Published in final edited form as:

Genet Epidemiol. 2015 September ; 39(6): 456–468. doi:10.1002/gepi.21909.

A Fast Multiple-Kernel Method with Applications to Detect Gene-Environment Interaction

Rachel Marceau¹, Wenbin Lu¹, Shannon Holloway¹, Michèle M. Sale², Bradford B. Worrall³, Stephen R. Williams³, Fang-Chi Hsu⁴, and Jung-Ying Tzeng^{1,5,6,*}

¹Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America

²Center for Public Health Genomics, Department of Medicine, Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, United States of America

³Center for Public Health Genomics, Department of Public Health Sciences, Department of Neurology, University of Virginia, Charlottesville, Virginia, United States of America

⁴Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, North Carolina, United States of America

⁵Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America

⁶Department of Statistics, National Cheng-Kung University, Tainan, Taiwan

Abstract

Kernel machine (KM) models are a powerful tool for exploring associations between sets of genetic variants and complex traits. While most KM methods use a single kernel function to assess the marginal effect of a variable set, KM analyses involving multiple kernels have become increasingly popular. Multi-kernel analysis allows researchers to study more complex problems, such as assessing gene-gene or gene-environment interactions, incorporating variance-component based methods for population substructure into rare-variant association testing, and assessing the conditional effects of a variable set adjusting for other variable sets. The KM framework is robust, powerful, and provides efficient dimension reduction for multi-factor analyses, but requires the estimation of high dimensional nuisance parameters. Traditional estimation techniques, including regularization and the EM algorithm, have a large computational cost and are not scalable to large sample sizes needed for rare variant analysis. Therefore, under the context of gene-environment interaction, we propose a computationally efficient and statistically rigorous “fastKM” algorithm for multi-kernel analysis that is based on a low-rank approximation to the nuisance-effect kernel matrices. Our algorithm is applicable to various trait types (e.g., continuous, binary, and survival traits) and can be implemented using any existing single-kernel analysis software. Through extensive simulation studies, we show that our algorithm has similar performance to an EM-based KM approach for quantitative traits while running much faster. We also apply our method to the

*Correspondence to: Jung-Ying Tzeng, Department of Statistics and Bioinformatics Research Center, North Carolina State University, Campus Box 7566, Raleigh, NC 27695, USA. jytzeng@stat.ncsu.edu.

Vitamin Intervention for Stroke Prevention (VISP) clinical trial, examining gene-by-vitamin effects on recurrent stroke risk and gene-by-age effects on change in homocysteine level.

Keywords

multiple-kernel analysis; kernel machine regression; exon level association test; gene-environment interaction; gene-gene interactions

Introduction

Kernel machine (KM) based approaches [Kwee et al., 2008; Lin et al., 2011; Liu et al., 2007, 2008; Wu et al., 2010, 2011] provide a powerful and popular strategy for evaluating associations between a set of genetic variants and complex traits of various types. The KM method uses a kernel function to quantify the pairwise genetic similarity for individuals based on multiple genetic variants; it then assesses the gene-trait association by examining if the genetic similarity of a pair of individuals is associated with their trait similarity [Lin et al., 2011; Wu et al., 2013]. Many other marker-set methods, e.g., variance component tests [Goeman et al., 2004; Lin et al., 2013] and similarity regressions [Tzeng et al., 2009, 2011, 2014; Wang et al., 2014a; Zhao et al., 2015], are closely related to KM tests under a random effects model. Therefore, although this paper primarily focuses on KM methods, the proposed approach and discussions are applicable to other variance component and similarity-based tests.

While the popular KM methods, e.g., SKAT [Wu et al., 2011], focus on single-kernel analysis (i.e., using one kernel function to model a single variable set), analyses involving multiple kernels (i.e., using separate kernels to simultaneously model multiple variable sets) are also frequently encountered in genomic research. Multi-kernel approaches include tests for gene-environment (G×E) interactions [Lin et al., 2013; Tzeng et al., 2011; Wang et al., 2015a; Zhao et al., 2015], tests for gene-gene interactions [Larson and Schaid, 2013; Wang et al., 2014a], conditional tests for evaluating the effect of a single variable set adjusting for other variable sets [Pang et al., 2014; Wang et al., 2015a, b], and SKAT analysis coupled with a variance component method [Kang et al., 2010] to account for population substructure. The number of explanatory variables in these analyses is much higher than in a single marker-set analysis. In these situations, KM offers efficient dimension reduction and yields higher power to evaluate the effects of interest compared to other alternatives such as burden-based methods (e.g., Madsen and Browning [2009]; Price et al. [2010]). In addition, KM methods can model nonlinear/non-additive effects, accommodate variables with different direction and magnitude of effects, and are more robust than burden-based methods because they impose fewer assumptions on the underlying effects. The latter is particularly important in multi-kernel analysis — for example, a KM G×E test is more robust against misspecification of the main effects of G and E than the burden-based G×E test [Wang et al., 2015a].

However, the merits of KM approaches come with high computational cost for multi-kernel analyses, which substantially limits their practical utility. In a multi-kernel model, each effect is modeled via a $n \times n$ kernel matrix and an n -dimensional parameter vector (where n

is the number of individuals). Computing the test statistic to evaluate the effect of interest in a multi-kernel analysis requires estimation of at least one set of n -dimensional nuisance parameters. For example, performing a KM G×E test, even under the null hypothesis of no G×E effects, requires the estimation of nuisance genetic main effects. Current attempts to overcome the dimensionality challenges include treating the n -dimensional parameters as random and using the EM algorithm to estimate its variance component [Tzeng et al., 2011; Wang et al., 2014a, 2015a; Zhao et al., 2015], or imposing penalization on these parameters [Lin et al., 2013]. While both techniques have proven to be valid, the estimation procedures are usually phenotype-specific (e.g., the algorithms developed for quantitative traits cannot be applied to binary or survival traits) and computationally intensive (e.g., requiring the inversion of a n -dimensional matrix at each iteration of the EM algorithm or the tuning of a regularization parameter), making them not scalable to the large samples considered in rare variant studies.

Using KM tests for G×E interactions as an example, we illustrate our solution to resolve these computational challenges: a computationally efficient and statistically rigorous algorithm for performing KM tests in multi-kernel analyses. Our algorithm is motivated by the fact that the $n \times n$ kernel matrix is often not full rank — its rank is generally much less than the minimum of the number of individuals and the number of variables (e.g., SNPs) in a variable set. Thus, by decomposing the kernel matrices of nuisance effects, we can reduce the dimensionality to a manageable size so that a random-effect treatment or penalization is not necessary, and consequently a fixed effect null model can be fit to obtain nuisance parameters. The proposed method is fast, scalable to larger n , and applicable to a variety of trait types; most importantly, it can be implemented using any existing software for single KM analysis. For example, our algorithm would allow one to perform a KM G×E test using the existing software for main effect KM tests such as SKAT.

We explore the performance of our method through an in depth simulation study for quantitative, binary, and survival traits. We also apply our method to the Vitamin Intervention for Stroke Prevention (VISP) clinical trial [Toole et al., 2004]. Focusing on the nine candidate genes within the homocysteine pathway, we examine the gene-by-vitamin effects on recurrent stroke risk [Hsu et al., 2011; Tzeng et al., 2014] and gene-by-age effects on change in homocysteine level [Tzeng et al., 2011]. In such, we are able to see the true flexibility and unifying features of our method, in its ability to perform multi-kernel analysis for different trait types in a computationally efficient manner.

Materials and Methods

The KM Model for G×E Interactions

Consider a study with n individuals. For individual i , $i = 1, \dots, n$, let Y_i denote the phenotype of interest and $G_i = (g_{i1}, \dots, g_{iM})^T$ denote the genetic markers. For now we assume an additive genetic effect where g_{im} is the count of minor alleles that individual i has at marker m , $m = 1, \dots, M$, though it is straightforward to extend this to recessive or dominant modes of inheritance. In addition, let $X_i = (1, X_{i1}, \dots, X_{iq})^T$ denote the baseline covariates that have no interaction with genetic markers and E_i the covariate interacting with genetic markers. For simplicity, we assume that E_i is a scalar.

We consider a generalized linear model for Y_i

$$g(\mu_i) = X_i^T \beta_X + E_i \beta_E + h_G(G_i) + h_{GE}(G_i, E_i), \quad (1)$$

where $\mu_i = E(Y_i | X_i, E_i, G_i)$, $g(\mu_i)$ is a canonical link function, for example, $g(\mu_i) = \mu_i$ for

quantitative traits and $g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right)$ for binary traits, $h_G(\cdot)$ and $h_{GE}(\cdot)$ are two nonparametric smooth functions representing the main effect of genetic markers (i.e., G effect) and interaction effect between the E covariate and genetic markers (i.e., G×E effect). Model (1) can also be expressed in matrix notation as

$$g(\mu) = X\beta_X + E\beta_E + h_G(G) + h_{GE}(G, E), \quad (2)$$

where, for example, $X = (X_1, \dots, X_n)^T$, $G = (G_1, \dots, G_n)^T$, $E = (E_1, \dots, E_n)^T$, $g(\mu) = (g(\mu_1), \dots, g(\mu_n))^T$, and $h_G(G) = (h_G(G_1), \dots, h_G(G_n))^T$.

Under this model, we can test for G×E interaction using the null hypothesis $H_0 : h_{GE}(\cdot) = 0$. Since $h_G(\cdot)$ and $h_{GE}(\cdot)$ are smooth functions which lie in a Hilbert space, by the representer theorem we can write $h_G(G)$ and $h_{GE}(G, E)$ in dual form expressions [Kimeldorf and Wahba, 1971]: $h_G(G) = K_G \alpha_G$ and $h_{GE}(G, E) = K_{GE} \alpha_{GE}$, where $K_G = \{K_G(G_i, G_j) : 1 \leq i, j \leq n\}$ is an $n \times n$ kernel matrix for the G effect, $K_{GE} = \{K_{GE}(\{G_i, E_i\}, \{G_j, E_j\}) : 1 \leq i, j \leq n\}$ is an $n \times n$ kernel matrix for the G×E effect, and α_G and α_{GE} are $n \times 1$ vectors of unknown parameters. One commonly adopted kernel function for the G main effect is the weighted identity by state (IBS) kernel [Kwee et al., 2008; Wu et al., 2010], i.e.,

$$K_G^{IBS}(G_i, G_j) = \frac{\sum_{m=1}^M w_m \{2I(g_{im}=g_{jm}) + I(|g_{im} - g_{jm}|=1)\}}{\sum_{m=1}^M w_m}$$
. The IBS kernel quantifies genetic similarity using the weighted average number of alleles for which two individuals have in common in the marker set. The weights w_m 's are prespecified to upweight or downweight a variant based on certain features. For example, one can weight against the minor allele frequency of marker m so as to upweight similarities that are contributed by rare variants. Given the genetic main effect kernel K_G , one possible way to construct the interaction kernel K_{GE} is to take the element-wise product of the genetic main effect kernel and the environmental kernel K_E , as described in Larson and Schaid [2013] and Wang et al. [2015a]. When the environmental covariate is a scalar, this simplifies to $K_{GE} = D_E K_G D_E$, where D_E is a diagonal matrix with elements E [Tzeng et al., 2011]. However, caution must be taken when using this direct product kernel to avoid duplicating the main effect terms in the G×E kernel [Wang et al., 2015a].

The smooth functions $h_{GE}(G, E)$ can be viewed as random effects [Liu et al., 2007, 2008] and modeled through a multivariate normal distribution with mean zero and variance-covariance $\tau_{GE} K_{GE}$, i.e., $h_{GE}(G, E) \sim N(0, \tau_{GE} K_{GE})$. This is equivalent to

$\alpha_{GE} \sim N(0, \tau_{GE} K_{GE}^{-1})$ since $h_{GE}(G, E) = K_{GE} \alpha_{GE}$. Using this representation, testing $H_0 : h_{GE}(\cdot) = 0$ is equivalent to testing the null hypothesis $H_0 : \tau_{GE} = 0$ via a variance component score test [Liu et al., 2007, 2008].

Lin [1997] showed that the variance component score test is locally most powerful for testing the genetic main effect and only requires fitting the model under the null hypothesis, i.e., a standard generalized linear model. For interaction tests however, fitting the null model requires estimation of the nonparametric function $h_G(\cdot)$. There are two main approaches for fitting the null model: the first strategy treats $h_G(G)$ as random effects following a multivariate normal distribution with mean zero and variance-covariance $\tau_G K_G$, and uses an EM algorithm to estimate the nuisance variance component τ_G (e.g., Tzeng et al. [2011]; Wang et al. [2014a, 2015a]; Zhao et al. [2015]). The second strategy uses penalization techniques like ridge regression to estimate the $n \times 1$ vector of parameters α_G (e.g., Lin et al. [2013]). Both strategies, however, are difficult computationally. The random effect approach is time consuming due to inverting an $n \times n$ matrix at each iteration of the EM algorithm, and has difficulties with estimation on the boundary of the parameter space. On the other hand, penalization methods require selection of a proper tuning parameter.

FastKM Test for G×E Interactions

We propose to take advantage of the low-rank structure of the kernel matrix K_G to enhance computational efficiency. Clearly, the weighted IBS kernel matrix is almost never full rank—typically, $\text{rank}(K_G) \ll \min(n, M)$ (i.e., less than the number of individuals and the number of markers of interest). Since K_G is a symmetric matrix, it can be decomposed using eigendecomposition as $K_G = Q\Lambda Q^T$, where Q is the matrix of eigenvectors of K_G and Λ is a diagonal matrix of eigenvalues of K_G . Removing the near-zero eigenvalues or taking only the leading eigenvalues that capture a high percentage of the total variation results in a low-rank decomposition $K_G = Z_{n \times r} Z_{r \times n}^T$, where $r \ll n$ is the number of positive eigenvalues kept. The null model, then, reduces to the form

$$g(\mu) = X\beta_X + E\beta_E + ZZ^T\alpha_G \triangleq X\beta_X + E\beta_E + Z\gamma. \quad (3)$$

where $\gamma = Z^T\alpha_G$. Model (3), referred to as the fastKM null model, is a standard GLM with low dimensional parameters. The fastKM null model can be rewritten in terms of an augmented design matrix $A = (X, E, Z) \equiv (A_1, \dots, A_n)^T$ and corresponding parameter vector

$\theta = (\beta_X^T, \beta_E^T, \gamma^T)^T$ as $g(\mu) = A\theta$; the parameter θ can be directly estimated by the maximum likelihood estimation using standard software. In the same spirit, we can rewrite Model (2) and obtain the following fastKM model: $g(\mu) = X\beta_X + E\beta_E + Z + h_{GE}(G, E)$. We then construct the score test statistic for $H_0: \tau_{GE} = 0$ as $U_n = n^{-1}(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n) K_{GE}(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)^T$, where $\hat{\theta}$ is the maximum likelihood estimator of θ under the null and $\hat{\varepsilon}_i = Y_i - g^{-1}(A_i^T \hat{\theta})$ ($i = 1, \dots, n$) are fitted residuals.

Note that our score test statistic shares the same form of the KM score test statistic for genetic main effects (except that K_{GE} is involved instead of K_G). Therefore, the KM G×E test can be conducted using any existing testing software for genetic main effects, such as SKAT [Wu et al., 2011], providing the augmented design matrix A and the G×E kernel K_{GE} as input. Moreover, as with main effect KM tests, the limiting distribution of the fastKM test

statistic under the null can also be represented as $\sum_{i=1}^d \lambda_i \chi_{1,i}^2$, where $\chi_{1,i}^2$ ($i = 1, \dots, d$) are independent Chi-squared random variables with one degree of freedom and weights λ_i 's are

the positive eigenvalues of a nonnegative definite matrix Σ . Here matrix Σ is the variance-covariance matrix of the limiting distribution of $n^{-1/2} \sum_{i=1}^n \hat{\varepsilon}_i Z_{GE,i}$, where $Z_{GE,i}$ is the i -th row of matrix Z_{GE} with $Z_{GE} Z_{GE}^T = K_{GE}$ [Lin, 1997; Zhang and Lin, 2003]. The associated p-value can be calculated using moment matching method [Duchesne and Lafaye De Micheaux, 2010], the Davies method [Davies, 1980], or empirically using resampling techniques. For a resampling method, one can generate many sets of independent Chi-squared random variables with one degree of freedom and calculate $U_{n,b} = \sum_{i=1}^d \hat{\lambda}_i \chi_{1,i}^2$, $b = 1, \dots, B$, where $\hat{\lambda}_i$ is a consistent estimator of λ_i . Then the estimated p-value is

$$B^{-1} \sum_{b=1}^B I(U_n < U_{n,b}).$$

Extension to Survival Traits

As of yet, no methods have been developed for testing marker-set G×E effects for survival traits. Our fastKM method, however, can be naturally extended to include survival traits. For individual i , let T_i denote the event time of interest and C_i the censoring time. Further, define $\tilde{T}_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. As usual, we assume $T_i \perp C_i$ given X_i, E_i and G_i . For simplicity, we consider the proportional hazards model, though other survival models can also be used following the derivations in Tzeng et al. [2014]. Under the null, we fit the proportional hazards model with the augmented covariates A_i as defined previously. Let $\hat{\theta}$ denote the maximum partial likelihood estimator of θ and $\hat{\Lambda}(\cdot)$ denote the Breslow estimator of the baseline cumulative hazard function. Our proposed test statistic for the null hypothesis $H_0 : h_{GE}(\cdot) = 0$ is given by $U_n = n^{-1}(\varepsilon_1, \dots, \varepsilon_n) K_{GE}(\varepsilon_1, \dots, \varepsilon_n)^T$, where

$\hat{\varepsilon}_i = \delta_i - \hat{\Lambda}(\tilde{T}_i) \exp(A_i^T \hat{\theta})$ ($i = 1, \dots, n$) is a martingale residual. The fastKM test statistic again shares the same form as the score test statistic for the genetic main effect tests discussed in Lin et al. [2011] and Tzeng et al. [2014], with our augmented design matrix A as the covariate design matrix and G×E kernel K_{GE} as the kernel matrix. Consequently, the fastKM test statistic and its p-value can be obtained by using the existing software for KM main-effect tests with survival traits (e.g., Lin et al. [2011]; Tzeng et al. [2014]).

Implementation Detail

In practice, the low rank decomposition may contain several near-zero eigenvalues, which would result in the Z matrix (and hence also the augmented design matrix A) containing multiple near-zero columns and lead to unstable parameter estimation. We suggest performing kernel principal component analysis (kPCA) to further reduce dimensionality and improve stability [Cai et al., 2011; Schölkopf et al., 1998]. In practice, choosing to keep the top eigenvalues which collectively explain $p = 95\%$ or 99% of the total variability can give good empirical results, especially for continuous traits. However, when the variants of interest are rare and there are many near-zero eigenvalues, a smaller p may be necessary for achieving estimation stability with binary and survival traits. In our numerical studies (simulation and real data analysis), we identify a suitable $p\%$ by starting with a high percentage (e.g., 99%) and then gradually reducing p until the null model can be fitted reasonably well (e.g., no warning messages in the GLM fits or no extremely large coefficients). This is also the rule we implemented in the fastKM R function.

Simulation Study

Data Generation—We generated a set of 10,000 haplotypes using the COSI software of Schaffner et al. [2005] with a coalescent model mimicking the linkage disequilibrium and population history of the European population. We then formed a marker-set of M rare ($\text{MAF} < 0.05$) loci, of which the first 40 were considered causal. We generated a sample of n individual genotypes by randomly sampling two haplotypes with replacement. We set $M = 100$ loci and $n = 5000$ individuals. We also considered $M = 200$ and $n = 1000$ in a subset of scenarios to investigate the impact of M and n on the performance of fastKM. We considered a single environmental covariate E that is either continuous (generated from a $\text{Normal}(0,1)$ distribution) or binary (generated from a $\text{Bernoulli}(0.5)$ distribution). We assumed no confounding covariates in the simulations.

We evaluate the performance of the fastKM $G \times E$ tests using data generated from a fixed effect model, where the genetic main and interaction effects depend on mutational burden. Specifically, given each individual's environmental covariate and genotypes, define

$\eta(E_i, G_i) = \gamma_0 + \gamma_E E_i + \gamma_G \sum_{m=1}^{M_c} g_{im} + \gamma_{GE} E_i (\sum_{m=1}^{M_c} g_{im})$, where $M_c = 40$ represents the number of causal variants. We considered three types of traits: quantitative, binary, and survival. Quantitative responses were generated from the model $Y_i = \eta(E_i, G_i) + \varepsilon$ where $\varepsilon \sim N(0, 1)$. Binary responses were generated in a case-control framework from a Bernoulli

distribution with $P(Y_i = 1 | E_i, G_i) = \frac{e^{\eta(E_i, G_i)}}{1 + e^{\eta(E_i, G_i)}}$. Finally, survival traits were generated from a Cox proportional hazards model: $\log(T_i) = -\eta(E_i, G_i) + \varepsilon_i^*$, where $\gamma_0 = 0$ and ε_i^* follows a standard extreme value distribution. For survival traits, censoring times were generated from a uniform distribution on $[0, c]$, where c was chosen to yield censoring proportions of 15% and 40%.

Examining Power and Type I Error of FastKM—In each simulation scenario, we perform 2000 replicates for type I error analysis and 1000 replicates for power analysis. We compared our fastKM $G \times E$ test to the burden-based $G \times E$ test. For quantitative traits, we also compare with the traditional EM-based KM $G \times E$ test (referred to as “originalKM”). We calculated the IBS kernel for K_G as described in Tzeng et al. [2011] and set the variant-specific weight as $w_m = (1 - q_m)^{24}$ [Wu et al., 2011]. For certain settings under quantitative traits, we also calculated the polynomial kernel (i.e., $K_G^{\text{poly}}(G_i, G_j) = (1 + G_i^T G_j)^d$) with $d = 2$ and $d = 3$. After performing the eigenvalue decomposition of K_G , we rounded those eigenvalues with magnitude $< 10^{-10}$ to zero, and kept the top eigenvalues to explain $p\%$ of the variability ($p\%$ kPCA); the value of p is chosen to retain the maximum amount of variation while still yielding stable GLM estimates.

For the fastKM algorithm, we fit the fastKM null model for each trait with linear predictor $\eta(A) = \gamma_0 + \gamma_A A$ where A is the augmented design matrix composed of a standardized covariate and the reduced Z matrix. P-values were calculated using the Davies method [Davies, 1980] as implemented in Duchesne and Lafaye De Micheaux [2010]. For the burden-based test, we fit the models with the covariate effects

$\eta(G_i, E_i) = \gamma_0 + \gamma_E E_i + \gamma_G \sum_{m=1}^M g_{m,i} w_m + \gamma_{GE} E_i \sum_{m=1}^M g_{m,i} w_m$, and conducted a Wald test of $H_0 : \gamma_{GE} = 0$. Finally, for the original KM method for quantitative traits, we used the G×E test of Tzeng et al. [2011], which estimates the variance component τ_G via an EM algorithm.

Application to VISP Study

We apply our method to the Vitamin Intervention for Stroke Prevention (VISP) clinical trial data [Toole et al., 2004]. The VISP trial was a multi-center study in which 3680 ischemic stroke patients, with informed consent, were randomly assigned to one of two vitamin dosage arms and were followed up until they experienced a subsequent stroke, or alternatively suffered from a myocardial infarction or death. Two vitamin dosages were administered: the low-dose arm consisted of 200 μg B_6 , 6 μg B_{12} , and 20 μg folic acid; the high-dose arm consisted of 25 mg B_6 , 0.4 mg B_{12} , and 2.5 mg folic acid. The genetic sub-study enrolled and consented 2,164 individuals [Hsu et al., 2011; Tzeng et al., 2014]. Following the studies of Hsu et al. [2011] and Tzeng et al. [2014], we focus on the nine candidate genes within the homocysteine (Hcy) pathway: *BHMT*, *BHMT2*, *CBS*, *CTH*, *MTHFR*, *MTR*, *MTRR*, *TCN1* and *TCN2*, treating each as a recessive gene.

Our primary interest was to test whether dosage level significantly interacts with any of the genes in the Hcy pathway to determine the time until subsequent stroke. Individuals were considered censored if they dropped out of the study or did not have another stroke before the end of the study. Approximately 91% of the patients were censored. Hsu et al. [2011] and Tzeng et al. [2014] previously examined this problem respectively using single SNP and gene-level genetic main effect tests stratified by treatment dosage. We extend the analysis by conducting a G×E aggregation test with intervention group as the environmental variable.

To perform the G×E test, we fit a Cox proportional hazards model looking at the effect of gene × intervention interaction on time until stroke, adjusting for age, sex, and race. We assume any missingness is at random and exclude all individuals with missing covariate values or genotype and all loci with too much missingness. The final data set consisted of 1914 total subjects and 74 loci. This data consisted of a mixture of common and rare variants, so the weight $w_m = q_m^{-3/4}$ was used in calculating the weighted IBS kernel [Pongpanich et al., 2012; Tzeng et al., 2014]. For quantitative traits, we also considered the polynomial kernel with $d = 2$. To improve stability and efficiency we consider dimension reduction via kernel PCA. We find 85% kPCA to be sufficient reduction.

As a secondary research question we consider whether a patient's age interacts with their genotype in affecting the change in homocysteine during a 2 hour fasting methionine load test performed at baseline (pre-randomization), similar to Tzeng et al. [2011]. The change in total Hcy was analyzed as both a continuous trait and a binary trait using the sample 90th percentile as a cut-off to dichotomize patients (i.e., the phenotype is 1 for individuals whose change in total Hcy is in the top 10%, and 0 for all others). A 90% cut off has been used in the past as an indicator for possible hyperhomocysteinaemia (e.g., van der Griend et al. [2002]). Sex and race were included as covariates in both models. As with the primary outcome, we used weight $w_m = q_m^{-3/4}$, and used 95% kPCA for dimension reduction. P-values were calculated via Davies' method [Davies, 1980] as implemented in Duchesne and

Lafaye De Micheaux [2010]. The analysis of continuous outcomes was also compared to the originalKM method.

Results

Quantitative Traits

Performance evaluation with IBS kernel, $M = 100$ loci and $n = 5000$ individuals

—Figure 1 shows how fastKM and the originalKM method had very similar type I error levels for the G×E test. The type I error rates of both tests were close to the nominal level of 0.05 when the E variable was Gaussian distributed, but were slightly conservative when E was binary. The burden-based test appeared to only be valid for the null models with no genetic main effect (i.e., $\gamma_G = 0$); it had inflated type I error rates for all simulated scenarios with $\gamma_G > 0$, and the magnitude of inflation increased with γ_G . This is consistent with previous findings that fixed effects G×E tests, including burden-based tests, may have inflated type I error when the genetic main effect is modeled incorrectly [Voorman et al., 2011; Wang et al., 2015a].

Figure 2 shows that fastKM had almost identical power to originalKM for quantitative traits, quickly increasing with larger interaction sizes γ_{GE} . The burden-based tests had the lowest power among the three tests, despite having inflated type I error rates for non-zero main effects. The power of all three tests is reduced when the E variable was generated from a Bernoulli distribution.

Tables I and II show the computational time required for the different approaches' G×E tests for type I error analysis and power analysis, respectively. Computations were carried out on one processor of the IBM dual-Xeon (HS21 blade) computer nodes (2.66 GHz) with 4 GB RAM. We found that the burden-based test was the quickest method, but this speed came at the cost of poor overall performance (i.e., inflated type I error and lower power). On the other hand, while fastKM and originalKM had very similar empirical performance, it is evident that fastKM was much more efficient, taking just over 2 minutes per run for a sample size of $n = 5000$ individuals regardless of simulation scenario. The originalKM method was slower, ranging from taking just over three times as long to over 10 times as long per run. It took the longest when there was no genetic main effect, since the EM algorithm has extremely slow convergence when $\tau_G = 0$, i.e., when the nuisance variance component is at the boundary of the parameter space.

Impact of kernel choices and (M, n) on method performance—Figure 3 (type I error rates) and Figure 4 (power) show the impact of marker-set size and sample size, (M, n) , and of kernel selection on the performance of fastKM. In particular, we compare the performance of $(M = 200, n = 1000)$ to $(M = 100, n = 5000)$ under the IBS kernel and polynomial kernels with $d = 2$ and $d = 3$. By comparing the left panel of $(M = 100, n = 5000)$ and right panel of $(M = 200, n = 1000)$ in Figure 3 and Figure 4, we see that, when sample size n is large and marker set size is relatively lower (e.g., $(M = 100, n = 5000)$), fastKM performed similarly to originalKM for all kernel types (i.e., IBS and polynomial with $d = 2$ and $d = 3$). When $M = 200$ and $n = 1000$, while fastKM with IBS kernel still has similar performance to originalKM, fastKM with the more complex polynomial kernels yields

slightly more conservative than originalKM and reduced power. We also observed that the IBS kernel tends to have slightly higher power than the polynomial kernel even with originalKM, which is not unexpected given the effect mechanism considered in the simulation.

Binary and Survival Traits

For binary traits and survival traits, we compared the performance of the fastKM G×E test with burden-based G×E tests. The relative performance of the two approaches was similar to what was observed in quantitative traits. The results of type I error analyses are shown in Figure 5 (for binary traits) and Figure 6 (for survival traits). The fastKM G×E tests had type I error rates around the nominal level, except in the case of binary traits with a continuous E variable, where the test was slightly conservative. The burden-based G×E test was valid for small genetic main effect sizes, but its type I error rate increased with larger γ_G .

The binary trait power analysis (Figure 7) shows that fastKM was more powerful than the burden-based test, which is similar to the quantitative traits results. For survival traits (Figure 8), fastKM had similar or higher power compared to the burden-based test when the censoring proportion was low (i.e., 15%). The power difference between fastKM and burden-based G×E test became more obvious when the censoring proportion was 40%. Overall, when the trait is continuous, binary, or survival, fastKM is a valid test and has pretty good power which scales quickly with increasing interaction effect γ_{GE} .

VISP Study

We first performed a survival analysis of gene-by-intervention interaction on time until subsequent stroke (Table III) using fastKM. We found no significance at the Bonferroni correction threshold of $0.05/9=0.0056$. However, the two genes with the lowest p-values, i.e., *TCN2* (p-value 0.0408) and *CTH* (p-value 0.0171), were also the most significant found in a previous study [Tzeng et al., 2014]. Tzeng et al. [2014] performed a stratified analysis by vitamin intervention and found these two genes had the smallest p-values in the low-dose intervention group using a KM genetic main effect test. Both *TCN2* and *CTH* are members of the folate one-carbon metabolism (FOCM) pathway. The FOCM mediates many key biological processes in the cell including methionine metabolism, Hcy synthesis, B-vitamin utilization and provision of de novo cellular methyl group availability through conversion of S-adenosyl-methionine (SAM) to S-adenosyl-homocysteine (SAH). *TCN2* is the primary plasma facilitator of cellular uptake of B12 [Seetharam et al., 1999], while *CTH* is responsible for converting cystathionine into cysteine. While they did not meet or exceed our Bonferroni corrected p-value, previous work shows that *TCN2* was found to be associated with recurrent stroke in VISP participants randomized to the low-dose B-vitamin arm of the trial [Hsu et al., 2011]. Related to these findings, *CTH* has been found to be associated with Hcy levels [Wang et al., 2004], a well-recognized risk factor for stroke, and *TCN2* has been associated with Hcy levels in healthy individuals [Lievers et al., 2002] and among subjects with low B12 [Stanisawska-Sachadyn et al., 2010].

In the analyses of gene × age effects on Hcy change treated as continuous trait, fastKM and originalKM had similar p-values. Both tests found *MTR* to significantly interact with age to

influence change in homocysteine when treated as a continuous trait, even after correcting for multiple testing. A similar result was obtained when treating the change in Hcy as binary. *MTR* is a member of the FOCM and is responsible for the methylation of Hcy in the resynthesis of methionine. As an essential component of the FOCM, the *MTR* enzyme is activated by *MTRR*, utilizing B12 as a necessary component of the Hcy methylation reaction. Mutations in *MTR* have previously been identified as the underlying cause of multiple metabolic disorders including cases of hyperhomocysteinemia, as a dysfunctional *MTR* enzyme would lead to the inability to convert Hcy in to methionine [Mellman et al., 1979]. Also, mutations in *MTRR* can lead to hyperhomocysteinemia through the inability to activate *MTR* [Leclerc et al., 1998; Rosenblatt et al., 1985; Schuh et al., 1984; Zavadáková et al., 2005]. Through genome-wide association studies, a common *MTR* functional variant, A2756G (D919G) has been found to be associated with a modest, but significant, increase in plasma Hcy levels [Dekou et al., 2001; Harmon et al., 1999; Tsai et al., 2000; Wang et al., 1999]. However, no consistent effect on risk of developing vascular disease has been found [Dekou et al., 2001; Hyndman et al., 2000].

Given the fact that Hcy levels rise with age [Nygård et al., 1995] and that *MTR* is essential in conversion of Hcy to methionine, it is plausible that the effect between change in Hcy and an inherited heterozygous mutation of *MTR* in combination with other FOCM mutations, including *MTR*, may change over time. Additionally, it must be noted that there are other clinical factors that should be taken into account, including kidney and liver function which are also known to diminish with age and are both related to B-vitamin utilization and FOCM function [Spence et al., 1999].

Discussion

While multi-kernel analyses are frequently encountered, the practical utility of existing approaches may be limited due to the computational cost and method complexity. The intensive computation makes the KM analysis unscalable to large samples. The method complexity, mainly arising from the estimation of nuisance variance components in the null model, has lead the majority of multi-kernel approaches to focus on the analysis of quantitative traits. Although a few methods currently exist for binary traits (e.g. Lin et al. [2013]; Zhao et al. [2015]), there is still a lack of multi-kernel methods for survival traits to the best of our knowledge.

In this work, we use the G×E interaction test to illustrate our solution to address these issues based on a low-rank approximation to the kernel matrix for genetic main effects. We demonstrate that the proposed low-rank fastKM framework, when coupled with the IBS kernel, can retain the power and validity of the robust G×E KM test based on random-effects models for quantitative traits. The fastKM method greatly enhances the feasibility of the robust G×E KM test in several aspects. First, fastKM speeds up computation — in some cases our algorithm is up to ten times faster than the robust G×E KM test. The reduction in computation time of the fastKM is expected to be even larger for binary and survival traits. This decrease in computation increases the scalability of multi-kernel approaches to larger sample sizes, as are required for rare variant analyses, and to larger number of SNP sets, making whole genome interactive analysis much more feasible. Secondly, fastKM is

applicable to general trait types, from continuous, binary, to survival traits. By creating an augmented covariate-genotype matrix, fastKM transforms the interaction test into a single-kernel analysis framework, allowing one to perform an interaction test using any existing main-effect testing software. Specifically, single-kernel analysis softwares, e.g. SKAT [Wu et al., 2011], require the input of a covariate design matrix and a kernel matrix, so one can perform fastKM by providing the augmented matrix as the required covariate design matrix and the kernel matrix for the effect to be tested (e.g., $G \times E$) as the required kernel matrix. We provide the R functions that carries out these fastKM steps on the authors' website at <http://www4.stat.ncsu.edu/~sthollow/JYT/fastKM/>.

Our simulation studies suggest that the burden $G \times E$ test can have inflated type I errors and power loss when the genetic main effect is not appropriate modeled. Because the burden $G \times E$ test can also be performed under the KM framework, i.e., by using the burden kernel

for the genetic main effect, $K_G^{burden}(G_i, G_j) = \left(\sum_{m=1}^M w_m g_{im} \right) \left(\sum_{m=1}^M w_m g_{jm} \right)$, the corresponding $G \times E$ kernel test would also have inflated type I error rate even with original KM methods. The results show the importance of choosing an appropriate kernel to accommodate the genetic main effect when performing kernel $G \times E$ test.

We note that a fundamental presumption of fastKM is that the kernel matrix of the genetic main effect can be well approximated by a low rank matrix decomposition; therefore the null model can be fitted using augmented covariates by including the leading components of the low rank matrix decomposition. The low-rank structure of the kernel matrix may depend on the choice of kernels, sample size (n) and the number of variants that are jointly analyzed (M). In our simulation studies, we found that fastKM with the IBS kernel performed appropriately with varying M and n . When fastKM is coupled with a more complex kernel (e.g., higher order of polynomial kernels), it would have appropriate performance when n is large relative to M (e.g., $M = 100$ and $n = 5000$). However, cautions are needed when a complex kernel is applied with a moderate sample size relative to M (e.g., $M = 200$ and $n = 1000$). Based on these findings, we would recommend to use fastKM with IBS kernel. If a complex kernel is needed and the sample size n is moderate relative to M , one may consider to perform the original KM method with EM algorithm or penalization, which tends to be become less intractable with moderate sample sizes.

In this paper, we focused our investigation of fastKM on rare variants. It would be of interest to examine if the findings can be extended to common variants. In our limited investigation (described in the supplementary information), we considered a randomly selected 50-SNP region from HapMap3 CEU data and generated 1000 individuals. The results suggest fastKM with IBS kernel and polynomial kernel has similar performance to their EM counterpart. Nevertheless, further studies would be needed to fully understand the performance of fastKM when apply to common variant analyses.

The proposed low-rank KM framework has broad impact on KM modeling and beyond. It greatly enhances the computational efficiency of KM tests that contain multiple kernel components and involve high-dimensional nuisance parameters, e.g., the $G \times G$ kernel tests [Larson and Schaid, 2013] and the conditional kernel tests [Wang et al., 2015a,b]. It can be

generalized to study CNVs, for example to extend burden-based CNV tests [Raychaudhuri et al., 2010], which simultaneously model multiple CNV features, to the framework of KM tests (e.g., testing for a CNV dosage effect while adjusting for length and gene interruption status) [Tzeng et al., 2015]. In addition, it can also be extended to perform KM interaction tests for multivariate-phenotype analysis (e.g., Davenport et al. [2015]; Maity et al. [2012]). Lastly, because KM has a generalized linear mixed model (GLMM) representation, our low-rank framework can also benefit other GLMM-equivalent methods, for example the GLMM-based $G \times E$ test [Lin et al., 2013] and SimReg $G \times G$ and $G \times E$ tests [Tzeng et al., 2011; Wang et al., 2014a; Zhao et al., 2015].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by NIH grants 5T32GM081057 (to RM; PI Muse), R01 CA140632 (to WL), U01 HG005160 (to MMS, BBW, SRW, and FCH; PIs MMS and BBW), R01 NS34447 (to MMS and FCH; PI Toole), R01 MH084022 (to JYT), and P01 CA142538 (to WL, SH and JYT; PIs Kosorok, Davidian, George).

References

- Cai T, Tonini G, Lin X. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics*. 2011; 67(3):975–986. [PubMed: 21281275]
- Davenport C, Maity A, Sullivan P, Tzeng JY. A powerful test for snp effects on multivariate binary outcomes using kernel machine regression. *Genet Epidemiol*. 2015 Under Revision.
- Davies R. The distribution of a linear combination of chi-square random variables. *J Roy Stat Soc C-App*. 1980; 29(3):323–333.
- DeKou V, Gudnason V, Hawe E, Miller G, Stansbie D, Humphries S. Gene-environment and gene-gene interaction in the determination of plasma homocysteine levels in healthy middle-aged men. *Thromb Haemost*. 2001; 85(1):67–74. [PubMed: 11204591]
- Duchesne P, Lafaye De Micheaux P. Computing the distribution of quadratic forms: further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Comput Stat Data An*. 2010; 54(4):858–862.
- Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*. 2004; 20(1):93–99. [PubMed: 14693814]
- Harmon D, Shields D, Woodside J, McMaster D, Yarnell J, Young I, Peng K, Shane B, Evans A, Whitehead A. Methionine synthase D919G polymorphism is a significant but modest determinant of circulating homocysteine concentrations. *Genet Epidemiol*. 1999; 17(4):298–309. [PubMed: 10520212]
- Hsu FC, Sides E, Mychaleckyj J, Worrall B, Elias G, Liu Y, Chen WM, Coull B, Toole J, Rich S, et al. Transcobalamin 2 variant associated with poststroke homocysteine modifies recurrent stroke risk. *Neurology*. 2011; 77(16):1543–1550. [PubMed: 21975197]
- Hyndman M, Bridge P, Warnica J, Fick G, Parsons H. Effect of heterozygosity for the methionine synthase 2756 A → G mutation on the risk for recurrent cardiovascular events. *Am J Cardiol*. 2000; 86(10):1144–1146. [PubMed: 11074217]
- Kang HMM, Sul JHH, Service SK, Zaitlen NA, Kong SYY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*. 2010; 24:348–354. [PubMed: 20208533]
- Kimeldorf G, Wahba G. Some results on Tchebycheffian spline functions. *J Math Anal Appl*. 1971; 33(1):82–95.

- Kwee L, Liu D, Lin X, Ghosh D, Epstein M. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet.* 2008; 82(2):386–397. [PubMed: 18252219]
- Larson N, Schaid D. A kernel regression approach to gene-gene interaction detection for case-control studies. *Genet Epidemiol.* 2013; 37(7):695–703. [PubMed: 23868214]
- Leclerc D, Wilson A, Dumas R, Gafuik C, Song D, Watkins D, Heng H, Rommens J, Scherer S, Rosenblatt D, et al. Cloning and mapping of a cDNA for methionine synthase reductase, a flavoprotein defective in patients with homocystinuria. *Proc Natl Acad Sci USA.* 1998; 95(6):3059–3064. [PubMed: 9501215]
- Lievers K, Afman L, Kluijtmans L, Boers G, Verhoef P, den Heijer M, Trijbels F, Blom H. Polymorphisms in the transcobalamin gene: association with plasma homocysteine in healthy individuals and vascular disease patients. *Clin Chem.* 2002; 48(9):1383–1389. [PubMed: 12194912]
- Lin X. Variance component testing in generalised linear models with random effects. *Biometrika.* 1997; 84(2):309–326.
- Lin X, Cai T, Wu M, Zhou Q, Liu G, Christiani D, Lin X. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genet Epidemiol.* 2011; 35(7):620–631. [PubMed: 21818772]
- Lin X, Lee S, Christiani D, Lin X. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics.* 2013; 14(4):667–681. [PubMed: 23462021]
- Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics.* 2007; 63(4):1079–1088. [PubMed: 18078480]
- Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics.* 2008; 9(1):292. [PubMed: 18577223]
- Madsen B, Browning S. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5(2):e1000384. [PubMed: 19214210]
- Maity A, Sullivan P, Tzeng JY. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet Epidemiol.* 2012; 36(7):686–695. [PubMed: 22899176]
- Mellman I, Lin PF, Ruddell F, Rosenberg L. Genetic control of cobalamin binding in normal and mutant cells: assignment of the gene for 5-methyltetrahydrofolate: L-homocysteine S-methyltransferase to human chromosome 1. *Proc Natl Acad Sci USA.* 1979; 76(1):405–409. [PubMed: 284356]
- Nygård O, Vollset S, Refsum H, Stensvold I, Tverdal A, Nordrehaug J, Ueland P, Kvåle G. Total plasma homocysteine and cardiovascular risk profile: the Hordaland Homocysteine Study. *JAMA.* 1995; 274(19):1526–1533. [PubMed: 7474221]
- Pang H, Kim I, Zhao H. Random effects model for multiple pathway analysis with applications to type II diabetes microarray data. *Stat Biosci.* 2014
- Pongpanich M, Neely M, Tzeng JY. On the aggregation of multimarker information for marker-set and sequencing data analysis: genotype collapsing vs. similarity collapsing. *Front Genet.* 2012; 2:110. [PubMed: 22303404]
- Price A, Kryukov G, de Bakker P, Purcell S, Staples J, Wei LJ, Sunyaev S. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 2010; 86(6):832–838. [PubMed: 20471002]
- Raychaudhuri S, Korn J, McCarroll S, International Schizophrenia Consortium. Altshuler D, Sklar P, Purcell S, Daly M. Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet.* 2010; 6(9):e1001097. [PubMed: 20838587]
- Rosenblatt D, Schmutz S, Cooper B, Zaleski W, Casey R. Prenatal vitamin B12 therapy of a fetus with methylcobalamin deficiency (cobalamin E disease). *Lancet.* 1985; 325(8438):1127–1129. [PubMed: 2860337]
- Schaffner S, Foo C, Gabriel S, Reich D, Daly M, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 2005; 15(11):1576–1583. [PubMed: 16251467]

- Schölkopf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* 1998; 10(5):1299–1319.
- Schuh S, Rosenblatt D, Cooper B, Schroeder ML, Bishop A, Seargeant L, Haworth J. Homocystinuria and megaloblastic anemia responsive to vitamin B12 therapy: an inborn error of metabolism due to a defect in cobalamin metabolism. *New Engl J Med.* 1984; 310(11):686–690. [PubMed: 6700644]
- Seetharam B, Bose S, Li N. Cellular import of cobalamin (vitamin B-12). *J Nutr.* 1999; 129(10):1761–1764. [PubMed: 10498744]
- Spence J, Cordy P, Kortas C, Freeman D. Effect of usual doses of folate supplementation on elevated plasma homocyst(e)ine in hemodialysis patients: no difference between 1 and 5 mg daily. *Am J Nephrol.* 1999; 19(3):405–410. [PubMed: 10393379]
- Stanisawska-Sachadyn A, Woodside JV, Sayers C, Yarnell J, Young I, Evans A, Mitchell L, Whitehead A. The transcobalamin (TCN2) 776C>G polymorphism affects homocysteine concentrations among subjects with low vitamin B(12) status. *Eur J Clin Nutr.* 2010; 64(11):1338–1343. [PubMed: 20808328]
- Toole J, Malinow MR, Chambless L, Spence J, Pettigrew L, Howard V, Sides E, Wang CH, Stampfer M. Lowering homocysteine in patients with ischemic stroke to prevent recurrent stroke, myocardial infarction, and death: the Vitamin Intervention for Stroke Prevention (VISP) randomized controlled trial. *J Am Med Assoc.* 2004; 291(5):565–575.
- Tsai M, Bignell M, Yang F, Welge B, Graham K, Hanson N. Polygenic influence on plasma homocysteine: association of two prevalent mutations, the 844ins68 of cystathionine β -synthase and A2756 G of methionine synthase, with lowered plasma homocysteine levels. *Atherosclerosis.* 2000; 149(1):131–137. [PubMed: 10704624]
- Tzeng JY, Zhang D, Chang SM, Thomas D, Davidian M. Gene-trait similarity regression for multimarker-based association analysis. *Biometrics.* 2009; 65(3):822–832. [PubMed: 19210740]
- Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy M, Sale M, Worrall B, Hsu FC, Thomas D, Sullivan P. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *Am J Hum Genet.* 2011; 89(2):277–288. [PubMed: 21835306]
- Tzeng JY, Lu W, Hsu FC. Gene-level pharmacogenetic analysis on survival outcomes using gene-trait similarity regression. *Ann Appl Stat.* 2014; 8(2):1232–1255. [PubMed: 25018788]
- Tzeng JY, Magnusson PKE, Sullivan PF, The Swedish Schizophrenia Consortium. Szatkiewicz J. A new method for detecting associations with rare copy-number variants. Submitted.
- van der Griend R, Biesma D, Banga JD. Postmethionine-load homocysteine determination for the diagnosis hyperhomocysteinaemia and efficacy of homocysteine lowering treatment regimens. *Vasc Med.* 2002; 7(1):29–33. [PubMed: 12083731]
- Voorman A, Lumley T, McKnight B, Rice K. Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PloS ONE.* 2011; 6(5):e19416. [PubMed: 21589913]
- Wang J, Huff A, Spence J, Hegele R. Single nucleotide polymorphism in CTH associated with variation in plasma homocysteine concentration. *Clin Genet.* 2004; 65(6):483–486. [PubMed: 15151507]
- Wang X, Duarte N, Cai H, Adachi T, Sim A, Cranney G, Wilcken D. Relationship between total plasma homocysteine, polymorphisms of homocysteine metabolism related enzymes, risk factors and coronary artery disease in the Australian hospital-based population. *Atherosclerosis.* 1999; 146(1):133–140. [PubMed: 10487496]
- Wang X, Epstein M, Tzeng JY. Analysis of gene-gene interactions using gene-trait similarity regression. *Hum Hered.* 2014; 78(1):17–26. [PubMed: 24969398]
- Wang X, Zhang D, Tzeng JY. Pathway-guided identification of gene-gene interactions. *Ann Hum Genet.* 2014; 78(6):478–491. [PubMed: 25227508]
- Wang Z, Maity A, Luo Y, Neely ML, Tzeng JY. Complete effect-profile assessment in association studies with multiple genetic and multiple environmental factors. *Genet Epidemiol.* 2015; 39:122–133. [PubMed: 25538034]
- Wang Z, Maity A, Hsiao CK, Voora D, Kaddurah-Daouk R, Tzeng JY. Module-based association analysis for omics data with network structure. *PLoS One.* 2015; 10(3):e0122309. [PubMed: 25822417]

- Wu M, Kraft P, Epstein M, Taylor D, Chanock S, Hunter D, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet.* 2010; 86(6):929–942. [PubMed: 20560208]
- Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011; 89(1):82–93. [PubMed: 21737059]
- Wu M, Maity A, Lee S, Simmons E, Harmon Q, Lin X, Engel S, Molldrem J, Armistead P. Kernel machine SNP-set testing under multiple candidate kernels. *Genet Epidemiol.* 2013; 37(3):267–275. [PubMed: 23471868]
- Zavadáková P, Fowler B, Suormala T, Novotna Z, Mueller P, Hennermann J, Zeman J, Vilaseca M, Vilarinho L, Gutsche S, et al. cblE type of homocystinuria due to methionine synthase reductase deficiency: functional correction by minigene expression. *Hum Mutat.* 2005; 25(3):239–247. [PubMed: 15714522]
- Zhang D, Lin X. Hypothesis testing in semiparametric additive mixed models. *Biostatistics.* 2003; 4(1):57–74. [PubMed: 12925330]
- Zhao G, Marceau R, Zhang D, Tzeng JY. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. *Genetics.* 2015; 199:695–710. [PubMed: 25585620]

Quantitative Trait Type I Error

Figure 1.

Type I error for fastKM, originalKM, and the weighted counting burden-based G×E test for quantitative traits with $M = 100$ loci, $n = 5000$ individuals, and varying main effect parameter γ_G . Models with a continuous environmental E covariate are on the left and those with a binary environmental E covariate are on the right. The KM tests are based on the IBS kernel.

Figure 2.

Power for fastKM, originalKM, and the weighted counting burden-based G×E test for quantitative traits with $M = 100$ loci and $n = 5000$ individuals over varying interaction effect sizes γ_{GE} . The left panel shows the results of no genetic main effect (i.e., $\gamma_G = 0$) and the right panels shows the results of nonzero main effect (i.e., $\gamma_G = 1$). For each plot, continuous E covariates are on the left and binary E covariates are on the right. The KM tests are based on the IBS kernel.

Figure 3.

Type I error for fastKM, originalKM, and the weighted counting burden-based $G \times E$ test for quantitative traits with continuous environmental E covariate and varying main effect parameter γ_G . The left panel shows the results of $M = 100$ loci and $n = 5000$ individuals. The right panel shows the results of $M = 200$ loci and $n = 1000$ individuals. The KM tests are based on the IBS kernel and the polynomial kernels with $d = 2$ and $d = 3$.

Figure 4.

Power for fastKM, originalKM, and the weighted counting burden-based $G \times E$ test for quantitative traits with continuous E covariate over varying main effect size γ_G ($\gamma_G = 0$ for zero main effect, and $\gamma_G = 1$ for nonzero main effect) and interaction effect size γ_{GE} . The left panel shows the results of $M = 100$ loci and $n = 5000$ individuals. The right panel shows the results of $M = 200$ loci and $n = 1000$ individuals. The KM tests are based on the IBS kernel and the polynomial kernels with $d = 2$ and $d = 3$.

Binary Trait Type I Error

Figure 5.

Type I error for fastKM and the weighted counting burden-based G×E test for binary traits for $M = 100$ loci, $n = 5000$ individuals, and varying main effect parameter γ_G . Models where E is generated from a Gaussian distribution are displayed on the left, and those where E is from a Bernoulli distribution are on the right. The KM tests are based on the IBS kernel.

Survival Trait Type I Error

Figure 6.

Type I error for survival traits for fastKM and the weighted counting burden-based G×E test with $M = 100$ loci, $n = 5000$ individuals, $c=15\%$ and 40% censoring proportions, and varying main effect parameter γ_G . Models where E is generated from a Gaussian distribution are displayed on the left, and those where E is from a Bernoulli distribution are on the right. The KM tests are based on the IBS kernel.

Figure 7.

Power for fastKM and the weighted counting burden-based $G \times E$ test for binary traits with $M = 100$ loci and $n = 5000$ individuals over varying interaction effect sizes γ_{GE} . The left panel shows the results of no genetic main effect (i.e., $\gamma_G = 0$) and the right panels shows the results of nonzero main effect (i.e., $\gamma_G = 1$). For each plot, continuous E covariates are on the left and binary E covariates are on the right. The KM tests are based on the IBS kernel.

Figure 8.

Power for fastKM and the weighted counting burden-based $G \times E$ test for survival traits with $M = 100$ loci and $n = 5000$ individuals for varying interaction parameter γ_{GE} over two censoring proportions ($c=15\%$ and 40%). The left panel shows the results of no genetic main effect (i.e., $\gamma_G = 0$) and the right panels shows the results of nonzero main effect (i.e., $\gamma_G = 1$). For each plot, continuous E covariates are on the left and binary E covariates are on the right. The KM tests are based on the IBS kernel.

Table I

Average run time in minutes (and corresponding standard error) for quantitative traits when the G×E effect is zero with a sample size of $n = 5000$ individuals. Results are based on the IBS kernel.

Covariate	Main Effect	FastKM	OriginalKM	Burden
Continuous	0	2.30 (0.011)	31.2 (0.348)	0.0037 (0.001)
	1	2.31 (0.009)	7.9 (0.023)	0.0022 (5e-04)
	1.5	2.38 (0.018)	7.8 (0.023)	0.0028 (7e-04)
Binary	0	2.10 (0.006)	31.5 (0.354)	0.0013 (2e-05)
	1	2.09 (0.005)	7.5 (0.016)	0.0013 (2e-05)
	1.5	2.14 (0.006)	7.4 (0.016)	0.0014 (6e-05)

Average run time in minutes (and corresponding standard error) for quantitative traits when the G×E effect is nonzero with a sample size of $n = 5000$ individuals. Results are based on the IBS kernel.

Table II

Covariate	Main Effect	Interaction	FastKM	OriginalKM	Burden
Continuous	0	0.1	2.32 (0.012)	29.6 (0.476)	0.0021 (5e-05)
		0.2	2.33 (0.011)	29.1 (0.479)	0.0022 (5e-05)
	1	0.1	2.62 (0.014)	7.8 (0.046)	0.0015 (1e-05)
		0.2	2.52 (0.014)	7.8 (0.046)	0.0015 (1e-05)
Binary	0	0.1	2.42 (0.011)	26.5 (0.488)	0.0021 (5e-05)
		0.2	2.45 (0.010)	17.1 (0.386)	0.0021 (5e-05)
	1	0.1	2.62 (0.013)	7.5 (0.046)	0.0015 (1e-05)
		0.2	2.63 (0.013)	8.1 (0.049)	0.0015 (1e-05)

P-values for the fastKM analyses of VISP study data, including (a) testing gene $\times$ age interaction on post-methionine change in total Hcy, treating change as continuous, with a IBS kernel or polynomial kernel (d=2); (b) testing gene $\times$ age interaction on post-methionine change in total Hcy, treating change as binary using the 90th sample percentile as a cut off; (c) testing gene $\times$ intervention interaction on time to recurrent stroke.

Table III

Gene Name		BHMT	BHMT2	CBS	CTH	MTHFR	MTR	MTRR	TCN1	TCN2
Trait	Method	5	3	6	10	7	20	5	3	15
Hcy change (Quantitative)	originalKM, IBS	0.9997	0.7762	0.8572	0.7882	0.2932	0.0009*	0.3904	0.3970	0.8396
	originalKM, d2	0.9089	0.9117	0.5616	0.4553	0.2410	0.0008*	0.5364	0.2447	0.8739
	fastKM, IBS	0.9998	0.7953	0.8373	0.8046	0.3163	0.0008*	0.4000	0.3940	0.8504
	fastKM, d2	0.8856	0.9351	0.5570	0.4700	0.2322	0.0007*	0.5648	0.2569	0.8673
Hcy change (Binary)	fastKM, IBS	0.9675	0.5212	0.3893	0.9362	0.4229	0.0007*	0.3617	0.2703	0.2967
Time to stroke (Survival)	fastKM, IBS	0.8436	0.6934	0.9106	0.0176*	0.2230	0.4918	0.0545	0.9973	0.0430*

* indicates p-values which are < 0.05.