HAL
open science

Eigen-Epistasis for detecting Gene-Gene interactions

Virginie Stanislas, Cyril Dalmasso, Christophe Ambroise

To cite this version:

Virginie Stanislas, Cyril Dalmasso, Christophe Ambroise. Eigen-Epistasis for detecting Gene-Gene interactions. 2016. hal-01275624v1

HAL Id: hal-01275624

https://hal.science/hal-01275624v1

Preprint submitted on 17 Feb 2016 (v1), last revised 14 Feb 2017 (v6)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Eigen-Epistasis for detecting Gene-Gene interactions

Virginie Stanislas[1], Cyril Dalmasso[1] and Christophe Ambroise[1]

[1] Laboratoire de Mathématiques et Modélisation d'Evry (LaMME), Université d'Evry Val d'Essonne, UMR CNRS 8071, ENSIIE, USC INRA

Corresponding author: Virginie Stanislas, Laboratoire de Mathématiques et Modélisation d'Evry, 23 bvd de France, 91 037 Évry Cedex, France. E-mail: virginie.stanislas@math.cnrs.fr

Abstract

A large amount of research has been devoted to the detection and investigation of epistatic interactions in genome-wide association study (GWAS). Most of the litterature focuses on low-order interactions between single-nucleotide polymorphisms (SNPs) with significant main effects.

In this paper, we propose an original approach for detecting epistasis at the gene level, without systematically filtering on significant genes. We first compute interaction variables for each gene pair by finding its Eigen-epistasis Component defined as the linear combination of Gene SNPs having the highest correlation with the phenotype. The selection of the significant effects results from a penalized regression method based on group Lasso controlling the False Discovery Rate.

The method is compared to three recent alternative proposals from the literature using synthetic data and exhibits high performance in different settings. Using a genome-wide association study on ankylosing spondylitis cases, we demonstrate the power of the approach by detecting new gene-gene interactions.

Key words : Genome-wide association study, Gene-gene interactions, Epistasis, Group Lasso

Introduction

Genome Wide Association Studies (GWAS) aim at finding genetic markers associated with a phenotype of interest. Typically, hundreds of thousands of single nucleotide polymorphism (SNP) are studied for a limited number of individuals using high density genotyping arrays. The association between each SNP and the phenotype is usually tested by single marker approaches. Multiple markers may also be considered but are typically selected with simple forward selection methods. GWAS represent a powerful tool to investigate the genetic architecture of complex diseases and have shown success in identifying hundreds of variants. However, they have explained only a small part of the phenotypic variations expected from classical family studies [Manolio et al., 2009]. Many reasons for this missing heritability have been proposed, among which the inadequate accounting for shared environment among relatives or the idea that much larger numbers of variants of small effect are yet to be found. Rare variants, which can hardly be captured by existing genotyping arrays [Manolio et al., 2009], seem to be important causal factors, as well as structural variations. But complex diseases may also partly result from complex genetic structures such as multiple interactions between markers, known as epistasis. Indeed, the genetic effect on phenotype appears as part of the additive genetic variance in pedigree studies but as an unmeasured gene-gene interaction in GWAS [Haig, 2011]. For example Zuk et al. [2012] proposed a model that takes into account epistatic interaction on Crohn's disease. They found that 80% of the missing heritability could be due to genetic interactions.

In past years, numerous methods have been proposed for studying epistasis and have been reported in various reviews [Wei et al., 2014; Steen, 2012]. They vary in terms of data analysis (genome-wide or filtering) and statistical methodology (Bayesian, frequentist, machine learning or data mining). Most of them focus on single-locus interactions, but considering interactions at gene level may offer many advantages. Firstly, as genes are the functional unit of genome, results can be more biologically interpretable. Furthermore, genetic effects can be more easily detected when SNP effects are aggregated together. Finally, gene based analysis simplifies the multiple testing problem by decreasing the number of variables. Several gene-gene methods have thereby been proposed. Basically, they rely on a summarizing step which is used to obtain information at the gene level. Then, for the most recent methods, filters or penalized models are used to make the method applicable to a large number of genes while oldest methods are only applicable to a couple or a reduced number of genes. For the summarizing step, most methods resort to a principal components (PC) approach but each of them presents its own specificity. We describe below some of them.

Chatterjee et al. [2006] developed Tukey's one df method to investigate interaction between two genes. This method assumes that the SNPs included in each gene region act as surrogates for an underlying biological phenotype. The genotypic information of the gene region is extracted as a single component by a weighted sum of all SNPs. The weights are determined according to the SNP's correlation with the

trait. Then, the product of the two sums is introduced in a logistic model as the gene-gene interaction term with marginal effects represented by the respective sums. Following this idea, Wang et al. [2009] compared two different interaction tests. On the one hand, they used principal component analysis (PCA) to summarize SNPs information within a gene, on the other hand they used partial least squares (PLS) to extract components that summarize both the information among SNPs in a gene and the correlation between SNPs and the outcome of interest. Then, they proposed an interaction test based on either the first PC or the PLS component for each gene. They showed that the PCA and PLS methods often had better performance than the Tukey 1-df method. But it is worth noting that the main objective of these three methods was more to increase the power to detect association in the presence of gene-gene interactions than to identify interactions themselves. Other approaches based on principal component analysis have then been proposed for epistasis detection. Li et al. [2009] proposed to select PCs that explain at least 80% of the variation as the gene representation. He et al. [2011] proposed another approach using linkage disequilibrium information to weight genotype scores which are then aggregated using principal components. Other approaches are based on linkage disequilibrium. As an example, Rajapakse et al. [2012] developed a gene-based test of interactions for case-control studies which compares LD patterns between cases and controls. In other respects, Wang et al. [2014a] proposed the Gene-Trait similarity regression (Simreg). This method does not resort to a PC approach to summarize gene information but to a genetic similarity measure calculated for each gene across the individuals.

However, most of theses methods are only applicable to a reduced number of genes. Indeed, directly modeling all gene-gene interactions would be inefficient due to computational challenge and lack of power. One possibility is to reduce the gene-gene search space by eliminating unimportant genes. Hence, two-step procedures that first filter out specific genes or SNPs through genome wide search before testing for interactions have been developed. One example is the model-based kernel machine method (3G-SPA) proposed by Li and Cui [2012] which first performs a search for gene pairs contributing to the overall phenotypic variations. Then, significant pairs are tested for interaction effects.Another attractive alternative is offered by penalized regression methods that allow to select a subset of important predictors from a large number of potential ones. These methods operate by shrinking the size of the coefficients. The coefficients of predictors with little or no apparent effect are pushed on a trait down toward zero, allowing to reduce the effective degrees of freedom and in many cases to performs model selection. A few approaches using penalized models have been proposed. Thus, D'Angelo et al. [2009] combined principal component analysis and the penalized regression LASSO. Wang et al. [2014b] also used a principal component analysis combined to a L1 penalty with adaptive weights based on gene size, pathway support and effect size.

Here we propose a Group LASSO approach [Yuan and Lin, 2006] that takes into account the group structure of each gene in order to detect epistasis. We introduce the Gene-Gene Eigen Epistasis (G-GEE)

as a new approach to compute the gene-gene interaction part of the model and we compare G-GEE with three different interaction variable modeling approaches inspired by previous literature proposals: Principal Component Analysis (PCA), Partial Least squares (PLS) and Canonical-Correlation Analysis (CCA). An adaptive ridge cleaning approach is then used in order to compute p-values for each group. In the next section, we detail the different models. Then, in Section 3, we present results from a simulation study performed to compare the performance of the different approaches. In section 4, we apply our proposed G-GEE Group LASSO method on a real data set on Ankylosing Spondylitis. Finally, the proposed approach and the results are discussed in Section 5.

Methods

We consider n individuals where $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ denotes the vector of trait values. For each individual, genetic variants among G genes are considered. Each gene is described by a given number of SNPs p_g where $\sum_g p_g = p$. The SNP matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ considers an additive coding scheme in which the genotype value of each SNP j from individual i is denoted $X_{ij} \in \{0, 1, 2\}$. $\mathbf{X}_i$ is a p-dimensional vector of covariates for observation i and for $j \in \{1, \ldots, p\}$. $\boldsymbol{X}^g$ denotes the submatrix of $\boldsymbol{X}$ whose columns are the p_m SNPs of gene g. In the following, we assume a linear model where the phenotype is considered as a random variable y_i whose conditional expectation can be written as a function of the covariates $\boldsymbol{X}_i$ and their interactions $\boldsymbol{Z}_i$,

$$E[y_i|X] = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{\gamma},$$

where

$$\boldsymbol{\beta} = \left(\underbrace{\beta_{1,1}, \beta_{1,2}, ..., \beta_{1,p_1}}_{gene_1}, ..., \underbrace{\beta_{G,1}, ..., \beta_{G,p_G}}_{gene_G} \right)^T,$$

and $\boldsymbol{Z}_i$ is the ith line of the matrix of interactions and $\boldsymbol{\gamma}$ a parameter vector of appropriate dimension.

The main effect of each gene is thus modeled through the sum of all its SNPs effects. Concerning interaction effects, we compute new variables representing interaction for two specific genes and define as a group all the interaction variables related to a given pair of genes. The matrix of interaction is thus structured into $G(G-1)/2$ submatrices:

$$\boldsymbol{Z} = [\boldsymbol{Z}^{11} \cdots \boldsymbol{Z}^{rs} \cdots \boldsymbol{Z}^{G(G-1)/2}]$$

where $\boldsymbol{Z}^{rs}$ describes the interactions between the two genes r and s. The parameter vector $\boldsymbol{\gamma}$ is accordingly structured into sub-vectors $\boldsymbol{\gamma}^{rs}$. In the following, we present and compare four different approaches for modeling gene-gene interactions.

Modeling Gene-Gene interaction

Let consider two genes r and s described respectively by p_r and p_s SNPs. A possible interaction term describing the epistasis between the two genes is

$$\boldsymbol{Z}_i^{rsT}\boldsymbol{\gamma}^{rs} = \sum_{jk} \gamma_{jk}^{rs} X_{ij}^r X_{ik}^s. \tag{1}$$

In that case the submatrix coding interactions would be $\boldsymbol{Z}^{rs} = \boldsymbol{W}^{rs} = \{X_{ij}^r X_{ik}^s\}_{i=1\cdots n}^{j=1,\cdots,p_r;k=1,\cdots,p_s}$ and $\boldsymbol{\gamma}^{rs} = \{\gamma_{jk}^{rs}\}$ a vector of size $\sum_{rs} p_r p_s$. The number of parameters of such a model is obviously too large for being reliably estimated. Many papers in the literature thus consider reducing the dimension of $\boldsymbol{\gamma}$.

In the remaining of the paper we do consider four different dimension reduction methods: principal component analysis (PCA), canonical-correlation analysis (ACC), partial least squares (PLS) and our proposed approach named G-GEE for Gene-Gene EigenEpistasis.

Principal Component Analysis

Principal Component Analysis (PCA) allows to reduce the number of variables describing each gene r from p_r to $q_r < p_r$. Considering gene r described by p_r SNPs, we compute the matrix of the q first principal components

$$C^r = \boldsymbol{X}^r U^r,$$

where U^r is the matrix of the first q_r principal axis. Using C^r and C^s instead of $\boldsymbol{X}^r$ and $\boldsymbol{X}^s$ in the computation of the interaction allows to control the number of parameters relative to each interaction. This control is achieved by choosing the number of principal components q. The PCA model that we described is related to the ideas of previously published work by Zhang and Wagener [2008]. In this context the interaction term takes the form

$$\boldsymbol{Z}_i^{rsT}\boldsymbol{\gamma}^{rs} = \sum_{j=1}^{q}\sum_{k=1}^{q} \gamma_{jk}^{rs} C_{ij}^r C_{ik}^s.$$

Relating this expression to the general form of the interaction term $\boldsymbol{W}_i^{rs}$ presented in the previous section, it appears that performing PCA prior to computing the interactions is a way to constrain the linear interaction term of Equation 1.

The matrix of interactions is $\boldsymbol{Z} = \{C_{ij}^r C_{ik}^s\}_{i=1\cdots n}^{jk=1\cdots \sum_{rs} q_r q_s}$ and $\boldsymbol{\gamma} = \{\gamma_{jk}\}$ is a vector of size $\sum_{rs} q_r q_s$ with successive chunks of $q_r \times q_s$ coefficients, each describing an interaction between genes r and s. In particular if a unique principal component is chosen, there will be only one parameter to estimate per interaction.

Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) aims at finding linear combination of groups of variable which have maximum correlation. In our setting we do consider each gene as a group of SNPs. For two genes r and s, we define new variables $\boldsymbol{A}^r$ and $\boldsymbol{B}^s$ which are linear combination of the original variables $\boldsymbol{X}^r$ and $\boldsymbol{X}^s$:

$$\begin{cases} \boldsymbol{A}^r = \boldsymbol{X}^r U^r, \\ \boldsymbol{B}^s = \boldsymbol{X}^s V^s \end{cases}$$

where U^r and V^s are the matrices whose columns define the weight vectors, which are solution of the CCA. We propose to code the interaction of a couple of genes (r, s) by the first q component couples of a CCA:

$$\boldsymbol{Z}_i^{rsT} \boldsymbol{\gamma}^{rs} = \frac{\text{ffl}}{\text{ffl}}$$

Partial Least Square

Wang et al. [2009] proposed an alternative method for integrating interactions using a partial least squares approach (PLS). Let $(\boldsymbol{X}^r, \boldsymbol{X}^s)$ be the genotypic matrix for the given pair of gene (r, s). The approach of Wang et al. [2009] computes components maximizing $cov^2(\boldsymbol{X}^r \boldsymbol{u}, \boldsymbol{T}\boldsymbol{v})$, with $\boldsymbol{T} = (\boldsymbol{y}, \boldsymbol{X}^s)$ and $(\boldsymbol{u}, \boldsymbol{v})$ the weight vectors. This approach allows to keep the phenotypic information in the construction of the interaction variables.

The G-GEE group LASSO model

We propose an original approach in order to model interaction. The general idea is to consider the interaction variable between the two genes r and s as a function $f_{\boldsymbol{u}}(\boldsymbol{X}_i^r, \boldsymbol{X}_i^s)$ parametrized by $\boldsymbol{u}$. One way to estimate $\boldsymbol{u}$ is to maximize to correlation between the interaction function and the phenotype:

$$\hat{\boldsymbol{u}} = arg \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} cor(\boldsymbol{y}, f_{\boldsymbol{u}}(\boldsymbol{X}_i^r, \boldsymbol{X}_i^s)).$$

If we consider the function f to be linear our problem becomes easily tractable with a unique solution. Setting

$$\boldsymbol{Z}^{rs} = f_{\boldsymbol{u}}(\boldsymbol{X}_i^r, \boldsymbol{X}_i^s) = \boldsymbol{W}^{rs} \boldsymbol{u},$$

where $\boldsymbol{W}^{rs} = \{X_{ij}^r X_{ik}^s\}_{i=1\cdots n}^{j=1\cdots,p_r;k=1,\cdots,p_s}$ and $\boldsymbol{u} \in \mathbb{R}^{p_r p_s}$ we get the following problem:

$$\max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} ||c\hat{o}r[\boldsymbol{W}^{rs}\boldsymbol{u}, \boldsymbol{y}]||^2 = \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} ||\boldsymbol{u}^T \boldsymbol{W}^{rsT} \boldsymbol{y}||^2 = \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} \boldsymbol{u}^T \boldsymbol{W}^{rsT} \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{W}^{rs} \boldsymbol{u} \quad . \tag{2}$$

The solution $\boldsymbol{u}$ is the eigenvector associated to the largest eigenvalue of the matrix $\boldsymbol{W}^{rsT} \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{W}^{rs}$. We then use the projection of the matrix $\boldsymbol{W}^{rs}$ on $\boldsymbol{u}$ as the interaction variable. The resulting Eigen-epistasis

vector $\boldsymbol{Z}$ is the linear combination of the all SNP-SNP interactions which is the most correlated with the phenotype. As in the PLS approach, this method allows to take into account the phenotypic information in the construction of the interaction variables.

Coefficients estimation

We propose a group lasso model for estimating the parameters of all models. A group is either made of the SNPs of a given gene or of interaction terms relative to a given gene-pair interaction.

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\operatorname{argmin}} \left(\sum_i (y_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{\gamma})^2 + \lambda \left[\sum_g \sqrt{p_g} ||\boldsymbol{\beta}^g||_2 + \sum_{rs} \sqrt{p_r p_s} ||\boldsymbol{\gamma}^{rs}||_2 \right] \right) \quad ,$$

The parameter λ is selected by cross-validation.

In order to improve estimation accuracy and to obtain p-values for each selected groups, we use the adaptive ridge cleaning approach proposed by Bécu et al. [2015]. This screen and clean procedure is a two-stage method. The group lasso model is first fitted on half of the data. The coefficient of the candidate groups selected by the model are then introduced in a ridge regression model fitted on the second half of the data with a specific penalty allowing to take into account the group structure. Significances of the regression coefficients for each group are then estimated by permutation tests.

Simulation study

To evaluate the performance of the proposed approach, we conduct a simulation study. The proposed G-GEE model is compared to the three other interaction variable modeling through data which is generated in order to mimic realistic genotypes and phenotypes.

Design

Genotypes

The n lines of the genotype matrix are an i.i.d. sample from a multivariate random vector $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. The correlation matrix $\boldsymbol{\Sigma}$ is block diagonal, each block corresponding to a gene. Two variables belonging to the same gene are correlated at level $\rho = 0.8$ while all other correlations are null. Each SNP (column of the genotype matrix) is randomly assigned a minor allele frequency (MAF) p from a uniform distribution between 0.05 and 0.5. A MAF value of 0.2 is assigned to all causal SNPs. The genotype frequencies derived from the Hardy-Weinberg equation were then used to discretized X_{ik} values to 0, 1 or 2. In practice, X_{ik} is set to 0 if $X_{ik} < q_{p^2;N(0,1)}$, X_{ik} is set to 3 if $X_{ik} < q_{(1-p)^2;N(0,1)}$ and X_{ik} is set to 2 otherwise.

Phenotypes

Phenotype vectors are generated following two different schemes. We first considered the model proposed by Wang et al. [2014b] :

$$Y_i = \beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g \right) + \sum_{rs} \gamma_{rs} \left(\sum_{(j,k) \in \mathcal{C}^2} X_{ij}^r X_{ik}^s \right) + \epsilon_i, \tag{3}$$

where $\mathcal{C}$ and $\mathcal{C}^2$ are respectively the set of causal SNPs and causal interactions, and ϵ_i a random Gaussian variable. For each causal gene g, we consider two causal SNPs and a coefficient β_g is assigned to the standardized sum of these causal SNPs. Following the same idea for the interactions, all the causal SNPs from a causal pair (r, s) are pairwise multiplied and a coefficient γ_{rs} is assigned to the standardized sum of the product.

The second phenotype simulation model relies on the following model:

$$Y_i = \beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g \right) + \sum_{rs} \gamma_{rs} \left(\sum_{(j,k) \in \mathcal{C}^2} C_{ij}^r C_{ik}^s \right) + \epsilon_i. \tag{4}$$

The difference with the previous model concerns the simulation of the interaction effect. In the last model the interaction effect for a causal couple (r, s) is defined as the product of the first PCA component $\boldsymbol{C}_{.1}^r$ of gene r and the first PCA component $\boldsymbol{C}_{.1}^s$ of gene s.

In both model, β_0 is set to 0, and ϵ_i are generated independently from a $\mathcal{N}(0, \sigma^2)$ with σ^2 determined from the coefficient of determination R^2 that calibrates the strength of the association. Both simulation model can be written as $y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{\gamma} + \epsilon_i$ where $\boldsymbol{X}$ the marginal effect genotype matrix and $\boldsymbol{Z}$ the interaction effect matrix.

Let us denote $\boldsymbol{Q\phi} = [\boldsymbol{X}, \boldsymbol{Z}] \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}$ and

$$\begin{aligned} R^2 &= \frac{\sum (\boldsymbol{Q_i \phi} - \bar{y})^2}{\sum (\boldsymbol{Q_i \phi} + \epsilon_i - \bar{y})^2} \\ &= \frac{\sum (\boldsymbol{Q_i \phi} - \bar{y})^2}{\sum (\boldsymbol{Q_i \phi} - \bar{y})^2 + \sum \epsilon_i^2 + \sum 2(\epsilon_i (\boldsymbol{Q_i \phi} - \bar{y}))} \\ &= \frac{\sum (\boldsymbol{Q_i \phi} - \bar{y})^2}{\sum (\boldsymbol{Q_i \phi} - \bar{y})^2 + \text{n } \hat{\text{var}}(\epsilon_i) + 2\text{n } \hat{\text{cov}}(\epsilon_i, \boldsymbol{Q_i \phi} - \bar{y})}. \end{aligned}$$

Let us remark that:

$$\begin{aligned}
2\text{n cov}(\epsilon_i, \boldsymbol{Q_i\phi} - \bar{y}) &= 2\text{n cov}(\epsilon_i, \boldsymbol{Q_i\phi} - \frac{\sum_j y_j}{n}) \\
&= 2\text{n cov}(\epsilon_i, \boldsymbol{Q_i\phi}) - \sum_j \frac{2\text{n}}{\text{n}}\text{cov}(\epsilon_i, y_j) \\
&= 0 - 2\text{cov}(\epsilon_i, \epsilon_i) = -2\sigma^2
\end{aligned}$$

Thus replacing $\hat{\text{var}}(\epsilon_i)$ by σ^2 and $\hat{\text{cov}}(\epsilon_i, \boldsymbol{Q_i\phi} - \bar{y})$ by $-\sigma^2/n$ we get $R^2 \approx \frac{\sum(\boldsymbol{Q_i\phi} - \bar{y})^2}{\sum(\boldsymbol{Q_i\phi} - \bar{y})^2 + \text{n}\sigma^2 - 2\sigma^2}$. This relation between R^2 and σ^2 allows us to get an expression for σ^2 depending on R^2, $\sigma^2 = \frac{(R^2 - 1)\sum(\boldsymbol{Q_i\phi} - \bar{y})^2}{R^2(2 - n)}$.

In order to investigate the respective roles of main and interactions effects in the model we decide to examine the part of the coefficient of determination R^2 that is explained by one or the other. With similar motivation, Wang et al. [2014b] control the part of the partial R^2 due to interactions effects when they simulate phenotype. The coefficient values are selected so that 30% of the partial R^2 was explained by interactions effects. Li and Cui [2012] didn't use the R^2 directly but simulated data assuming different proportions of interaction effects among the total genetic variance. Once the phenotype y has been set for each simulated design matrix, we computed the part of the R^2 that can be attributed to interaction or main effects as respectively $p_{R_I^2} = \frac{R_I^2}{R_T^2}$ and $p_{R_M^2} = \frac{R_M^2}{R_T^2}$ with R_I^2 being the R-square value for the model containing only simulated interaction effects, R_M^2 the R-square value when there are only simulated main effects and R_T^2 the R-square value for the model containing both simulated main and interaction effects.

Scenarios

We first consider a simple scenario where we have 6 genes composed each of 6 SNPs for 600 subjects. We define one causal interaction between genes and two causal genes with main effects and consider two different simulation settings:

- a first setting where the main effects and interaction effects involve the same genes,
- a second setting where interaction genes are different from main effect genes.

Both main effects and interaction effects are weighted with the same coefficient values ($\beta_g = \gamma_{rs} = 2, \forall g, r, s$). For these two settings, different coefficients of determination, from 0.05 to 0.7, are considered.

To evaluate the performance of the different methods with a more complicated scenario, we also consider a third setting where we simulate 25 genes with four causal interactions between genes and two genes with causal main effects. In these simulations, interaction genes are different from main effect genes. We only consider the case where $R^2 = 0.7$.

For each of these scenarios, we perform 1000 simulations. For each interaction, the power is estimated as the proportion of detected interactions over the total number of simulations. In the last setting, where four interactions are present, we consider the averaged power over the four interactions.

Results

Figure 1 displays results obtained for the first and second settings in which six genes are considered. The first two columns show heatmap matrices reflecting proportion of significant values for each variable and each method over the 1000 simulations for different R^2 values. The third column presents the estimated power to detect the gene interaction as a function of the R^2 values.

In the first setting (Fig. 1(A,B)), we consider the two first genes both having main and interaction effects. When the phenotype is simulated under the model proposed by Wang et al. [2014b] (Fig. 1(A)), G-GEE and PLS methods have a better power to detect the interaction effect than PCA and CCA methods which tend to capture only the two main effects of the two genes (Fig. 1(A)). While the power is non-decreasing with R^2 for CCA, PCA and PLS, we obtain a U-shaped curve for G-GEE. Indeed, for the smallest R^2 values, which correspond to the most difficult cases, the power of G-GEE to detect the interaction tends to decrease. When R^2 values reach 0.4 the G-GEE power to detect the interaction starts to increase. The situation is different for the main effects as the power of G-GEE method to detect them increases continuously with R^2 (data not shown). For PLS method, the power to detect the interaction effect is continuously non decreasing. However note that for this method one of the two main effects (here gene 1) is detected to the detriment of the second regardless the R^2 value. Under the PCA phenotype simulation model (Fig. 1(B)), G-GEE method has a better power than the other methods to detect interaction effects while keeping a good specificity, whatever the R^2 value is. The reasonably high power of the PCA method can be explained by the similitude between the phenotype simulation model and the estimation model. It is worth noting that in this first setting, only few variables are falsly significant, what reflects a good specificity for all methods (the worst being for the gene 3 $\times$ gene 4 interaction variable under Wang et al. [2014b] model and $r^2 = 0.1$ with a false discovery rate value of 0.068).

In the second setting (Fig. 1(C,D)), the two first genes have only main effects and the third and fourth genes have only an interaction effect. When the phenotype is simulated using the model proposed by Wang et al. [2014b], the interaction power of G-GEE method is uniformly greater than the one of the other methods (Fig. 1(C)). For all R^2 values, PCA and CCA methods tend to detect false main effects for genes 3 and gene 4 but not the interaction. Under the PCA phenotype simulation (Fig. 1(D)), PCA method leads to a large power to detect interaction effects, but once again these good performances can be explained by the similitude between the simulation model and the estimation model. The interaction power for G-GEE method is lower but still large. Power of PLS and CCA methods are almost null. Gene 3 and gene 4 are not detected neither as main nor as interaction effects for both CCA and PLS method. Under this model, only G-GEE method tends to attributes false main effect to the third and fourth genes. In this second setting, whatever the phenotype simulation model is, PLS method only identifies the first gene as having a main effect while the effects of gene 3 and gene 4 are not detected, neither as main nor

as interaction effects. Moreover, PLS method tends to attribute a false interaction effect between the two first genes.

The results obtain with these two first settings point out a certain phenomenon of confusion between main and interaction effects. This is easier to see in the second setting where the false discovery rate is higher. In this setting the false discoveries refer whether to main effects of genes which are simulated to have only an interaction effect with an other gene or to interaction variables which refer to gene with main effects. This phenomenon can be observed for all methods but is more or less marked depending of the scenarios. That could explain the U-shaped power curve for G-GEE in the first setting when effect are simulated using the Wang et al. [2014b] model: as the problem becomes harder the genetic effects of both genes are preferentially assigned to the interaction effect contributing to the better power to detect interaction for small R^2 values.

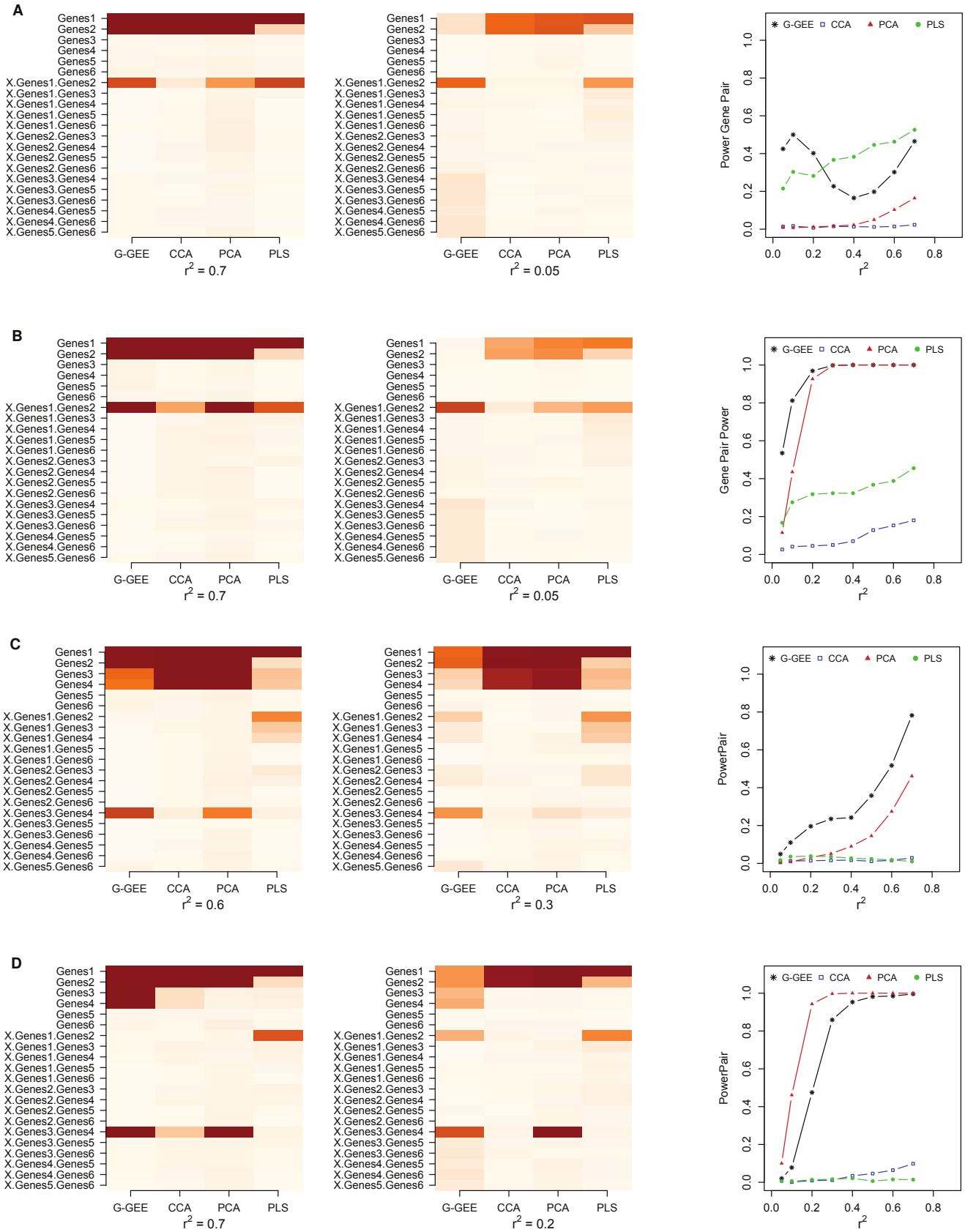

Figure 1: The figures present on the first two columns show the ratio of the number of times where each variable was significant on the total number of simulations for a given r^2. On the last column is represented the power of the four methods depending on the r^2. (A) and (B) correspond to the setting where gene 1 and gene 2 are simulated with main and interaction effects, (C) and (D) to the one where gene 1 and gene 2 are simulated with main effect and gene 3 and gene 4 with interaction effect. In (A) and (C) the phenotype was simulated under Wang et al. [2014b] model and under the PCA model in (B) and (D).

	A		B		C		D		NE
	r^2=0.7	r^2=0.05	r^2=0.7	r^2=0.05	r^2=0.6	r^2=0.3	r^2=0.7	r^2=0.2	r^2=0.7
pR_I^2	97.73	92.08	33.11	32.80	33.32	33.47	33.51	33.57	66.69
pR_M^2	98.84	95.57	66.42	66.97	66.60	66.57	66.70	66.56	33.62

TABLE I: Average part of the R^2 attributable to interaction or main effects by setting. (A) and (B) correspond to the setting where gene 1 and gene 2 are simulated with main and interaction effects, (C) and (D) to the one where gene 1 and gene 2 are simulated with main effect and gene 3 and gene 4 with interaction effect. In (A) and (C) the phenotype was simulated under Wang et al. [2014b] model and under the PCA model in (B) and (D). (NE) corresponds to the setting with numerous effects.

In the third setting, we consider 25 genes with two main effects for the two first genes and four interaction effects between gene 3 and gene 4, gene 5 and gene 6, gene 7 and gene 8, gene 9 and gene 10. The phenotype was simulated under the model proposed by Wang et al. [2014b] with a coefficient of determination R^2 set to 0.7. In this setting, the G-GEE method has a good power to detect all the simulated effects. The power varies between 0.50 and 0.53 for the four interactions. PCA and CCA methods detect the two main effects but tend to attribute false main effects for genes simulated with only interaction effects (Fig. 2). Only the first gene is detected with a high frequency for PLS method while the effects of other genes are less often detected as main or interaction effects. Moreover, the method tends to attributes a false interaction effect between the two first genes.

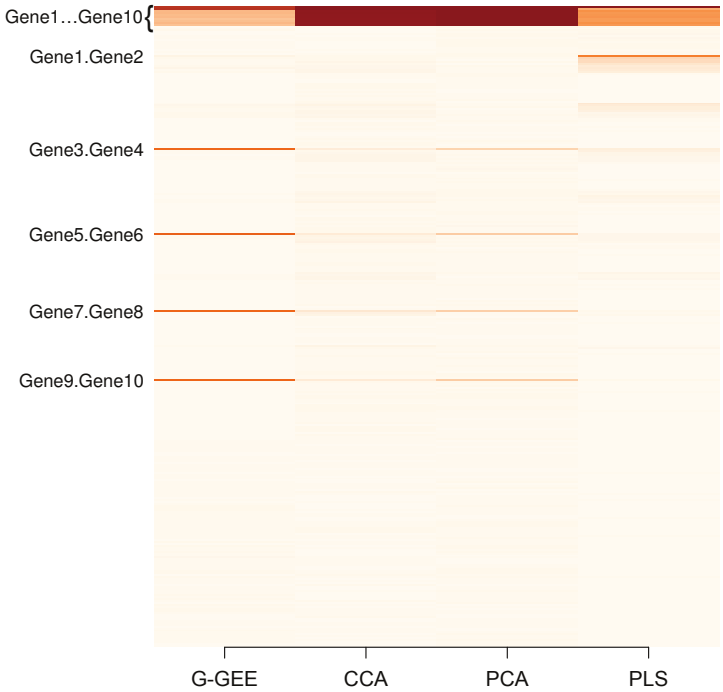

Figure 2: Matrix of the ratio of the number of times where each variable was significant on the total number of simulations for $r^2 = 0.7$ under Wang et al. [2014b] model for the phenotype simulation

Under each setting we determine the pR_I^2 and pR_M^2 average values that correspond to the proportion of the R^2 attributed to interaction and main effects, respectively. For most settings, the pR_I^2 depends on the number of simulated effects. With one interaction and two main effects simulated the R^2 part attributable to interaction effects is around 33% (Tab. I). Under the setting with numerous effect the average pR_I^2 is

67% because we consider four interaction effects for only two main effects. The R^2 distribution between main and interaction effects is not distinguishable in the setting where the phenotype is simulated under Wang et al. [2014b] model with the same main and interaction effects. The pR_I^2 and pR_M^2 values are respectively 98% and 99% (Tab. I).

Application

Ankylosing spondylitis (AS) is a common form of inflammatory arthritis predominantly affecting the spine and pelvis. It occurs with a prevalence of 0.1% to 1.4% depending on the considered population [Sieper et al., 2002]. Genetic factors contribute for more than 90% to the susceptibility risk to AS. Human leukocyte antigen (HLA) class I molecule HLA B27, belonging to the Major Histocompatibility Complex (MHC) region, was the first genetic risk factor identified as associated with ankylosing spondylitis in the 1970's [Schlosstein et al., 1973; Woodrow and Eastmond, 1978] and remains the most important risk locus for this pathology. Despite the strong association only a small portion of HLA-B27 carriers develop the disease. Furthermore studies in families suggest that less than 50% of the overall genetic risk is due to HLA-B27, what suggests that other genetic factors are involved [Thomas and Brown, 2010]. A number of updated reviews on AS genetics, including genome-wide association study (GWAS) results, identified new ankylosing spondylitis-susceptibility genes outside of the MHC region [Tsui et al., 2014; Reveille et al., 2010].

We applied all previously described methods to the AS dataset. The data contain 408 cases and 358 controls, and each individual was genotyped for 116, 513 SNPs with Immunochip technology. For each SNPs we obtained detailed genetic information, as gene affiliation, with the NCBI2R package [Melville, 2015] which annotates lists of SNPs with current information from NCBI. We considered only SNPs located within a single gene in order to form gene groups without overlap. We focused our analysis on a list of 29 genes previously identified as having a main effect in GWAS.

The four tested methods lead to different results and only PLS and G-GEE methods identify interactions. Indeed, none effects is detected by CCA whereas PCA detect only the main effect HLA-B without any interaction. PLS detects the main effect HLA-B but also identifies one interaction effect between the genes EOMES and BACH2. Our method G-GEE does not detect any main effect, but it exhibits two significant interactions, the first between the genes HLA-B and SULT1A1 and the second between IL23R and ERAP2.

Discussion

In this paper we compared different approaches to model gene gene epistasis in a penalized regression framework. Our first concerned was the detection of interaction effects. We thus defined a general model

and tested different interaction terms. We focused our analysis at the gene scale and compared four ways to design the interaction term. Some methods were inspired by previous proposed approaches based on dimensional reduction methods as principal component analysis (PCA), canonical component analysis (CCA) or partial least square analysis (PLS). We additionally proposed a new interaction modeling approach we called Gene-Gene Eigen Epistasis (G-GEE), allowing to build one interaction variable for each couple. The interaction variable was defined based on a criterion that maximizes the correlation between the phenotype and the pairwize SNP products matrix of the two genes. The interaction components were then introduced in a group LASSO penalized regression model which allows to take the gene structure into account and to simultaneously consider an important number of genes.

The power study of the different methods evaluated from simulated data provides us with rich information. In different papers, similar methods have been compared with different phenotype simulation settings. In this work we compared two simulation models, the first from a previous study [Wang et al., 2014b] that simulated the interaction component of each couple in a SNP pairwise product fashion. The second that defined interaction component as a pairwise product of representative variables of each gene. The G-GEE, PCA and CCA approaches performed better in the second setting whereas the PLS method was not very sensitive to the difference. Overall the G-GEE method performs well to detect interactions in all tested settings. The PLS method is characterized by a lack of power in detecting interactions. Favorable contexts for the PLS method appear only when the related main effect are also present. When the simulated main and interaction effects do not concern the same genes the detection performances of the PLS approach drastically collapses. The other methods primarily detect the main effects but encounter problem with the interactions which are often considered as main effects. This confusion phenomenon is mainly visible in the Wang et al. [2014b] simulation where the main and interaction effects concern different genes.

The gene scale dimension of the proposed method drastically reduces the number of interaction variables to consider for a genetic region compared to SNP-SNP interaction approaches. This reduction of problem size allows to deal with larger problems. Moreover using a penalized regression method allows to consider a true multivariate approach on a larger number of genes. Notice that it also extends other proposed gene scale approaches such as the one presented by Wang et al. [2009]. The advantage to simultaneously consider a relatively important number of genes gives us the possibility to detect interactions between various genetic regions. The method can thus be used to point out whole genetic regions. It can be viewed as a first step before using SNP-SNP interaction methods which may provide more accurate information.

As the G-GEE method is not able yet to consider all the human genes at the same time, it is necessary to specify a gene list, which we want to explore for potential interactions. As the method is not powerful for main effect detection, it is safer to use previously acquired knowledge of the genetic effects, or to

use a pre-step method to detect main effects. Another limitation of the method is the gene size. The computation of the Gene-Eigen epistasis vector of two genes of size p_r and p_s requires the computation and eigen decomposition of $(p_r p_s) \times (p_r p_s)$ matrix.

The perspective of this work is to increase the performance of the G-GEE method by optimizing the computational cost and explore new interaction functions to plug in the G-GEE criterion.

References

Bécu JM, Grandvalet Y, Ambroise C, Dalmasso C. 2015. Beyond Support in Two-Stage Variable Selection. ArXiv150507281 Stat .

Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. 2006. Powerful Multilocus Tests of Genetic Association in the Presence of Gene-Gene and Gene-Environment Interactions. Am J Hum Genet 79:1002–1016.

D'Angelo GM, Rao D, Gu CC. 2009. Combining least absolute shrinkage and selection operator (lasso) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. BMC Proc 3 Suppl 7:S62.

Haig D. 2011. Does heritability hide in epistasis between linked SNPs? Eur J Hum Genet 19:123.

He J, Wang K, Edmondson AC, Rader DJ, Li C, Li M. 2011. Gene-based interaction analysis by incorporating external linkage disequilibrium information. Eur J Hum Genet 19:164–172.

Li J, Tang R, Biernacka JM, de Andrade M. 2009. Identification of gene-gene interaction using principal components. BMC Proc 3:S78.

Li S, Cui Y. 2012. Gene-centric gene–gene interaction: A model-based kernel machine method. Ann Appl Stat 6:1134–1161.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753.

Melville S. 2015. NCBI2R: Queries and Annotates SNPs, Genes and Pathway from NCBI.

Rajapakse I, Perlman MD, Martin PJ, Hansen JA, Kooperberg C. 2012. Multivariate detection of gene-gene interactions. Genet Epidemiol 36:622–630.

Reveille JD, Sims AM, Danoy P, Evans DM, Leo P, Pointon JJ, Jin R, Zhou X, Bradbury LA, Appleton LH, Davis JC, Diekman L, Doan T, Dowling A, Duan R, Duncan EL, Farrar C, Hadler J, Harvey D, Karaderi T, Mogg R, Pomeroy E, Pryce K, Taylor J, Savage L, Deloukas P, Kumanduri V, Peltonen L, Ring SM, Whittaker P, Glazov E, Thomas GP, Maksymowych WP, Inman RD, Ward MM, Stone MA, Weisman MH, Wordsworth BP, Brown MA. 2010. Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. Nat Genet 42:123–127.

Schlosstein L, Terasaki PI, Bluestone R, Pearson CM. 1973. High Association of an HL-A Antigen, W27, with Ankylosing Spondylitis. N Engl J Med 288:704–706.

Sieper J, Braun J, Rudwaleit M, Boonen A, Zink A. 2002. Ankylosing spondylitis: an overview. Ann Rheum Dis 61:iii8–iii18.

Steen KV. 2012. Travelling the world of gene-gene interactions. Brief Bioinformatics 13:1–19.

Thomas GP, Brown MA. 2010. Genetics and genomics of ankylosing spondylitis. Immunol Rev 233:162–180.

Tsui F, Tsui HW, Akram A, Haroon N, Inman R. 2014. The genetic basis of ankylosing spondylitis: new insights into disease pathogenesis. Appl Clin Genet :105.

Wang T, Ho G, Ye K, Strickler H, Elston RC. 2009. A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. Genet Epidemiol 33:6–15.

Wang X, Epstein MP, Tzeng JY. 2014a. Analysis of gene-gene interactions using gene-trait similarity regression. Hum Hered 78:17–26.

Wang X, Zhang D, Tzeng JY. 2014b. Pathway-Guided Identification of Gene-Gene Interactions. Annals of Human Genetics 78:478–491.

Wei WH, Hemani G, Haley CS. 2014. Detecting epistasis in human complex traits. Nat Rev Genet 15:722–733.

Woodrow JC, Eastmond CJ. 1978. HLA B27 and the genetics of ankylosing spondylitis. Ann Rheum Dis 37:504–509.

Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B 68:49–67.

Zhang F, Wagener D. 2008. An approach to incorporate linkage disequilibrium structure into genomic association analysis. Journal of Genetics and Genomics 35:381–385.

Zuk O, Hechter E, Sunyaev SR, Lander ES. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. PNAS 109:1193–1198.