Published in final edited form as:

Genetica. 2012 December ; 140(10-12): 421–427. doi:10.1007/s10709-012-9690-5.

Power of a reproducing kernel-based method for testing the joint effect of a set of single-nucleotide polymorphisms

Hong He,

Department of Epidemiology and Biostatistics, Norman J Arnold School of Public Health,
University of South Carolina, Columbia, SC 29201, USA, heh@email.sc.edu

Hongmei Zhang,

Department of Epidemiology and Biostatistics, Norman J Arnold School of Public Health,
University of South Carolina, Columbia, SC, 29201, USA, hzhang@mailbox.sc.edu

Arnab Maity,

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA,
amaity@ncsu.edu

Yubo Zou,

Department of Epidemiology and Biostatistics, Norman J Arnold School of Public Health,
University of South Carolina, Columbia, SC 29201, USA, zou@mailbox.sc.edu

James Hussey, and

Department of Epidemiology and Biostatistics, Norman J Arnold School of Public Health,
University of South Carolina, Columbia, SC, 29201, USA, jhussey@mailbox.sc.edu

Wilfried Karmaus

Department of Epidemiology and Biostatistics, Norman J Arnold School of Public Health,
University of South Carolina, Columbia, SC, 29201, USA, karmaus@mailbox.sc.edu

Abstract

This study explored a semi-parametric method built upon reproducing kernels for estimating and testing the joint effect of a set of single nucleotide polymorphisms (SNPs). The kernel adopted is the identity-by-state (IBS) kernel that measures SNP similarity between subjects. In this article, through simulations we first assessed its statistical power under different situations. It was found that in addition to the effect of sample size, the testing power was impacted by the strength of association between SNPs and the outcome of interest, and by the SNP similarity among the subjects. A quadratic relationship between SNP similarity and testing power was identified, and this relationship was further affected by sample sizes. Next we applied the method to a SNP-lung function data set to estimate and test the joint effect of a set of SNPs on forced vital capacity, one type of lung function measure. The findings were then connected to the patterns observed in simulation studies and further explored via variable importance indices of each SNP inferred from a variable selection procedure.

Keywords

Reproducing kernels; SNP; Mixed linear models; Testing power; Variable selection

1 Introduction

The past decade has seen a surge of interest in disease-SNP and disease-haplotype association studies because of the availability of densely situated single nucleotide polymorphisms (SNPs) throughout the genome [Guo and Lin, 2009].

The widely utilized statistical methods to analyze SNP effects on health outcomes are parametric regression models such as linear regression, logistic regression, or analysis of variance (ANOVA). Sadeghnejad et al. [2009] used a linear regression model and detected significant effects of 6 SNPs that were associated with Chronic Obstructive Pulmonary Disease (COPD) on pulmonary function. Hansel et al. [2009] identified significant associations (p -value < 0.05) between lung function decline and 21 SNPs also using linear regression models. In a lung cancer case-control study, Chin et al. [2008] used a logistic regression model to infer the odds ratio and confidence interval for the KRAS LCS6 SNP effect in moderate and heavy smokers with adjustment for selected covariates. Yeager et al. [2007] demonstrated an association of prostate cancer with SNP rs6983267 in the centromeric locus via logistic regression as well. Most studies like the above examples drew their conclusions based on evaluations of one SNP effect at a time.

However, SNPs do not necessarily function individually, rather, they work in concert with other SNPs to manifest a disease condition. There are some limitations if we simply include multiple SNPs in one traditional regression model [Gianola and van Kaam, 2008, Gonzalez-Recio et al., 2008]. These models are additive models that oversimplify the biological relationship between different SNPs and can potentially generate misleading conclusions. Another problem is that the number of SNPs plus the interaction between SNPs can far exceed the number of observations in the data.

There are several possible solutions to deal with the two limitations, such as spline regression [Eilers and Marx, 1996, Cook et al., 2004, Lin et al., 2008], the random forest method [Jiang et al., 2009, Breiman, 2001, Huebner et al., 2008], and the reproducing kernel-based (RKB) approach [Wu et al., 2010, Christiani et al., 2010]. The method of spline regression has the ability to model any underlying nonadditive effects. However, it cannot overcome the identifiability problem resulting from a large number of covariates and is not appropriate for discrete variables like SNPs. The random forest approach can address both problems noted earlier, but it is not able to determine the strength of association between SNPs and the outcome. The RKB approach, on the other hand, not only addresses the limitations but also provides inferences on the strength of associations. However, it is unknown how well the method performs in terms of statistical power on the testing of associations, especially with regard to SNP similarity among subjects. In this work, we use the identity-by-state (IBS) kernel in the RKB method to evaluate the association of SNPs with a continuous outcome of interest.

The road map of the remainder of the article is as follows. Section 2 discusses the RKB method and related definitions. In section 3 we conduct simulation studies under several scenarios to examine the strength of the method and its limitations. In section 4 this method is applied to analyze the association between lung function and 12 SNPs based on data of a cohort study. Section 5 summarizes the main contributions of the article.

2 Methods

2.1 The model

The RKB method is adapted from Liu et al. [2007], where a Gaussian kernel is applied to examine the relationship between prostate-specific antigen and genes in a genetic pathway.

To describe the relationship between a continuous outcome of interest $\mathbf{Y}$ and a set of SNPs, we consider a semi-parametric linear regression model,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\varepsilon}. \quad (1)$$

The matrix $\mathbf{X}$ ($n \times p$) is the design matrix composed of covariates of interest, where n is the number of subjects and $p - 1$ is the number of covariates, $\boldsymbol{\beta}$ ($p \times 1$) is a vector composed of the intercept and regression coefficients, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, and $\mathbf{h}$ ($n \times 1$) is defined as $\mathbf{h} = \mathbf{K}\boldsymbol{\alpha}$, a linear combination of reproducing kernels measuring the joint effect of a set of SNPs. Here $\boldsymbol{\alpha}$ is a vector of unknown parameters, and $\mathbf{K}$ ($n \times n$) is a positive definite matrix representing SNP similarity between every pair of subjects. Each entry of $\mathbf{K}$ is determined by a pre-specified kernel function discussed in the next section. The kernel matrix $\mathbf{K}$ or simply the $\mathbf{K}$ matrix acts as a correlation matrix [Liu et al., 2007].

It has been shown that model (1) is equivalent to a liner mixed model by assuming $\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{\tau}\mathbf{K}^{-1})$ [Liu et al., 2007]. With this assumption, it is straightforward that $\mathbf{h} \sim N(\mathbf{0}, \boldsymbol{\tau}\mathbf{K})$. In this case, $\mathbf{h}$ is a vector of random effects and parameter $\boldsymbol{\tau}$ is a regularization parameter measuring the strength of joint SNP effects.

2.2 Definition of the kernel matrix $\mathbf{K}$

There are a number of strategies for characterizing the similarity of individuals with respect to their genetic variations. The identity-by-state (IBS) kernel is one of those many examples [Wesse and Schork, 2006] and will be used in our study due to its specificity to SNP data. The IBS kernel assesses the genomic similarity between two individuals on the basis of genotype data. The (i, j) entry of the IBS kernel matrix, $S_{i,j}$, for individuals i and j ($i, j = 1, \dots, n$) is $S_{i,j}^{IBS} = \frac{\sum_{l=1}^L S_{i,j}^l(g_i^l, g_j^l)}{2L}$, where L is the number of SNP loci considered in the

calculation; g_i^l and g_j^l are the genotype of individuals i and j , respectively, at the l th locus ($l = 1, \dots, L$); and $S_{i,j}^l(g_i^l, g_j^l)$ is a function mapping the genotype information, for individuals i and j at locus l , to a particular numeric value and, for our purposes, has a value of 0 (lowest similarity level) if at locus l individuals i and j are homozygous for different SNP alleles (e.g., $g_i^l = AA$ and $g_j^l = TT$), a value of 1 if they share one allele (e.g., $g_i^l = AA$ and $g_j^l = AT$), and a value of 2 (highest similarity level) if they share both alleles (e.g., $g_i^l = AA$ and $g_j^l = AA$). The similarity is evaluated for all pairs of n individuals in a sample to formulate an $n \times n$ kernel matrix.

2.3 The method to infer $\boldsymbol{\tau}$

When we evaluate the effect of a set of SNPs, the first task is to estimate $\boldsymbol{\tau}$. The parameter $\boldsymbol{\tau}$ and the residual variance σ^2 can be treated as variance components and estimated simultaneously using restricted maximum likelihood (REML).

The restricted log-likelihood function under the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\varepsilon}$ can be written as

$$l_r(\sigma^2, \boldsymbol{\tau}) = -\frac{1}{2} \log |\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2} \log |\mathbf{X}^T \mathbf{V}^{-1}(\boldsymbol{\theta}) \mathbf{X}| - \frac{1}{2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1}(\boldsymbol{\theta}) (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \quad (2)$$

where $\mathbf{V}(\boldsymbol{\theta}) = \sigma^2 \mathbf{I} + \boldsymbol{\tau}\mathbf{K}$. The score equations of $(\boldsymbol{\tau}, \sigma^2)$ are

$$-\frac{1}{2} \text{tr}(\mathbf{K}\mathbf{P}) + \frac{1}{2} (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T \mathbf{V}^{-1} \mathbf{K} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = 0, \quad (3)$$

$$-\frac{1}{2}\text{tr}(\mathbf{P}) + \frac{1}{2}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T \mathbf{V}^{-1} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0, \quad (4)$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1}$ [Liu et al., 2007].

To evaluate whether the effect of a set of SNPs is significant, we simply test $H_0: \tau = 0$. Liu et al. [2007] proposed a score test for testing $H_0: \tau = 0$. From (3), it can be shown that the score statistic of τ under $H_0: \tau = 0$ is $Q_T(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) - \text{tr}(\mathbf{P}_0 \mathbf{K})$, where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are the maximum likelihood estimators (MLEs) of $\boldsymbol{\beta}$ and σ^2 under the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. The

projection matrix $\mathbf{P}_0$ is $\mathbf{P}_0 = \mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, and $Q_T(\boldsymbol{\beta}, \sigma^2) = \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{K} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Q_T has an approximate scaled Chi-square distribution κX_v^2 under H_0 . The scale parameter, $\kappa = \tilde{I}_{\tau\tau}/2\tilde{\epsilon}$, and the degrees of freedom, $\tilde{v} = 2\tilde{\epsilon}^2/\tilde{I}_{\tau\tau}$, where $\tilde{I}_{\tau\tau} = I_{\tau\tau} - I_{\tau\sigma^2} I_{\sigma^2\sigma^2}^{-1} I_{\sigma^2\tau}^T$, $I_{\tau\tau} = \text{tr}(\mathbf{P}_0 \mathbf{K})^2/2$, $I_{\tau\sigma^2} = \text{tr}(\mathbf{P}_0 \mathbf{K} \mathbf{P}_0)/2$, and $I_{\sigma^2\sigma^2} = \text{tr}(\mathbf{P}_0^2)/2$. $\tilde{\epsilon} = \text{tr}(\mathbf{P}_0 \mathbf{K})/2$. Computation of the score test is quite simple, because one only needs to fit the simple linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$.

3 Power-based evaluations of the methods via simulations

We evaluate the RKB method via an evaluation of testing power based on the following scenarios:

1. The effect of sample size. We expect to see a positive impact of sample size on testing power.
2. The effect of the sparsity of the $\mathbf{K}$ matrix. Since each entry in the kernel matrix provides information on the similarity of SNPs between every two subjects, the amount of such information is expected to affect the testing power, although how it behaves exactly is unclear and deserves analytical proofs.
3. The effect of different levels of similarity of SNPs between subjects. It is relatively straightforward that low similarity of SNPs between subjects is likely to result in a low testing power. However, it is unclear how the testing power changes if the similarity among subjects increases and at the same time the sample size also increases.

These scenarios will be evaluated using simulated data. Our simulations start from the generation of a $\mathbf{K}$ matrix. In this section, most often a $\mathbf{K}$ matrix is constructed by scaling a covariance matrix, generated from a Wishart distribution, to its corresponding correlation matrix. To mimic the property of similarity between subjects, we take the absolute values of the correlations and at the same time we ensure the $\mathbf{K}$ matrix is positive definite. At other times, the $\mathbf{K}$ matrix was formed by extending the identity matrix. The data are then simulated assuming $\boldsymbol{\beta} = 1$ and $\sigma^2 = 1$ based on (1). One thousand simulated data sets are generated for each combination of τ , $\mathbf{K}$, and n to facilitate power estimation. In all the simulations, the significance level is set at $\alpha = 0.05$. All simulation studies presented below are programmed in R.

3.1 Power changes with the sample size and τ

We select sample sizes of 50, 100, 150 and 200, and consider the value of τ being 0, .01, .05, 1.0, 1.5, and 2.0. Here, the matrix $\mathbf{K}$ is fixed indicating a fixed population with certain SNP characteristics.

We observed that the testing power increases in a non-linear pattern with an increase in the sample size and with the increase of τ (Figure 1). As expected, when $\tau = 0$ all the powers are very small and close to the 0.05 type I error rate. When $\tau = 1.5$ or $\tau = 2.0$ and sample size is 200, the power exceeds 0.8. With $\mathbf{K}$ matrix fixed and τ varying, the simulated data sets mimic data generated from different association studies in one population. The results above indicated that the method has the ability to test the SNP effects on disease occurrence in a population, as long as the association exists and we have sufficient information in the data (i.e., large enough sample).

If we do not fix the $\mathbf{K}$ matrix, but the parameters used to generate the $\mathbf{K}$ matrix are kept the same, then we mimic populations with similar features (such as race, etc). In this case, similar patterns of power changes are observed compared to the results with $\mathbf{K}$ fixed.

3.2 Power changes with the sparsity of $\mathbf{K}$ matrix

Extreme case If all the individuals are different in terms of SNP genotypes, the $\mathbf{K}$ matrix becomes an identity matrix (a diagonal matrix with 1's on the diagonal). Such an extreme case can be approximately achieved when the number of subjects is small while the number of SNPs is large. When $\mathbf{K}$ is an identity matrix, then

$I_{\tau\tau} = \text{tr}(\mathbf{P}_0 \mathbf{K})^2 / 2 = \text{tr}(\mathbf{P}_0^2) / 2$, $I_{\tau\sigma^2} = \text{tr}(\mathbf{P}_0 \mathbf{K} \mathbf{P}_0) / 2 = \text{tr}(\mathbf{P}_0^2) / 2$, and $I_{\sigma^2\sigma^2} = \text{tr}(\mathbf{P}_0^2) / 2$. $\tilde{e} = \text{tr}(\mathbf{P}_0 \mathbf{K}) / 2 = \text{tr}(\mathbf{P}_0) / 2$. So $\tilde{I}_{\tau\tau} = I_{\tau\tau} - I_{\tau\sigma^2} I_{\sigma^2\sigma^2}^{-1} I_{\tau\sigma^2}^T = 0$, and the degree of freedom $\tilde{\nu} = 2\tilde{e}^2 / \tilde{I}_{\tau\tau}$ goes to infinity. That means the Chi-square distribution is very flat, which results in critical values (i.e., the 95th percentile) approaching infinity. This implies that τ is not identifiable if all individuals' SNPs are completely different. This conclusion can also be drawn from our model assumption discussed in section 2.1 by taking $\mathbf{K}$ as an identity matrix. Therefore, this method will completely lose its power if all subjects have different SNP genotypes.

Sparse $\mathbf{K}$ matrices Now we know that if all the subjects' SNP genotypes are completely different ($\mathbf{K}$ is a diagonal matrix), then this is an indication of no SNP effect or SNP effects are not identifiable. What if we only have a few people with similar SNPs? That means only a few non-zero values in $(0, 1]$ appear in the off-diagonal of the $\mathbf{K}$ matrix, and other off-diagonal elements are zeros. To generate $\mathbf{K}$ matrices satisfying this criterion, we randomly select entries in the upper off-diagonal of the identity matrix and fill them with randomly generated numbers between 0 and 1 at these positions. The lower off-diagonal entries then mirror their counterparts. To simulate data representing different levels of sparsity, we choose different numbers of non-zero entries determined by the size of $\mathbf{K}$ matrices with each number being relatively small compared to the size of $\mathbf{K}$. For example, with a sample size of 150 at sparsity level 1 (Figure 2), we choose 150 non-zero off-diagonal entries with 75 in the upper triangle of the matrix and the same 75 in the lower triangle of the matrix. This is 0.67% of all the elements in $\mathbf{K}$ excluding diagonal elements.

Testing powers change along with the change of sparsity levels in the $\mathbf{K}$ matrix (Figure 2). The testing power increases with the increase of τ for all different sparsity levels of $\mathbf{K}$. At the lowest sparsity level (*SL1*), the testing power increases quickly from .07 to .38 as the τ increases from .1 to 1.5. This implies for the same sparsity level of $\mathbf{K}$ the power will increase quickly when the SNP effect becomes stronger. From Figure 2 we also can find that for the same τ the power increases with an increase in the number of non-zero values in the $\mathbf{K}$ matrix. For $\tau = 1.5$ the testing power changes from .11 to .38 as the sparsity level decreased from *SL4* to *SL1*. This implies that the testing power will increase quickly even if a small portion of subjects share some SNPs.

3.3 Power changes with the SNP similarity level among subjects

How does the testing power change if the similarity becomes stronger? An increase of similarity can be the result of an inbreeding over time in a population of interest.

To simulate different levels of similarity, for each $\mathbf{K}$ matrix, we limit the range of correlations at different levels. Specifically, we consider the following seven ranges: [0, .02], [.02, .05], [.05, .1], [.2, .3], [.5, .6], [.7, .8], and [.9, 1]. The medians of the ranges are used to label the seven similarity levels. Sample sizes 50, 100, 150, and 200 are selected. The value of τ is fixed at 2.0.

The results are shown in Figure 3. Two points are noteworthy. First, the association of testing power with similarity level is not monotonic; the testing power increases initially, then decreases with the increase of similarity level regardless of the sample size. Weak or strong similarity among subjects will cause loss of power in detecting SNP effect. This is intuitively understandable; SNPs of all subjects being the same or completely different make it inconclusive about the effect of a set of SNPs on a disease. Second, although more detailed simulations may be needed, the optimum similarity for the highest power is related to the sample size; a lower similarity level is favored for larger sample sizes while higher similarity levels are needed for smaller sample sizes. At each similarity level the testing power increases with the increase of sample size, which is consistent with earlier results and is as expected. The pattern shown in Figure 3 seems informative for researchers to decide in which level of similarity and with what sample size this method may provide satisfactory power.

4 Real data application

4.1 Testing the SNP set effect

In the previous sections, we examined the features of the method through simulations. In this section, we apply this method to a real data set to test the effect of a set of SNPs on a continuous outcome, and discuss the findings.

The chosen data were taken from the Isle of Wight (IOW) birth cohort study [Arshad and Hide, 1992]. For our study, the lung function marker forced vital capacity (FVC) is of particular interest due to its potential connection to asthma and other lung diseases. Since *CHIA* (GeneID-27159) is a potential asthma candidate gene [Chatterjee et al., 2008], 12 SNPs in *CHIA* gene were chosen to evaluate their overall effect on FVC. The rs numbers of the 12 SNPs are rs61756687, rs7411387, rs4240530, rs3818822, rs3806446, rs2820072, rs2275254, rs2275253, rs2256721, rs12036618, rs10776724, and rs10494132.

Data from 657 subjects were used to examine the effect of the 12 SNPs. An un-weighted IBS kernel matrix was calculated to examine SNP similarity between subjects. Among the off-diagonal elements, 95% of them are between 0.4167 and 0.9583. The median is 0.7083. These statistics suggest relatively high similarity of SNPs among the 657 subjects.

Based on a grid search applied to the restricted log-likelihood function (2) with $\mathbf{K}$ being a weighted IBS kernel matrix and weights determined by minor allele frequencies, the MLE of τ is 0.00031, and that of σ^2 is 0.87. The results are consistent with the pattern shown in the contour plot (Figure 4), and further confirmed by the score test described in section 2 (p-value is 0.41). This implies that the set of these 12 SNPs is possibly not related to lung function.

4.2 Selecting informative SNPs

Our simulation studies showed that the testing power depends on sample size, the sparsity of the $\mathbf{K}$ matrix, the strength of SNP effect, i.e., the value of τ , and the similarity of SNPs between subjects. It is possible that there is no relationship between these SNPs and the lung function. However, recall that the calculated IBS kernel matrix indicates that many subjects have very similar SNPs. As shown in Figure 3, when the similarity level is high, the testing power drops quickly with the increase of similarity levels even though the effect of some SNPs is strong. When the SNP effect is small, as indicated in Figure 1, the power would be even worse than these shown in Figure 3.

Besides the similarity levels, another possibility for the estimate and test results of τ is that there exist redundant SNPs in the SNP set and those SNPs bring in extraordinary error to the evaluation of the whole SNP set effect. This motivated us to utilize a variable selection strategy to filter out unimportant SNPs. We used a marker selection algorithm built upon reproducing kernels for SNP-SNP interactions to exclude unimportant SNPs [Maity and Lin, 2011, Maity et al., 2012]. The algorithm identifies SNPs based on their association with the outcome and provides an importance index ranged from 0 to 1 for each SNP. The importance index is calculated based on random permutations and is an estimate of the probability that the SNP should be kept in the model [Maity et al., 2012]. The higher the index, the more important the SNP is. We used 200 random permutations. Among the 12 SNPs, there is one SNP (rs10776724) with importance index 1 and additional 4 SNPs (rs4240530, rs3818822, rs3806446, and rs2256721) with importance indices higher than 1/12 (Figure 5), which is the probability of selecting a SNP completely at random. We further tested the effect of SNP rs10776724. With SNP rs10776724, the estimate of τ is 0.0088 with p-value being 0.032, much improved compared to the result using the whole set of 12 SNPs (Table 1). When adding the additional 4 SNPs, the estimate of τ is reduced to 0.0056 with p-value increased to 0.16, which is still reasonably informative to geneticists (Table 1). Combined with the results from simulation studies and our assessment on similarities between subjects on SNPs, this finding indicates that the lack of statistical significance of the SNP set effect is likely due to the high similarity in SNPs between subjects and the existence of redundant SNPs.

5 Conclusion

A semi-parametric method built upon reproducing kernels is used to test the SNP effects on a continuous outcome. Through simulation studies, we examined the performance of the method under varying scenarios.

Our results demonstrate that the method has the ability to test the SNP effects on health outcomes, as long as the association exists and we have sufficient information in the data. Furthermore, if all the individuals are completely different, the method cannot detect any SNP effect. However, as long as we have a few people with similar SNPs, the testing power increases quickly along with the decrease of sparsity. Finally, the testing power is quadratically related to the similarity level of SNPs among subjects.

The non-effective conclusion in the real data application is likely due to the high similarity of SNPs among the subjects and redundant SNPs. This is supported by the simulation results and SNP selection results based on a semi-parametric variable selection approach.

The findings from simulation studies deserve future analytical works. This includes studying the association of testing power with sparsity level and that with SNP similarity among subjects. It will be beneficial to estimate the optimal sample sizes for different similarity levels to achieve desired power, which is one part of our on-going work.

Acknowledgments

This work was supported by the National Institute of Allergy and Infectious Diseases (H He, H Zhang, and W Karmaus, grant number R01 AI091905) and by the National Institute of Environmental Health Sciences (A Maity, grant number R00 ES017744).

References

- Arshad SH, Hide DW. Effect of environmental factors on the development of allergic disorders in infancy. *Journal of Allergy and Clinical Immunology*. 1992; 90:235–241. [PubMed: 1500628]
- Breiman L. Random forests. *Machine Learning*. 2001; 45:5–32.
- Chatterjee R, Batra J, Das S, Sharma SK, Ghosh B. Genetic association of acidic mammalian chitinase with atopic asthma and serum total ige levels. *Journal of Allergy Clinical Immunology*. 2008; 122:202–208.
- Chin LJ, Ratner E, Leng S, Zhai R, Nallur S, Babar I, et al. A SNP in a let-7 microRNA complementary site in the KRAS 3' untranslated region increases non-small cell lung cancer risk. *Cancer Research*. 2008; 68:8535–8540. [PubMed: 18922928]
- Christiani DC, Sheu C, Chen F, Su L, Bajwa E, et al. A large-scale genotyping study identifies five genes associated with ARDS development. *American Journal of Respiratory and Critical Care Medicine*. 2010; 181:A1024.
- Cook NR, Zee RYL, Ridker PM. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Statistics In Medicine*. 2004; 23:1439–1453. [PubMed: 15116352]
- Eilers PHC, Marx BD. Flexible smoothing with b-splines and penalties. *Statistical Science*. 1996; 11:89–121.
- Gianola D, van Kaam JBCHM. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*. 2008; 178:2289–2303. [PubMed: 18430950]
- Gonzalez-Recio O, Gianola D, Long N, Weigel KA, Rosa GJM, Avendano S. Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. *Genetics*. 2008; 178:2305–2313. [PubMed: 18430951]
- Guo W, Lin S. Generalized linear modeling with regularization for detecting common disease rare haplotype association. *Genetic Epidemiology*. 2009; 13:308–316. [PubMed: 19025789]
- Hansel NN, Gao L, Rafaels NM, Mathias RA, Neptune ER, Tankersley C, Grant AV, Connett J, Beaty TH, Wise RA, Barnes KC. Leptin receptor polymorphisms and lung function decline in COPD. *European Respiratory Journal*. 2009; 34:103–110. [PubMed: 19196818]
- Huebner M, Kim DY, Ewart S, Karmaus W, Sadeghnejad A, Arshad SH. Patterns of GATA3 and IL13 gene polymorphisms associated with childhood rhinitis and atopy in a birth cohort. *Journal of Allergy Clinical Immunology*. 2008; 121:408–414.
- Wesse J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. *The American Journal of Human Genetics*. 2006; 79:792–806.
- Jiang R, Tang W, Wu X, Fu W. A random forest approach to the detection of epistatic interactions in case-control studies. *Bioinformatics*. 2009; 10(Suppl 1):S65. [PubMed: 19208169]
- Lin HY, Wang W, Liu YH, Soong SJ, York TP, Myers L, Hu JJ. Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP-SNP interactions and their application in prostate cancer. *Journal of Human Genetics*. 2008; 53:802–811. [PubMed: 18607530]
- Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*. 2007; 63:1079–1088. [PubMed: 18078480]
- Maity A, Lin X. Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines. *Biometrics*. 2011; 67(4):1271–1284. [PubMed: 21504419]
- Maity A, Zhang H, Karmaus W, Ewart S. A marker selection algorithm using kernel machine regression. 2012 Unpublished manuscript.

- Sadeghnejad A, Ohar JA, Zheng SL, Sterling DA, Hawkins GA, Meyers DA, Bleecker ER. Adam33 polymorphisms are associated with COPD and lung function in long-term tobacco smokers. *Respiratory Research*. 2009; 10
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*. 2010; 86:929–942.
- Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, et al. Genomewide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics*. 2007; 39:645–649. [PubMed: 17401363]

Fig. 1.
Power changes with τ for fixed K matrix

Fig. 2.

Power changes with sparsity of $\mathbf{K}$ matrix (sample size=150)^a

^a SL: Sparsity Level. SL1: 150 (0.67% of off-diagonal entries) non-zero values in $\mathbf{K}$; SL2: 46 (0.21%); SL3: 30 (0.13%); SL4: 16 (0.072%).

Fig. 3.
Power changes with similarity level of K matrices for fixed $\tau = 2.0$

Fig. 4.
Contour Plot for Lung function with CHIA gene Data

Fig. 5.
Plot of importance indices of the SNPs. The dashed line indicates the level had the snps were chosen completely at random

Table 1

SNP set effects based on the whole SNP set and selected SNP sets (IOW data)

SNP set	τ	p-value
the whole SNP set (12 SNPs)	0.00031	0.41
rs10776724, rs4240530, rs3818822, rs3806446, rs2256721	0.0056	0.16
rs10776724	0.0088	0.032