BIOINFORMATICS

Vol. 28 ECCB 2012, pages i375–i381
doi:10.1093/bioinformatics/bts406

Nonlinear dimension reduction with Wright–Fisher kernel for genotype aggregation and association mapping

Hongjie Zhu[1,*] Lexin Li[2] and Hua Zhou[2]

[1]Department of Psychiatry and Behavior Science, Duke University, Durham, NC 27710 and
[2]Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA

ABSTRACT

Motivation: Association tests based on next-generation sequencing data are often under-powered due to the presence of rare variants and large amount of neutral or protective variants. A successful strategy is to aggregate genetic information within meaningful single-nucleotide polymorphism (SNP) sets, e.g. genes or pathways, and test association on SNP sets. Many existing methods for group-wise tests require specific assumptions about the direction of individual SNP effects and/or perform poorly in the presence of interactions.

Results: We propose a joint association test strategy based on two key components: a nonlinear supervised dimension reduction approach for effective SNP information aggregation and a novel kernel specially designed for qualitative genotype data. The new test demonstrates superior performance in identifying causal genes over existing methods across a large variety of disease models simulated from sequence data of real genes. In general, the proposed method provides an association test strategy that can (i) detect both rare and common causal variants, (ii) deal with both additive and interaction effect, (iii) handle both quantitative traits and disease dichotomies and (iv) incorporate non-genetic covariates. In addition, the new kernel can potentially boost the power of the entire family of kernel-based methods for genetic data analysis.

Availability: The method is implemented in MATLAB. Source code is available upon request.

Contact: hongjie.zhu@duke.edu

1 INTRODUCTION

Genome-wide association studies (GWASs) based on single-nucleotide polymorphism (SNP) chips have enjoyed varying degrees of success in identifying genes associated with complex diseases or traits (Easton and Eeles, 2008; Frazer *et al.*, 2009; Lettre and Rioux, 2008). It has now been widely accepted that standard GWAS explains at most a small fraction of the population variation of most complex traits (Frazer *et al.*, 2009). Recently, deep resequencing is emerging as a new and potent means for mapping complex trait genes. Resequencing delivers orders of magnitude more variants than SNP chips and include both common variants with minor allele frequency (MAF) >10%, as well as rare variants with MAF <1%. Availability of rare variant information presents unique opportunities to evaluate the 'common disease rare variants' hypothesis. This hypothesis states that a complex disease can be attributed to multiple rare variants with relatively high risks, and it has attracted much attention in recent studies. Attesting to this hypothesis, a number of deleterious or protective rare variants have been identified for low plasma levels of high-density lipoprotein cholesterol (Cohen *et al.*, 2004), hypertension (Ji *et al.*, 2008) and Type-I diabetes (Nejentsev *et al.*, 2009).

Identifying disease associated rare variants, however, is challenging, because a particular rare disease predisposing allele may be present in only a handful of patients. Henceforth, traditional single marker tests that capture only marginal effects are doomed to have low power. A useful strategy to address this challenge is to effectively merge information in SNP variants by some meaningful SNP sets, for instance, genes or pathways, and then to identify disease-associated genes or pathways rather than disease variants. Following this idea, several aggregation-based association test approaches have been developed. Li and Leal (2008) proposed a group-wise test exploiting both multivariate and collapsing strategies that possess higher power than a simple multivariate test or simple collapsing. Madsen and Browning (2009) extended the method by incorporating weights that depend on MAF into the group-wise statistics and approximating *P*-values by permutations within each group. Both methods consider rare variants with MAF falling below a pre-specified threshold and exclude common variants from analysis. This separate treatment seems counterproductive because in reality both common and rare variants can be informative. The pooling strategy of Price *et al.* (2010) circumvents the issue of arbitrarily chosen frequency threshold by calculating a group-wise statistic under a variety of thresholds. However, this strategy also has several limitations. First, environmental predictors are excluded from analysis even though they may contribute significantly to a disease. Second, interactions among SNPs cannot be effectively detected. Third, the solution is sensitive to the classification of variants: if all types of variants, deleterious, protective or neutral, coexist, then various signals can cancel one another during the pooling and thus can potentially compromise statistical power. Liu and Leal (2010) proposed a SNP set genotype-based statistic for rare variants and declared that the common variants and environment factors can be modeled together with the rare variant statistic in a logistic regression (LR) model. However, interactions between rare and common variants still cannot be explicitly modeled in this way, and the method handles only dichotomous traits.

The field of statistical dimension reduction (DR) offers a useful and appealing means for genotype aggregation. It is based on the belief that high-dimensional data can be effectively summarized in a low-dimensional space, and the subsequent modeling can concentrate on the reduced space. The most commonly used dimension reduction approach is principal components analysis (PCA). Chen *et al.* (2010) applied PCA to combine SNP information within pathways, generated the so-called eigen-SNPs, and used eigen-SNPs in subsequent association mapping. However, PCA has at least two limitations. First, PCA aggregates SNPs regardless of

*To whom correspondence should be addressed.

the phenotypic trait information. Since mapping traits to associated genes is of ultimate interest, it is intuitively desirable to aggregate SNPs under the guidance of trait information such as disease status or quantitative traits. In statistical terms, PCA is an 'unsupervised' DR solution, whereas a 'supervised' DR solution is preferred. Second, the eigen-SNPs, or principle components, are 'linear' combinations of the SNPs. As a consequence, such summary measures may fail to capture complex interacting effects among the individual SNPs, and in turn reduce the power of subsequent association mapping.

In this article, we develop a powerful association mapping approach based on next-generation sequencing data. Two key ingredients of the proposed method are a statistical DR that achieves supervised and nonlinear reduction, and a new kernel that is based on Markov chain theory and particularly suitable for qualitative SNP data. Our contributions are mainly 2-folds. First, the proposed association mapping approach simultaneously takes into account (i) both rare and common variants, (ii) both additive and interaction effect, (iii) quantitative traits as well as disease dichotomies and (iv) non-genetic covariates. Second, the commonly used kernels, such as Gaussian, polynomial and spline, work successfully with continuous attributes, but may perform poorly for discrete genetic data. The proposed new kernel is specially designed for discrete attributes and can effectively capture the similarity between individual genotypes. Moreover, the new kernel is novelly derived from powerful Markov chain theory and can benefit many kernel-based learning methods in general. We compare our proposal with some state-of-the-art aggregation and mapping solutions and find that our method clearly achieves superior power in a variety of different genetic model scenarios.

The remainder of the article is organized as follows. Section 2 describes our proposed association mapping, including kernel-based nonlinear DR and construction of new kernels. Section 3 presents numerical studies comparing various aggregation and association mapping solutions. Section 4 concludes the article with a discussion and suggests potential future extensions.

2 METHODS

Suppose n study individuals are genotyped at a SNP set (e.g. a gene) that is composed of p SNPs denoted by $X_1,\ldots,X_p$, and the trait Y can be either binary (case–control study) or quantitative. Potential non-genetic covariates, such as sex, age smoke, and are denoted by $C_1,\ldots,C_t$. There can be multiple SNP sets, and we treat 'one set at a time'. The goal is to test the association of the trait and all markers in the SNP-set 'jointly', after adjusting for non-genetic covariates. In sequence studies, the number of markers p is potentially large and can outnumber the number of subjects n. Our proposed group-wise association test consists of two key elements: a nonlinear supervised DR method that aggregates markers and produces summary features and a novel kernel that encodes genomic similarity. The nonlinear DR method was first proposed in Zhu and Li (2011) for gene pathway analysis and for completeness, we review the method here. The new kernel is constructed under the guidance of Markov chain theory for discrete genotype. Using Markov chains to build kernels for non-standard data is novel and the resulting kernels can potentially benefit the entire family of kernel methods (e.g. support vector machines) for genetic data analysis.

2.1 Nonlinear supervised DR

We begin with a brief review of PCA, which has been an extremely popular tool in analysis of genetic and genomic data. For instance, PCA has been used to adjust for population stratification (Price *et al.*, 2006) in GWAS or to produce a set of eigen-SNPs for association mapping (Chen *et al.*, 2010). Given p SNPs, PCA seeks linear combinations of SNPs that have maximal variances. It is solved by an eigen decomposition of SNP covariance matrix. Then, the eigenvectors with leading eigenvalues give the coefficients of linear combinations being sought. The linear transformed SNPs form the eigen-SNPs used in the subsequent analysis. Despite its widespread applications, however, PCA conducts DR without utilization of the phenotypic trait information, and thus there is no guarantee that the top extracted principle components are relevant to the traits. Consider a simple illustrative example. Suppose two SNPs X_1 and X_2 are in linkage disequilibrium (LD). Then, the variance of the linear combination X_1+X_2 is larger than that of X_1-X_2 and thus the eigen-gene found by PCA will be closer to the direction X_1+X_2 than to X_1-X_2. If in truth these two SNPs have opposite effects—one deleterious the other protective—then the trait depends on the SNPs through X_1-X_2 and the eigen-gene would contain no signal for association.

Intuitively, it is natural to incorporate the trait information during the phase of DR, and this leads us to the family of 'supervised' sufficient dimensional reduction (SDR) approaches. For a regression of a response Y given a p-dimensional predictor X, SDR seeks a minimum number of linear combinations, $\eta_1^\tau X,\ldots,\eta_d^\tau X$, such that

$$Y \perp\!\!\!\perp X | (\eta_1^\tau X,\ldots,\eta_d^\tau X). \quad (1)$$

That is, Y depends on X only through those linear combinations, and one can replace the original p-dimensional X by now d-dimensional $\eta^\tau X$. In practice, d is often much smaller than p, and thus DR is achieved. We call $(\eta_1^\tau X,\ldots,\eta_d^\tau X)$ the 'linear sufficient predictors', which will serve as the induced summary features in subsequent analysis. There have been many methods proposed for SDR, many of which can be formulated as a generalized eigen decomposition problem. Specifically, a reduction estimate can be obtained by the first d eigenvectors η_js that correspond to the nonzero eigenvalues λ_js in a descending order from the decomposition: $\Omega_x\eta_j=\lambda_j\Sigma_x\eta_j,\quad j=1,\ldots,d$, where $\Sigma_x=\mathrm{Cov}(X)$ and Ω_x is a method-specific $p\times p$ semi-positive definite matrix (Zhu and Li, 2011). For instance, sliced inverse regression (SIR) (Li, 1991) is a widely used SDR estimator, where $\Omega_x=\mathrm{Cov}\{\mathrm{E}(X|Y)-\mathrm{E}(X)\}$. This family of DR methods differ from PCA in that the response information is used in the DR phase. It is also interesting to note that all those SDR methods impose no parametric assumption on $Y|X$. Instead, they require the marginal distribution of X to be elliptically symmetric. This is often viewed as a mild condition, since it holds approximately when p goes to infinity. We assume the condition holds since we are dealing with a very large p.

The above SDR methods yield 'linear' DR, because the reduction admits the form of linear combinations of X. This could have some limitations. Consider an illustrative example, where $X=(X_1,\ldots,X_6)$ and $Y=X_1+X_2X_3+X_4^2+X_5X_6+\varepsilon$, with an independent error ε. In this case, no linear reduction is possible. Another potential limitation is that one needs to invert a $p\times p$ covariance matrix Σ_x, whereas its sample estimator is not invertible when p exceeds the sample

size n. These observations motivate us to consider a 'nonlinear' DR strategy.

The basic idea is to use a function $\phi(\cdot)$, with an associated kernel matrix K, to map X to $\phi(X)$. One then performs a linear DR in the space of $\phi(X)$, which in effect results in a nonlinear DR in the original predictor space $\mathcal{X}$. The well-known kernel trick turns the primal problem that depends on the dimension of the space of $\phi(X)$, which is high or even infinite, to a dual problem that only depends on the sample size. Consequently, the method works for $n<p$. Specifically, in analogy to linear reduction in Equation (1), nonlinear DR seeks

$$Y \perp\!\!\!\perp X | (\langle \beta_1, \phi(X)\rangle, \ldots, \langle \beta_{\tilde{d}}, \phi(X)\rangle). \quad (2)$$

Comparing with Equation (1), the linear combinations $(\eta_1^\top X, \ldots, \eta_d^\top X)$ are replaced by the inner products $(\langle \beta_1, \phi(X)\rangle, \ldots, \langle \beta_{\tilde{d}}, \phi(X)\rangle)$, and Y depends on X only through those inner products. We refer them as the 'nonlinear sufficient predictors' and assume the number $\tilde{d}$ of inner products $\leq \min(n,p)$. In terms of estimation, conceptually, one can estimate β_js in a way analogous to linear reduction, i.e. through the eigen decomposition: $\Omega_\phi \beta_j = \rho_j \Sigma_\phi \beta_j, j=1,\ldots,\tilde{d}$, where $\Sigma_\phi = \mathrm{Cov}\{\phi(X)\}$ and Ω_ϕ is defined similarly as Ω_x except we replace X with $\phi(X)$. Given $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, estimation of β_js is obtained by substituting in the corresponding sample counterparts: $\hat{\Omega}_\phi \beta_j = \rho_j \hat{\Sigma}_\phi \beta_j$, $j=1,\ldots,\tilde{d}$. On the other hand, the dimension of the induced mapping $\phi(X)$ can be very high, sometimes even infinite. As such, a direct decomposition is not feasible computationally.

The problem can be solved by noting that the target of nonlinear DR estimation are the inner products $\langle \beta_j, \phi(X)\rangle$, rather than β_j themselves. Then, given a pre-specified kernel function k, these inner products can be obtained by solving a dual problem: $\bar{K} J \bar{K} \alpha_j = \rho_j \bar{K}^2 \alpha_j$, $j=1,\ldots,\tilde{d}$, where $\bar{K} \in \mathbb{R}^{n\times n}$ is the centered kernel matrix and J is a method-specific $n\times n$ matrix. Zhu and Li (2011) gave for the specification of J matrix for different SDR methods, including the one for kernel SIR that will be used in the numerical studies of this article. Then, for a new observation $x \in \mathcal{X}$, $\langle \beta_j, \phi(x)\rangle = \alpha_j^\top [\tilde{k}(x_1,x),\ldots,\tilde{k}(x_n,x)]^\top$, where $\tilde{k}(x_i,x) = k(x_i,x) - n^{-1}\sum_{l=1}^n k(x_l,x), i=1,\ldots,n$. So the inner product $\langle \beta_j, \phi(x)\rangle$ can be obtained from the kernel k and α_js. It is noted that the proposed nonlinear DR approach only involves decomposition of an $n\times n$ matrix, so it can handle $n<p$. Its flexible reduction form beyond the linear combination is also expected to facilitate DR. For the illustrative example considered above, if one uses a quadratic kernel, then only one linear combination in the mapped feature space is needed to summarize all regression information, and thus substantial reduction is achieved.

2.2 Kernel-based on Markov chain

Critical to success of any kernel-based methods for genotype data analysis is the design of kernels that can effectively capture genomic similarity (Schaid, 2010a,b). The most popular Gaussian kernel works well for continuous predictors but can perform poorly on categorical predictors such as SNPs. Some specialized kernels have been crafted for SNP data. For instance, the identity-by-state (IBS) kernel (Wessel and Schork, 2006) calculates the distance between two individuals with genotype vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ coded by numbers of minor alleles as

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \frac{\sum_{s=1}^p 2I(x_{is}=x_{js}) + I(|x_{is}-x_{js}|=1)}{2p}.$$

The more general weighted IBS kernel (Kwee *et al.*, 2008; Wu *et al.*, 2010) takes the form

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j) = \frac{\sum_{s=1}^p w_s(2I(x_{is}=x_{js}) + I(|x_{is}-x_{js}|=1))}{2\sum_s^p w_s},$$

which offers flexibility of incorporating variant specific weights into kernel. Kwee *et al.* (2008) and Wu *et al.* (2010) use $w_s = 1/\sqrt{f_s}$ where f_s is the MAF of the variant, up-weighting the importance of rare variants.

IBS and weighted IBS kernels can be regarded as sums of simple similarities evaluated at each individual SNP, which, however, may result in loss of power when complex interactions dominate the genetic effect of the SNP set. This motivates us to propose new kernels for genotype data.

We first summarize some useful devices for forming Mercer (symmetric and positive definite) kernels. Suppose k is a possibly asymmetric kernel and K is the corresponding kernel matrix with all eigenvalues positive, then

1. $(K+K^T)/2$ is symmetric and positive definite;
2. KK^T is a symmetric and positive definite;
3. $K \circ K^T$ is symmetric and positive definite, where $\circ$ denotes the element-wise Hadamad product of two matrices;
4. If another Q is a kernel matrix with positive eigenvalues, then KQ is a kernel with positive eigenvalues.

Fact 1 follows from an inequality due to Fan (Marshall *et al.*, 2011, Theorem F.1, p. 324). Fact 2 is trivial. Fact 3 can be shown based on the Shur's inequality (Scholkopf and Smola, 2001, Proposition 13.2). Fact 4 follows from the inequality H.1.i of Marshall *et al.* (2011). These rules will be used as we construct Markov kernels below. Reader are referred to Scholkopf and Smola (2001) Chapter 13 for more basic tricks for constructing kernels.

Since SNP values are all dichotomous or trichotomous, we next focus on kernels that are built upon the discrete state space $\mathcal{X} = \{0,1,2\}^p = \{\boldsymbol{x} = (x_1,\ldots,x_p) : x_j \in \{0,1,2\}\}$. Our new kernel for genotype data is derived from Markov chains on $\mathcal{X}$. The key idea is that the transition kernel of many Markov chains with state space $\mathcal{X}$ defines a Mercer kernel after appropriate transformations. This opens the door to create more informative kernels for data in a non-standard space $\mathcal{X}$. Specifically, our new kernel is based on the well-known multi-allele Wright–Fisher (WF) process (Ewens, 2004) in population genetics. Therefore, we call this the WF kernel. Each locus is coded by the number of 'major' alleles and the genotype vector $\boldsymbol{x} \in \mathcal{X}$ for a SNP set is modeled by a Dirichlet-Multinomial distribution with parameter $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_p)$, which can be efficiently estimated from genotypes of all individuals (Zhou and Lange, 2010). The transition kernel of WF process between two genotype vectors $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$ is

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j) = \binom{\sum_{t=1}^p x_{jt}}{\boldsymbol{x}_j} \prod_{t=1}^p \pi(x_{it})^{x_{jt}},$$

where $\pi(x_{it})=(x_{it}+\alpha_t)/N+|\boldsymbol{\alpha}|$ and $|\boldsymbol{\alpha}|=\sum_{t=1}^{p}\alpha_t$. k is an irreversible Markov transition kernel and its stationary distribution is not known explicitly but can be well approximated by a Dirichlet-multinomial distribution with parameter $\boldsymbol{\alpha}$. It is well known that k has positive eigenvalues $\lambda_i=(2p)_{[i]}/(2p)^i=2p(2p-1)\cdots(2p-i+1)/(2p)^i$ for $i=0,\ldots,2p-1$ (Cannings, 1974). Therefore, we aim to form a Mercer kernel by scaling and symmetrizing k. For large p, the entries of the corresponding kernel matrix K is small. To achieve better scaling, we normalize each column of K by dividing its ℓ_2 norm. The column-scaled matrix is denoted by $\tilde{K}$. By Fact 4 above, we know $\tilde{K}$ has positive eigenvalues. Then, by Facts 1 and 2, either the additive or multiplicative symmetrization gives a symmetric kernel matrix with all positive eigenvalues: $K=\tilde{K}+\tilde{K}^T/2$ or $K=\tilde{K}\tilde{K}^T$. The simulation study in Section 3.3 shows the promise of the WF kernel combined with nonlinear DR method proposed in Section 2.

2.3 A joint association testing strategy

Based on the nonlinear DR methods and the new kernel, we propose a joint strategy to identify SNP sets that are associated with a trait of interest in Algorithm 1. Several remarks are in order. First, as has been mentioned above, the nonlinear DR methods can be performed even if the sample size, n, is smaller than the number of predictors, p. Therefore, the entire strategy handles the situation when n is smaller than the number of SNPs in any candidate SNP sets. Second, the non-genetic covariates, if any, can be naturally adjusted in the GLM modeling step. Third, both the DR and the GLM modeling step permit different types of trait response. For a continuous trait, GLM reduces to a linear regression, and for a typical case–control study, it becomes a binomial model with a logit link, which leads to a LR.

Designate the number of permutations, B
Divide the entire dataset into G SNP-sets by genes or pathways
for $i=1\rightarrow G$ **do**
  1. Conduct nonlinear dimension reduction of the *real* trait given all the SNPs in the i-th SNP-set, and obtain the nonlinear sufficient predictors, $Z_{i,1},\ldots,Z_{i,\tilde{d}_i}$
  2. Fit a generalized linear model (GLM) for the *real* trait with $Z_{i,1},\ldots,Z_{i,\tilde{d}_i}$ as predictors; obtain a F-statistic, F_i^{Real}, for the nonlinear sufficient predictors
  for $j=1\rightarrow B$ **do**
    Permute the *real* trait
    Run steps 1 and 2 above for the permuted trait, and obtain a F-statistic, F_j^{Permu}, for the j-th permutation
  end for
  Count the number of F_j^{Permu}'s larger than F_i^{Real}, B_c. The empirical p-value for the i-th SNP-set is B_c/B
  Declare significance of the i-th SNP-set if $B_c/B<0.05/G$
end for

Algorithm 1 A joint association testing strategy for whole-genome sequencing data

3 RESULTS

3.1 Data description

We have performed simulation studies to illustrate the promise of the information aggregation method using the nonlinear DR and the WF kernel discussed above. These studies use real genotype data of 697 individuals compiled from the 1000 Genome Project (2010) by the Genetic Analysis Workshop 17 (GAW17). We investigate the empirical power and Type-I error of our method on the basis of the LD structure of two genes, *TG* and *COL6A3*. The two genes are used because mutations in these genes have been found to cause certain diseases. *TG* encodes a protein called thyroglobulin; mutations in this gene have been found related to congenital hypothyroidism and autoimmune disorders (`http://ghr.nlm.nih.gov/gene/TG`). *COL6A3* encodes one component of Type-VI collagen; mutations in this gene have been found related to Bethlem myopathy and Ullrich congenital muscular dystrophy (`http://ghr.nlm.nih.gov/gene/COL6A3`). In the dataset, *TG* contains 146 SNPs, among which 10 are common variants (MAF>10%) and 113 are rare variants (MAF<1%); *COL6A3* contains 187 SNPs, among which 10 are common variants and 143 are rare variants. In the two genes, only a few pairs of SNPs have high LD (Fig. 1). Figure 2 provides a comparison of the MAF distribution of *TG* and *COL6A3* with that of the entire dataset. The three distributions agree well except that *TG* and *COL6A3* have slightly higher percentage of common variants.

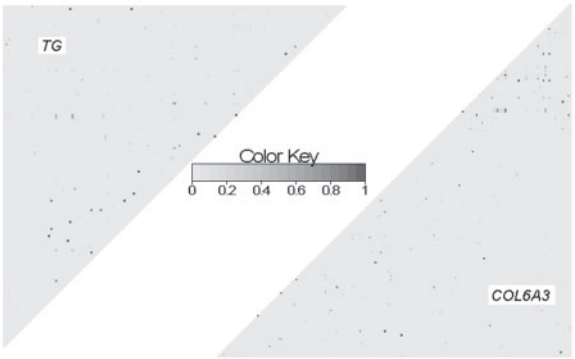

Fig. 1. LD structures of the 146 SNPs in *TG* (top left) and 187 SNPs in *COL6A3* (bottom right)

3.2 Simulation setup

In order to make a comprehensive performance comparison between the proposed methods and existing ones, a variety of true genetic effects are examined in different simulation studies (Table 1). For each study, a total of 1000 replicates are simulated. In each replicate, a quantitative trait is first simulated under the null model $Q_0=\epsilon$, where ϵ is a standard normal noise. The top 50% of the distribution of Q_0 are then declared affected, by which we define a binary disease status. It is easy to see that, under the null model, the quantitative trait and disease status do not depend on genotypes.

Then, under the alternative model, a quantitative trait is generated according to

$$Q_1=f(X_{\{1\}},\ldots,X_{\{p\}})+Q_0, \quad (3)$$

where $X_{\{j\}}$'s are SNPs in descending order according to their MAFs and f is the true genetic effect model, which differs among simulation scenarios. A binary disease status based on Q_1 is defined in a similar way as Q_0. For reference purpose, if in Equation (3) $f=\sum_{j=1}^{p}\beta_j X_{\{j\}}$ is a linear additive model, then conditional on the fact that this is an half-affected-half-control sample and assuming

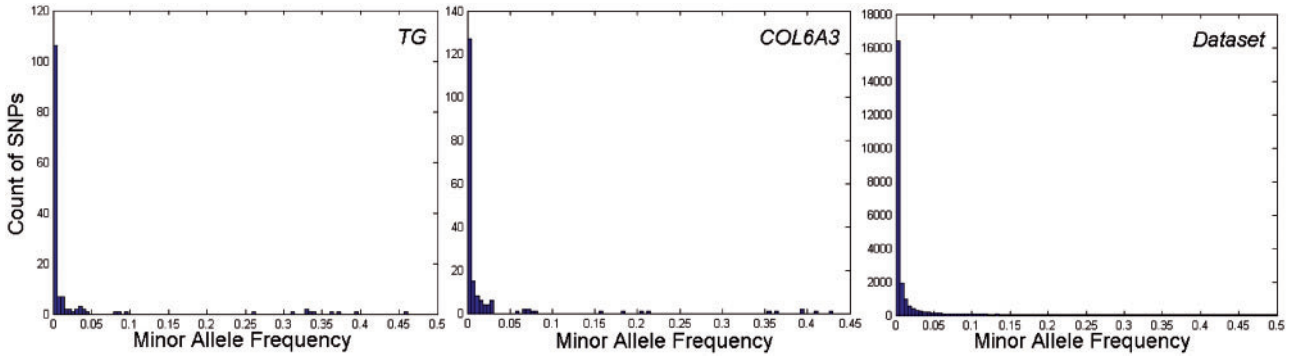

Fig. 2. Histograms of MAFs for SNPs in *TG* (left), *COL6A3* (middle) and the entire GAW17 dataset (right)

that existence of causal SNPs in the gene does not heavily influence the general cutoff to Q_1 (which is the case for most of our simulation studies shown below), there is an approximate correspondence between coefficient β_j and odds ratio of the SNP $X_{\{j\}}$: odds ratio $\approx (1-\Phi(-\beta_j;0,1))/\Phi(-\beta_j;0,1)$, where $\Phi(x;\mu,\sigma)$ represents cumulative normal distribution function with mean μ and standard deviation σ. Therefore, a coefficient β from 0.2 to 1 corresponds to an odds ratio from 1.4 to 5.3. On the contrary, a coefficient β from -0.2 to -1 reduces odds ratio to 0.7–0.2.

3.3 Simulation results

We compared our method of DR followed by LR model (a special case of GLM for binary trait) with kernel-based adaptive cluster (KBAC) method of Liu and Leal (2010) and variable threshold (VT) test of Price *et al.* (2010). For KBAC, if a SNP set contains only rare variants, we adopted a permutation test for the KBAC statistic; otherwise, following Liu and Leal (2010), a LR model was used to incorporate both the common variants and a variable for the kernel weight generated by KBAC. Standard permutation procedure was then applied to evaluate the significance. VT test can be used regardless of the existence of common variants. We also compared some variants at the step of DR, namely PCA, LR for testing all SNPs, SIR and kernel SIR (kSIR) with various kernel functions (Gaussian, IBS and WF). Multiplicative symmetrization was performed for WF kernel. For each of the approaches under comparison, 10 000 permutations were used to generate a null distribution for the test statistic. For each DR-based method, the leading summary variable was used to represent the SNP set. For each method under comparison, an empirical power was then obtained by counting how many of empirical P-values over 1000 replications are less than a nominal significance level of $\alpha = 0.05$ under the alternative model using the designated causal SNPs. An empirical Type-I error was evaluated in a similar fashion except that the trait was generated under the null model. As mentioned above, we designed various scenarios to evaluate methods. Results are summarized in Table 1.

3.3.1 Empirical Type-I error The last two rows of Table 1 show that the empirical Type-I errors of all the methods are relatively close to the nominal significance level of 0.05.

3.3.2 Empirical power comparison Simulations 1–4 represent main effect models. In Simulations 1 and 2, informative SNPs are all common variants, whose MAFs all fall into the range of 0.32–0.43. When all SNPs are deleterious (Simulation 1), VT outperforms KBAC, linear and nonlinear DR methods. However, when two of the informative alleles are protective (Simulation 2), the performance of VT drops dramatically, while the performance of the other methods does not. In this case, WF works the best and IBS is the second. In Simulation 3 and 4, informative SNPs are all rare variants, whose MAFs are between 0.005 and 0.0086. Note that few people have totally more than one minor alleles over these 10 SNPs. No matter there are protective rare variants or not, WF and IBS have the best and second best performance, respectively. VT has a comparable performance when there are only deleterious effect (Simulation 3), but the existence of protective variants destroys its performance again (Simulation 4). Note that under these linear genetic models, the WF and IBS kSIR's still work better than the LR and linear SIR, while the Gaussian kSIR does not. This reflects the superiority of these kernels for sequence data.

Simulation 5 represents a model with both main and epistasis effects. For each gene, both of the two informative SNPs are common variants with very low LD. In this case, WF works better than IBS, which then outperforms KBAC, VT and other DR-based methods. Simulations 6 and 7 are pure epistasis models among three common SNPs with low LD between each other, while Simulations 8 and 9 represent epistasis between common and rare variants. It is clear that nonlinear DR methods with WF or IBS kernel outperform linear dimension reduction methods as well as KBAC and VT. KBAC loses most of its power in detecting common and rare variant interactions. VT performs poorly when there are protective effects.

Weighted IBS kernel using weights $w_j = 1/\sqrt{f_j}$ is also evaluated as a candidate kernel for kSIR. The performance of weighted IBS, however, is found worse than that of IBS for most of the studies in Table 1 (results not shown).

We further evaluated the performance of kSIR with WF kernel at different signal-to-noise ratios (SNRs) using gene *TG*. For Model (3), SNR is defined as the ratio of the variance of the signal f and the error variance. A number of SNR values were simulated by multiplying f with different coefficients. Figure 3 shows the empirical power for the nine scenarios that are given in Table 1 with the SNRs ranging from 0.02 to 0.12. It is seen that the empirical power exceeds 0.8 for additive effects with 0.1 SNR, including the ones composed of protective and/or rare variants. The complex interaction effects are less detectable than simple additive

Table 1. Results of simulation studies based on sequence data of genes *TG* and *COL*6A3

Study	Gene	Genetic effect	Current methods		Linear DR			Nonlinear DR (kSIR)		
			KBAC	VT	PCA	LR	SIR	Gauss	IBS	WF
1	*TG*	$.2(X_{\{2\}}+X_{\{3\}}+X_{\{4\}}+X_{\{7\}})$	0.306	0.923	0.115	0.640	0.598	0.415	0.671	0.782
	COL6A3	$.2(X_{\{1\}}+X_{\{2\}}+X_{\{5\}}+X_{\{6\}})$	0.205	0.898	0.209	0.680	0.609	0.416	0.741	0.853
2	*TG*	$.2(X_{\{2\}}-X_{\{3\}}-X_{\{4\}}+X_{\{7\}})$	0.314	0.055	0.379	0.738	0.684	0.513	0.767	0.858
	COL6A3	$.2(X_{\{1\}}+X_{\{2\}}-X_{\{5\}}-X_{\{6\}})$	0.213	0.000	0.643	0.751	0.722	0.517	0.834	0.912
3	*TG*	$\sum_{i=36\sim45}X_{\{i\}}$	0.000	0.807	0.056	0.805	0.763	0.676	0.843	0.896
	COL6A3	$\sum_{i=46\sim55}X_{\{i\}}$	0.003	0.868	0.051	0.729	0.702	0.631	0.865	0.929
4	*TG*	$\sum_{i=36,37,43\sim45}X_{\{i\}}-\sum_{i=38\sim42}X_{\{i\}}$	0.349	0.111	0.131	0.805	0.748	0.694	0.841	0.917
	COL6A3	$\sum_{i=46,47,53\sim55}X_{\{i\}}-\sum_{i=48\sim52}X_{\{i\}}$	0.255	0.043	0.096	0.611	0.610	0.498	0.716	0.813
5	*TG*	$(X_{\{6\}}+X_{\{7\}}+X_{\{6\}}\times X_{\{7\}})/8$	0.193	0.179	0.257	0.301	0.292	0.211	0.339	0.434
	COL6A3	$(X_{\{1\}}+X_{\{5\}}+X_{\{1\}}\times X_{\{5\}})/8$	0.108	0.149	0.221	0.287	0.279	0.186	0.353	0.437
6	*TG*	$(X_{\{3\}}\times X_{\{4\}}+X_{\{3\}}\times X_{\{7\}}+X_{\{4\}}\times X_{\{7\}})/6$	0.379	0.695	0.140	0.624	0.594	0.539	0.639	0.770
	COL6A3	$(X_{\{1\}}\times X_{\{2\}}+X_{\{1\}}\times X_{\{5\}}+X_{\{2\}}\times X_{\{5\}})/6$	0.258	0.500	0.035	0.567	0.525	0.438	0.708	0.780
7	*TG*	$(X_{\{3\}}\times X_{\{4\}}+X_{\{3\}}\times X_{\{7\}}-X_{\{4\}}\times X_{\{7\}})/6$	0.142	0.208	0.055	0.195	0.192	0.245	0.211	0.258
	COL6A3	$(X_{\{1\}}\times X_{\{2\}}+X_{\{1\}}\times X_{\{5\}}-X_{\{2\}}\times X_{\{5\}})/6$	0.155	0.073	0.041	0.235	0.233	0.267	0.338	0.339
8	*TG*	$X_{\{2\}}\times\sum_{i=36\sim40}X_{\{i\}}+X_{\{3\}}\times\sum_{i=41\sim45}X_{\{i\}}$	0.003	0.456	0.052	0.438	0.419	0.423	0.573	0.645
	COL6A3	$X_{\{1\}}\times\sum_{i=46\sim50}X_{\{i\}}+X_{\{5\}}\times\sum_{i=51\sim55}X_{\{i\}}$	0.009	0.532	0.103	0.362	0.337	0.306	0.448	0.546
9	*TG*	$X_{\{2\}}\times\sum_{i=36\sim40}X_{\{i\}}-X_{\{3\}}\times\sum_{i=41\sim45}X_{\{i\}}$	0.118	0.044	0.260	0.468	0.448	0.396	0.583	0.673
	COL6A3	$X_{\{1\}}\times\sum_{i=46\sim50}X_{\{i\}}-X_{\{5\}}\times\sum_{i=51\sim55}X_{\{i\}}$	0.181	0.032	0.077	0.401	0.382	0.286	0.428	0.545
Type-I error	*TG*	Null model	0.061	0.051	0.053	0.048	0.054	0.054	0.056	0.047
	COL6A3	Null model	0.049	0.047	0.042	0.050	0.051	0.049	0.048	0.047

Each study focuses on one genetic model mimicking a specific type of true genetic effect. Under 'Genetic Effect' are the true genetic effects that generate the quantitative trait, where $X_{\{j\}}$'s are SNPs in descending order according to their MAFs. The common variants ($j \leq 10$) are selected to have low pairwise LD. The binary trait is determined from the quantitative trait and serves as the response variable in the simulation studies. The numbers under the names of different methods are their empirical power in different studies or Type-I error.

effects. Nevertheless, when the SNR reaches 0.12, the empirical power exceeds 0.7 and 0.6 for common–common and rare–common interactions, respectively.

In conclusion, VT is extremely vulnerable to protective variants and less sensitive to epistasis effects. KBAC does not capture interactions between rare and common variants. Even if the true genetic model is additive, linear DR-based methods do not outperform nonlinear DR methods with kernels specifically designed for genotype data. Among the different kernel methods, the proposed WF kernel always outperforms IBS kernel, which then works better than Gaussian kernel.

4 DISCUSSION

Ideally, an association test should be able to handle: (i) high dimensionality of genomic dataset, which typically far exceeds the sample size; (ii) both rare and common variants; (iii) additive, recessive and dominant models of gene action; (iv) both quantitative traits and disease dichotomies and (v) non-genetic covariates. Most of existing solutions could not deliver across the board judged by all those criteria. To bridge the gap, we have proposed a strategy based on nonlinear supervised DR and a new kernel to effectively aggregate genotype information. Such an aggregation increases the likelihood of detecting multiple causal variants, whereas nonlinear reduction permits complex interactive relationship among genetic variants. Moreover, the GLM framework based on the aggregated features can naturally handle both quantitative and categorical traits, and incorporate non-genetic predictors and/or environmental variables. Finally, the proposed WF kernel based on Markov chain theory for genotype data has been proven useful compared with some existing kernels and can potentially benefit a wide range of kernel methods, such as kernel PCA, support vector machines and nonparametric and semiparametric regressions.

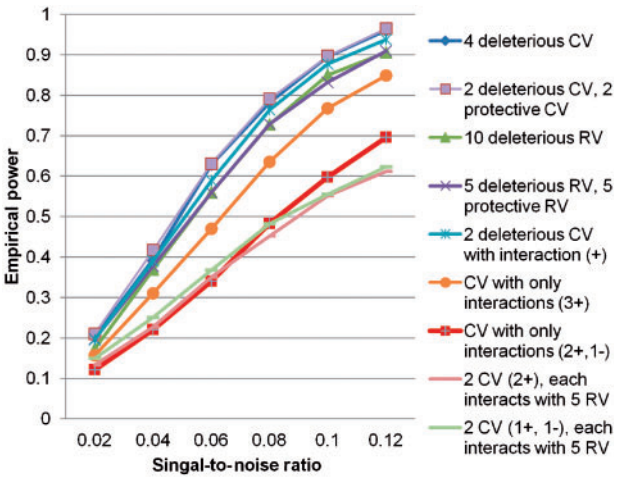

Fig. 3. Empirical power of kSIR with WF kernel varies with SNR. Gene *TG* is used for simulation. The nine scenarios in Table 1 are examined. CV, common variant; RV, rare variant

There are a number of possible avenues for future extensions. First, a number of Markov chain kernels in addition to the WF kernel can be explored, for instance, Ehrenfest kernels, hypergeometric kernels and Dirichlet-Multinomial kernels (Khare and Zhou, 2009; Zhou and Lange, 2009). Second, our current aggregation strategy has not taken advantage of any available information regarding the relationships among SNPs within a SNP set, e.g. their genetic distances. Incorporating those information can potentially facilitate the design of more effective kernels.

ACKNOWLEDGMENT

The authors thank the three reviewers for their insightful and constructive comments.

Funding: National Science Foundation (DMS-1106668 to Li), and the National Institutes of Health (HG006139 to Hua Zhou). Preparation of the Genetic Analysis Workshop 17 Simulated Exome Data Set was supported by the National Institutes of Health (MH059490, in part) and used sequencing data from the 1000 Genomes Project (`www.1000genomes.org`).

Conflict of Interest: none declared.

REFERENCES

Cannings,C. (1974) The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Adv. Appl. Probab.*, **6**, 260–290.

Chen,L.S. *et al.* (2010) Insights into colon cancer etiology via a regularized approach to gene set analysis of gwas data. *Am. J. Hum. Genet.*, **86**, 860–871.

Cohen,J.C. *et al.* (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869–872.

Easton,D.F. and Eeles,R.A. (2008) Genome-wide association studies in cancer. *Hum. Mol. Genet.*, **17**, ddn287+.

Ewens,W. (2004) *Mathematical Population Genetics. I*, Vol. 27 of *Interdisciplinary Applied Mathematics*, 2nd edn. Springer-Verlag, New York, Theoretical introduction.

Frazer,K.A. *et al.* (2009) Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.*, **10**, 241–251.

1000 Genome Project. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Ji,W. *et al.* (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.*, **40**, 592–599.

Khare,K. and Zhou,H. (2009) Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions. *Ann. Appl. Probab.*, **19**, 737–777.

Kwee,L.C. *et al.* (2008) A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.*, **82**, 386–397.

Lettre,G. and Rioux,J.D. (2008) Autoimmune diseases: insights from genome-wide association studies. *Hum. Mol. Genet.*, **17**, ddn246+.

Li,B. and Leal,S. (2008) Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *AJHG*, **83**, 311–321.

Li,K.-C. (1991) Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.*, **86**, 316–342. With discussion and a rejoinder by the author.

Liu,D.J. and Leal,S.M. (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.*, **6**, e1001156.

Madsen,B.E. and Browning,S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, e1000384.

Marshall,A.W. *et al.* (2011) *Inequalities: Theory of Majorization and its Applications*, 2nd edn. Springer Series in Statistics. Springer, New York.

Nejentsev,S. *et al.* (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387–389.

Price,A. *et al.* (2010) Pooled association tests for rare variants in exon-resequencing studies. *AJHG*, **86**, 832–838.

Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Schaid,D. (2010a) Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. *Hum. Hered.*, **70**, 109–131.

Schaid,D. (2010b) Genomic similarity and kernel methods II: methods for genomic information. *Hum. Hered.*, **70**, 132–140.

Scholkopf,B. and Smola,A.J. (2001) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Wessel,J. and Schork,N.J. (2006) Generalized genomic distance-based regression methodology for multilocus association analysis. *Am. J. Hum. Genet.*, **79**, 792–806.

Wu,M.C. *et al.* (2010) Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 929–942.

Zhou,H. and Lange,K. (2009) Composition markov chains of multinomial type. *Adv. Appl. Probab.*, **41**, 270–291.

Zhou,H. and Lange,K. (2010) Mm algorithms for some discrete multivariate distributions. *J. Comput. Graph. Stat.*, **19**, 645–665.

Zhu,H. and Li,L. (2011) Biological pathway selection through nonlinear dimension reduction. *Biostatistics*, **12**, 429–444.