ARTICLE

A Powerful and Flexible Multilocus Association Test for Quantitative Traits

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

Association mapping of complex traits typically employs tagSNP genotype data to identify a trait locus 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 as a result of incomplete linkage disequilibrium between the genotyped tagSNP and the trait locus. Alternatively, one can jointly test all tagSNPs simultaneously within the region (by 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 by 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 by using standard software packages like SAS and R). Using simulated SNP data based on real data from the International HapMap Project, we demonstrate that 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, because 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 dissection of the genetic mechanisms of complex traits. Using panels of single-nucleotide polymorphisms (SNPs), association methods seek to identify those genetic markers that either are a trait locus or are in linkage disequilibrium (LD) with a trait locus. 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 trait locus, a study could genotype and subsequently analyze all polymorphic SNPs in the genetic interval. However, the probable 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—called tagSNPs—within the region 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[1–3] to SNP genotype data from existing public databases of human genetic variation, such as the International HapMap Project.[4]

In this article, we focus on the use of tagSNP data to identify genetic regions that influence a quantitative trait of interest by 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 (via regression or ANOVA methods) followed by inference on the maximum of the resulting single-tagSNP statistics. Because of 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,[5] or a Bonferroni correction based on the effective number of independent tests within the region.[6,7]

Although 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) quantitative-trait locus (QTL). 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.[8–10] 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).

Because these multivariate approaches combine genetic information from multiple tagSNPs simultaneously into analysis, they intuitively should provide greater power to

[1]Department of Biostatistics, Emory University, Atlanta, GA 30322, USA; [2]Department of Human Genetics, Emory University, Atlanta, GA 30322, USA; [3]Center for Statistical Sciences, Brown University, Providence, RI 02912, USA; [4]Department of Biostatistics, Harvard University, Boston, MA 02115, USA; [5]Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA
*Correspondence: mepstein@genetics.emory.edu
DOI 10.1016/j.ajhg.2007.10.010.

detect QTLs than do 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[11–13] unless the trait originates from the effect of a specific haplotype rather than a specific SNP.[13] 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.[10] 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[14] or principal components[15] 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.[16,17] One can measure such tagSNP similarity by using a "kernel" function that reduces a comparison of multiple tagSNPs for a pair of subjects into a single scalar factor. Because of this phenomenon, resulting statistics using kernel functions typically have small degrees of freedom; for example, Schaid et al.[16] 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 because 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.[17]

In this article, we propose a novel approach for association mapping of quantitative traits that 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[18] 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 QTLs. Unlike existing dimension-reduction techniques, we will show that our approach permits us to incorporate valuable prior information in the analysis via the kernel function. Unlike existing similarity approaches, we will show that 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 by using a flexible high-dimensional technique called least-squares kernel machines (LSKM).[19,20] Previously, LSKM methods have been applied to continuous variables, such as expression data from microarray analysis.[20] Here, we propose the novel use of kernel functions that are designed for categorical tagSNP data. The kernels we discuss incorporate relevant weights as well as appropriate measures of genetic similarity between subjects. Although LSKM fitting of a semiparametric model appears complicated, Liu et al.[20] noted that one can represent the LSKM procedure by using a specific form of a linear mixed model, such that one can estimate and test the nonparametric function of the tagSNP data by 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 by using the LSKM maximization approach of Liu et al.[20] 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,[4] 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.

Material 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 ($j = 1, \ldots, N$). We assume that each subject is genotyped at S tagSNPs within the region of interest. We let $G_{j,s}$ denote the genotype of subject j at tagSNP s ($s = 1, \ldots, S$) and let $G_j = (G_{j,1}, G_{j,2}, \ldots, G_{j,S})$ denote 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 ($p \times 1$) 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:

$$Y_j = X_j^T \beta + h(G_j) + e_j \tag{1}$$

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 ($p \times 1$) 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 Equation 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. Because we are using a semiparametric framework in Equation 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 flexible 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.,[20] we obtain the following estimates of h and β in Equation 1:

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

$$\hat{\beta} = \left[X^T(K + \lambda I)^{-1}X\right]^{-1} X^T(K + \lambda I)^{-1}Y \tag{3}$$

Here, $Y = (Y_1, \ldots, Y_N)^T$ is an ($N \times 1$) vector of the trait values for all subjects and X is an ($N \times p$) matrix of environmental covariates for all subjects. Further, I denotes an (N x N) identity matrix. Finally, there are two additional terms in Equations 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 Equations 2 and 3 is K, which denotes an ($N \times N$) 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. Because $k(G_j, G_l)$ 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 Equations 2 and 3 must satisfy the conditions of Mercer's Theorem,[21] which includes the condition that the K matrix must be positive semidefinite (i.e., the eigenvalues of K must be positive).

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.[17] The IBS kernel takes the form

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

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 a weighted-IBS kernel based on Equation 4 as the following:

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

We focus on two potentially valuable weights for use in the IBS kernel in Equation 5. First, we consider a weight that upweights tagSNPs with a rare minor-allele frequency (MAF) and downweights tagSNPs with more common MAFs. Such a weight could be valuable because of the potential for the information from tagSNPs with rare MAFs to be smoothed over by the information from surrounding tagSNPs with more common MAFs. To upweight tagSNPs with rare MAFs, we apply the weight $w_s = 1/\sqrt{q_s}$, where q_s denotes the MAF of tagSNP s ($s = 1, \ldots, S$). 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 that 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 $\hat{h}$ in Equation 2 shows that the nonparametric function in Equation 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 by using Equation 1. To facilitate the construction of such a test statistic, Liu et al.[20] noted that LSKM-based estimation of $\hat{h}$ and $\hat{\beta}$ is analogous to the estimation of random and fixed effects, respectively, within a specific linear mixed model.

Therefore, rather than employ complicated procedures to directly test H_0: $h = 0$, we can exploit the LSKM relationship with a mixed model to apply a likelihood framework to construct an appropriate test statistic for inference. Additionally, the use of a linear mixed model for inference is appealing because it allows implementation of our approach with any common software package for mixed-model analysis (e.g., SAS PROC MIXED).

To apply the results from Liu et al.[20] and develop the mixed-model representation of the LSKM analysis by using the semiparametric model in Equation 1, we consider the following linear mixed model:

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

where Y denotes the earlier trait vector and X denotes the earlier matrix of fixed environmental covariates with related regression-coefficient vector β. Within Equation 6, we denote h as a $(N \times 1)$ 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 h 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 commonly applied to linear mixed models to estimate (β, λ, σ^2). After applying REML procedures, we can show, following Liu et al.,[20] that the best-linear unbiased estimators of the random effects h and the fixed effects β in the linear mixed model are

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

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

where λ can be estimated with REML procedures. One can see that the estimates of $\hat{h}$ and $\hat{\beta}$ in Equations 7 and 8 are exactly the same as the estimates of $\hat{h}$ and $\hat{\beta}$ in Equations 2 and 3, respectively, derived via LSKM estimation of the semiparametric model in Equation 1. The equivalence of these estimates shows that we can perform our LSKM multilocus analysis by using a straightforward linear mixed model that is easy to implement with existing statistical software packages for mixed models.

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 h in the linear mixed model in Equation 6. As noted earlier, we assume that h follows a multivariate-normal distribution with mean vector 0 and covariance matrix $\frac{\sigma^2}{\lambda}K$. Assume $\tau = \sigma^2/\lambda$ such that we rewrite the covariance matrix as τK. If $\tau = 0$, then this directly implies that $h = 0$. Because K must be positive semidefinite under the LSKM model[21] (with diagonal elements equaling 1 with any of the suggested kernel functions), it also follows that $h = 0$ only when $\tau = 0$. Therefore, a test of H_0: $\tau = 0$ in the linear mixed model (Equation 6) is equivalent to testing H_0: $h = 0$ in the semiparametric model (Equation 1).

To test H_0: $\tau = 0$, we propose the use of the score statistic of Liu et al.[20] The score statistic takes the form

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

where $\hat{\beta}$ and $\hat{\sigma}^2$ are the maximum-likelihood estimates of β and σ^2 under H_0, which are obtained from the linear-regression model $Y = X\beta + E$. Because $\tau \geq 0$, 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 A) 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 Centre d'Etude du Polymorphisme Humain (CEU) genotypes from build 35 of the International HapMap Project.[4] We based our simulations on the LD structure of two genes: *CHI3L2* (MIM 601526) and *NAT2* (MIM 243400). *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 by using the Tagger program.[3] 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 ten tagSNPs for *CHI3L2* and seven tagSNPs for *NAT2*. We show the LD structure of the tagged and nontagged SNPs within *CHI3L2* and *NAT2* in Figures 1 and 2, respectively. Within each gene, we applied PHASE[22–24] 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 by using these estimated haplotype frequencies under the assumption of Hardy-Weinberg equilibrium.

To ensure that 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.[25] We assumed that 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 QTL. We allowed the QTL to be either a typed tagSNP or an untyped SNP but required the variant to have MAF greater than 0.05 (as done elsewhere[10,12,14]). Within *CHI3L2*, 30 of the 37 polymorphic SNPs fulfilled this criteria, with six of these 30 polymorphisms being tagSNPs. Within *NAT2*, 17 of the 20 polymorphic SNPs fulfilled this criteria, with three of the 17 polymorphisms being tagSNPs. Denoting the QTL as S^*, we generated the trait outcome for subject j with the following model:

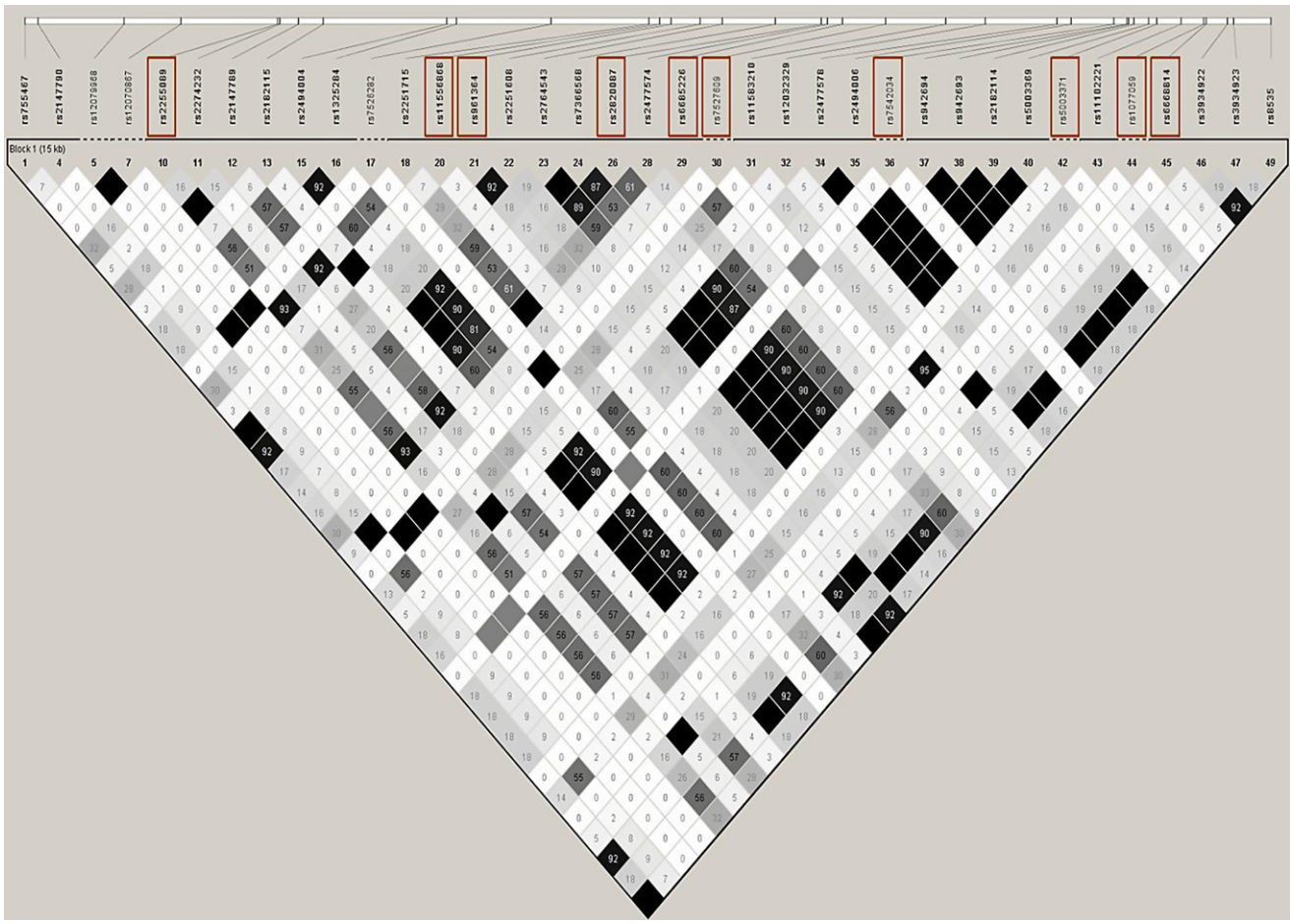

Figure 1. LD Plot of 37 Polymorphic SNPs within the *CHI3L2* Gene
Results based on the CEU sample from the International HapMap Project. TagSNPs are denoted by a box surrounding the relevant SNP label.

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

Here, $X_{G_{j,S^*}}$ denotes the coding of the genotype at QTL *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 QTL *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 that we did not observe genotypes at untyped SNPs (even though such untyped SNPs may be QTLs). We analyzed each dataset by using both our proposed semiparametric approach and, as a benchmark, traditional single-tagSNP statistics (modeled under an additive model of allelic effect).

For our semiparametric approach, we analyzed the data three times. First, we used the unweighted IBS kernel in Equation 4. Next, we used the weighted IBS kernel in Equation 5 with weights based on the MAF of the tagSNP. Finally, we used a weighted IBS kernel with weights 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 both under the same genetic model as and under a different genetic model than that used in 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 with the null model in Equation 10 but generating the independent dataset with the alternative model (Equation 11) assuming a particular SNP as the QTL.

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 by 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 $\alpha = 0.05$.

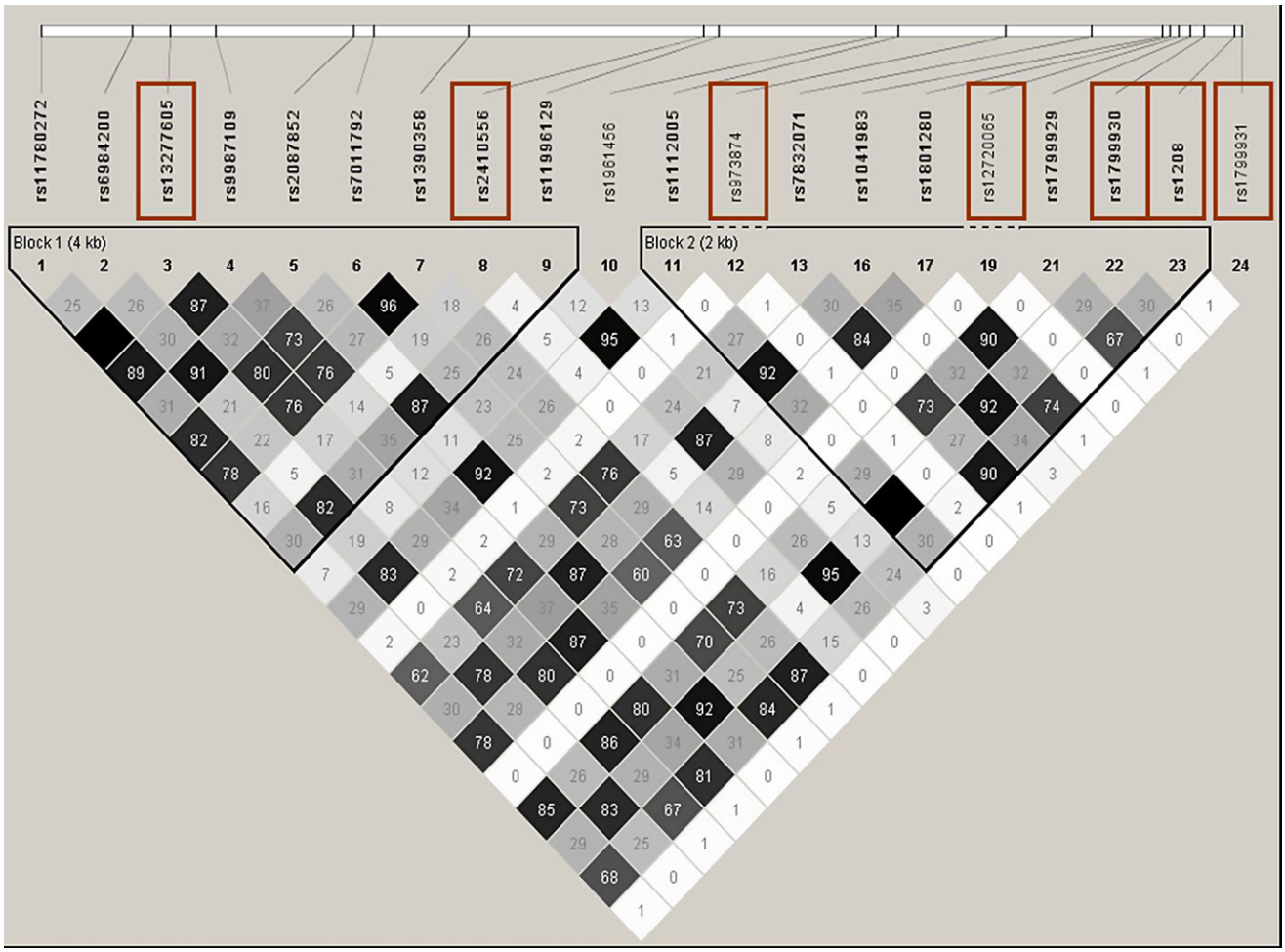

Figure 2. LD Plot of 20 Polymorphic SNPs Residing within the *NAT2* Gene
Results based on the CEU sample from the International HapMap Project. TagSNPs are denoted by a box surrounding the relevant SNP label.

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 Material and 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 by using a dataset that is generated under a different model (i.e., is genetically heterogeneous) from that used for 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

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 QTL SNP explained 3% of the trait variation. For simulations based on *CHI3L2*, the QTL SNP was rs961364 (MAF = 0.293). For simulations based on *NAT2*, the QTL SNP was rs1799930 (MAF = 0.292).

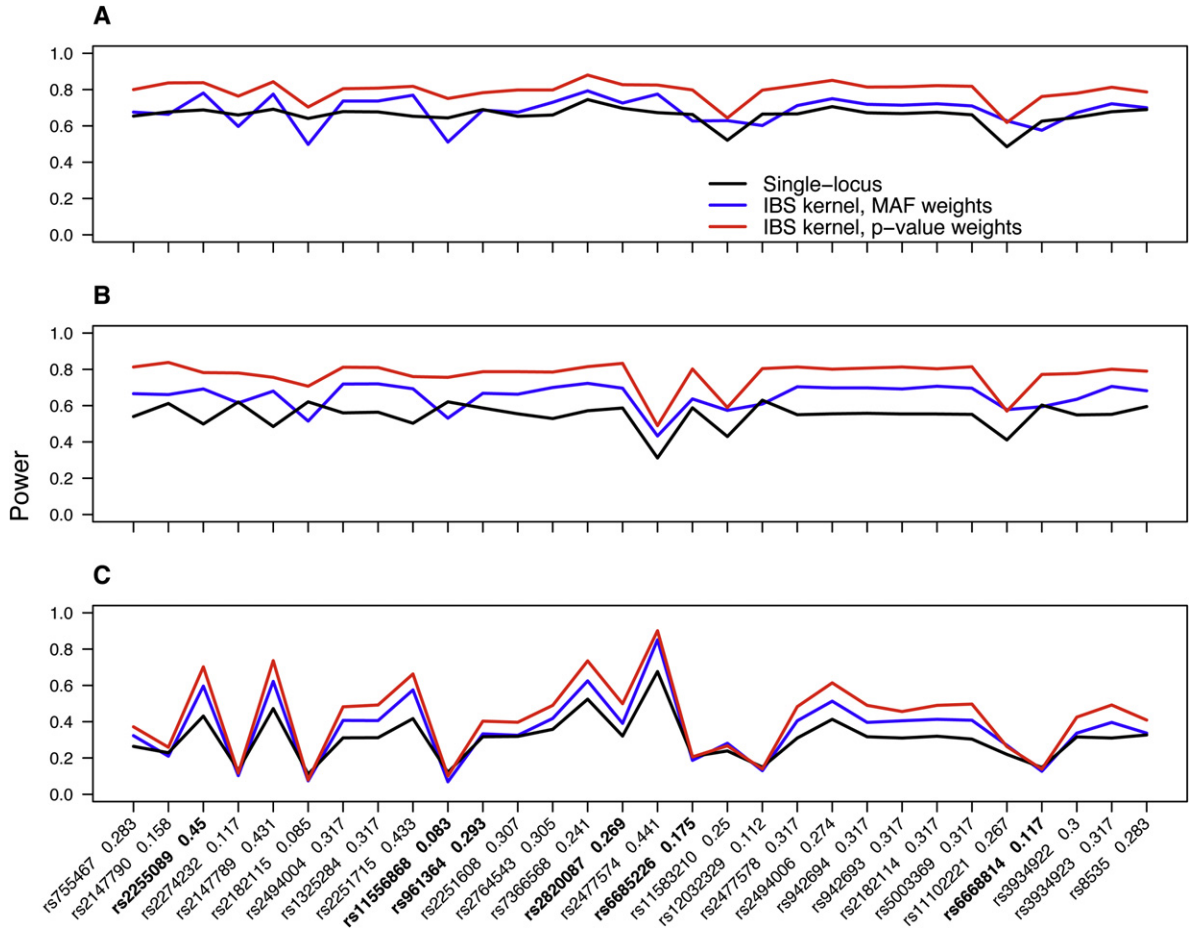

Figure 3. Power Results for Simulations Based on the *CHI3L2* Gene

Power results at $\alpha = 0.05$ for simulations based on the *CHI3L2* gene under additive (A), dominant (B), and recessive (C) mechanisms of allelic effect for the QTL SNP. The x axis labels show the name and minor-allele frequency of the QTL 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.

and, hence, does not affect the validity of our semiparametric approach. For comparison, we analyzed the same datasets by 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 *CHI3L2* SNP used as the QTL 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 SNP used as the QTL (i.e., tagSNP or untyped SNP), and the SNP's MAF. This increased power is hardly surprising, given that the approach using a kernel weighted by p values is the only one of the three shown that uses additional information from an independent dataset to assist in inference.

Although the IBS kernel weighted by MAF displays lower power than the IBS kernel weighted by p values, Figure 3 shows that the former kernel is still generally more powerful than the maximum of the single-tagSNP statistics across QTLs 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 QTL 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

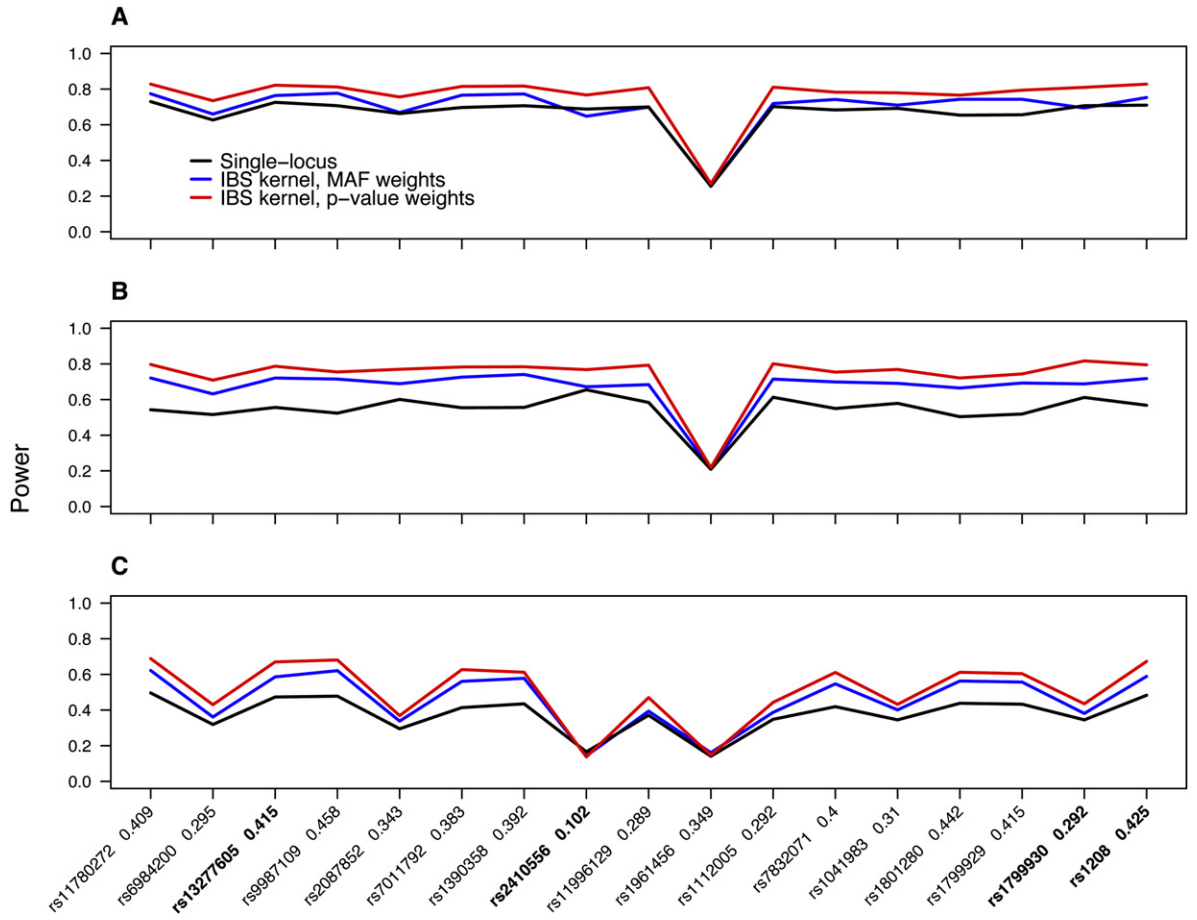

Figure 4. Power Results for Simulations Based on the *NAT2* Gene

Power results at $\alpha = 0.05$ for simulations based on the *NAT2* gene under additive (A), dominant (B), and recessive (C) mechanisms of allelic effect for the QTL SNP. The x axis labels show the name and minor-allele frequency of the QTL 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.

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 anticipate this finding because the semiparametric approach uses a nonparametric approximation of the tagSNP effect [via $h(\cdot)$ in Equation 1] that makes the approach robust to the effects of model misspecification (unlike traditional tag-SNP tests that typically assume a parametric additive model). We also 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 (Equation 4) in Figures 3 and 4. Overall, the performance of the unweighted IBS kernel was similar to that of the IBS kernel weighted by MAF with a few notable differences. For QTL 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 QTL SNPs with MAF $<$ 0.10, we found that 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 QTL SNP was rs2182115 (MAF = 0.085) in *CHI3L2*, we found that 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 QTL SNPs with rare MAFs may be smoothed over by information from surrounding SNPs with more common MAFs. Because 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.

Although primary interest focuses on the testing of the nonparametric function h, secondary interest may focus on the estimation and testing of environmental covariate

Table 2. Parameter Estimates of Environmental Covariates with the Semiparametric Approach

	NAT2		*CHI3L2*	
Genetic Model	$\hat{\beta}_{E,Bin}$	$\hat{\beta}_{E,Cont}$	$\hat{\beta}_{E,Bin}$	$\hat{\beta}_{E,Cont}$
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 QTL SNP was rs1799930 (MAF = 0.292). For simulations based on *CHI3L2*, the QTL SNP was rs961364 (MAF = 0.293). For all simulations, we analyzed replicates by using our semiparametric approach and assuming a IBS kernel weighted by MAF.

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. Because of 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 that uses genotype data from multiple tagSNPs within a region of interest. Using simulated genetic data based on real data from the International HapMap Project,[4] we demonstrated that 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 because 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 nonparametrically with the function h in Equation 1). By maximizing the semiparametric model with LSKM, we show that we can fit the model easily by using common maximization procedures—available in a variety of software packages —for linear mixed models. The approach is computationally efficient to implement; analysis of 1000 replicates of simulated data (with the design described in the Simulations section) took only 5 min 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).

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 by 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 significance of a particular statistic (taking into account the adjustment for multiple correlated tests) by using either permutations or a more computationally efficient approach based on adjustment of correlated p values.[26,27] We will investigate this latter approach in a subsequent paper.

As with traditional multilocus genotype and haplotype analyses, we were primarily interested in applying our semiparametric approach to regions of modest size containing tagSNPs in various degrees of LD with one another and, presumably, the QTL of interest. Nevertheless, we conducted additional simulations examining the stability and performance of our semiparametric approach in situations where the region of interest (and the number of modeled tagSNPs) was considerably larger. For example, using the HapMap CEU sample, we conducted simulations using 33 tagSNPs contained within the 74 kb HNF4α gene (MIM 600281) and found that our approach always converged properly and had appropriate type I error (results not shown). Regarding power, we found that the performance of our semiparametric approach using p value weights was still improved over the single-locus approach as the number of tagSNPs and the length of region considered increased. However, using MAF weights, we found that the performance of our method became quite similar to the single-locus method as the length of the region of interest (and the number of tagSNPs) increased. We explain this result by noting that, as the size of the region of interest increases, the chance of including tagSNPs that are uncorrelated with the true QTL also increases. Such uncorrelated tagSNPs only introduce noise into our method, which makes the true signal from the QTL more challenging to find. In these situations, we recommend applying our

approach within a sliding-window framework, described in the previous paragraph, that considers smaller sets of tagSNPs and thereby decreases the chance of including tagSNPs uncorrelated with the QTL within analysis.

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 trait-influencing regions. Within this article, we considered both MAF weights and p value weights for inference. Other weights are certainly possible (e.g., when gene information is used) and, further, such weights could actually be composite weights that combine information from different sources (e.g., MAF and p values). In this situation, we would first normalize the separate weights to be on the same scale and then develop the composite weight as an average of these scaled weights. We could further modify these composite weights to emphasize one particular source (e.g., p values) over the others in analysis, if so desired.

Of the weights we considered, the most appealing choice 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 on the basis of 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 National Institutes of Health [NIH] funding of such projects). An example of such a database is the NIH-sponsored dbGaP, which will eventually contain information on at least ten whole-genome association studies of complex traits. Also, if a study happens to have p value weights available for certain tagSNPs but not others, then one can apply imputation procedures[28,29] to obtain p values for these untyped variants by using information from nearby SNPs coupled to LD patterns from references sampled from the HapMap project.[4] Finally, 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).

In implementing our approach, we assumed no missing genotype data for the tagSNPs in the region of interest. Although 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 by using either observed genotype data from the study sample[30] or appropriate genotype data from the International HapMap project.[31] Once we impute missing genotypes, we can then incorporate them within our nonparametric function and proceed with analysis as we previously described.

Although 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. For such analyses, we would consider a semiparametric logistic-regression model for a binary outcome ($Y_j = 1$ and 0 for cases and controls, respectively) with the form $log(\mu_j/1-\mu_j) = X_j^T\beta + h(G_j)$, where $\mu_j = P[Y_j = 1|G_j, X_j]$ and G_j, X_j, β, $h(\cdot)$ are defined previously as in Equation 1. Maximization of parameters in this semiparametric logistic-regression model requires the use of a modified LSKM algorithm that is similar to Liu et al.[20] but correctly models the categorical nature of the disease data. As we will describe more thoroughly in a subsequent paper, we can conduct this LSKM analysis analogously by using a logistic mixed model with the form $log(\mu_j^h/1-\mu_j^h) = X_j^T\beta + h$, where X_j, β, and h are defined as previously and $\mu_j^h = E[Y_j|X_j, h]$. We assume that the random tagSNP effects in h follow a multivariate normal distribution with mean vector 0 and variance-covariance matrix $\lambda^{-1}K$, where λ denotes the smoothing parameter and K denotes the chosen kernel matrix. Under these conditions, we can maximize this nonlinear mixed model with a corrected penalized quasi-likelihood algorithm[32] and estimate the nonparametric function by $\hat{h}$ in the LSKM model by $\hat{h}$ in the logistic mixed model. We can then apply a score statistic similar to that of Liu et al.[20] to test the nonparametric function of the genotype data. Although the iterative nature of the penalized quasi-likelihood algorithm will increase the numerical complexity of the semiparametric analysis, it should still be computationally efficient for candidate-gene or whole-genome association analysis.

Our approach fits a semiparametric regression model using LSKM, which we show corresponds to inference via a specific linear mixed model. Although mixed-modeling procedures often are connected to pedigree analysis,[33–35] 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[36] have 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 use of SNP-based haplotypes in their covariance structure and does not 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.[37,38] 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[39] or

available pathway information.[40,41] Such second-level information can improve the precision and accuracy of SNP-based risk estimates.

Because our semiparametric approach is implemented in a linear mixed model, we implicitly assume that the trait data follow or can be transformed to follow approximate normality. With mixed-model-based linkage analysis of quantitative traits,[34] violation of this normality assumption can yield inflated type I error rates to detect linkage if the trait distribution is leptokurtic in nature.[42] To examine whether our semiparametric approach is similarly sensitive to nonnormality of the trait outcome, we conducted additional type I error simulations that generated trait data under various nonnormal distributions (e.g., gamma and log-normal distributions) with large kurtosis values. In all trait simulations, we found that our semiparametric approach had appropriate type I error under the null hypothesis (results not shown) and hence does not appear to be sensitive to nonnormality of the trait data.

Appendix A

Approximate Distribution of the Score Statistic S_τ in Equation 9

We consider the linear mixed model described previously in Equation 6:

$$Y = X\beta + h + E,$$

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

Using the mixed model in Equation 6, we seek to determine the distribution of the score statistic in Equation 9 for testing H_0: $\tau = 0$. Zhang and Lin[43] noted that, because $\tau \geq 0$, we are testing the parameter on its boundary value, and, as a result, the distribution of S_τ follows a mixture of ${\chi_1}^2$ distributions. To facilitate inference, the authors showed that one can approximate this complicated mixture distribution with a scaled χ^2 distribution $\delta{\chi_\nu}^2$, 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 Equation 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 Equation 9. Therefore, we replace the mean e with $\tilde{e} = \mathrm{tr}(P_0K)/2$, where $P_0 = I - X(X^TX)^{-1}X^T$ 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 $I_{\tau\tau} = tr(P_0K)^2/2, I_{\tau\sigma^2} = tr(P_0KP_0)/2$, and $I_{\sigma^2\sigma^2} = tr(P_0^2)/2$.

Once we obtain $\tilde{e}$ and $\tilde{I}_{\tau\tau}$, 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 $\delta = \tilde{I}_{\tau\tau}/2\tilde{e}$ and calculate the degrees of freedom as $\nu = 2\tilde{e}^2/\tilde{I}_{\tau\tau}$. We can then compare the value of the resulting scaled score statistic, S_τ/δ, to a chi-square distribution with ν degrees of freedom in order to assess significance of the test of H_0: $\tau = 0$.

Acknowledgments

This work was sponsored by NIH grants GM074909 (to L.C.K.), HG003618 (to L.C.K and M.P.E.), and CA76404 (to X.L.).

Received: July 2, 2007
Revised: October 4, 2007
Accepted: October 16, 2007
Published online: February 7, 2008

Web Resources

The URLs for data presented herein are as follows:

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/

References

1. Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., and Nickerson, D.A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet. *74*, 106–120.
2. Stram, D.O. (2004). Tag SNP selection for association studies. Genet. Epidemiol. *27*, 365–374.
3. de Bakker, P.I.W., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. Nat. Genet. *37*, 1217–1223.
4. The International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.
5. Lin, D.Y. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics *21*, 781–787.
6. Nyholt, D.R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am. J. Hum. Genet. *74*, 765–769.
7. Li, J., and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity *95*, 221–227.

8. Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M., and Poland, G.A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am. J. Hum. Genet. *70*, 425–434.
9. Zaykin, D.V., Westfall, P.H., Young, S.S., Karnoub, M.A., Wagner, M.J., and Ehm, M.G. (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum. Hered. *53*, 79–91.
10. Tzeng, J.Y., Wang, C.H., Kao, J.T., and Hsiao, C.K. (2006). Regression-based association analysis with clustered haplotypes through use of genotypes. Am. J. Hum. Genet. *78*, 231–242.
11. Chapman, J.M., Cooper, J.D., Todd, J.A., and Clayton, D.G. (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.
12. Roeder, K., Bacanu, S.A., Sonpar, V., Zhang, X., and Devlin, B. (2005). Analysis of single-locus tests to detect gene/disease associations. Genet. Epidemiol. *28*, 207–219.
13. Rosenberg, P.S., Che, A., and Chen, B.E. (2006). Multiple hypothesis testing strategies for genetic case-control association studies. Stat. Med. *25*, 3134–3149.
14. Wang, T., and Elston, R.C. (2007). Improved power by use of a weighted score test for linkage disequilibrium and mapping. Am. J. Hum. Genet. *80*, 353–360.
15. Gauderman, W.J., Murcray, C., Gilliland, F., and Conti, D.V. (2007). Testing association between disease and multiple SNPs in a candidate gene. Genet. Epidemiol. *31*, 383–395.
16. Schaid, D.J., McDonnell, S.K., Hebbring, S.J., Cunningham, J.M., and Thibodeau, S.N. (2005). Nonparametric tests of association of multiple genes with human disease. Am. J. Hum. Genet. *76*, 780–793.
17. Wessel, J., and Schork, N.J. (2006). Generalized Genomic Distance–Based Regression Methodology for Multilocus Association Analysis. Am. J. Hum. Genet. *79*, 792–806.
18. Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). Semiparametric Regression (Cambridge, UK: Cambridge University Press).
19. Rasmussen, C.E., and Williams, C.K.I. (2006). Gaussian Processes for Machine Learning (Cambridge, MA: MIT Press).
20. Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least squares kernel machines and linear mixed models. Biometrics *63*, 1079–1088.
21. Cristianini, N., and Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods) (Cambridge: Cambridge University Press).
22. Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. *68*, 978–989.
23. Stephens, M., and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am. J. Hum. Genet. *76*, 449–462.
24. Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., and Donnelly, P., for the International HapMap Consortium. (2006). A comparison of phasing algorithms for trios and unrelated individuals. Am. J. Hum. Genet. *78*, 437–450.
25. Valle, T., Tuomilehto, J., Bergman, R.N., Ghosh, S., Hauser, E.R., Eriksson, J., Nylund, S.J., Kohtamaki, K., Toivanen, L., Vidgren, G., et al. (1998). Mapping genes for NIDDM: design of the Finland-United States Investigation of NIDDM Genetics (FUSION) study. Diabetes Care *21*, 949–958.
26. Zaykin, D.V., Zhivotovskty, L.A., Westfall, P.H., and Weir, B.S. (2002). Truncated product method for combining p-values. Genet. Epidemiol. *22*, 170–185.
27. Conneely, K.N., and Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated traits. Am. J. Hum. Genet. *81*, 1158–1168.
28. Nicolae, D.L. (2006). Testing untyped alleles (TUNA)-Applications to genome-wide association studies. Genet. Epidemiol. *30*, 718–727.
29. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. *39*, 906–913.
30. Scheet, P., and 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.
31. Zaitlen, N., Kang, H.K., Eskin, E., and Halperin, E. (2007). Leveraging the HapMap correlation structure in association studies. Am. J. Hum. Genet. *80*, 683–691.
32. Lin, X., and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. J. Am. Stat. Assoc. *91*, 1007–1016.
33. Amos, C.I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. Am. J. Hum. Genet. *54*, 535–543.
34. Almasy, L., and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. *62*, 1198–1211.
35. Abecasis, G.R., Cardon, L.R., and Cookson, W.O.C. (2000). A general test of association for quantitative traits in nuclear families. Am. J. Hum. Genet. *66*, 279–292.
36. Tzeng, J.Y., and Zhang, D. (2007). Haplotype-based association analysis via variance components score test. Am. J. Hum. Genet. *81*, 927–938.
37. Witte, J.S. (1997). Genetic analysis with hierarchical models. Genet. Epidemiol. *14*, 1137–1142.
38. Witte, J.S., Greenland, S., Kim, L., and Arab, L. (2000). Multilevel modeling in epidemiology with GLIMMIX. Epidemiology *11*, 684–688.
39. Conti, D.V., and Witte, J.S. (2003). Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. Am. J. Hum. Genet. *72*, 351–363.
40. Hung, R.J., Brennan, P., Malaveille, C., Porru, S., Donato, F., Borreta, P., and Witte, J.S. (2004). Using hierarchical modeling in genetic association studies with multiple markers: Application to a case-control study of bladder cancer. Cancer Epidemiol. Biomarkers Prev. *13*, 1013–1021.
41. Chen, G.K., and Witte, J.S. (2007). Enriching the analysis of genomewide association studies with hierarchical modeling. Am. J. Hum. Genet. *81*, 397–404.
42. Allison, D.B., Neale, M.C., Zannolli, R., Schork, N.J., Amos, C.I., and Blangero, J. (1999). Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am. J. Hum. Genet. *65*, 531–544.
43. Zhang, D., and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. Biostatistics *4*, 57–74.