z \geq c_{.25}\}, \{Z < c_{.50}\}, \{Z \geq c_{.50}\}, \{Z < c_{.75}\}, \{Z \geq c_{.75}\}$, where c_q is the qth quantille of the covariate distribution. While we used step-function or subgroup type basis functions on continuous variables, the method allows other more smooth basis functions and/or binary clinical factors if they had been of interest.

Figure 3, panel 1, shows the increased values of maximal subgroup test statistics compared to marginal test statistics for each SNP. While the maximum subgroup SNP test statistic is increased from 3.29 to 4.27, a permutation test based on 1,000 simulations does not indicate promise of this simple maximal score statistic ($p = 0.339$). The permutation test considered all possible test statistics based on the recessive and dominant coding and the maximum statistic was calculated across all SNPs; hence, the P-values we present address the multiple comparisons with respect to SNPs. Panel 2 shows a pairwise scatter plot of stage-wise statistics involving both variables serum β2 microglobulin and creatinine (using a fixed tuning parameter of 2.5) versus maximal subgroup statistics. Here one SNP (rs4809960) test statistic is increased to 4.76 and the corresponding permutation P-value is $p = 0.013$. Figure 4 gives another representation of the increased magnitude of the different weighted association test statics. Yellow indicates the magnitude of the marginal statistics, green indicates the increase in magnitude of using the maximal test statistics, and blue is the further increase using stage-wise weighting.

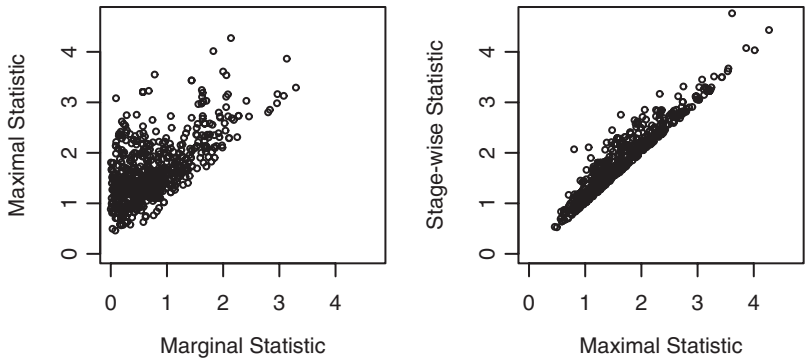

Fig. 3. Plot of SNP test statistics by methods: marginal, maximal and stage-wise statistics.

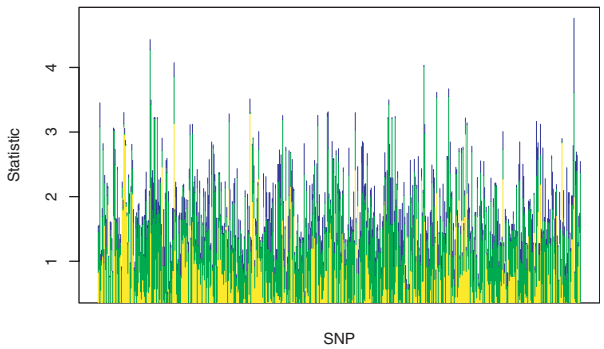

Fig. 4. SNP test statistics. Lightest gray (yellow in the online version) indicates the magnitude of the marginal statistic, medium gray (green online) indicates the increase in magnitude of using the maximal test, and darkest gray (blue online) is the further increase using stage-wise method.

As shown in the panels, more complex statistics must be at least as large as the less adaptive statistics. For instance, for the stage-wise methods restricted to a single parameter the test statistic will be equivalent to the maximal subgroup statistic. We have shown that resampling techniques are needed to evaluate relative performance of this more flexible testing.

SIMULATIONS

To evaluate the performance of adaptively weighed test statistics more generally, we simulated hypothetical SNP association case-control studies. While typically hundreds or thousands of SNPs would be evaluated in Genome Wide Association Studies (GWAS), the key components impacting the performance of the method relate to the outcome probability model linking the SNP and environmental factor to outcome, the allele frequency and the potential correlation of the observed SNP to causal SNP. The performance of the methods is evaluated in the presence of one or more continuous environmental variables potentially modifying the association between causal SNP and disease status.

The following logistic model was used to generate the binary phenotype,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 G + \beta_2 h(Z) + \beta_3 G h(Z),$$

where $h(Z)$ is a function of one or more environmental or clinical factors. We assume binary SNPs with allele frequency 0.1 and 0.2; the modifying variable is assumed to be continuous but is linked to outcome by a subset function $\{Z<c\}$. We consider two cases: (A) the single environmental variable is associated with the outcome variable and (B) multiple (five) observed environmental variables are observed and linked to the causal association through two of the variables. In the first case the single continuous variable is transformed into basis functions $(1, \{Z_1<c_{.25}\}, \{Z_1<c_{.50}\}, \{Z_1<c_{.75}\})$ and the linking function is $h(Z) = \{Z_1<c_{.50}\}$. The second model uses five variables and corresponding basis functions $(1, \{Z_j<c_{.25}\}, \{Z_j<c_{.50}\}, \{Z_j<c_{.75}\})$, $j = 1, \ldots, 5$, and the linking function depends on the linear combination of two variables $h(Z) = \{Z_1<c_{.50}\} - \{Z_2<c_{.50}\}$. Given that the second model is a linear combination, we expect a method that allows multiple environmental factors, such as the stage-wise test statistic, to perform better than themarginal, subgroup, or maximal test statistics. We evaluated four test statistics, the marginal, maximal, stage-wise, and a global (full) test statistic which uses least squares to estimate the weights for all basis functions in the score test statistic. Type I error was controlled by the Lin [2005] method applied to the weighted scores using 4,000 samples. The 4,000 simulations were used to calculate the mean and variance the null distribution of the test statistics. For each scenario, 1,000 simulations were used to calculate the alternative mean and variance of the test statistics and the power was calculated based on the assumption of large sample normality. This provides an approximate method to calculate power when using low type I error rates without conducting a huge number of simulations. Clearly, much larger numbers of simulations would be needed to calculate type I error rates for α 0.001 and 0.00001 using simple empirical counting. Figures 5 and 7 show estimates for a type I error of 0.001. The analyses were repeated for a type I error of 0.00001 and results are shown in Figures 6 and 8. For Figures 5–8 the true null cases only correspond to the interaction parameter $\beta_3 = 0$ on the left side of the each of top row of panels. Note, for other panels, even if $\beta_3 = 0$, power will in general deviate from type I error rates due to the impact of the non-zero main effects.

For Model A (Fig. 5) involving a single association modifier, one can observe that the marginal test statistic performs at least as well as the other test statistics if the subgroup effect small relative to the main effect (plots given in second row of the panel). However, for a wide

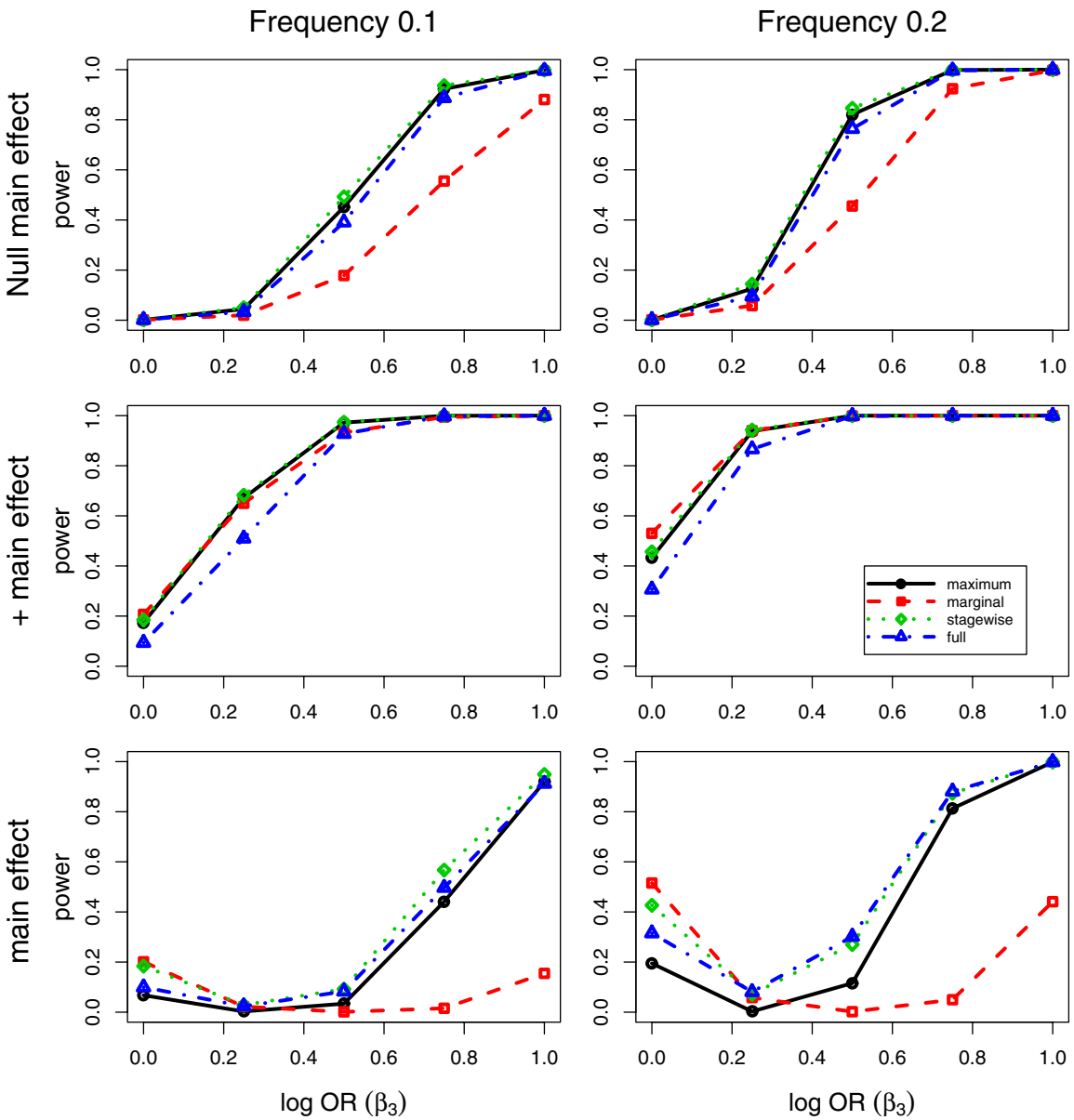

Fig. 5. Model A: single modifier. Power for testing using 2,000 cases and controls. The four statistics include: maximum (maximum of marginal and subgroups), marginal (marginal test), stage-wise (stage-wise weighted test), full (unconstrained weighted test) with $\alpha = 0.001$.

range of values away from the additive model, the maximum test statistic performs as well or better than the marginal test statistic. The case of negative main effect but positive subgroup effect leads to the worst performance for the marginal test statistic. While the power is enhanced for the maximal statistic, the global test using all four basis functions and the stage-wise method lead to improvements over the maximal statistic, with respect to power for this more complex interaction. The analyses for a type I error of 0.00001 in are given in Figure 6.

For model B, as shown in Figure 7, the marginal statistic performs best when there is a main effect model. However, as we move toward more complex interaction models, it has very limited power (right sides of rows 2 and 3). The maximal statistic can recover some of the power in a number of cases; however, given that the statistic depends at most only on a single environmental covariate, it too has limitations. The full model weighting (here depending on all five variables—corresponding to 16 basis functions) suffers from increased variance. The stage-wise estimated test statistic is the overall winner in the range of models. The analyses for a type I error of 0.00001 in are given in Figure 8.

Given that it is difficult to view the low type I error rates in given in the upper row of panels (at $\beta_3 = 0$) for each of

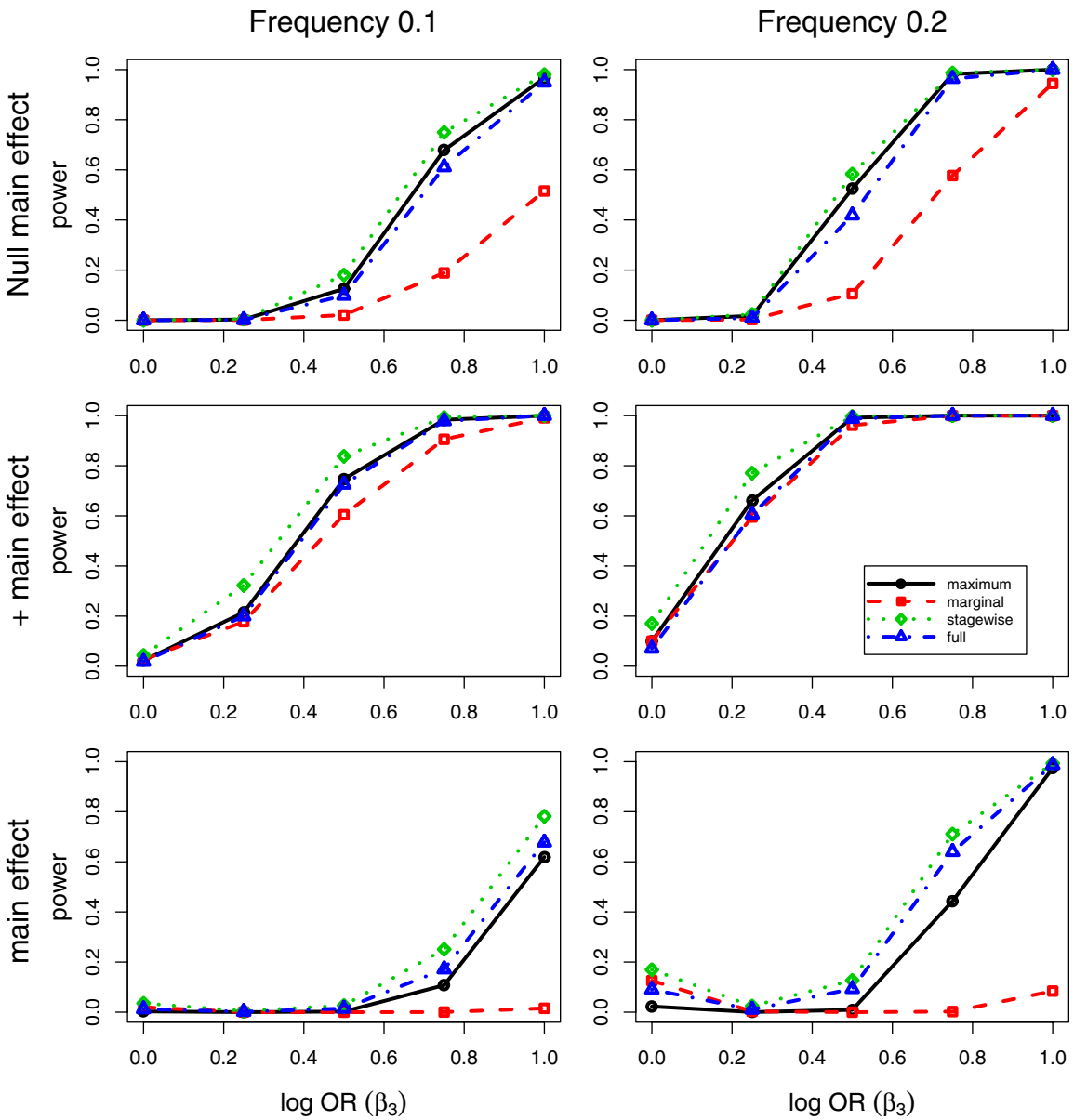

Fig. 6. Model A: single modifier. Power for testing using 2,000 cases and controls. The four statistics include: maximum (maximum of marginal and subgroups), marginal (marginal test), stage-wise (stage-wise weighted test), full (unconstrained weighted test) with $\alpha = 0.00001$.

the plots, we have presented them in Table I. Note, while the specified low type I error rates based on our simulation based estimates using the normal approximation show some variability in Table I, the alternative of using strictly empirical counting would require 100s of times more simulations.

Only the simple case with a stage-wise model selection with a tuning parameter equal to 2.5 was presented (although using other tuning parameters such as 3.5 lead to similar conclusions). One could also evaluate several tuning parameters, in the range 2–5. While the effect modification was determined by two of the five analyzed variables, we believe more generally that only a small number of effect modifiers should be considered in the score test weighting to control variance.

Additional simulations where the causal SNP was only correlated with the observed SNP and where the environmental factor $h(Z)$ was hidden by a small amount of independent error were conducted and yielded similar shaped power profiles, albeit with reduced power. Furthermore, our simulations above only considered categorical SNP coding. One could also investigate additive coding in number of copies of a minor allele versus categorical coding of the three genotypes. While the absolute power will depend on true genetic association, since the choice of coding represents the genetic effect in the hypothetical gene-environmental interaction model,

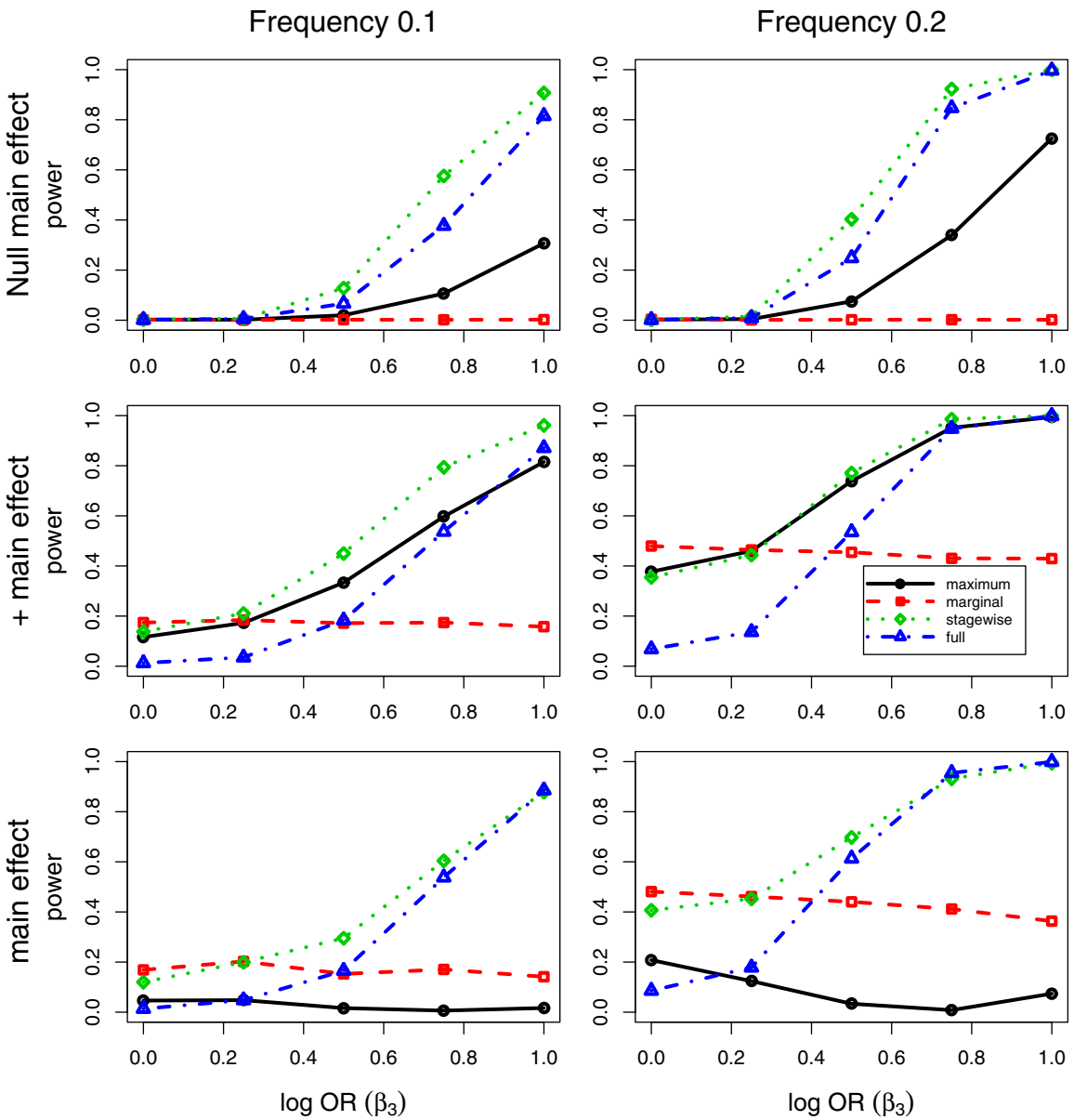

Fig. 7. Model B: two variable modifier. Power for testing 2,000 cases and 2,000 controls. The four statistics include: maximum (maximum of marginal and subgroups), marginal (marginal test), stage-wise (stage-wise weighted test), full (unconstrained weighted test) with $\alpha = 0.001$.

not the nature of the interaction, we would anticipate the results to parallel those which we have presented.

DISCUSSION

In this article we have developed an adaptive selection or weighting strategy to improve the power of testing the association of genetic factors with disease outcome. The goal is not to specifically address the question of whether or not there is an interaction between some measured clinical or environmental factor, but rather to modestly expand the association search space to better identify associations with outcome. The new method is based on the often plausible assumption that genetic associations may be stronger within specific subgroups of subjects in clinical or epidemiologic studies. Simulation results support that the strategy can lead to substantially improved power in situations where stronger genetic associations exist in subgroups of subjects.

Our approach constructs statistics using only the marginal scores on the genetic factors. While our interest was primarily in SNP association studies, such a strategy may also be of use in studying gene expression to outcome. Given the construction works on marginal scores, either approximate permutation sampling or resampling methods applied to the scores, such as Lin [2005], can be used to calibrate inference for the multiple testing and control the type I error. A natural extension to adaptively choosing the maximal statistic is a stage-wise

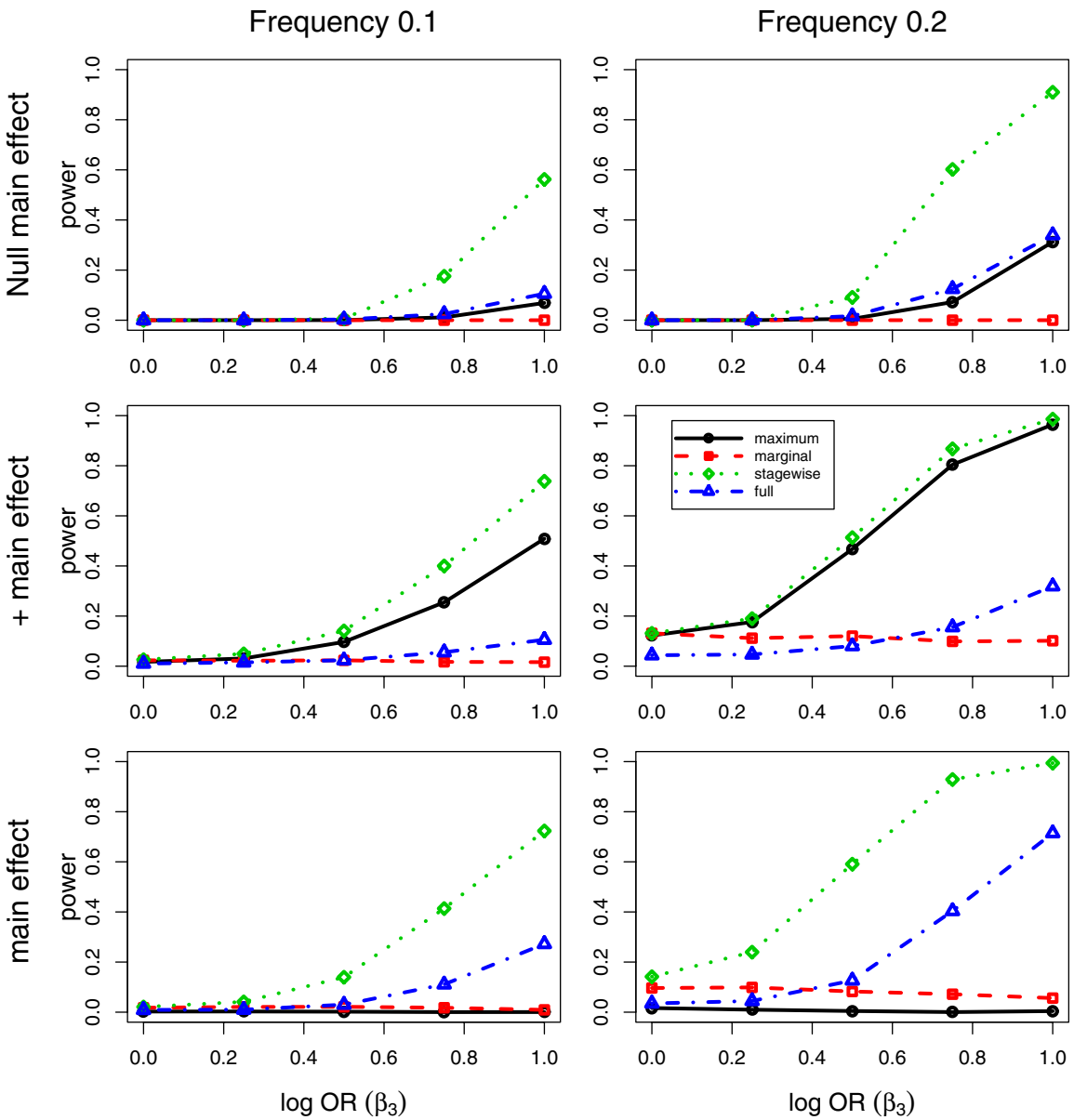

Fig. 8. Model B: two variable modifier. Power for testing 2,000 cases and 2,000 controls. The four statistics include: maximum (maximum of marginal and subgroups), marginal (marginal test), stage-wise (stage-wise weighted test), full (unconstrained weighted test) with $\alpha = 0.00001$.

method which maximizes the test statistic analogous to stage-wise boosting of regression models. The adaptive stage-wise statistic appears to perform well in a range of true underlying gene-environment interaction models.

We considered a set of predefined binary basis functions as the environmental terms in the weighted test statistic; however, they could more generally be either the originally coded or some transformed version of the environmental variables. In our simulation studies we used subgroup basis functions representing at least 25% of the sample size. Subgroups or binary environmental variables with lower frequencies could be considered, but very small frequencies could lead to increased variance and reduced power. While a rule is difficult to state in general, we anticipate lower bounds of 5–10% frequency for binary factors or subgroups would be appropriate for reasonably sized association studies.

In addition, in some settings one could use tree-based partitioning of the environmental variables to determine basis functions to be used in the test statistic weighting. In the case where the basis functions are not predefined, the original boosting version of the weighted test statistic may have to be used rather than the computationally efficient LAR weighting. However, with a large number of genetic variables, this strategy would not typically be feasible.

If the goal is to directly detect interactions in high dimensional settings, other approaches are useful. An important strategy is to control the search of possible interaction models by utilizing a well-defined filtering

TABLE I. Estimated type 1 errors for full null models A and B

Model A	target $\alpha = 10^{-3}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-5}$
Method	Allele Freq (0.1)		Allele Freq (0.2)	
Maximum	0.00204	2.13e−05	0.00212	1.02e−05
Marginal	0.00191	5.52e−05	0.00248	1.20e−05
Stage-wise	0.00288	1.93e−05	0.00084	5.00e−06
Full	0.00220	1.93e−05	0.00147	1.91e−05
Model B	target $\alpha = 10^{-3}$	$\alpha = 10^{-5}$	$\alpha = 10^{-3}$	$\alpha = 10^{-5}$
Method	Allele Freq (0.1)		Allele Freq (0.2)	
Maximum	0.00060	2.80e−06	0.00115	1.20e−05
Marginal	0.00089	5.70e−06	0.00089	1.31e−05
Stage-wise	0.00074	7.60e−06	0.00158	2.12e−05
Full	0.00142	1.38e−05	0.00083	6.70e−06

procedure on marginal effects [for instance see Kooperberg and LeBlanc, 2008]. In addition, in randomized trials an assumption of independence of the environmental and genetic factors can lead to greater efficiency when used as part of the testing procedure [for instance see Chatterjee and Carroll, 2005]. Multistage sampling and independence assumptions for testing for interactions were also considered by Dai et al. [2008].

We also note that subgroup weighting has potential use in testing treatment efficacy in randomized clinical trials. Suppose the treatment is more efficacious in tumors with differing gene expression levels. Given that the primary question involves treatment effect, the testing could be adaptively weighted based on single or multiple gene expression factors.

ACKNOWLEDGMENTS

The authors thank John Crowley for use of the Myeloma data set. This work was funded in part by grants CA 90998, CA 125489, CA 74781, and CA 53996 from the National Institutes of Health. The authors also thank John Crowley and Li Hsu for helpful discussions and Mark Blitzer for article review and preparation.

REFERENCES

Barlogie B, Tricot G, Rasmussen E, Anaissie E, van Rhee F, Zangar M, Fassas A, Hollmig K, Pineda-Roman M, Shaughnessy J, Epstein J, Crowley J. 2006. Total therapy 2 without thalidomide in comparison with total therapy 1: role of intensified induction and posttransplantation consolidation therapies. Blood 107:2633–2638.

Breiman L, Friedman J, Olshen R, Stone C. 1984. Classification and Regression Trees. Belmont CA: Wadsworth, 1984.

Chatterjee N, Carroll R. 2005. Semiparametric maximum-likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 92:399–418.

Chatterjee N, Zeynep K, Moslehi R, Peters U, Wacholder S. 2006. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Gen 79:1012–1016.

Clark C. 1961. The greatest of a finite set of random variables. Oper Res 9:145–162.

Dai JY, LeBlanc M, Kooperberg C. 2008. Semiparametric estimation exploiting covariate independence in two-phase randomized trials. Biometrics Published online, May 12.

Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression. Ann Stat 32:407–451.

Friedman JH, Hastie T, Tibshirani R. 2000. Additive logistic regression: a statistical view of boosting. Ann Stat 28:337–407.

Freund Y, Shapire T. 1996. Experiments with a new boosting algorithm . In: *Machine Learning: Proceedings of the Thirteenth International Conference;* pp 148–156.

Greipp PR, San Miguel J, Durie BG, Crowley JJ, Barlogie B, Bladé J, Boccadoro M, Child JA, Avet-Loiseau H, Kyle RA, Lahuerta JJ, Ludwig H, Morgan G, Powles R, Shimizu K, Shustik C, Sonneveld P, Tosi P, Turesson I, Westin J. 2005. International staging system for multiple myeloma. J Clin Oncol. 23:3412–3420.

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

Hauser ER, Watanabe RM, Duren WL, Bass MP, Langefeld CD, Boehnke M. 2004. Ordered subset analysis in genetic linkage mapping of complex traits. Genet Epidemiol 27:53–63.

Kooperberg C, LeBlanc M. 2008. Increasing the power of identifying gene x gene interactions in genome-wide association studies. Genet Epidemiol 32:255–263.

Lin DY. 2005. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics 21:781–787.

Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37:413–417.

Tibshirani T. 1996. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B (Methodol) 58:267–288.