Harvard University
Harvard University Biostatistics Working Paper Series

Year 2008 *Paper* 82

A Powerful and Flexible Multilocus Association Test for Quantitative Traits

Lydia Coulter Kwee* Dawei Liu† Xihong Lin‡
Debashis Ghosh** Michael P. Epstein††

*Emory University
†Brown University, daweiliu@stat.brown.edu
‡Harvard University, xlin@hsph.harvard.edu
**Penn State University, debashis.ghosh@ucdenver.edu
††Emory University, mepstein@genetics.emory.edu

http://biostats.bepress.com/harvardbiostat/paper82

A Powerful and Flexible Multilocus Association Test for Quantitative Traits

Running Title: Multilocus Association Test for QTL Mapping

Lydia Coulter Kwee[1], Dawei Liu[3], Xihong Lin[4], Debashis Ghosh[5], and Michael P. Epstein[2]

Departments of Biostatistics[1]and Human Genetics[2], Emory University, Atlanta, GA

[3]Center for Statistical Sciences, Brown University, Providence, RI

[4]Department of Biostatistics, Harvard University, Boston, MA

[5]Department of Statistics, Pennsylvania State University, State College, PA

Address for Correspondence: Michael P. Epstein, Ph.D.

Department of Human Genetics, Emory University School of Medicine

615 Michael Street, Suite 301, Atlanta, GA 30322

Phone: (404)712-8289

Fax: (404)727-3949

E-mail: mepstein@genetics.emory.edu

SUMMARY

Association mapping of complex traits typically employs tagSNP genotype data to identify functional variation within a region of interest. However, considerable debate exists regarding the most powerful strategy for utilizing such tagSNP data for inference. A popular approach tests each tagSNP within the region individually, but such tests could lose power due to incomplete linkage disequilibrium between the genotyped tagSNP and the functional variant. Alternatively, one can jointly test all tagSNPs simultaneously within the region (using genotypes or haplotypes), but such multivariate tests have large degrees of freedom that can also compromise power. Here, we consider a semiparametric model for quantitative-trait mapping that uses genetic information from multiple tagSNPs simultaneously in analysis but produces a test statistic with reduced degrees of freedom compared to existing multivariate approaches. We fit this model using a dimension-reducing technique called least-squares kernel machines, which we show is identical to analysis using a specific linear mixed model (which we can fit using standard software packages like SAS and R). Using simulated SNP data based on real data from the International HapMap Project, we demonstrate our approach often has superior performance for association mapping of quantitative traits compared to the popular approach of single-tagSNP testing. Our approach is also flexible, as it allows easy modeling of covariates and, if interest exists, high-dimensional interactions among tagSNPs and environmental predictors.

INTRODUCTION

The arrival of improved high-throughput genotyping technology has accelerated the use of association methods for dissecting the genetic mechanisms of complex traits. Using panels of single-nucleotide polymorphisms (SNPs), association methods seek to identify those genetic markers that are either functional or are in linkage disequilibrium (LD) with a functional variant. In the process of association mapping of a complex trait, interest will eventually focus on regions or genes that are identified either from interesting signals from previous gene-mapping work or from perceived biological relevance to the trait of interest. To examine whether such a region harbors a functional variant, a study could genotype and subsequently analyze all polymorphic SNPs in the genetic interval. However, the likely existence of LD in the region will induce correlation among such SNPs such that many of the genetic markers provide redundant information for association analysis. Therefore, many association studies instead genotype a reduced set of SNPs within the region (called tagSNPs) that effectively captures the genetic variation from all SNPs within the region but substantially reduces the genotype cost. Studies can identify relevant tagSNPs by applying existing selection algorithms (Carlson et al. 2004; Stram 2004; de Bakker et al. 2005) to SNP genotype data from existing public databases of human genetic variation, such as the International HapMap Project (2005).

In this article, we focus on the use of tagSNP data to identify genetic regions that influence a quantitative trait of interest using samples collected under a population-based study design. Currently, considerable debate exists regarding the most powerful manner by which to utilize such tagSNP data in association analysis. A simple and popular approach considers association testing of each individual tagSNP with the quantitative trait of interest (using regression or ANOVA methods) followed by inference on the maximum of the resulting single-tagSNP statistics. Due to the testing of multiple correlated tagSNPs within a region, one must implement an appropriate multiple-testing

procedure to ensure appropriate significance levels. Such multiple-testing corrections may include permutation procedures, efficient Monte Carlo procedures (Lin 2005), or a Bonferroni correction based on the effective number of independent tests within the candidate region (Nyholt 2004).

While the testing of individual tagSNPs is simple to implement, such methods may have low power if each tested tagSNP is in incomplete LD with the (untyped) functional variant. This potential liability of single-tagSNP approaches led to the development of novel statistical approaches that consider the joint effects of tagSNPs simultaneously within analysis. Such multivariate tagSNP analyses of quantitative traits typically apply multilinear regression to model a subject's trait as a function of a vector of covariates corresponding to either the subject's genotypes at the various tagSNPs or the subject's pair of tagSNP-based haplotypes (Schaid et al. 2002; Zaykin et al. 2002; Tzeng et al. 2006). Such regression procedures produce omnibus test statistics that follow a χ^2 distribution with degrees of freedom equal to either the number of modeled tagSNPs (for a genotype-based analysis) or the number of observed haplotypes minus one (for a haplotype-based analysis)

As these multivariate approaches combine genetic information from multiple tagSNPs simultaneously into analysis, they intuitively should provide greater power to detect functional variants compared to tests of individual tagSNPs. However, many simulation studies have found the opposite result to be true: multivariate approaches typically have similar or reduced power relative to single-SNP procedures (Chapman et al. 2003; Roeder et al. 2005; Rosenberg et al. 2006) unless the trait originates from the effect of a specific haplotype rather than a specific SNP (Rosenberg et al. 2006). An explanation for this surprising finding is that multivariate procedures produce test statistics with degrees of freedom that will increase substantially (particularly in the situation of haplotype analysis) with the number of modeled tagSNPs within the region (Tzeng et al. 2006). As the degrees of freedom of the test statistic increases, it follows

that the power of the omnibus test will decrease. Therefore, it is likely that any information gained from joint consideration of multiple tagSNPs in association analysis of a quantitative trait will subsequently be lost by dealing with test statistics with large degrees of freedom.

Given these results, we seek to develop a novel statistical approach for association mapping of quantitative traits that incorporates all tagSNPs (and, hence, all valuable genetic information) within a region into the association analysis but produces test statistics with smaller degrees of freedom than the multivariate approaches described earlier. Existing statistical work in this area generally approaches the problem in one of two broad ways. The first strategy applies a dimension-reduction procedure such as a Fourier transformation (Wang and Elston 2007) or principal components (Gauderman et al. 2007) to the tagSNP data in the region to produce a reduced set of orthogonal genetic predictors that contain the majority of information found in the original tagSNPs. One then models this reduced set of genetic predictors within a multilinear regression framework and constructs appropriate omnibus tests for inference (which should have smaller degrees of freedom than a standard multivariate test). The second strategy calculates a measure of average tagSNP similarity for each pair of subjects and compares the pairwise genetic similarity with the pairwise trait similarity (Schaid et al. 2005; Wessel and Schork 2006). One can measure such tagSNP similarity using a 'kernel' function that reduces a comparison of multiple tagSNPs for a pair of subjects into a single scalar factor. Due to this phenomenon, resulting statistics using kernel functions typically have small degrees of freedom; for example, Schaid et al. (2005) constructed a kernel-based U-statistic for case-control association analysis that has only 1 degree of freedom. In addition, the use of a kernel function is appealing since it allows for the inclusion of prior information (such as bioinformatic relevance or association signals from tagSNPs in an independent study) in the form of weights to assist in the evaluation of the tagSNP similarity. One drawback of these existing similarity-based approaches is that they do not

easily allow for covariates and sometimes require computationally-intensive permutation procedures to establish significance (Wessel and Schork 2006).

In this article, we propose a novel approach for association mapping of quantitative traits which uses all tagSNP data simultaneously in analysis but produces test statistics with smaller degrees of freedom than multivariate tagSNP approaches. We base our approach on a semiparametric-regression framework (Ruppert et al. 2003) that regresses the quantitative trait of interest on a smooth nonparametric function of the tagSNP genotypes within the region, adjusting for the parametric effects of any covariates of interest. As we will show, we can model this nonparametric function of the tagSNP data in a reduced-dimension space that is induced by a user-defined kernel function. As a result, statistics that test for association between the trait and the nonparametric function of the tagSNP effects should have reduced degrees-of-freedom compared to existing multivariate tests and, hence, should have improved power to detect functional variants. Unlike existing dimension-reduction techniques, we will show our approach permits us to incorporate valuable prior information in the analysis via the kernel function. Unlike existing similarity approaches, we will show our approach can easily allow for covariates and interaction terms. Further, we can rely on asymptotic theory to establish significance of the resulting tests, avoiding computationally-intensive permutation procedures.

We estimate the parameters in our proposed semiparametric model using an elegant dimension-reduction technique called least-squares kernel machines (LSKM) (Rasmussen and Williams 2006; Liu et al. in press), which has been applied previously to high-dimensional microarray analysis. While LSKM fitting of a semiparametric model appears complicated, Liu et al. (in press) noted that one can represent the LSKM procedure using a specific form of a linear mixed model, such that one can estimate and test the nonparametric function of the tagSNP data using simple restricted-maximum-

likelihood procedures that are typically applied to mixed models and are available in common statistical software packages such as SAS and R.

In subsequent sections, we develop our semiparametric model and show how we can estimate model parameters using the LSKM maximization approach of Liu et al. (in press). We then show how one can represent the LSKM approach in terms of a linear mixed model that facilitates testing of the nonparametric function of the tagSNP genotype data. Using simulated tagSNP data based on real data from the International HapMap Project (2005), we show that our proposed semiparametric approach often has improved power to detect an association between a genetic region and a quantitative trait compared to the popular single-tagSNP testing approach. We also describe a variety of valuable gene-mapping extensions of our semiparametric approach in the Discussion.

MATERIALS AND METHODS

Notation Using a population-based study design, we assume a sample of N unrelated subjects. Let Y_j denote the quantitative trait value for subject j ($j = 1, ..., N$). We assume each subject is genotyped at S tagSNPs within a defined candidate gene or region of interest. We let $G_{j,s}$ denote the genotype of subject j at tagSNP s ($s = 1, ..., S$) and let $G_j = (G_{j,1}, G_{j,2}, \ldots, G_{j,S})$ denotes an (S x 1) vector of all tagSNP genotypes for subject j. For tagSNP s, we code $G_{j,s}$ to be the number of copies of the minor allele that the subject j possesses at the tagSNP such that the predictor takes values of 0, 1, or 2. These values correspond to an additive model of allelic effect; we can consider alternative coding scenarios for $G_{j,s}$ under dominant and recessive models, if desired. Finally, we let X_j denote a (p x 1) vector of measured environmental covariates for subject j.

Semiparametric Regression Model We propose the use of semiparametric regression to model the relationship between the outcome Y_j and the tagSNPs G_j,

adjusting for potential covariates in X_j. We can write this semiparametric model as the following:

$$Y_j = X_j^T \beta + h(G_j) + e_j \quad (1)$$

Here, $h(G_j)$ denotes a nonparametric function of the tagSNP genotype data G_j that resides in some function space κ. β is a (p x 1) vector of regression coefficients describing the effects of X_j, which are modeled parametrically. Finally, e_j is a random subject-specific environmental effect, which we assume to be normally-distributed with mean 0 and variance σ^2.

Within the model in (1), interest focuses primarily on the estimation of the nonparametric function of the tagSNP data h and its relationship to the trait outcome Y_j. Secondary interest focuses on the estimation and testing of β to assess the effects of the covariates in X_j on Y_j. As we are using a semiparametric framework in (1), traditional maximization procedures for linear regression models are not applicable in this setting. To estimate h and β, we instead propose the use of the dimension-reducing LSKM procedure to analyze our high-dimensional data (which, in our context, refers to the tagSNP genotype data in G_j). Using the LSKM approach of Liu et al. (in press), we show in Appendix A that we obtain the following estimates of h and β in (1):

$$\widehat{h} = K(K + \lambda I)^{-1}(Y - X\widehat{\beta}) \quad (2)$$

$$\widehat{\beta} = [X^T(K + \lambda I)^{-1}X]^{-1}X^T(K + \lambda I)^{-1}Y \quad (3)$$

Here, $Y = (Y_1, ..., Y_N)^T$ is an (N x 1) vector of the trait values for all subjects and X is an (N x p) matrix of environmental covariates for all subjects. Further, I denotes an (N x N) identity matrix. Finally, there are two additional terms in (2) and (3) that are important to discuss. The first term is the parameter λ, which denotes a scalar smoothing parameter. As we will show in subsequent sections, λ plays an important role in

constructing appropriate test statistics to assess whether the nonparametric function h of the tagSNP genotype data influences Y.

The second important term in (2) and (3) is K, which denotes an (N x N) kernel matrix that is a function of the tagSNP genotype data in the region. In particular, the $(j,l)^{th}$element of K denotes a kernel $k(G_j, G_l)$ that is a scalar function of the tagSNP genotypes of subjects j and l. Broadly speaking, $k(G_j, G_l)$ will often be a measure of pairwise tagSNP genotype similarity across the region. As $k(G_j, G_l)$ is scalar, the kernel intuitively serves as a dimension-reducing function as it collapses the comparison of the multidimensional tagSNP vectors G_j and G_l into a simple scalar factor.

A variety of choices exist for the kernel function $k(G_j, G_l)$. However, the choice of kernel is not arbitrary. In particular, the kernel function in K within (2) and (3) must satisfy the conditions of Mercer's Theorem (Cristianini and Shawe-Taylor 2000), which includes the condition that the K matrix must be positive semidefinite (i.e. the eigenvalues of K must be positive). While many common kernels fulfill Mercer's Theorem, we note that certain kernels in the literature which may be intuitively appealing for SNP data, such as the quadratic kernel in the U-statistic approach of Schaid et al. (2005), fail to meet this criteria and cannot be used in the proposed LSKM procedure.

For this article, we focus on kernel functions that are based on the number of alleles shared identical by state (IBS) by subjects j and l at the tagSNPs within the region (Wessel and Schork 2006). The IBS kernel takes the form

$$k(G_j, G_l) = \frac{\sum_{s=1}^{S} IBS(G_{j,s}, G_{l,s})}{2S} \qquad (4)$$

where $IBS(G_{j,s}, G_{l,s})$ denotes the number of alleles shared IBS (0, 1, or 2) by subjects j and l at tagSNP s. An appealing feature of the IBS kernel is that we can augment it to include tagSNP-specific weights that can incorporate valuable prior information into

analysis to potentially improve performance. Define w_s as a scalar weight for tagSNP s. We can then define an weighted-IBS kernel based on (4) as the following:

$$k(G_j, G_l) = \frac{\sum_{s=1}^{S} w_s IBS(G_{j,s}, G_{l,s})}{\sum_{s=1}^{S} w_s} \quad (5)$$

We focus on two potentially-valuable weights for use in the IBS kernel in (5). First, we consider a weight that upweights tagSNPs with a rare minor-allele frequency (MAF) and downweights tagSNPs with more common MAF. Such a weight could be valuable due to the potential for the information from tagSNPs with rare MAF to be smoothed over by the information from surrounding tagSNPs with more common MAF. To upweight tagSNPs with rare MAF, we apply the weight $w_s = 1/\sqrt{q_s}$, where q_s denotes the MAF of tagSNP s $(s = 1, ..., S)$. Other MAF weights are certainly possible, such as $w_s = 1/q_s$, but there is concern that such stronger weights may substantially diminish the information provided by those tagSNPs with common MAF.

In addition to weights based on MAF, we can use weights based on prior evidence of association between the tagSNP and the trait (or a related trait of interest) in an independent dataset. Here, we let $w_s = -log_{10}(p_s)$ where p_s is the p-value for the test of tagSNP s with the trait in the independent dataset. Intuitively, such weights will upweight SNPs showing stronger prior evidence of association and downweight SNPs that demonstrate weaker prior evidence of association. As noted in the Discussion, we feel such weights are, or will be, readily available from relevant genetic literature or public release of data from whole-genome association studies.

Relationship to Linear Mixed Models Inspection of $\widehat{h}$ in (2) shows that the nonparametric function in (1) models the tagSNP genotype data in a reduced-dimension space κ induced by the chosen kernel function in K. Next, we focus on constructing an appropriate test statistic to evaluate whether the function h of the tagSNP genotype data

is associated with the trait of interest. That is, we wish to construct a test statistic to evaluate the null hypothesis $H_0 : h = 0$, where we model h using equation (1). To facilitate the construction of such a test statistic, Liu et al. (in press) noted that LSKM-based estimation of $\widehat{h}$ and $\widehat{\beta}$ is analogous to the estimation of random and fixed effects, respectively, within a specific linear mixed model. Therefore, we can use a likelihood framework based on a mixed model to construct an appropriate test statistic for inference.

To develop the mixed-model representation of the LSKM analysis using the semiparametric model in (1), we consider the following linear mixed model:

$$Y = X\beta + U + E \tag{6}$$

where Y denotes the earlier trait vector and X denotes the earlier matrix of fixed environmental covariates with related regression-coefficient vector β. Within (6), we denote U as a vector of random effects belonging to the tagSNP genotype data and denote E as a vector of random effects due to subject-specific environment.

Suppose we assume that the random tagSNP effects in U follow a multivariate normal distribution with mean 0 and variance-covariance matrix $\frac{\sigma^2}{\lambda}K$, where K is our kernel matrix, λ denotes the smoothing parameter discussed earlier, and σ^2 denotes the variance due to subject-specific environment. Further, suppose we assume that E also follows a multivariate normal distribution with mean vector 0 and variance-covariance matrix $\sigma^2 I$, where I denotes the identity matrix. Under these assumptions, we can use restricted maximum-likelihood (REML) procedures to show that the best-linear unbiased estimators of the random effects U and the fixed effects β in the linear mixed model are

$$\widehat{U} = K(K + \lambda I)^{-1}(Y - X\widehat{\beta}) \tag{7}$$

$$\widehat{\beta} = [X^T(K + \lambda I)^{-1}X]^{-1}X^T(K + \lambda I)^{-1}Y \tag{8}$$

One can see that the estimates of $\widehat{U}$ and $\widehat{\beta}$ in (7) and (8) are exactly the same as the estimates of $\widehat{h}$ and $\widehat{\beta}$ in equations (2) and (3), respectively, using LSKM estimation of the

semiparametric model in (1). This important result shows that we can perform our LSKM multilocus analysis using a straightforward linear mixed model that is easy to implement using existing statistical software packages for mixed models (such as SAS PROC MIXED).

Testing the Nonparametric Function The relationship between LSKM and the linear mixed model implies that we can test $H_0 : h = 0$ in the semiparametric model by appropriate testing of the existence of the random tagSNP effect U in the linear mixed model in (6). As noted earlier, we assume U follows a multivariate-normal distribution with mean vector 0 and covariance matrix $\frac{\sigma^2}{\lambda}K$. Assume $\tau = \frac{\sigma^2}{\lambda}$ such that we rewrite the covariance matrix as τK. As elements of K will be non-zero, it is straightforward to show that the random tagSNP effect U does not exist when $\tau = 0$. Therefore, the test of $H_0 : h = 0$ in the semiparametric model (1) corresponds to the test of $H_0 : \tau = 0$ in the linear mixed model (6).

To test $H_0 : \tau = 0$, we propose the use of the score statistic of Liu et al. (in press). The score statistic includes estimates of the unknown parameters $\widehat{\beta}$ and $\widehat{\sigma}^2$ under H_0, which are obtained using REML procedures. The score statistic then takes the form

$$S_\tau = \frac{1}{2\widehat{\sigma}^2}(Y - X\widehat{\beta})^T K (Y - X\widehat{\beta}) \qquad (9)$$

Since $\tau \geq 0$, we are testing the parameter of interest on its boundary value. As a result, S_τ does not follow a standard χ^2_1 distribution under H_0 and, instead, follows a complicated mixture of χ^2_1 distributions. To simplify inference, we use a Satterthwaite procedure (described in Appendix B) to approximate the distribution of S_τ.

Simulations We used simulations to assess the performance of our semiparametric approach in a typical candidate-gene study. For genetic data, we used simulated tagSNP data based on the CEU genotypes from build 35 of the International HapMap Project (2005). We based our simulations on the LD structure of two genes: *CHI3L2* and *NAT2*. *CHI3L2* is 15.8 kb long, with 37 polymorphic SNPs in the CEU

sample. *NAT2* spans 9.9 kb, with 20 polymorphic SNPs in the same sample. Within each gene, we selected tagSNPs using the Tagger program (de Bakker et al. 2005). We allowed for multimarker tagging and captured all polymorphic markers in each gene with $R^2 > 0.8$, regardless of the marker's minor allele frequency. Using these criteria, we identified 10 tagSNPs for *CHI3L2* and 7 tagSNPs for *NAT2*. We show the LD structure of the tagged and non-tagged SNPs within *CHI3L2* and *NAT2* in Figures 1 and 2, respectively. Within each gene, we applied PHASE (Stephens et al. 2001; Stephens and Scheet 2005; Marchini et al. 2006) to the genetic data to estimate haplotype frequencies for the encompassed SNPs. We then generated relevant SNP genotype data at each gene for each subject using these estimated haplotype frequencies under the assumption of Hardy-Weinberg equilibrium.

To ensure our semiparametric approach had appropriate size, we first considered simulations under null models where none of the SNPs within the gene had an effect on our trait of interest. However, we did allow for trait-influencing effects from environmental predictors. Therefore, we simulated trait data under the following null model

$$Y_j = X_{E_j}\beta_E + e_j \tag{10}$$

Here, X_{E_j} denotes the coding vector of environmental covariates for subject j with respective effect-size vector β_E. We assumed that X_{E_j} contained both a binary covariate (with frequency of exposure of 0.506) and a continuous covariate (assumed to be normally distributed with mean 29.2 and variance 21.1). The assumed parameterization for the covariates closely mirrored those of relevant covariates in the FUSION study of type 2 diabetes (Valle et al. 1998). We assumed the effect size was 0.50 for the binary covariate and 0.03 for the continuous covariate. Finally, we let e_j denote a random subject-specific error term for subject *j*, which we generated under a normal distribution with mean 0 and variance 1.

We next considered simulations under alternative models where we selected one of the SNPs within the gene to serve as the functional variant. We allowed the functional variant to be either a typed tagSNP or an untyped SNP, but required the variant to have MAF greater than 0.05 (as done elsewhere, such as Roeder et al. 2005). Within *CHI3L2*, 30 of the 37 polymorphic SNPs fulfilled this criteria with 6 of these 30 polymorphisms being tagSNPs. Within *NAT2*, 17 of the 20 polymorphic SNPs fulfilled this criteria with 3 of the 17 polymorphisms being tagSNPs. Denoting the functional SNP as S^*, we generated the trait outcome for subject j using the following model

$$Y_j = X_{G_{j,S^*}}\beta_{S^*} + X_{E_j}\beta_E + e_j \quad (11)$$

Here, $X_{G_{j,S^*}}$ denotes the coding of the genotype at functional SNP S^* for subject j with respective effect size β_{S^*}. We considered additive, dominant, and recessive effects of the minor allele and chose β_{S^*} in each case such that the functional SNP S^* explained 3% of the trait variation, which is reasonable given that many complex traits originate from the effects of multiple genes each with small effect. We assumed values for X_{E_j} and β_E that were the same as those used in the null simulations.

For a given simulation design, we generated either 5000 datasets (for null models) or 1000 datasets (for alternative models), each consisting of 300 unrelated subjects. Each dataset contained trait data on all subjects, genotype data for the tagSNPs in the candidate gene, and environmental data on the covariates mentioned earlier. We assumed we did not observe genotypes at untyped SNPs (even though such untyped SNPs may be functional). We analyzed each dataset using both our proposed semiparametric approach and, as a benchmark, single-tagSNP statistics.

For our semiparametric approach, we analyzed the data three times. First, we used the unweighted IBS kernel in (4). Next, we used the weighted IBS kernel in (5) with weights based on the MAF of the tagSNP. Finally, we used a weighted IBS kernel with weights determined based on single-tagSNP p-values from an independently-generated

dataset. We wished to evaluate the performance of this last kernel when we simulated the independent dataset under both the same and different genetic model as our dataset under study. The primary purpose of a independent-dataset simulation under a different genetic model than the one used for the dataset of interest was to address whether inappropriate prior p-value weights from an independent dataset affected the size of our semiparametric approach. We investigated this issue by generating the dataset under study using the null model in (10) but generating the independent dataset using the alternative model (11) assuming a particular functional SNP.

For the single-tagSNP tests, we performed least-squares regression at each tagSNP in the gene under an additive model (allowing for the binary and continuous covariates) and tested the effect of the tagSNP using a Wald statistic. We retained the largest Wald statistic across the tested tagSNPs and used 5000 permutations of the data to establish the significance of this maximum statistic. We examined type-I error and power of the semiparametric and single-tagSNP approaches assuming a nominal significance level of $\alpha = 0.05$.

RESULTS

Table 1 provides the empirical type-I error results at nominal $\alpha = 0.05$ for our semiparametric method assuming the different IBS-based kernels described in the Methods. These results suggest our semiparametric approach has appropriate size regardless of the choice of kernel. In particular, we note that our semiparametric approach using p-value weights has appropriate size when we select weights using a dataset that is generated under a different model (i.e. is genetically heterogeneous) compared to the dataset under study . This result is important because it suggests that the choice of inappropriate p-value weights does not affect the size of our score statistic and, hence, does not affect the validity of our semiparametric approach. For comparison, we

analyzed the same datasets using the maximum of the single-tagSNP statistics, which also had appropriate size.

Figure 3 shows power results for simulations based on the *CHI3L2* gene. The x-axis of the figure shows the functional *CHI3L2* SNP used in the simulation, as well as the SNP's MAF. The y-axis shows the power of our semiparametric approach using IBS kernels weighted by either the tagSNPs' MAFs or the tagSNPs' p-values from an independently-generated dataset. The y-axis also shows the power of the maximum of the single-tagSNP statistics, which serves as a benchmark for our proposed semiparametric approaches. The plots show that our proposed semiparametric approach using a weighted IBS kernel based on tagSNPs' p-values clearly has optimal performance relative to the other approaches shown in the figure, regardless of the genetic model used to simulate the data, the nature of the functional SNP (i.e. tagSNP or untyped SNP), and the SNP's MAF. This result is hardly surprising, given this approach is the only one of the three shown that uses additional information from an independent dataset to assist in inference.

While the IBS kernel weighted by MAF displays lower power than the IBS kernel weighted by p-values, Figure 3 shows the former kernel is still generally more powerful than the maximum of the single-tagSNP statistics across functional SNPs and genetic models. There are a few situations where this condition does not hold, however. In particular, under an additive model, results show that the maximum of single-tagSNP statistics is more powerful than the weighted IBS kernel based on MAF for functional SNPs with MAF < 0.10 (e.g. SNP rs2182115, MAF=0.085). However, this power difference between the two approaches substantially decreases for dominant and recessive genetic models.

Figure 4 shows analogous power results for simulations based on the *NAT2* gene. Overall, we observed similar power results for this gene compared to that of the *CHI3L2* gene. Our semiparametric method using the IBS kernel weighted by p-values substantially outperformed the other competing approaches across all genetic models

tested, although the difference was most pronounced under a dominant model. The semiparametric approach weighted by MAF generally exhibited greater power than the maximum of the single-tagSNP statistics across the tested SNPs and genetic models. The differences in power were most pronounced under dominant and recessive models. We note the low power observed for all methods at one particular marker, rs1961456. As seen in Figure 2, this marker displays comparatively weak LD with the other SNPs in the gene, which leads to relatively low power by all methods to detect the association between the trait and this particular SNP.

To simplify presentation, we did not show power results for the unweighted IBS kernel (4) in Figures 3 and 4. Overall, the performance of the unweighted IBS kernel was similar to the IBS kernel weighted by MAF with a few notable differences. For functional SNPs with MAF >0.10, we found that the unweighted IBS kernel had equivalent or slightly improved power compared to the IBS kernel weighted by MAF. However, for functional SNPs with MAF < 0.10, we found the unweighted IBS kernel could have substantially-reduced power relative to the IBS kernel weighted by MAF. For example, assuming an additive model where the functional SNP was rs2182115 (MAF=0.085) in *CHI3L2*, we found the power of the unweighted IBS kernel was 0.327 compared to 0.498 for the IBS kernel weighted by MAF. This result suggests that, without weighting, the effects of functional SNPs with rare MAF may be smoothed over by information from surrounding SNPs with more common MAF. Since the IBS kernel weighted by MAF appears to have better performance averaged across the range of MAF compared to the unweighted IBS kernel, we recommend the use of the former kernel over the latter in association analysis.

While primary interest focuses on the testing of the nonparametric function h, secondary interest may focus on the estimation and testing of environmental covariate effects. Table 2 shows estimates of the mean and standard deviation, along with the empirical standard deviation, of the regression parameters related to the binary and

continuous covariates used in our simulations. Due to the large number of SNPs and models examined, we display results only for one representative configuration of both the *NAT2* and *CHI3L2* genes. These examples show that the semiparametric regression method produces unbiased estimates of the covariate effects with empirical standard deviations that closely match the LSKM-based standard deviations. We observed similar results for other simulation models (results not shown).

DISCUSSION

In this article, we have proposed a flexible semiparametric-regression framework for association mapping of quantitative traits using genotype data from multiple tagSNPs within a region of interest. Using simulated genetic data based on real data from the International HapMap Project (2005), we demonstrated our approach often has superior performance compared to tests of individual tagSNPs, which is the most common approach for association mapping of complex traits. Our method's improved performance results from modeling the effects of multiple tagSNPs within a reduced-dimension function, thereby using more genetic information in analysis but producing test statistics (based on the function) with smaller degrees of freedom than typical multivariate methods. In addition to improved power, our approach is also quite flexible as it can easily adjust for the effects of potential confounders (such as subpopulation assignment in a stratified population) and further can evaluate interaction effects among tagSNPs and environmental factors (by modeling such interactions parametrically or non-parametrically using the function h in (1)). By maximizing the semiparametric model using LSKM, we show that we can fit the model easily using common maximization procedures for linear mixed models, which are available in a variety of software packages. The approach is computationally efficient to implement as analysis of 1000 replicates of simulated data (using the design described in the Simulations section) took

only 5 minutes to run on a Dell Latitude D810 with a 2.26 GHz processor. We provide SAS and Fortran code for implementing the approach on our website (Epstein Software).

An appealing feature of our semiparametric approach is that it can utilize prior information (in the form of weights) to improve one's ability to detect functional regions. For a specific tagSNP, a natural choice of weight is to use the strength of evidence for association between that tagSNP and the trait of interest (or a correlated trait) from an independent study. We quantify this strength based on the $-\log_{10}$ of the relevant p-value. To obtain such p-values, one could conduct an exhaustive literature search of relevant genetic studies of interest. However, we note that such p-value weights will become increasingly available with the public release of tagSNP genotype and phenotype data from whole-genome association studies into free databases (often a requirement for NIH funding of such projects). An example of such a database is the NIH-sponsored dbGaP (see Web Resources), which will eventually contain information on at least 10 whole-genome association studies of complex traits. On a related issue, we strongly recommend against using p-value weights based on single-tagSNP analysis of the same dataset upon which one intends to apply the proposed semiparametric approach. Such an application will lead to anticonservative tests (results not shown).

We applied our semiparametric approach to the problem of testing whether a specific region influenced a quantitative trait of interest. However, with some effort, we can extend our approach to create a multilocus association test for genome-wide association studies. Specifically, we can implement our approach using a sliding-window process that considers overlapping or nonoverlapping sets of tagSNPs across each chromosome. Within a particular window, we can apply our approach to the genotype data from the multiple tagSNPs and produce a statistic for testing whether the tagSNPs within the given window are associated with the trait of interest. After constructing test statistics for each window across the genome, we can establish empirical significance of a particular statistic (taking into account the mandatory adjustment for multiple

correlated tests) using the computationally-efficient Monte Carlo approach of Lin (2005). We will investigate this valuable extension in a subsequent paper.

We note that the nonparametric function within the semiparametric framework implicitly models both the main and higher-order interaction effects of the specified predictors within a space induced by a dimension-reducing kernel. This appealing feature makes our semiparametric approach potentially valuable for genetic pathway analysis, where one is interested in evaluating the trait-influencing effects of SNPs, environmental factors, and their complex interactions that reside within a putative biochemical pathway. Here, one would model all relevant pathway predictors within the nonparametric function in (1) and test whether that function has an effect on the trait of interest. This idea is appealing since the development of parametric approaches that explicitly model all of the pathway effects is challenging because the underlying genetic model (likely containing many high-dimensional interactions) is typically unknown. Our semiparametric approach would circumvent this serious issues of parametric methods in pathway analysis.

In implementing our approach, we assumed no missing genotype data for the tagSNPs in the region of interest. While our approach doesn't naturally accommodate missing genotype data within the nonparametric function, we note that we can use existing statistical procedures for imputing genotype data for a given subject to resolve this issue. Such imputation procedures can rely on the LD structure of nearby SNPs to predict a subject's missing genotype using either observed genotype data from the study sample (Scheet and Stephens 2006) or appropriate genotype data from the International HapMap project (Zaitlen et al. 2007). Once we impute missing genotypes, we can then incorporate them within our nonparametric function and proceed with analysis as we previously described.

While we have developed our approach for association analysis of quantitative trait data, we note that we can extend our approach to conduct similar multi-SNP association analysis in case-control studies of disease. Implementation of our approach

for disease data requires minimization of a Lagrangian function related to that shown in equation (A3) of Appendix A. Estimating relevant parameters using this function is analogous to model fitting using a non-linear mixed model with a corrected penalized quasi-likelihood algorithm (Lin and Breslow 1996). While the iterative nature of the algorithm will increase the numerical complexity of the semiparametric analysis, it should still be computationally efficient for candidate-gene or whole-genome association analysis. We will explore this approach in a future work.

Our approach fits a semiparametric regression model using LSKM, which we show corresponds to inference using a specific linear mixed model. While mixed-modeling procedures often are connected to pedigree analysis (Amos 1994; Almasy and Blangero 1998; Abecasis et al. 2000), we note that their elegance and flexibility make them increasingly popular tools for association mapping in population-based or case-control studies. Tzeng and Zhang (personal communication) proposed a powerful mixed model for SNP-based haplotype analysis of complex traits that models the covariance of the outcomes among a pair of subjects as a function of their (inferred) haplotype similarity along a region of interest. The distribution of the authors' random effect has similarity to the distribution of the random tagSNP effect in our linear mixed model, although the authors' approach is not based on the use of reproducing kernels in a LSKM framework. Further, their approach focuses primarily on using SNP-based haplotypes in their covariance structure and doesn't consider the use of influential and valuable prior weights in analysis. Another mixed-model tool for such a study consists of a two-level hierarchical model (Witte 1997; Witte et al. 2000). The first level of the hierarchical model regresses the trait outcome on the SNPs of interest (and potential confounders) whereas the second level models the SNP-related risk parameters as a function of influential covariates including the underlying haplotype structure (Conti and Witte 2003) or available pathway information (Hung et al. 2004). Such second-level information can improve the precision and accuracy of SNP-based risk estimates.

ACKNOWLEDGEMENTS

This work was sponsored by National Institutes of Health grants GM074909 (to L.C.K.) and HG003618 (to L.C.K and M.P.E.).

WEB RESOURCES

dbGaP, http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap

Epstein software, http://www.genetics.emory.edu/labs/epstein/software

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim

APPENDIX A: ESTIMATION OF SEMIPARAMETRIC MODEL PARAMETERS IN (1) USING LSKM

In this Appendix, we use the work of Liu et al. (in press) to apply LSKM to obtain estimates of h and β in the semiparametric model shown in (1). Prior to implementation, we note again that we assume $h(G_j)$ exists in some function space κ. Therefore, we can rewrite the nonparametric function $h(G_j)$ for subject j as

$$h(G_j) = \sum_{i=1}^{\infty} \omega_i \phi_i(G_j) = \phi(G_j)^T \omega \tag{A1}$$

where $\{\phi_i(G_j)\}_{i=1}^{\infty}$ is an orthonormal basis of κ and $\{\omega_i\}_{i=1}^{\infty}$ denotes a series of weights.

Using the relationship in (A1), we apply LSKM to formulate the optimization problem as the minimization of the following function:

$$J(\omega, \beta, e) = \frac{1}{2}\sum_{j=1}^{N} e_j^2 + \frac{1}{2}\lambda \|\omega\|^2 \tag{A2}$$

subject to the constraint $e_j = Y_j - \left(X_j^T \beta + \phi(G_j)^T \omega\right)$ $(j = 1, \ldots, N)$. Here, λ denotes our previously-defined smoothing parameter.

Using a Lagrangian multiplier $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_N)$, we can modify $J(\omega, \beta, e)$ in (A2) to develop a Lagrangian function to be minimized as

$$f(\beta, e, \gamma, \omega) = \frac{1}{2}\sum_{j=1}^{N} e_j^2 + \frac{1}{2}\lambda \|\omega\|^2 - \sum_{j=1}^{N} \gamma_j \left(X_j^T \beta + \phi(G_j)^T \omega + e_j - Y_j\right) \tag{A3}$$

To determine the estimates of $(\omega,\ \gamma, e, \beta)$ that minimize this function in (A3), we obtain the following optimality conditions:

$$\nabla_\omega f = 0 \rightarrow \omega = \frac{1}{\lambda}\sum_{j=1}^{N} \gamma_j \phi(G_j) \tag{A4}$$

$$\frac{df}{de_j} = 0 \rightarrow e_j = \gamma_j \tag{A5}$$

$$\frac{df}{d\gamma_j} = 0 \rightarrow Y_j - X_j^T\beta - \phi(G_j)^T\omega - e_j = 0 \quad \text{(A6)}$$

$$\nabla_\beta f = 0 \rightarrow \sum_{j=1}^{N}\gamma_j X_j = 0 \quad \text{(A7)}$$

Using the optimality conditions shown in (A4) and (A5), we can rewrite the optimality condition in (A6) as

$$Y_j - X_j^T\beta - \frac{1}{\lambda}\sum_{l=1}^{N}\gamma_l\phi(G_j)^T\phi(G_l) - \gamma_j = 0 \quad \text{(A8)}$$

Therefore, we have rewritten the primal formulation of the optimization problem in (A3) into the dual formulation consisting of equations (A7) and (A8). However, due to the requirement of calculating $\phi(G_j)^T\phi(G_l)$ in (A8), we cannot use this dual formulation for inference since $\{\phi_i(G_j)\}_{i=1}^{\infty}$ is likely unknown. However, we can use the theory of reproducing kernels to help resolve this issue.

Reproducing kernels For two tagSNP vectors G_j and G_l, we can construct a reproducing kernel function $k(G_j, G_l)$ that, under certain regulatory conditions, can be written as an eigenvalue-eigenvector decomposition $k(G_j, G_l) = \sum_{i=1}^{\infty}\mu_i\psi_i(G_j)^T\psi_i(G_l)$, where $\{\psi_i(\,\cdot\,)\}_{i=1}^{\infty}$ forms an orthonormal system. We can use the theory of reproducing kernels to help solve the dual formulation of the optimization problem in equations (A7) and (A8) by noting that we can rewrite a kernel function as $k(G_j, G_l)$=$\phi(G_j)^T\phi(G_l)$, where we define $\phi_i(\,\cdot\,)$=$\sqrt{\mu_i}\psi_i(\,\cdot\,)$. Therefore, we can solve the optimization problem in (A8) by specifying an appropriate kernel function $k(\,\cdot\,,\,\cdot\,)$ to represent $\phi(\,\cdot\,)^T\phi(\,\cdot\,)$. The function space induced by a given kernel function has many nice properties such that most functions $h(\,\cdot\,) = \phi(\,\cdot\,)^T\omega$ can be approximated well within the space. We describe examples of kernels within the Materials and Methods section of the article.

Parameter estimation Using a kernel function, we can rewrite the optimality condition in (A8) in matrix notation as

$$Y - X\beta - \left(\frac{1}{\lambda}K + I\right) = 0. \tag{A9}$$

$Y = (Y_1, Y_2, \ldots, Y_N)^T$ denotes the (Nx1) trait vector, X denotes the (N x p) covariate matrix, and I denotes a (N x N) identity matrix. Finally, K in (A9) is an $N\,x\,N$ kernel matrix with (j, l)-th element $k(G_j, G_l)$. We can also rewrite the optimality condition in (A7) in matrix notation as

$$X^T\gamma = 0 \tag{A10}$$

Using the matrix forms in (A9) and (A10), we can write the dual formulation of the optimization problem in one matrix equation as

$$\begin{bmatrix} 0 & X^T \\ X & \lambda^{-1}K + I \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \tag{A11}$$

By solving (A11), we can obtain estimates of β and γ, which we can then use to solve for the nonparametric function h. One can easily show that $\widehat{\beta} = [X^T(K + \lambda I)^{-1}X]^{-1}X^T(K + \lambda I)^{-1}Y$ and $\widehat{\gamma} = \lambda(K + \lambda I)^{-1}\left(Y - X\widehat{\beta}\right)$. To estimate $h(\,\cdot\,) = \phi(\,\cdot\,)^T\omega$, we use the optimality condition in (A4) and $\widehat{\gamma}$ together to obtain $\widehat{h} = K(K + \lambda I)^{-1}\left(Y - X\widehat{\beta}\right)$.

APPENDIX B: APPROXIMATE DISTRIBUTION OF THE SCORE STATISTIC S_τ IN (9)

We consider the linear mixed model described previously in (6):

$$Y = X\beta + U + E$$

where Y is the vector of quantitative trait values, X is the vector of fixed effects, U is the vector of random tagSNP effects which follows a multivariate normal distribution with mean 0 and variance-covariance matrix τK, and E is a vector of subject-specific random effects which follows a multivariate normal distribution with mean 0 and variance-covariance matrix $\sigma^2 I$.

Using the mixed model in (6), we seek to determine the distribution of the score statistic in (9) for testing $H_0 : \tau = 0$. Zhang and Lin (2003) noted that, since $\tau \geq 0$, we are testing the parameter on its boundary value and, as a result, the distribution of S_τ follows a mixture of χ^2_1 distributions. To facilitate inference, the authors showed that one can approximate this complicated mixture distribution with a scaled χ^2 distribution $\delta\chi^2_\nu$, where δ denotes the scale parameter and ν denotes the degrees of freedom. To estimate δ and ν, the authors suggested the use of the Satterthwaite method, which equates the mean and variance of the score statistic S_τ in (9) with the mean and variance of $\delta\chi^2_\nu$.

Let e denote the mean of S_τ and let $I_{\tau\tau}$ denote the variance of the score statistic. When calculating the mean and variance of S_τ, we must account for the fact that we use estimates of σ^2 and β instead of the true values of these parameters in (9). Therefore, we replace the mean e with $\tilde{e} = \text{tr}(P_0 K)/2$, where $P_0 = I - X(X^T X)^{-1} X^T$ is the projection matrix under the null hypothesis. Also, we replace the variance $I_{\tau\tau}$ with the efficient information $\tilde{I}_{\tau\tau}$ as follows:

$$\tilde{I}_{\tau\tau} = I_{\tau\tau} - I_{\tau\sigma^2} I^{-1}_{\sigma^2\sigma^2} I^T_{\tau\sigma^2}$$

where $I_{\tau\tau} = tr(P_0 K)^2/2$, $I_{\tau\sigma^2} = tr(P_0 K P_0)/2$, and $I_{\sigma^2\sigma^2} = tr(P_0^2)/2$.

Once we obtain $\widetilde{e}$ and $\widetilde{I}_{\tau\tau}$, we can set the former equal to $\delta\nu$ (the mean of a $\delta\chi^2_\nu$ random variable) and the latter equal to $2\delta^2\nu$ (the variance of a $\delta\chi^2_\nu$ random variable). After solving the system of equations, we calculate the scale parameter for the approximate distribution as $\delta = \widetilde{I}_{\tau\tau}/2\widetilde{e}$ and calculate the degrees of freedom as $\nu = 2\widetilde{e}^{\,2}/\widetilde{I}_{\tau\tau}$. We can then compare the value of the resulting scaled score statistic, S_τ/δ, to a chi-squared distribution with ν degrees of freedom in order to assess significance of the test of $H_0 : \tau = 0$.

REFERENCES

Abecasis GR, Cardon LR, Cookson WOC (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66:279-292

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62: 1198-1211

Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535-543

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120

Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18-31

Conti DV, Witte JS (2003) Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. American Journal of Human Genetics 72: 351-363

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines (and other kernel-based learning methods). Cambridge: Cambridge University Press

de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D (2005) Efficiency and power in genetic association studies. Nature Genetics. 37: 1217-1223

Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 75:35-43

Fan R, Knapp M (2003) Genome association studies of complex diseases by case-control designs. Am J Hum Genet 72: 850-868

Gauderman WJ, Murcray C, Gilliland F, Conti DV (2007) Testing association between disease and multiple SNPs in a candidate gene. Genet Epidemiol 31: 383-395

Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Borreta P, Witte JS (2004) Using hierarchical modeling in genetic association studies with multiple markers: Application to a case-control study of bladder cancer," Cancer Epidemiology, Biomarkers & Prevention, 13, 1013-1021.

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

Lin X, Breslow NE (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. J Am Stat Assoc 91: 1007-1016

Liu D, Lin X, Ghosh D. Semiparametric Regression of Multi-Dimensional Genetic Pathway Data: Least Squares Kernel Machines and Linear Mixed Models. Biometrics, in press

Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, Donnelly P; International HapMap Consortium (2006) A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet 78: 437-450

Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74: 765-769

Rasmussen CE, Williams CKI (2006). Gaussian Processes for Machine Learning. MA: MIT Press

Roeder K, Bacanu SA, Sonpar V, Zhang X, Devlin B (2005) Analysis of single-locus tests to detect gene/disease associations. Genet Epidemiol: 28: 207-219

Rosenberg PS, Che A, Chen BE (2006) Multiple hypothesis testing strategies for genetic case-control association studies. Stat Med 25: 3134-3149

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press

Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN (2005) Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet 76: 780-793

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70: 425-434

Scheet P, Stephens M (2006) A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. Am J Hum Genet 78:629-644

Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Stram DO (2004) Tag SNP selection for association studies. Genet Epidemiol 27: 365-374

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

Tzeng JY, Wang CH, Kao JT, Hsiao CK (2006) Regression-based association analysis with clustered haplotypes through use of genotypes. Am J Hum Genet 78: 231-242

Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, Eriksson J, Nylund SJ, Kohtamaki K, Toivanen L, Vidgren G, Tuomilehto-Wolf E, Ehnholm C, Blaschak J, Langefeld CD, Watanabe RM, Magnuson V, Ally DS, Hagopian WA, Ross E, Buchanan TA, Collins F, Boehnke M (1998) Mapping genes for NIDDM: design of the Finland-United States Investigation of NIDDM Genetics (FUSION) study. Diabetes Care 21:949-958

Wessel J, Schork NJ (2006) Generalized Genomic Distance–Based Regression Methodology for Multilocus Association Analysis. Am J Hum Genet 79:792-806

Witte JS, Greenland S, Kim L, Arab L (2000) Multilevel modeling in epidemiology with GLIMMIX. Epidemiology 11: 684-688

Witte JS (1997) Genetic analysis with hierarchical models. Genet Epidemiol 14: 1137-1142

Zaitlen N, Kang HK, Eskin E, Halperin E (2007) Leveraging the HapMap Correlation Structure in Association Studies. Am J Hum Genet 80: 683-691

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53: 79-91

Zhang D, Lin X (2003) Hypothesis testing in semiparametric additive mixed models. Biostatistics 4: 57-74.

Table 1: Empirical Type-I Error Rates at $\alpha = 0.05$

		Semiparametric Approach Using IBS Kernel			
Gene	Single-Locus Test	Unweighted	MAF Weights	(Same) P-value Weights	(Diff)P-value Weights
CHI3L2	0.0474	0.0458	0.0560	0.0518	0.0522
NAT2	0.0522	0.0486	0.0492	0.0494	0.0496

Results are based on 5000 replicates. (Same) P-value weights were based on an independent dataset generated under the same model as the dataset under study. (Diff) P-value weights were based on an independent dataset generated under an alternative model where the functional variant explained 3% of the trait variation. For simulations based on *CHI3L2*, the functional variant was rs961364 (MAF=0.293). For simulations based on *NAT2*, the functional variant was rs1799930 (MAF=0.292).

Table 2: Parameter Estimates of Environmental Covariates

Genetic Model		*NAT2*		*CHI3L2*	
		$\widehat{\beta}_{E,Bin}$	$\widehat{\beta}_{E,Cont}$	$\widehat{\beta}_{E,Bin}$	$\widehat{\beta}_{E,Cont}$
Additive	Mean	0.503	0.030	0.504	0.030
	Std. Dev.	0.117	0.013	0.117	0.013
	Est. Std. Dev.	0.118	0.013	0.118	0.013
Dominant	Mean	0.503	0.030	0.504	0.030
	Std. Dev.	0.118	0.013	0.118	0.013
	Est. Std. Dev.	0.118	0.013	0.118	0.013
Recessive	Mean	0.503	0.030	0.503	0.030
	Std. Dev.	0.117	0.013	0.116	0.013
	Est. Std. Dev.	0.118	0.013	0.118	0.013

$\beta_{E,Bin}$ and $\beta_{E,Cont}$ denote effect sizes for the binary and continuous covariates, respectively, described in the simulations. The true value of $\beta_{E,Bin}$ is 0.50 and the true value of $\beta_{E,Cont}$ is 0.03. Results are based on 1000 replicates generated under an alternative model. For *NAT2* simulations, the functional variant was rs1799930 (MAF=0.292). For simulations based on *CHI3L2*, the functional variant was rs961364 (MAF=0.293). For all simulations, we analyzed replicates using our semiparametric approach assuming a IBS kernel weighted by MAF.

FIGURE LEGEND

Figure 1: LD plot of 37 polymorphic SNPs within the *CHI3L2* gene. Results based on the CEU sample from the International HapMap Project. TagSNPs are denoted by a box surrounding the relevant SNP label.

Figure 2: LD plot of 20 polymorphic SNPs residing within the *NAT2* gene. Results based on the CEU sample from the International HapMap Project. TagSNPs are denoted by a box surrounding the relevant SNP label.

Figure 3: Power results at $\alpha = 0.05$ for simulations based on the *CHI3L2* gene under additive (top panel), dominant (middle panel), and recessive (bottom panel) mechanisms of allelic effect for the functional SNP. X-axis labels show the name and minor-allele frequency of the functional SNP used in the simulation (tagSNPs are shown in bold). For the IBS kernel with p-value weights, we obtained a relevant p-value for each tagSNP based on single-locus tests of an independent dataset simulated under the same model. We note that the range of the y-axis for the recessive model is different from the range for the additive and dominant models

Figure 4: Power results at $\alpha = 0.05$ for simulations based on the *NAT2* gene under additive (top panel), dominant (middle panel), and recessive (bottom panel) mechanisms of allelic effect for the functional SNP. X-axis labels show the name and minor-allele frequency of the functional SNP used in the simulation (tagSNPs are shown in bold). For the IBS kernel with p-value weights, we obtained a relevant p-value for each tagSNP based on single-locus tests of an independent dataset simulated under the same model. We note that the range of the y-axis for the recessive model is different from the range for the additive and dominant models

FIGURE 1

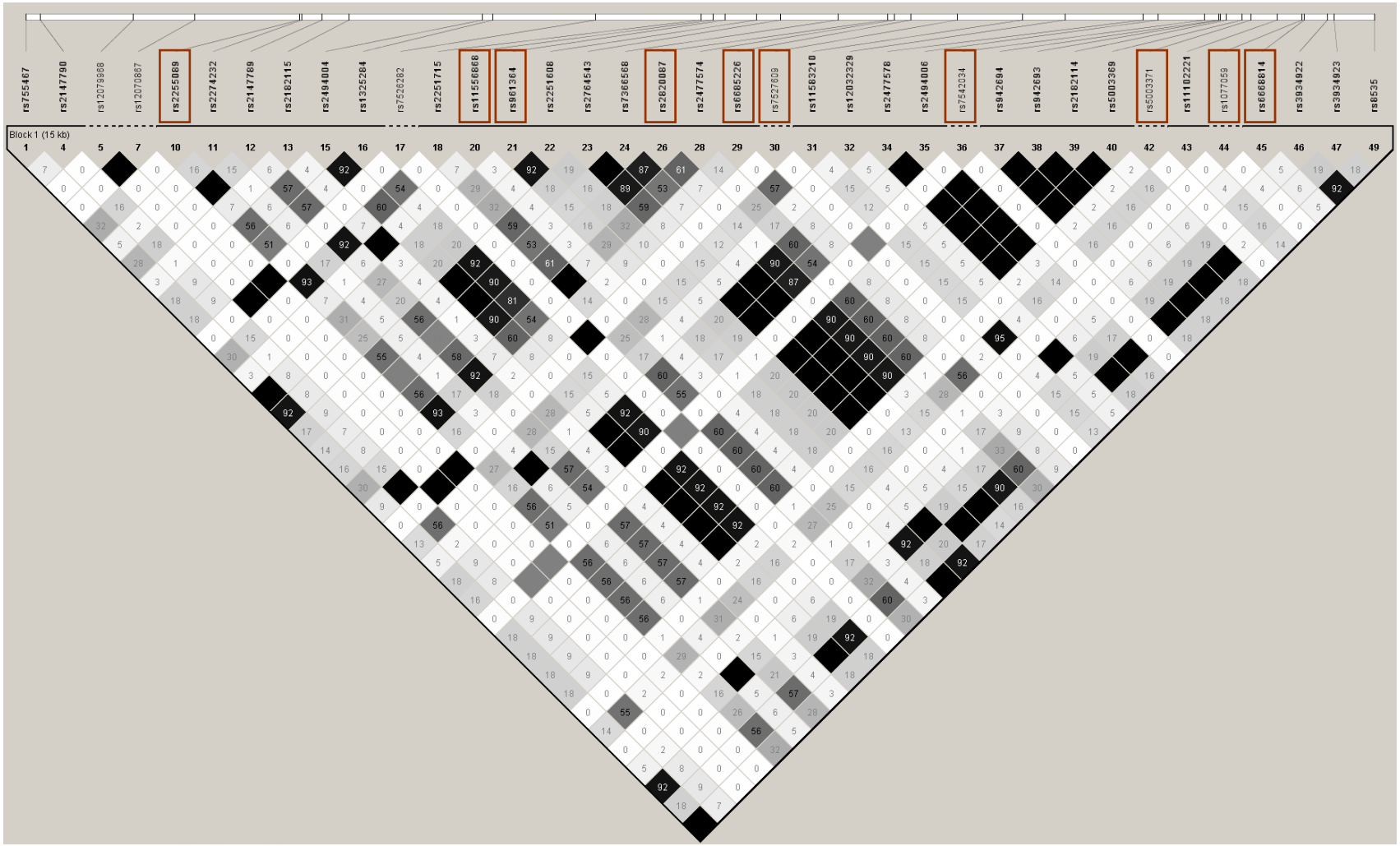

FIGURE 2

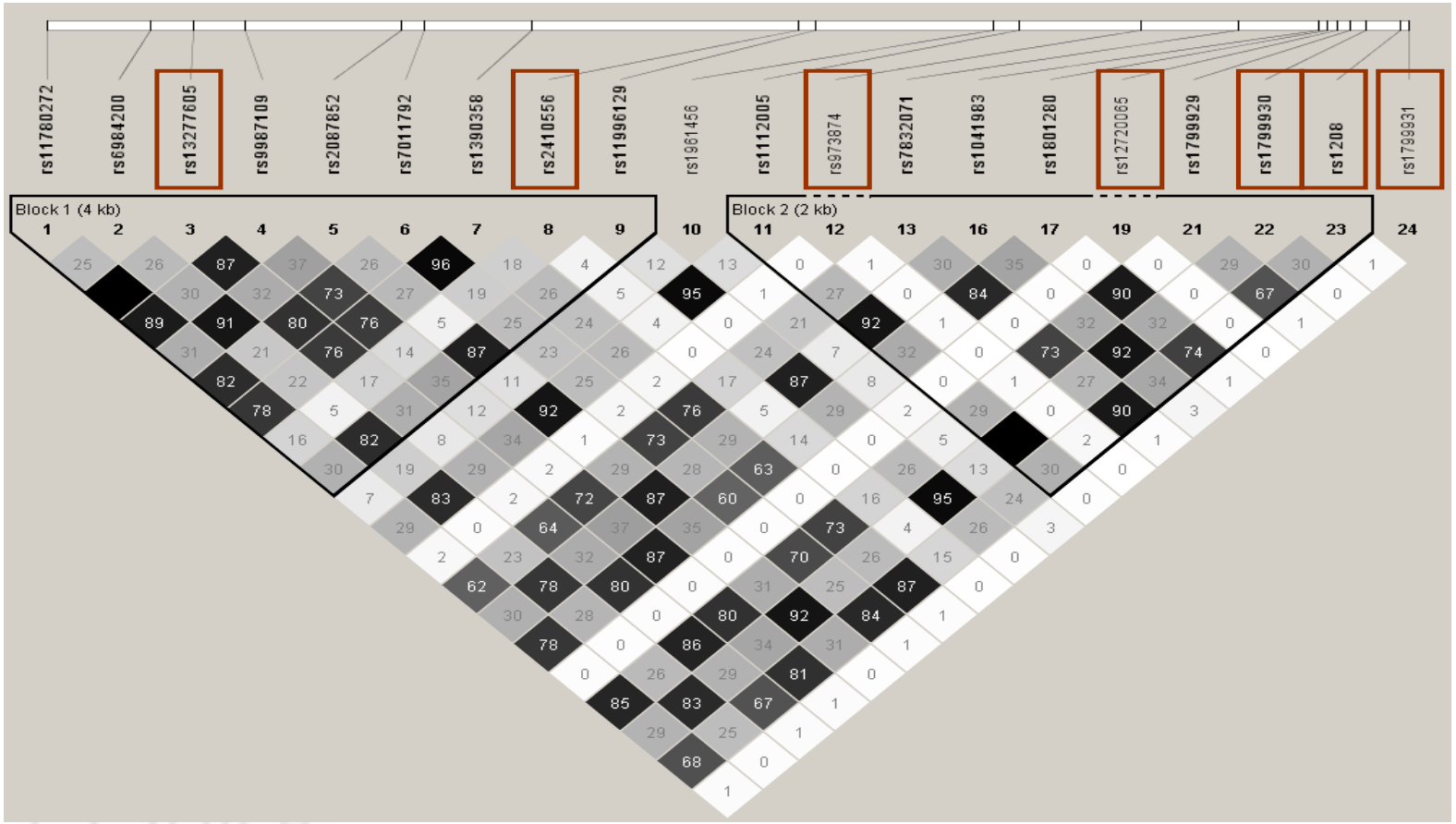

FIGURE 3

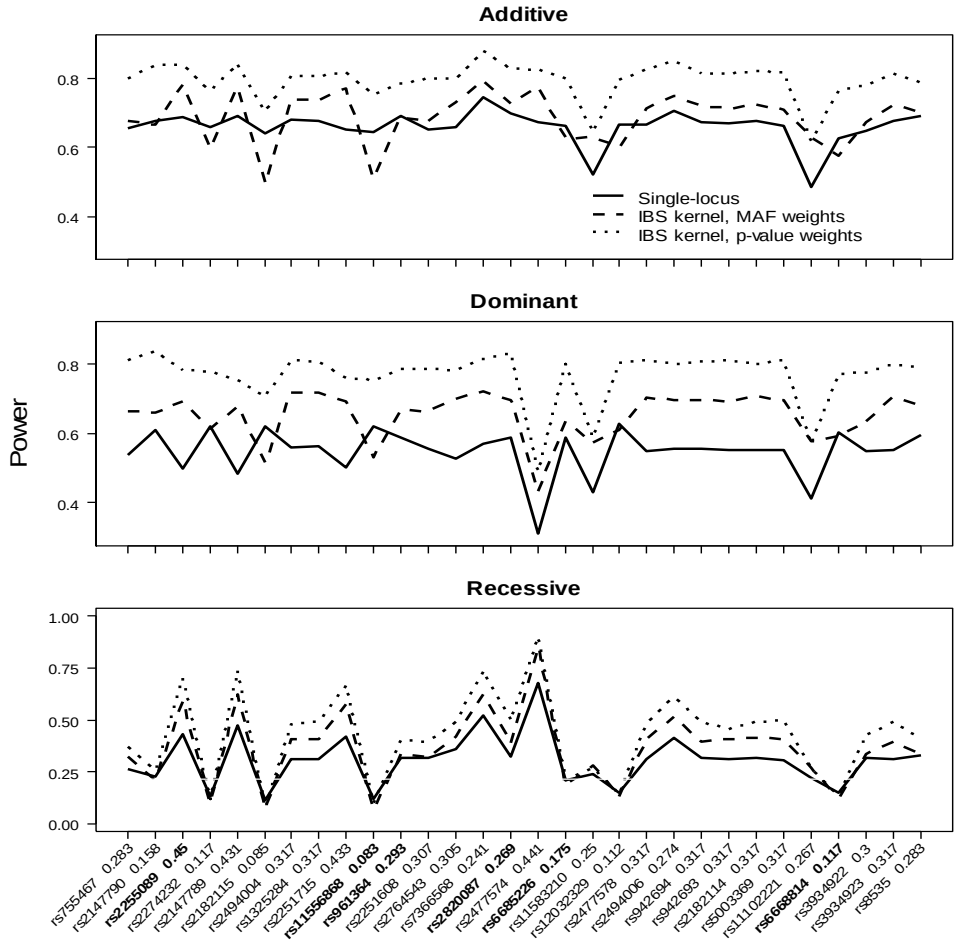

FIGURE 4