Title: Powerful SNP Set Analysis for Case-Control Genome Wide Association Studies

Running Title: Powerful SNP Set Analysis

Michael C. Wu[1], Peter Kraft[2,3], Michael P. Epstein[4], Deanne M. Taylor[2], Stephen J. Chanock[5], David J. Hunter[3], and Xihong Lin[2]

Department of Biostatistics[1], The University of North Carolina at Chapel Hill, Chapel Hill, NC.

Department of Biostatistics[2], Harvard School of Public Health, Boston, MA.

Department of Epidemiology[3], Harvard School of Public Health, Boston, MA.

Department of Human Genetics[4], Emory University, Atlanta, GA.

Division of Cancer Epidemiology and Genetics[5], National Cancer Institute, Bethesda, MD.

Address for Correspondence: Xihong Lin, Ph.D.
Department of Biostatistics, Harvard School of Public Health
655 Huntington Avenue, Boston, MA 02115
Phone: (617) 432-2914
Fax: (617) 432-5619
E-mail: xlin@hsph.harvard.edu

Abstract

Genome wide association studies (GWAS) have emerged as popular tools for identifying genetic variants that are associated with disease risk. Standard analysis of a case-control GWAS involves assessing the association between each individual genotyped SNP and disease risk. However, this approach suffers from limited reproducibility and difficulties in detecting multi-SNP and epistatic effects. As an alternative analytical strategy, we propose grouping SNPs together into SNP sets based on proximity to genomic features such as genes or haplotype blocks, and then testing the joint effect of each SNP set. Testing of each SNP set proceeds via the logistic kernel machine based test which is based on a statistical framework that allows for flexible modeling of epistatic and nonlinear SNP effects. This flexibility as well as the ability to naturally adjust for covariate effects are important features of our test that make it appealing compared to individual SNP tests and existing multi-marker tests. Using simulated data based on the International HapMap Project, we show that SNP set testing can have improved power over standard individual SNP analysis under a wide range of settings. In particular, we find that our approach has higher power than individual SNP analysis when the median correlation between disease susceptibility variant and the genotyped SNPs is moderate to high. When the correlation is low, both individual SNP analysis and the SNP set analysis tend to have low power. We apply SNP set analysis to analyze the CGEMS breast cancer GWAS discovery phase data.

1 Introduction

Identification of single nucleotide polymorphisms (SNPs) that are associated with risk for developing complex disease is an important goal of modern genetic studies. The hope is that such knowledge can ultimately be used both for understanding the biological mechanisms underlying these diseases and for generating individualized risk profiles that are useful in a public health context. To this end, genome wide association studies (GWAS) have emerged as a popular tool for identifying common genetic variants for complex disease. A standard case-control GWAS for identifying SNPs associated with disease susceptibility involves genotyping a large number of SNPs, on the order of hundreds of thousands, in thousands of individuals with the disease (cases) and thousands of healthy controls with the goal of identifying individual loci that are associated with the outcome. Such studies have been successfully used to identify SNPs associated with susceptability to diseases such as breast cancer[1,2] (MIM 114480), prostate cancer[3–5] (MIM 176807), and type II diabetes[6–8] (MIM 125853).

A typical GWAS consists a discovery phase in which an initial set of promising susceptibility loci are identified followed by a validation stage in which the SNPs identified in the initial discovery phase are replicated in a separate study cohort.[9] The standard approach for analyzing GWAS in the discovery phase involves individual SNP analysis. This mode of analysis often involves regressing the phenotype onto each individual typed SNP and generating a parametric p-value. The SNPs are then ranked based on their individual p-values and a threshold is set such that all SNPs with p-value less than that threshold will be pushed forward for validation. The threshold can be based on reaching a muliple-comparison adjusted significance level or a level based on non-analytical means.

Although use of individual SNP analysis has proved useful in identifying many disease susceptibility variants, this mode of analysis may be limited in some settings due to difficulty in reaching genome wide significance. More specifically, in order to control the overall type I error rate, the level at which each test is conducted must be adjusted. Due to the large number of considered hypotheses, the threshold for genome wide signficance can be very extreme and difficult to attain: for a GWAS examining the effects of 500,000 SNPs, each test is conducted at the $\alpha = 10^{-7}$ level, which is very stringent.

Additionally, individual-SNP analysis is often limited by poor reproduceability; many of the highly-ranked SNPs in the discovery phanes are false positives and cannot be validated. This is largely due to the restricted power to detect SNPs with small effects that are truly associated with the outcome. In particular, individual SNPs that are genotyped on GWAS platforms often show only modest effects. One explanation for this is that the true causal SNP is rarely genotyped, but there are typed SNPs which are in linkage disequilibrium (LD) with the causal SNP. In this case, using individual SNP analysis, the typed SNPs in LD with the causal SNP will each only show moderate effects since each typed SNP serves as an imperfect surrogate for the causal SNP. Thus, it could be advantageous to consider the joint effect of multiple SNPs in analysis[10] since it is probable that several of these markers are in LD with the causal SNP and could capture the true effect more effectively than individual-SNP analysis. Finally, individual SNP analysis only considers the marginal effect of each SNP and therefore fails to accommodate epistatic effects. Epistatic interactions between SNPs can contribute to disease susceptibility such that individual SNPs may show little individual effect, but their interactions can have a much larger effect. Individual SNP analysis will not be able to detect such effects which, more generally, are difficult to find due to the large number of potential interactions.[11]

As an alternative strategy for analysis, we propose grouping of SNPs together into SNP sets along the genome and perform genome-wide tests for individual SNP sets instead of individual SNPs. SNP set based analysis borrows information from different but correlated SNPs that are grouped based on prior biological knowledge and hence has the possibility of providing results with improved reproducibility and increased power, especially when individual SNP effects are moderate, and improve interpretability of the results. This mode of analysis proceeds via a two step procedure. First, SNPs are assigned to SNP sets based on some meaningful biological criteria (genomic features), e.g., genes. Then, tests for the association between each genomic feature and a disease phenotype are performed using a logistic kernel machine based multilocus test, across the genome.

SNP set analysis can prove advantageous over the standard analysis of individual SNPs. By forming SNP sets and testing each SNP set as a unit, we are reducing the number of hypotheses being tested and thus relaxing the stringent conditions for reaching genome-wide significance. Grouping SNPs together properly, we will have improved power in settings where SNPs are individually only moderately significant. In particular, though any single SNP may serve as a poor surrogate for an untyped causal SNP, by considering multiple typed SNPs, we will be better able to capture the true effect of the untyped causal SNP. Furthermore, if there are multiple independent causal SNPs, by considering their joint effects, we will have power to detect their joint activity.

To test each SNP set within a case-control GWAS, we propose a general semiparametric kernel based testing procedure which is tailored towards high-dimensional genetic data. Specifically, this test will combine the logistic kernel machine testing approach of Liu et al.[12] with the kernel framework suggested by Kwee et al.[13] As we will show, the logistic kernel machine has appealing features for SNP set analyses. The testing framework is powerful and allows for great flexibility in the functional relationship between the SNPs in a SNP set and the outcome. Thus, the method can easily account for complex SNP interactions and nonlinear effects. Combined with the ability to seamlessly adjust for covariate effects and the fast computational efficiency of our method, this flexibility gives the logistic kernel machine based test significant advantages over both individual

SNP tests and existing multi-marker tests.

Broadly speaking, our work advances the field in three important ways. First, we develop SNP set analysis as an alternative to standard individual SNP analysis and discuss principled approaches for forming SNP sets based on genomic features. Second, we develop a powerful statistical modeling and testing framework for genetic effects which has a number of practical advantages over other multi-marker tests: our approach is computationally efficient and naturally accommodates covariate adjustment, non-linear effects, and epistasis. Third, we will demonstrate through thorough numerical studies and data applications that our approach can have substantially improved power over standard individual SNP testing, and by extension, over the many multi-marker tests that individual SNP testing tends to dominate.

The remainder of this article is organized as follows. In the next section, we describe our proposed SNP set analysis framework including how to form SNP sets and how to subsequently test SNP sets. Then we will present simulation results comparing our approach to individual SNP analysis and two existing multi-SNP tests. Finally, we will apply logistic kernel machine based SNP set analysis to the CGEMS breast cancer data from the discovery phase. We will conclude with a brief discussion.

2 Materials and Methods

SNP set based analysis borrows information from different but correlated SNPs that are grouped based on prior biological knowledge and hence provides results with improved reproducibility and increased power, especially when individual SNP effects are moderate. This mode of analysis proceeds via a two step procedure. First, across the genome, SNPs are assigned to a SNP sets based on some meaningful biological criteria such as proximity to genomic features—SNP sets of a single SNP are possible. If we wished to perform genome wide SNP set analysis of a GWAS conducted on the Illumina HumanHap500 array by grouping SNPs based on genes, we could generate approximately 18000 SNP sets, each of which consisted of the SNPs within a single gene. For example, the 14 genotyped SNPs within the *ASAH1* (MIM 228000) gene could be assigned to a single SNP set and the 4 genotyped SNPs within the *NAT2* (MIM 612182) gene could be assigned to another SNP set, and so on. After the groupings are made, each of the 18000 SNP sets is tested using a multilocus test, and the genome-wide significance of SNP set, e.g. each gene, is calculated. Although a number of tests have been proposed,[14,15] we consider an extension of the logistic kernel machine test, which was developed in the gene expression profiling setting, that we tailor for analysis of genome wide association studies. In this section, we describe possible methods for grouping SNPs in a genome wide scan into SNP sets and then we present the logistic kernel machine test for evaluating the significance of each SNP set.

2.1 Forming SNP Sets

A key aspect of our proposed approach is the formation of meaningful SNP sets. In principle, a SNP set may be formed via any grouping of SNPs, and our testing approach is still valid in the sense that the type I error rate will always be protected. However, better groupings can be made on the basis of prior biological knowledge and if done properly, can lead to additional gains in power. In particular, the key advantages of our approach may be found in the ability to reduce the number of multiple comparisons, to harness correlation between SNPs, to measure the joint effect of independent SNPs, and to make direct inference on a biologically meaningful genomic feature. Some natural ways of forming SNP sets that can capitalize on these advantages include grouping SNPs on the basis of genomic features. We describe below some natural grouping structures.

A natural grouping strategy is to take all SNPs that are located in or near a gene, a fundamental unit of the genome, and group them to form a SNP set. In particular, one can take all SNPs between the start and end of transcription as well as SNPs that are upstream and downstream of the gene, in order to capture regulatory regions, as a single SNP set. In grouping based on known genes, we can significantly reduce the number of multiple comparisons. The SNPs on the Illumina HumanHap 500 array correspond to to approximately 17,800 genes in contrast to the original 530,000 SNPs. Since we take the entire gene region, not just exonic regions, we expect to have many typed SNPs that are correlated and thus the logistic kernel machine test will have good power to detect a significant SNP set effect. We could also expect multiple SNPs within a gene to be associated with disease risk and this grouping structure would allow us to detect this effect. Testing gene-based SNP sets also makes direct inference on the association between the gene and case-control status.

An extension of gene based SNP set analysis is to group SNPs based on whether they are located within a gene pathway from KEGG[16] or a Gene Ontology Consortium functional category.[17] Making inference on a pathway further reduces the number of multiple comparisons and still allows inference on a biologically meaningful unit. The logistic kernel machine test will be able to harness local LD to have power and will, additionally, be able to capture true pathway effects when several SNPs in multiple genes are related to the disease.

Although many variants associated with disease have been identified within gene regions, many lie outside of the boundaries of known genes (and hence pathways). To augment coverage of the genome, a possible strategy would be to group SNPs within evolutionarily conserved regions. Increased evolutionary conservation of a genomic region is suggestive of increased importance or functionality.[18] Significance of such a SNP set would potentially indicate that there is a genomic feature present that is related to disease risk, even if the feature is not well understood.

Finally, approaches to forming SNP sets that can achieve full coverage of the genome by placing all SNPs into SNP sets include grouping SNPs via a moving window or via haplotype blocks. For example, one could divide the genome into a fixed number of adjacent regions, purely based on length, and treat all SNPs within a region as a SNP set. Alternatively, one could build SNP sets based on haplotype blocks such as through Haploview.[19] Both approach will still allow us to harness local correlation to capture the effect of untyped SNPs.

An important limitation of employing a gene or pathway based approach is the omission of intergenic regions. However, use of additional grouping strategies, e.g. conserved regions, can augment coverage, and using the moving window and haplotybe block can provide comprehensive coverage of the entire genome. Although we wish to group SNPs that are near one another to harness correlation, this does not allow us to capture multi-SNP or epistatic effects among SNPs in separate SNP sets. Using gene pathway based SNP sets could ameliorate this issue since this looks across individual continuous regions. Groupings based on strategies beyond the ones that we have considered are also possible.

As noted above, we emphasize that while well formed SNP sets can optimize the power and interpretability of our SNP set testing strategy, our logistic kernel machine testing approach is statistically valid irrespective of the grouping scheme. For illustration, we will focus on SNP sets formed based on proximity to each of 18000 known genes.

2.2 Genome Wide SNP Set Testing

Although we propose our strategy as a genome wide approach, we will present the testing procedure by focusing on testing a single SNP set.

In this paper, we assume that a population based case-control GWAS was conducted in which n independent subjects were genotyped. To employ our SNP set analysis approach, we first group the SNPs into SNP sets across the genome. Then for a given SNP set containing p SNPs, let $z_{i1}, z_{i2}, \ldots, z_{ip}$ be genotype values for the SNPs in the SNP set for the i^{th} subject (i = 1,...,n). The case-control status for the i^{th} subject is denoted by y_i ($y_i = 1$ for cases and $y_i = 0$ for controls). We assume without loss of generality that the SNPs are coded in a trinary fashion with $z_{ij} = 0, 1, 2$ corresponding to homozygotes for the major allele, heterozygotes, and homozygotes for the minor allele respectively. This corresponds to the commonly employed additive model of allelic affect, but we note that alternative models, such as the dominant and recessive models, are also possible and can be tested within our framework. We further assume that for each individual, an additional set of m demographic, environmental, or other confounding variables is collected. For the i^{th} subject we let $x_{i1}, x_{i2}, \ldots, x_{im}$ denote the values of the covariates we would like to adjust for. The goal of SNP set analysis is then to test the global null of whether any of the p SNPs are related to the outcome while adjusting for the additional covariates.

In principle, many multi-locus testing approaches could be used for evaluating the significance of the SNPs in the SNP set, but to harness correlation and accommodate complex relationships between the SNPs and the outcome and epistatic effects, we propose a new approach to test the SNP set by modelling each SNP set's effect in a flexible fashion while adjusting for additional covariate effects. At the same time, to overcome the issue of the large number of degrees of freedom, our strategy will employ a test that adaptively estimates the degrees of freedom by accounting for correlation (LD) among the SNPs. Specifically, we will choose to use the logistic kernel machine regression modelling framework and a corresponding score test.[12]

2.2.1 Logistic Kernel Machine Model

In evaluating the significance of a SNP set, we need to employ a strategy that allows us to model, and subsequently test, the effects of multiple SNPs that have been grouped in a biologically meaningful fashion. The kernel machine framework has become very popular for modelling high-dimensional biomedical data due to its ability to allow for complex/nonlinear relationships between the dependent and independent variables[20,21] while adjusting for covariate effects. We consider a logistic kernel machine regression model for the joint effect of the SNPs in the SNP set and the additional covariates that we would like to adjust for. Under the notation above, for the i^{th} individual, we have the semiparametric model given by

$$\text{logit}\, P(y_i = 1) = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_m x_{im} + h(z_{i1}, z_{i2}, \ldots, z_{ip}) \tag{1}$$

where α_0 is an intercept term, $\alpha_1, \ldots, \alpha_m$ are regression coefficients corresponding to the environmental and demographic covariates. The SNPs, $z_{i1}, \ldots, z_{ip}$, influence y_i through the general function $h(\cdot)$ which is an arbitrary function that that has a form defined only by a positive semidefinite kernel function $K(\cdot, \cdot)$.

Our primary aim is to adequately model the SNPs and evaluate their effect, so $h(\cdot)$ is the model component in which we have primary interest because it fully determines the relationship between genotypes of the SNPs in the SNP set and disease risk. Leaving $h(\cdot)$ only generally specified permits a modelling framework that accommodates complex relationships between the SNPs and risk as well as epistatic effects.

We omit the mathematical details, but using the representer theorem,[22] we note that $h(z_{i1}, z_{i2}, \ldots, z_{ip})$ in Equation 1 is equal to $h_i = h(\mathbf{Z}_i) = \sum_{i'=1}^{n} \gamma_{i'} K(\mathbf{Z}_i, \mathbf{Z}_{i'})$ for some $\gamma_1, \ldots, \gamma_n$. This shows that $h(\cdot)$ is fully defined by the kernel function $K(\cdot, \cdot)$. Details on the mathematical relationships and estimation may be found in Liu et al.[12] and Cristianini et al.,[20] but the key is that by choosing different kernel functions, we can specify different, possibly complex, bases and corresponding models. For example, if we define $K(\cdot, \cdot)$ to be the *linear kernel* such that $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \sum_{j=1}^{p} z_{ij} z_{i'j}$ then we are implicitly assuming the simple logistic model defined by

$$\text{logit}\, P(y_i = 1) = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_m x_{im} + \beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_p z_{ip}$$

where β_j is a regression coefficient corresponding to the j^{th} SNP. To specify a more complicated model, we need only change our choice of $K(\cdot, \cdot)$.

From the above, it is apparent that the choice of kernel changes the underlying basis for the nonparametric function governing the relationship between case-control status and the SNPs in the SNP set. Essentially, $K(\cdot, \cdot)$ is a function that projects the genotype data from the original space to another space and then $h(\cdot)$ is modelled linearly in this new space, such that if one considers h on the original space, it can be highly nonlinear. More intuitively, however, $K(\mathbf{Z}_i, \mathbf{Z}_{i'})$ can be viewed as a function that measures the similarity between two individuals, the i^{th} and i'^{th} subject, based on the genotypes of the SNPs in the SNP set. Taking this perspective, many choices for K are possible. Some specific kernels functions that we can consider include the linear, identical-by-state (IBS), and weighted IBS kernels.

The linear kernel is: $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \sum_{j=1}^{p} z_{ij} z_{i'j}$ which is the usual inner product between the covariate vectors for subject i and i'. As described earlier, this kernel assumes a set of basis functions that spans the original covariate space such that one is implying a linear relationship between the logit of the probability of being a case and the genotypes of the SNPs in the SNP set, i.e. the usual multiple logistic regression model.

The gaussian kernel is: $K(\mathbf{Z}_i, \mathbf{Z}_{i'}; d) = \exp\{-\sum_{j=1}^{p}(z_{ij} - z_{i'j})^2/d\}$ and assumes the radial basis which is difficult to characterize using an explicit set of basis functions. The class of models generated by the gaussian kernel can be very broad and includes the linear model as a special case. Here d is a parameter that approximately controls area of influence of the kernel function such that larger values of d correspond to smoother h

functions.

The IBS kernel is: $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \frac{\sum_{j=1}^{p}\left\{2I(z_{ij}=z_{i'j})+I(|z_{ij}-z_{i'j}|=1)\right\}}{2p}$. In genetics, a possible metric for evaluating distance between individuals on the basis of genotype information is the number of alleles shared identical by state (IBS) by a pair.[15] As shown by Kwee et al.,[13] this may also be used a a valid kernel function.

The weighted IBS kernel is: $K(\mathbf{Z}_i, \mathbf{Z}_{i'}; \mathbf{w}) = \frac{\sum_{j=1}^{p} w_j\left\{2I(z_{ij}=z_{i'j})+I(|z_{ij}-z_{i'j}|=1)\right\}}{2p}$ where $w_j = 1/\sqrt{q_j}$ and q_j is the minor allele frequency (MAF) for the j^{th} SNP in the SNP set. The weighted IBS kernel is an extension of the IBS kernel that up-weights for similarity in rare alleles. The idea is that similarity in rare alleles is more informative than similarity in common alleles.

The ability to model data using the gaussian and IBS kernels are advantages of using the kernel machine framework since formulating an explicit set of basis functions can be difficult. Alternative kernel functions, such as those discussed in Wei and Schaid[23] and in Mukhopadhyay et al.[24] are possible and can be designed for specific data sets. To be a valid kernel function, $K(\cdot, \cdot)$ needs to be positive semi-definite and satisfy the conditions of Mercer's theorem.[20]

2.2.2 Logistic Kernel Machine Test

Here, our focus is on hypothesis testing for which only need to estimate α under the null hypothesis that $h(\mathbf{Z}_i) = 0$. Therefore, we omit the technical details on estimating the genetic effect, $h(\mathbf{Z})$, from the SNP set and refer the reader to Liu et al.[12]

The above modelling framework leads naturally to a powerful test for association between the SNPs in the SNP set and case-control status. Note that the probability that the i^{th} subject is case depends on the SNPs only through the function $h(\mathbf{Z}_i)$. Thus, in order to test whether there is a true SNP set effect, we can consider the null that

$$H_0 : \mathbf{h}(\mathbf{Z}) = 0 \tag{2}$$

against the general alternative. To test this hypothesis, Liu et al.[12] exploit the connection between the kernel machine framework and generalized linear mixed models (GLMM). Specifically, letting $\mathbf{K}$ be the $n \times n$ matrix with $(i, i')^{th}$ element equal to $K(\mathbf{Z}_i, \mathbf{Z}_{i'})$, then it is straightforward to see that $\mathbf{h} = \mathbf{K}\boldsymbol{\gamma}$, where $\mathbf{h} = [h_1, \ldots, h_n]'$. We can treat $\mathbf{h}$ as a subject specific random effect, then via the GLMM connection, $\mathbf{h}$ follows an arbitrary distribution F with mean zero and variance $\tau\mathbf{K}$. Note that τ indexes the effect of the SNPs in the SNP set such that

$$H_0 : h(\mathbf{Z}) = 0 \Leftrightarrow H_0 : \tau = 0.$$

Thus, we need only to test whether the indexing parameter τ is significantly different than zero. This can proceed via the variance component score test of Zhang and Lin[25] using the statistic:

$$Q = \frac{(\mathbf{y} - \widehat{\mathbf{p}}_0)'\mathbf{K}(\mathbf{y} - \widehat{\mathbf{p}}_0)}{2} \tag{3}$$

where logit $\widehat{p}_{0_i} = \widehat{\alpha}_0 + \widehat{\alpha}_1 x_{i1} + \widehat{\alpha}_2 x_{i2} + \cdots + \widehat{\alpha}_m x_{im}$. Since this is a score test, $\widehat{\alpha}_0$ and the $\widehat{\alpha}_j$ are estimated under the null model which does not contain $\mathbf{h}$, so we can use the standard estimate from the logistic regression model without the genotypes. To compute a p-value for significance, we can compare Q to a scaled χ^2 distribution with scale parameter, κ, and degrees of freedom, ν. Details on calculating κ and ν are found in the Appendix.

The adaptive estimation of the degrees of freedom, ν, constitutes a key advantage of the logistic kernel machine test. In particular, if the R^2 between the SNPs in the SNP set increases, then ν decreases such that if all the SNPs are perfectly correlated, $\nu \to 1$. It follows that for a given $h(\cdot)$, higher correlation is likely to lead to higher power, suggesting that the logistic kernel machine test improves the power for SNP set testing by harnessing the correlation between SNPs and adaptively estimating ν.

In general, it can be difficult to identify *a prior* whether it is the minor allele or the major allele that is associated with increased disease risk, and equivalently, whether the minor allele is protective or deleterious. The logistic kernel machine test is not affected by the directionality of effect and its power is robust to whether the minor alleles of the causal SNP are protective or deleterious (or a combination of both in settings with multiple causal variants).

The testing framework considered here has similarities to those of Schaid et al.,[10] Mukhopadhyay et al.,[24] and Wessel and Schork[15] which we describe below in that all three approaches are based on genetic distances among subjects. However, the kernel framework allows for improved flexibility in the functional relationship.

2.2.3 Existing Multi-SNP Tests

Although other multi-SNP tests could be used for evaluating the significance of each SNP set, the kernel machine has advantages over each of these. Here, we briefly discuss some alternative tests that fall into several different categories.

The first class of multi-SNP test encompasses the multi-marker methods that are based on individual SNP analysis. In particular, a common approach for evaluation the significance of a set of markers is to apply individual SNP analysis by testing the individual significance of each SNP, using the most significant p-value as the p-value for the set of loci, and then correct for having done multiple tests via monte carlo methods[26] or by estimating the effective number of tests.[27–29] Alternatively, the test statistics from each of the individual tests can be combined.[30] However, such tests still rely strongly on individual SNP analysis and when the individual SNPs are not in high LD with the causal variant, they may have low power, as they do not borrow information across SNPs which are frequently correlated. Furthermore, they cannot accommodate complex genetic effects and interactions. Our simulations will verify that the logistic kernel machine test often has improved power over this class of test.

Omnibus tests for multiple SNPs or haplotypes via multivariate regression[10,31] allow for simultaneous analysis of all SNPs, butstudies have shown that such methods often offer little benefit over individual SNP analysis based methods[32,33] as they are based on a large number of degrees of freedom. To reduce the degrees of freedom, a set of multimarker tests that compare pairwise genetic similarity with pairwise trait similarity were proposed by Schaid et al.,[14] Wessel and Schork,[15] and Mukhopadhyay et al.[24] All three approaches are attractive; however, as noted by Mukhopadyay et al., an important limitation of Schaid et al.'s approach is that it assumes all variants have the same direction of effect, i.e. all the minor alleles for each SNP increase risk or all minor alleles decrease risk. Although the methods of Wessel and Schork and Mukhopadhyay et al. are robust to directionality, both evaluate significance via computationally expensive permutation which may be impractical for some GWAS settings. None of the three similarity based methods allow for easy covariate adjustment. The logistic kernel machine test also considers pairwise similarity and shares the attractive nonparametric SNP effects model, but in addition to using a computationally efficient score test and being robust to directionality, the logistic kernel machine model naturally incorporates covariate effects, an important feature. Beyond adjusting for confounders and population structure, it is often necessary to adjust for highly significant SNPs in GWAS to distinguish between settings where a particular significant marker is the causal SNP (or a SNP in high LD with the causal SNP), versus setting where additional independent markers that are associated with disease are present. A third similarity based approach by Tzeng and Zhang[34] can be seen as a special case of the more general logistic kernel machine test that focuses exclusively on haplotype similarity. The need to phase sample haplotypes from genotype data incurs additional computational expense and variability—particularly for larger SNP sets.

A final class of multi-marker tests consists of methods that leverage explicit population genetic models to pinpoint the causal locus. Many involve reconstructing the sample phylogeny to guide the analysis and infer the causal mutation.[35,36] If the population genetics model assumed is realistic and correct, such problem specific methods should have high power. However, it is difficult to validate the assumed models and most procedures are computationally intensive such that in real applications the models need to be simplified. Once again, these models usually fail to allow for covariate adjustment. Computational efficiency and ease of covariate adjustment give a practical advantage to the logistic kernel machine regression test.

2.3 Simulations

To evaluate the performance of our SNP set analysis approach, we study the logistic kernel machine test in the genetics framework by considering its empirical performance under a variety of settings. For simplicity of implementation, all causal SNPs in our simulations are assumed to increase disease risk, but it is important to note that none of the methods we consider are affected by the direction of effect.

2.3.1 Simulations Based on the *ASAH1* Gene

We first investigate the size and power of the kernel machine testing framework under a setting in which the SNP set is generated based on the LD structure of a single gene which will allow us to better understand under which settings our SNP set analysis approach is most advantageous. We considered the *ASAH1*, *NAT2*, and *FGFR2* (MIM 176943) genes, but for clarity, we present only the simulation configurations and the results based on the *ASAH1* gene. The simulations and results from using the *NAT2* and *FGFR2* were qualitatively similar.

ASAH1, acid ceramidase 1, is a 28.5kb long gene with 86 HAPMAP SNPs and is located at 8p21.3-p22. Expression is associated with prostate cancer[37] and mutations in the gene are known to be associated with Farbers Disease[38] (MIM 228000). We based our gene specific simulations on the LD structure of the *ASAH1* gene and used HAPGEN[39] and the CEU sample of the International HapMap Project[40] to generate SNP genotype data at each of the 86 loci.[41] 14 out of 86 SNPs are genotyped using the Illumina HumanHap500 array. These will be the "typed" SNPs we use for our simulated analysis.

We first conducted simulations to verify that the logistic kernel machine test properly controls the type I error rate. To investigate the empirical size of our test, we conducted simulations in which we generated $n/2$ cases and $n/2$ controls under the null logistic model where disease risk does not depend on the genotype:

$$\text{logit}\, P(y_i = 1|\mathbf{X}_i) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i \tag{4}$$

where $\mathbf{X}_i$ is a vector of additional covariates that are independent of the simulated genotype data. We considered $n = 1000, 2000$ and also considered the use of the linear, IBS, and weighted IBS kernels. For each choice of n and kernel function we generated 5000 data sets using HAPGEN. To ensure that our simulations are realistic, our simulations generated all 86 HapMap SNPs, but we only apply our testing approach to the 14 typed SNPs. Specifically, we group the 14 SNPs as a SNP set based on the *ASAH1* gene and then we apply the logistic kernel machine test to compute a p-value evaluating the effect of the SNPs in the SNP set while adjusting for covariates in $\mathbf{X}$. For comparison, we also analyzed the 14 typed SNPs as we would have done under an individual SNP analysis: we tested the significance of each of the 14 SNPs individually, while again adjusting for covariates in $\mathbf{X}$, and then adjusted the individual p-values via a modified bonferroni correction where the effective number of tests was computed via two approaches. First, we used the method of Moskvina et al.;[29] second, we estimated the effective number of tests as the number of principal components necessary to account for 99% of the variability.[42] The two approaches were approximately concordant. The smallest p-value, corrected for the effective number of tests, was was taken as the p-value for the entire SNP set. Size for individual SNP analysis testing was again the proportion of p-values less than $\alpha = 0.05$.

To compute the empirical power for a SNP set, we generated data sets with $n/2$ cases and $n/2$ controls under the alternative logistic model:

$$\text{logit}\, P(y_i = 1|\mathbf{X}_i) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \beta^c z_i^c \tag{5}$$

where z_i^c is the genotype for the "causal" SNP, β^c is the log genetic odds ratio for the causal SNP, and $\mathbf{X}_i$ are a vector of additional covariates that are independent of z_i^c. Note that under each simulation configuration we allow only a single causal SNP. Each of the 86 HapMap SNPs was set to be the "causal" SNP in turn. Setting $\beta^c = 0.2$ which corresponds to a genetic odds-ratio of 1.22, we again considered sample sizes $n = 1000, 2000$. For each choice of n, and for each of the 86 causal SNPs, we generated 2000 data sets. We again apply our testing approach to each data set by grouping the 14 typed SNPs and computing a p-value for the significance of the SNP set, while adjusting for covariates in $\mathbf{X}$, via the logistic kernel machine test under a linear kernel. We emphasize that only the 14 typed SNPs were used so the causal SNP is unobserved under most configurations. For each configuration, we then computed the test power as the proportion of p-values less than the α level = 0.05. This was compared with the power based on the individual SNP analysis with modified bonferroni correction approach described above.

2.3.2 Simulations Based on Randomly Sampled Genes

We also evaluate the power of our approach under settings in which the LD structure of the simulated SNP sets varied across a wide range of possible genes. Specifically, we generated 20,000 SNP sets using HAPGEN where each SNP set is based on a real gene on chromosome 10. This allows for 670 possible SNP sets. Within each SNP set, we randomly selected one HapMap SNP to be the causal SNP and again generated $n/2$ cases and $n/2$ controls based on model given by Equation 5 with β^c again fixed at 0.2 (OR = 1.22). Again treating the SNPs on the Illumina HumanHap 500K array as the typed, we tested the significance of the SNP set using the logistic kernel machine test under a linear kernel. We also apply the individual SNP analysis testing procedure described above. Thus, for both our method and the competing individual SNP analysis test, we computed 20,000 p-values for significance.

2.3.3 Comparisons with Alternative Multi-SNP Tests

As discussed previously, in principle, any multi-SNP test can be used to test the significance of a SNP set. However, the kernel machine test is advantageous in that it adaptively finds the degrees of freedom of the test statistic in order to account for LD between genotyped markers, can permit complex relationships between the SNPs and the outcome, naturally allows for covariate adjustment, and is computationally efficient since no permutation is required. To provide additional empirical results, we compare the logistic kernel machine test to the similarity based testing approach of Mukhopadhyay et al[24] and the approach of Wessel and Schork,[15] which has been found to perform well relative to other multi-SNP tests.[23] We assessed the power under five models and the test size under two additional models. For each of the five models examining power, 500 simulations were conducted, and 1000 simulations were conducted under the two models examining the test size. For all seven models, we assumed sample sizes of 500 case and 500 controls, 1000 permutations were used to compute the p-values for the methods of Wessel and Schork and Mukhopadhyay et al., and power and size were computed as the proportion of p-values less than 0.05.

We first compare the power of the methods under four alternative models using SNP sets based on the *ASAH1* gene. Under Model 1, the data sets were simulated under the alternative logistic model based on Equation 5 in which the causal variant was fixed to be rs3810 (the third in the SNP set), one of the 14 typed SNPs, with β^c again fixed at 0.2 (OR = 1.22). Model 2 was similar to Model 1, except we change the causal SNP to be rs7825389 (the 69th SNP), an untyped SNP. Model 3 was again similar to the earlier models, except we allow for two causal variants, rs10105871 and rs7825389, which are the 63rd and 69th SNPs in the *ASAH1* SNP set, respectively. Both SNPs are in the same LD block and rs10105871 is typed while rs7825389 is untyped. The effect size for both causal SNPs was set at 0.2. Model 4 is identical to Model 3, but here we have two untyped causal SNPs which are in different LD blocks, rs4377998 and rs7825389, which are the 43rd and 69th SNPs respectively.

Under Model 5, we compared the power under the setting considered by Mukhopadhyay et al. in which 10 independent markers in Hardy-Weinberg equilibirium with MAF = 0.05 are simulated. Two of the ten markers were causal with relative risks of 1.25 under an additive model, and all 10 markers were considered to be genotyped. No additional covariates are present.

We compare the type I error rate control of the logistic kernel machine test and the approaches by Wessel and Schork and Mukhopadhyay et al. Specifically, under Model 6, we simulated null data sets based on Equation 2 and the *ASAH1* gene. We applied both approaches to each of the data sets to estimate p-values for the significance of the SNPs in the SNP set, and the size at the for each approach was estimated as the proportion of

p-values less than the 0.05 significance level. Under Model 7 is similar to Model 6, but we generate an additional demographic covariate that is correlated with rs3810 ($\rho = 0.065$), the third SNP in the SNP set.

2.4 CGEMS Breast Cancer Data

To demonstrate the applicability and power of our approach on real data, we apply SNP set analysis to real GWAS data and contrast our results with those found under individual SNP analysis.

The Cancer Genetic Markers of Susceptibility (CGEMS) breast cancer study[1] was conducted to identify individual SNPs associated with breast cancer risk. To this end, in the discovery phase, 1,145 cases with invasive breast cancer and 1,142 controls were genotyped at 528,173 loci using an Illumina HumanHap500 Array. All subjects were postmenopausal women of European ancestry recruited from the Nurses Health Study. The results of the top SNPs from the discovery phase are given in Table 2. In the initial validation study, the top 6 SNPs as well as two others in the *FGFR2* gene were genotyped in an independent set of 1,776 cases and 2,072 controls. A SNP within *FGFR2* was validated and found to be associated with risk of breast cancer. Note that the SNPs in *FGFR2* were not the top ranked variants and that the variants within *FGFR2* do not reach genome wide significance using either the bonferroni correction or an FDR correction in the initial scan.

To evaluate the performance of SNP set analysis with the logistic kernel machine based test by applying it to reanalyze the CGEMs Breast Cancer Data. Specifically, we formed SNP sets by grouping SNPs that lie within the same gene. To ensure that SNPs with possible gene regulatory roles were also included in the SNP sets, all SNPs from 20kb upstream of a gene to 20kb downstream of a gene were grouped. Using these criteria we were able to assemble a total of 17,774 SNP sets that consisted of 310,219 unique typed SNPs. We tested each of the gene based SNP sets using the logistic kernel machine test under the linear kernel, the IBS kernel, and the weighted IBS kernel. SNPs were coded in the additive mode and we adjusted for parametric effects of age group, whether the individual had hormone therapy, and the first four principal components of genetic variation to control for population stratification.[43]

3 Results

3.1 Empirical Size and Power Based on the *ASAH1* Gene

The size results for the logistic kernel machine test and individual SNP analysis are presented in Table 1. Based on our simulations, the logistic kernel machine test has correct size for the kernels and sample sizes corrected and therefore, our overall strategy of logistic kernel machine based SNP set analysis protects the type I error rate. Individual SNP analysis with modified bonferroni correction also has correct size. As expected, the average effective number of tests over the 5000 replicates was stable irrespective of sample size: 8.22 for $n = 1000$ and 8.23 for $n = 2000$.

We present the empirical power results for simulation based on the *ASAH1* gene in the top panel of Figure 1. The power for each testing approach and sample size is shown for each of the 86 HapMap SNPs acting as the causal SNP. Based on Figure 1, we can see that both methods have power when the causal SNP is in moderate or high LD with the 14 typed SNPs. In these settings, the power for our logistic kernel machine SNP set analysis approach tends to dominate individual SNP analysis for both considered sample sizes suggesting that our testing approach is an attractive alternative or auxiliary method to individual SNP analysis. For settings in which the causal SNP was not in LD with the typed SNPs, the power was approximately at the type I error rate as we would expect.

For the purpose of clarifying the optimal conditions for our testing approach, Figure

2 shows the power for each testing approach and sample size is again presented, but here the causal SNPs on the horizontal axis are ordered by the median R^2 of the causal SNP with the 14 typed SNPs. The median R^2 between the causal SNP and the 14 typed SNPs is plotted in the bottom panel. It is evident from the plots that the power for both testing approaches grows as a function of the median R^2 between the causal SNP and the typed SNPs. On the right side of the plot where the median R^2 is moderate to high, the kernel machine based testing tends to have dramatically improved power over individual SNP analysis even when the causal SNP is genotyped. When the median R^2 is low, neither approach has much power. We emphasize that we consider the *median* R^2 and not the maximum and note that the power for the kernel machine test is not necessarily the highest for situations in which the causal SNP is typed.

We repeated the size and power calculations based on the *ASAH1* gene for SNPs coded in a dominant model (results not shown). We also repeated power calculations for SNP sets with LD structure based on the *FGFR2* and *NAT2* genes. The size was again correct and power plots are qualitatively similar.

The empirical studies show that logistic kernel machine based SNP set analysis protects the type I error rate. Furthermore, except for SNPs in low LD with the genotyped SNPs (for which neither method has any power beyond the type I error rate and hence any differences in power are random), the kernel machine based SNP set analysis has greater power than individual SNP analysis.

3.2 Empirical Power Based on Randomly Sampled Genes

To summarize our results, we divide the 20,000 simulations into three groups based on p, the number of typed SNPs with the SNP set. Essentially, we compute power after binning the 20,000 simulations based on the SNP set size and then the median R^2 between the causal SNP and the typed SNPs. More specifically, we split the simulations in groups where $p \leq 10$, where $10 < p \leq 20$, and where $20 < p$. Then we further divided each of the three groups into subgroups by sorting the simulated SNP sets based on the median R^2 between the causal SNP and the typed SNPs and then splitting the group into 50 evenly sized subgroups. Within each subgroup, we estimated the power as the proportion of p-values less than $\alpha = 0.05$. For each of the groups, we plot the kernel density smoothed power against the median R^2 for the subgroups in Figure 3. We need to divide the SNP sets based on the number of SNPs because distantly located SNPs are uncorrelated such that the median R^2 decreases with increased numbers of typed SNPs.

The plots verify the earlier result we found that the power increases as a function of the median R^2 between the causal SNP and the typed SNPs. If the causal SNP is uncorrelated with most typed SNPs then we have little power to detect the SNP set effect, but if there is any power, then the kernel machine based SNP set analysis method again tends to have higher power than individual SNP analysis. Both the overall power and the relative power of our approach to individual SNP analysis increases as the number of typed SNP increases. This again indicates that our approach may be a better alternative to individual SNP analysis.

3.3 Multi-SNP Test Comparison Results

The results comparing the power and type I error rates of the logistic kernel machine test and the Wessel and Schork approach are presented in Figure 4. As expected, if the number of independent causal SNPs is increased, the power for both approaches increases. Across the first 4 models which compare the empirical power under practical settings based on the ASAH1 gene, the logistic kernel machine test tends to have higher power than both the Wessel and Schork method, with a gain of approximately 12-18%, and the approach of Mukhopadhyay et al., which improves little over the type I error rate. Under Model 5, which assumes common MAF and no LD among typed SNPs within a gene and

2 causal SNPs that are genotyped, the logistic kernel machine test and Mukhopadhyay et al.'s approach perform similarly, and both have considerably higher power than the Wessel and Schork method. Overall, these results suggest that the logistic kernel machine test has optimal power relative to other multi-SNP tests across different patterns of LD. More interesting are the simulations comparing the type I error rate. When the demographic and environmental covariates were simulated independently of the genotype information, the size for all three tests is correct. However, when we set correlation between the covariates, which is associated with the outcome and the genotypes to be modest (0.065), failing to account for the covariates using the Wessel and Schork and Mukhopadhyay et al. methods can possibly lead to an apparently inflated type I error rate of 25% and 10%, respectively. This illustrates the importance of evaluating the significance of SNP sets while in the presence of possible confounders.

Additional power simulations based on the ASAH1 gene in which as many as 4 causal SNPs were used did not yield qualitatively different results in that the logistic kernel machine test tended to have higher power. As this is unlikely to be a realistic situation, given the rarity of risk-associated common variants and the relatively small regions, these results are omitted.

We note that Mukhopadhyay et al.'s approach has similar power to the logistic kernel machine test under Model 5. This is a setting that is favors their approach. In particular, the method of Mukhopadhyay et al. is based on an ANOVA model that assumes that the effects of the modeled SNPs are constant and the residual correlation among kernel similarity scores is the same across all different pairs of cases or controls considered. Consequently, the method of Mukhopadhyay et al. will have excellent power when these modeling assumptions hold but may lose power when such assumptions are violated, such as under Models 1 through 4. The logistic kernel machine test does not make the same assumptions as the method of Mukhopadhyay et al.; for example, the effect sizes of the modeled SNPs and MAFs are allowed to vary in our approach.

Since the power of the logistic kernel machine tends to be comparable or higher, and given the difficulties posed by failing to adjust for demographic and environmental covariates and the additional computation cost incurred by permutation, the logistic kernel machine test appears to be an attractive approach for testing the significance of SNP sets.

3.4 CGEMs Breast Cancer Data Analysis Results

The results of our reanalysis may be found in Table 3. Using our approach and the linear kernel, we see that the SNP set formed of genetic variants close to the *FGFR2* is now the most highly ranked SNP set with p-value equal to 7.69$\times 10^{-7}$ and FDR q-value equal to 0.01. At that signficance level, it also reaches genome wide significance if we apply a bonferroni correction ($\alpha = 0.05/17,774 = 2.8 \times 10^{-6}$) or if we control the false discovery rate. Using a bonferroni correction, *FGFR2* again reaches genome wide significance if we apply use the IBS kernel, and if we control the FDR at 5% it reaches significance with the weighted IBS kernel as well.

4 Discussion

In this article, we propose logistic kernel machine based SNP set analysis as an approach for the analysis of case-control genome wide association studies. Our approach employs prior biological knowledge to group multiple SNPs that are located near genomic features into SNP sets and then tested as a single unit. Specifically, we choose to model the SNPs in the SNP set using a flexible semiparametric modelling framework which is based on kernel machines and we choose to test the effects of the SNP set via a powerful variance components test. We illustrate our approach using both data simulated from the International HapMap Project[40] as well as the CGEMS Breast Cancer GWAS study of Hunter et al.[1] and showed that our approach is an attractive alternative or auxiliary approach to individual SNP analysis.

The logic behind our analysis strategy is that we can borrow information between different SNPs to improve power to detect true effects. Thus the choice of grouping can influence the power of our approach. We focused on grouping SNPs based on their proximity to a known gene and noted that this allowed us to reduce multiple comparisons and harness local LD structure to improve power to capture untyped SNPs. Using genes as the genomic features of interest allows us to map approximately 310K SNPs to 18K SNP sets. However, it may be that the causal SNP lies far from a known gene in which case groupings based on genes (and pathways by extension) will fail to capture the effect of interest. To augment coverage of gene desert regions, we can group SNPs based on additional genomic features such as evolutionarily conserved regions. Such groupings again allow us to harness local correlation. The moving window approach will be useful for capturing all genotyped SNPs, but direct interpretation of SNP set analysis results are more difficult, though this may not be important. Groupings via haplotype blocks are attractive since they make explicit use of the LD information. Use of haplotype blocks will allow for comprehensive coverage of the entire genome and remove the need to explicitly predefine genomic features of interest.

Beyond harnessing local LD structure to boost power, another important feature of our approach is the ability to model the joint effect of multiple independent causal signals as well as possible epistatic effects. Practically, however, finding a SNP set formation strategy that optimizes for this can be difficult. Using a gene or moving window strategy can certainly capture multi-SNP and epistatic effects among SNPs that are located close to one another on the genome, but identification of such signals among SNPs that are distantly placed will not be possible. A potential strategy is to use existing prior biological knowledge. In particular, if multiple SNPs are expected to affect the disease risk, it is not unreasonable to expect them to lie within genes in the same pathway or genes with similar function; hence, forming SNP sets based on pathways can potentially capture such effects. Unfortunately, a systematic approach for identifying such grouping structures at the genome wide level is not obvious. To avoid bias in our testing procedure, any grouping strategy must be made without consideration of the case-control status of the subjects in the data set. Thus, groupings must be made using information from external sources, prior studies, or unsupervised statistical methods. As such, SNP set formation strategies will improve with advances in our knowledge of the genome and genomic structures.

Although we focused our power simulations on the linear kernel, our simulation results nevertheless suggest that our approach is as powerful as individual SNP analysis and our approach can often have improved power over both the individual SNP analysis strategy and other multi-SNP testing methods. In particular, we are able to show that when the causal SNP is correlated with multiple typed SNPs, our approach has higher power than individual SNP analysis. In settings where the causal SNP is not correlated with multiple typed SNPs, simulations show neither individual SNP analysis nor our approach will be able to detect an effect. Recall that, here, the term individual SNP analysis refers to correcting the smallest individual p-value for the SNPs in the SNP set for multiple comparisons and using the adjusted p-value as the p-value for the entire SNP set. The minimum uncorrected p-value for a SNP set may be smaller than the p-value from the logistic kernel machine test but would lead to significantly inflated type I error rate. Under several settings, we found the kernel machine test tended to have improved power over competing multi-SNP tests while naturally allowing for covariate adjustment to protect the type I error rate when confounders are present.

We noted earlier that the linear kernel corresponds to the usual simple logistic model whereas the IBS and weighted IBS are kernels tailored specifically to genetic data and the quadratic kernel is potentially useful for modelling epistatic effects. In fact, when epistatic effects are present, the IBS kernel can allow for dramatically improved power over the linear kernel. The ability to allow for complex relationships between the SNPs by just specifying a single distance metric is an attractive feature of our approach. In practice, however, one needs to choose a kernel *a priori*. Although our simulations demonstrated that the size of our test is correct irrespective of the kernel used, the power will be influenced by the choice of kernel. The best way to choosing a kernel to use is unclear since methods using the data to be tested are likely to overfit and simulations may reflect the process under which the data were simulated. Our experience in simulations and real data applications suggests using the linear kernel for testing SNP sets in which no epistatic effects are anticipated (such SNP sets based on short regions) and the IBS kernel, otherwise. Our experience is that there is a small loss in power for using the IBS kernel when the true effect is linear, but potentially a considerable loss in power when the true effect is complex/epistatic and the linear kernel is applied. Future research is necessary to study the power using other types of kernels.

Our numerical results lead us to recommend our kernel machine approach for performing multi-SNP analysis across a range of realistic settings. We have shown that it has more power compared to existing popular approaches. It also has the ability to adjust for covariates. This is particularly attractive since one usually needs to control for possible population stratification and additional confounders in association studies. As noted by Mukhopadhyay et al.,[24] the performance of individual multi-SNP tests can depend on a range of factors including the number of causal SNPs, effect size, and LD structure. Future research is needed for more comprehensive comparisons, e.g. in other settings and with other multi-SNP methods.

For a SNP set that is significantly associated with disease susceptibility, it is of great interest to subsequently perform fine mapping and identify the individual causal variants.

One strategy that can be used is to apply a variable selection procedure to select the "most important" SNPs. For instance, one could use a LASSO penalized logistic regression[44] to regress the case-control status on the 14 SNPs in the *ASAH1* SNP set. LASSO penalized logistic regression will cause some of the regression coefficients to be estimated as exactly zero, dropping the corresponding variables from the model. Such a strategy has been used by others.[45–47] However, existing variable selection literature does not allow for selection of features within the logistic kernel machine regression framework in the presence of SNP-SNP interactions. The optimal strategy for quantifying the contributions of individual SNPs remains an area of considerable interest.

In addition to being able to account for complex SNP effects and adjust for covariates, the key advantage of the logistic kernel machine test is the ability to adaptively estimate the degrees of freedom. As discussed earlier, when the genotyped SNPs are highly correlated, the degrees of freedom of the test remain approximately constant. As a result, the strength of our method can increase as progress in genotyping technology allows for denser screens.

Appendix

Approximating the Null Distribution of the Score Statistic for the Logistic Kernel Machine Test

The score statistic Q defined by Equation 3 tests the null hypothesis that $H_0 : \tau = 0$ and is based on the variance components tests developed by Zhang and Lin[48] and Lin[49] and adapted by Liu et al.[12] Note that this is a boundary case, so the null distribution for Q follows a complex mixture of χ^2. This can be approximated via the Satterthwaite method[50] as a scaled chi-squared distribution, $\kappa\chi^2_\nu$, where the scale parameter, κ, and the degrees of freedom, ν, are calculated via moment matching. Specifically, for $\mathbf{D}_0 = \text{diag}(\widehat{p}_{0_i}(1 - \widehat{p}_{0_i}))$ and $\mathbf{P}_0 = \mathbf{D}_0 - \mathbf{D}_0\mathbf{X}(\mathbf{X}'\mathbf{D}_0\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_0$, we define $\mu_Q = \text{tr}(\mathbf{P}_0\mathbf{K})/2$, $I_{\tau\tau} = \text{tr}(\mathbf{K}\mathbf{P}_0\mathbf{K}\mathbf{P}_0)/2$, $I_{\tau\sigma} = \text{tr}(\mathbf{P}_0\mathbf{K}\mathbf{P}_0)/2$, $I_{\sigma\sigma} = \text{tr}(\mathbf{P}_0\mathbf{P}_0')/2$, and $\widetilde{I}_{\tau\tau} = I_{\tau\tau} - I_{\tau\sigma}^2/I_{\sigma\sigma}$. Then κ can be estimated as $\kappa = \widetilde{I}_{\tau\tau}/(2\mu_Q)$ and we can calculate the p-value for significance by comparing Q/κ to a chi-square distribution of ν degrees of freedom, χ^2_ν, where $\nu = 2\mu_Q^2/\widetilde{I}_{\tau\tau}$. The original derivation of our score test can be found in Lin,[49] where the link function in Equation 2 of Lin is assumed to be the logit and the design matrix (**Z**) is set to be $\mathbf{K}^{1/2}$. Our score statistic, Q, in Equation 4 is identical to the first term of the score statistic, U, from Equation 8 of Lin (as $\dot{\mathbf{D}} = 1$ and $\Delta^{-1}\mathbf{W} = \mathbf{D}$ since the logit link is a canonical link).

Acknowledgements

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

Web Resources

The URLs for data presented herein are as follows:

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

HAPGEN program: `http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html`

R-functions for the logistic kernel machine test: `http://www.bios.unc.edu/~mwu/software/`

References

[1] Hunter, D., Kraft, P., Jacobs, K., Cox, D., Yeager, M., Hankinson, S., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., et al. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature Genetics, *39*, 870–4.

[2] Easton, D., Pooley, K., Dunning, A., Pharoah, P., Thompson, D., Ballinger, D., Struewing, J., Morrison, J., Field, H., Luben, R., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. Nature, *447*, 1087–1093.

[3] Yeager, M., Orr, N., Hayes, R., Jacobs, K., Kraft, P., Wacholder, S., Minichiello, M., Fearnhead, P., Yu, K., Chatterjee, N., et al. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nature Genetics, *39*, 645–649.

[4] Gudmundsson, J., Sulem, P., Manolescu, A., Amundadottir, L., Gudbjartsson, D., Helgason, A., Rafnar, T., Bergthorsson, J., Agnarsson, B., Baker, A., et al. (2007). Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nature Genetics, *39*, 631–637.

[5] Thomas, G., Jacobs, K., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., et al. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. Nature Genetics, *40*, 310–315.

[6] Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature, *445*, 881–885.

[7] Scott, L., Mohlke, K., Bonnycastle, L., Willer, C., Li, Y., Duren, W., Erdos, M., Stringham, H., Chines, P., Jackson, A., et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science, *316*, 1341–5.

[8] Saxena, R., Voight, B., Lyssenko, V., Burtt, N., de Bakker, P., Chen, H., Roix, J., Kathiresan, S., Hirschhorn, J., Daly, M., et al. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science, *316*, 1331–1336.

[9] Kraft, P. and Cox, D. (2008). Study designs for genome-wide association studies. Advances in Genetics, *60*, 465.

[10] Schaid, D., Rowland, C., Tines, D., Jacobson, R., and Poland, G. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. The American Journal of Human Genetics, *70*, 425–434.

[11] Hunter, D. and Kraft, P. (2007). Drinking from the fire hose–statistical issues in genomewide association studies. New England Journal of Medicine, *357*, 443–53.

[12] Liu, D., Ghosh, D., and Lin, X. (2008). 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, *9*.

[13] Kwee, L., Liu, D., Lin, X., Ghosh, D., and Epstein, M. (2008). A powerful and flexible multilocus association test for quantitative traits. The American Journal of Human Genetics, *82*, 386–397.

[14] Schaid, D., McDonnell, S., Hebbring, S., Cunningham, J., and Thibodeau, S. (2005). Nonparametric tests of association of multiple genes with human disease. The American Journal of Human Genetics, *76*, 780–793.

[15] Wessel, J. and Schork, N. (2006). Generalized Genomic Distance Based Regression Methodology for Multilocus Association Analysis. The American Journal of Human Genetics, *79*, 792–806.

[16] Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research, *27*, 29–34.

[17] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics, *25*, 25–29.

[18] McAuliffe, J., Pachter, L., and Jordan, M. (2004). Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. Bioinformatics, *20*, 1850–1860.

[19] Barrett, J., Fry, B., Maller, J., and Daly, M. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics, *21*, 263–265.

[20] Cristianini, N. and Shawe-Taylor, J. (2000). An introduction to support Vector Machines: and other kernel-based learning methods. (Cambridge Univ Pr).

[21] Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences, *97*, 262–267.

[22] Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions(Tchebycheffian spline functions, solving Hermite-Birkhoff interpolation as stochastic prediction and filtering). Journal of Mathematical Analysis and Applications, *33*, 82–95.

[23] Lin, W. and Schaid, D. (2009). Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. Genetic Epidemiology, *33*, 183–97.

[24] Mukhopadhyay, I., Feingold, E., Weeks, D., and Thalamuthu, A. (2009). Association tests using kernel-based measures of multi-locus genotype similarity between individuals. Genetic Epidemiology, *34*, 213–21.

[25] Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. Biostatistics, *4*, 57–74.

[26] Lin, D. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics, *21*, 781–787.

[27] Cheverud, J. (2001). A simple correction for multiple comparisons in interval mapping genome scans. Heredity, *87*, 52–58.

[28] Nyholt, D. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. The American Journal of Human Genetics, *74*, 765–769.

[29] Moskvina, V. and Schmidt, K. (2008). On multiple-testing correction in genome-wide association studies. Genetic Epidemiology, *32*, 567–73.

[30] Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. Nature Reviews Genetics, *4*, 701–709.

[31] Zaykin, D., Westfall, P., Young, S., Karnoub, M., Wagner, M., Ehm, M., and Inc, G. (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Human Hered, *53*, 79–91.

[32] Chapman, J., Cooper, J., Todd, J., and Clayton, D. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Human Hered, *56*, 18–31.

[33] Roeder, K., Bacanu, S., Sonpar, V., Zhang, X., and Devlin, B. (2005). Analysis of single-locus tests to detect gene/disease associations. Genetic Epidemiology, *28*, 207–219.

[34] Tzeng, J. and Zhang, D. (2007). Haplotype-based association analysis via variance-components score test. The American Journal of Human Genetics, *81*, 927–938.

[35] Minichiello, M. and Durbin, R. (2006). Mapping trait loci by use of inferred ancestral recombination graphs. The American Journal of Human Genetics, *79*, 910–922.

[36] Tachmazidou, I., Verzilli, C., and De Iorio, M. (2007). Genetic association mapping via evolution-based clustering of haplotypes. PLoS Genetics, *3*, e111.

[37] Saad, A., Meacham, W., Bai, A., Anelli, V., Elojeimy, S., Mahdy, A., Turner, L., Cheng, J., Bielawska, A., Bielawski, J., et al. (2007). The functional effects of acid ceramidase overexpression in prostate cancer progression and resistance to chemotherapy. Cancer Biology & Therapy, *6*, 1455–60.

[38] Li, C., Park, J., He, X., Levy, B., Chen, F., Arai, K., Adler, D., Disteche, C., Koch, J., Sandhoff, K., et al. (1999). The human acid ceramidase gene (ASAH): structure, chromosomal location, mutation analysis, and expression. Genomics, *62*, 223–231.

[39] Spencer, C., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genetics, *5*.

[40] Altschuler, D., Brooks, L., Chakravarti, A., Collins, F., Daly, M., and Donnelly, P. (2005). International HapMap Consortium. A haplotype map of the human genome. Nature, *437*, 1299–320.

[41] Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics, *39*, 906–913.

[42] Gao, X., Starmer, J., and Martin, E. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genetic Epidemiology, *32*, 361–9.

[43] Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, *38*, 904–909.

[44] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), *58*, 267–288.

[45] Devlin, B., Roeder, K., and Wasserman, L. (2003). Analysis of multilocus models of association. Genetic Epidemiology, *25*, 36–47.

[46] Croiseau, P. and Cordell, H. (2009). Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach. In BMC Proceedings, volume 3, BioMed Central Ltd, pp. S61.

[47] Szymczak, S., Biernacka, J., Cordell, H., González-Recio, O., K
"onig, I., Zhang, H., and Sun, Y. (2009). Machine learning in genome-wide association studies. Genetic Epidemiology, *33*, S51–S57.

[48] Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. Biostatistics, *4*, 57–74.

[49] Lin, X. (1997). Variance component testing in generalised linear models with random effects. Biometrika, *84*, 309–326.

[50] Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. Biometrics Bulletin, *2*, 110–14.

Table 1: Empirical type-I error rates at α=0.05 for the logistic kernel machine test and individual SNP analysis when applied to SNP sets simulated from the *ASAH1* gene.

	Individual	Logististic Kernel Machine Test		
n	SNP Analysis	Linear Kernel	IBS Kernel	Weighted IBS Kernel
1000	0.049	0.052	0.046	0.055
2000	0.048	0.047	0.053	0.052

Table 2: Top results from the discovery phase of the CGEMS breast cancer GWAS.

SNP	Chromosome	Gene	p-value
rs10510126	10		2.0×10^{-6}
rs12505080	4		8.0×10^{-6}
rs17157903	7	*RELN*	9.0×10^{-6}
rs1219648	10	*FGFR2*	1.2×10^{-5}
rs7696175	4	*TLR1—TLR6*	1.4×10^{-5}
rs2420946	10	*FGFR2*	1.5×10^{-5}
rs2107349	7	*AZGP1—AZGP1P2*	1.7×10^{-5}
rs6497337	16	*SYT17*	2.0×10^{-5}
rs1250255	2	*FN1*	3.4×10^{-5}
rs10804287	2		3.8×10^{-5}

Table 3: Top results from the logistic kernel machine based SNP set analysis of the CGEMs Breast Cancer Study data.

	Linear		IBS		Weighted IBS	
Gene	p-value	q-value	p-value	q-value	p-value	q-value
FGFR2	7.69×10^{-7}	0.01	2.53×10^{-6}	0.03	1.35×10^{-5}	0.05
CNGA3	5.59×10^{-6}	0.05	4.65×10^{-6}	0.03	3.25×10^{-6}	0.02
TBK1	1.30×10^{-5}	0.07	3.28×10^{-6}	0.03	5.48×10^{-6}	0.02
VWA3B	1.53×10^{-5}	0.07	7.84×10^{-6}	0.03	3.99×10^{-6}	0.02
PTCD3	5.50×10^{-5}	0.20	9.02×10^{-6}	0.03	3.78×10^{-6}	0.02
XPOT	6.60×10^{-5}	0.20	3.48×10^{-5}	0.09	4.91×10^{-5}	0.11
VAPB	9.79×10^{-5}	0.22	4.51×10^{-5}	0.10	8.11×10^{-5}	0.14
SHC3	1.01×10^{-4}	0.22	3.77×10^{-4}	0.34	1.61×10^{-3}	0.46
SFTPB	1.78×10^{-4}	0.31	1.38×10^{-4}	0.27	7.62×10^{-5}	0.14
SPATA7	1.90×10^{-4}	0.31	1.76×10^{-4}	0.28	1.39×10^{-4}	0.22

Figure 1: Empirical Power for SNP sets based on *ASAH1* and LD-plot for the 86 SNPs in the *ASAH1* gene based on the CEU sample from the International HapMap Project. The typed SNPs are denoted with a triangle and the bottom panel shows the LD-structure of the SNPs in the *ASAH1* gene.

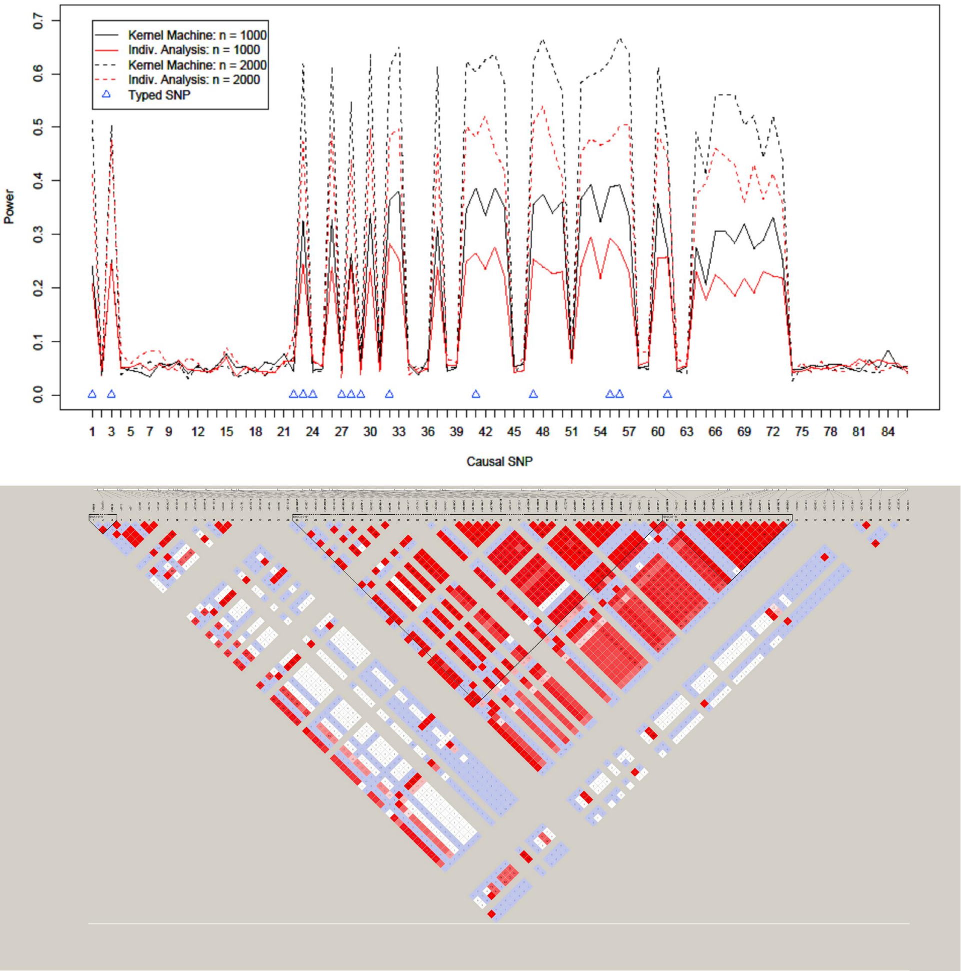


Figure 2: Empirical Power for SNP sets based on *ASAH1*. The SNPs on the x-axis are sorted by median R^2 with the 14 typed SNPs which is shown in the bottom plot.

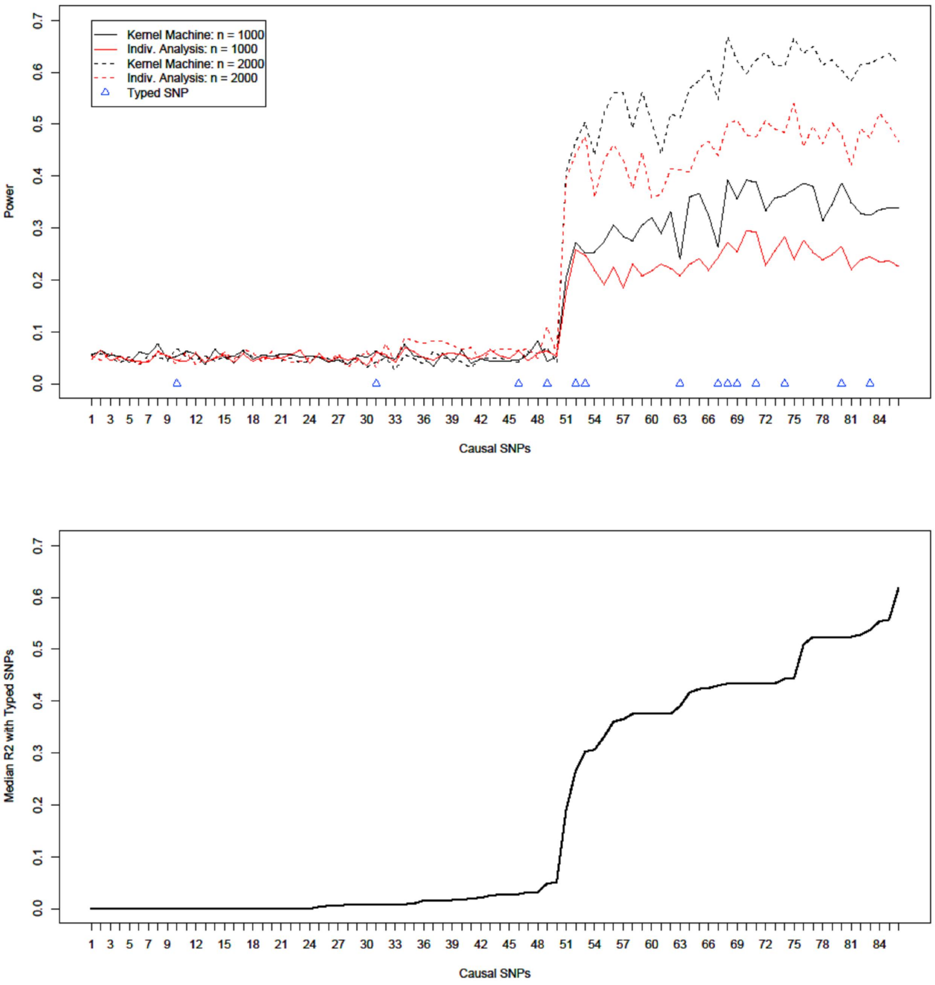
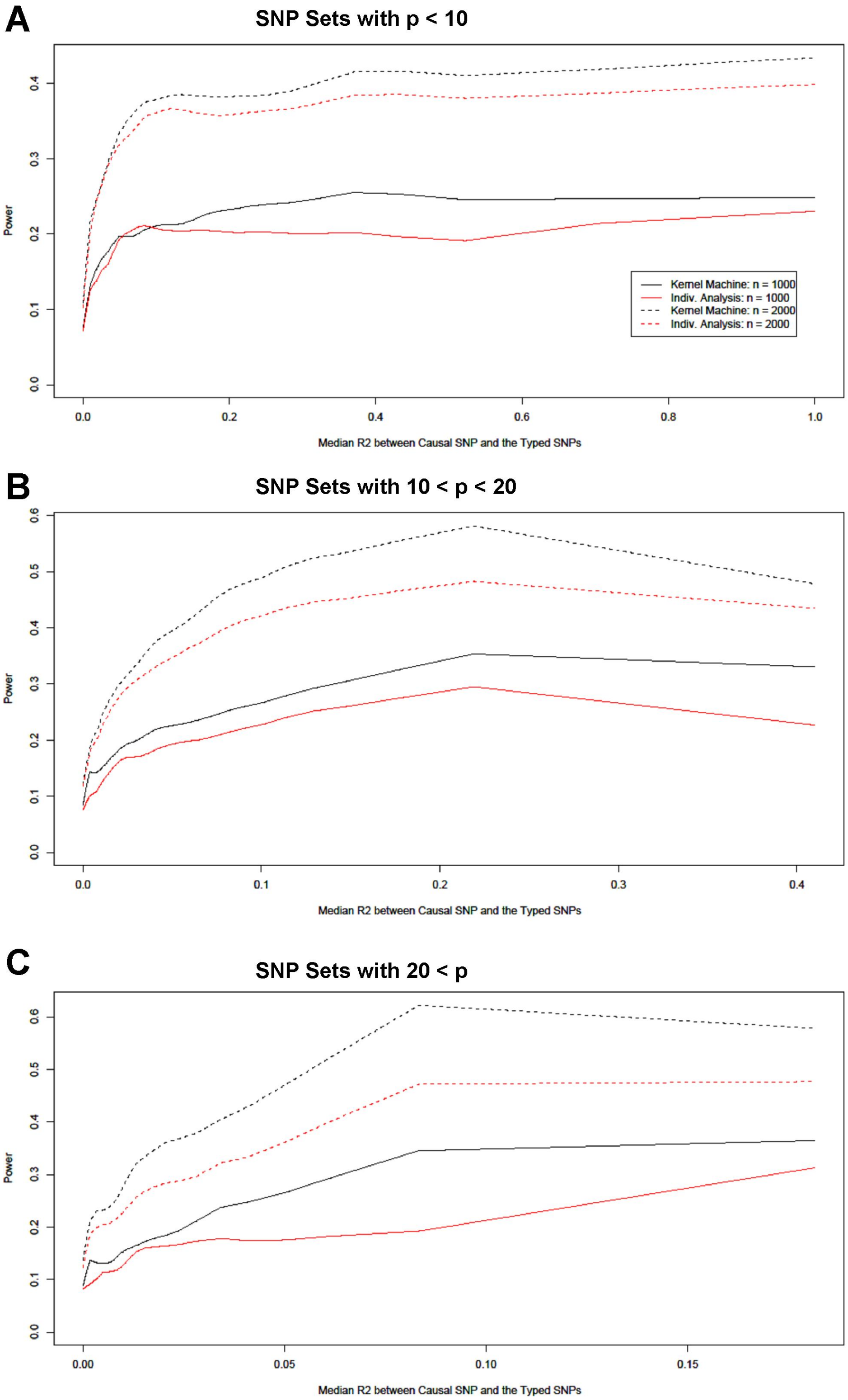


Figure 3: Smoothed Empirical Power Curves as a function of media R^2 between the causal SNP and the typed SNP for SNP sets based on a range of genes.

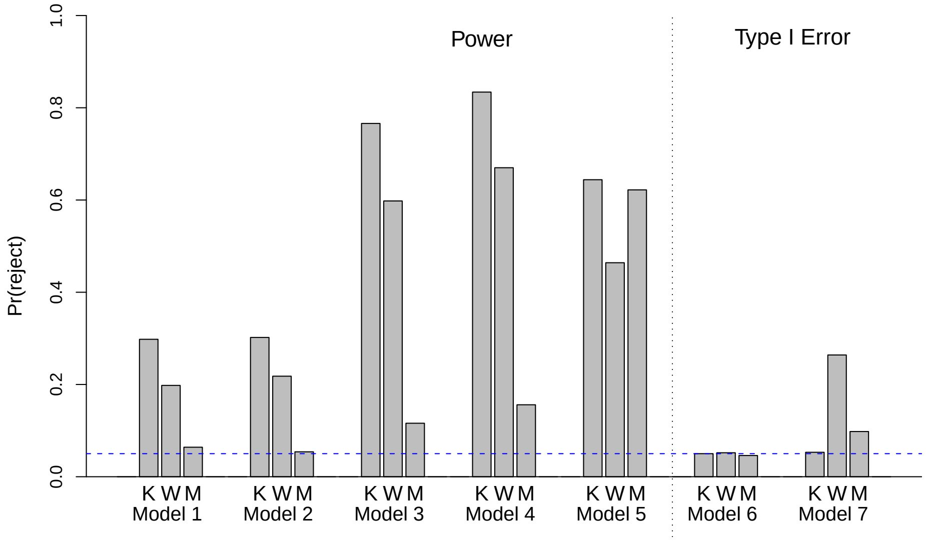


A SNP Sets with p < 10

B SNP Sets with 10 < p < 20

C SNP Sets with 20 < p

Figure 4: Comparison of the power and type I error of the Logistic Kernel Machine Test, the Wessel and Schork method, and Mukhopadhyay et al.'s approach, denoted by K, W, and M respectively, under seven models. Power and size estimates are based on 500 and 1000 simulations, respectively. The blue line shows the expected type I error rate.


Power
Type I Error
Pr(reject)
K W M
Model 1
K W M
Model 2
K W M
Model 3
K W M
Model 4
K W M
Model 5
K W M
Model 6
K W M
Model 7