Performance of Model Based Multifactor Dimensionality Reduction Methods for Epistasis Detection by Controlling Population Structure

Fentaw Abegaz[1,*], François Van Lishout[1], Jestinah M Mahachie John[1], Kridsadakorn Chiachoompu[1], Archana Bhardwaj[1], Diane Duroux[1], Elena S Gusareva[1], Zhi Wei[2], Hakon Hakonarson[3,4], and Kristel Van Steen[1,5]

* To whom correspondence should be addressed. Tel: +31 621331831; Email: fentawabegaz@yahoo.com

Abstract

Background: In genome-wide association studies the extent and impact of confounding due population structure has been well recognized. Inadequate handling of such confounding is likely to lead to spurious associations, hampering replication and the identification of causal variants. Several strategies have been developed for protecting associations against confounding, the most popular one being based on Principal Component Analysis. In contrast, the extent and impact of confounding due to population structure in gene-gene interaction association epistasis studies is much less investigated and understood. In particular, the role of nonlinear genetic population substructure in epistasis detection is largely under-investigated, especially outside a regression framework.

Methods: In order to identify causal variants in synergy, to improve interpretability and replicability of epistasis results, we introduce three strategies based on model-based multifactor dimensionality reduction approach for structured populations namely: MBMDR-PC, MBMDR-PG and MBMDR-GC.

Results: Simulation results comparing the performance of various approaches show that in the presence of population structure MBMDR-PC and MBMDR-PG consistently better control type I error rate at the nominal level than MBMDR-GC. Moreover, our proposed three methods of population structure correction outperform MDR-SP in terms of statistical power.

Conclusion: We demonstrate through extensive simulation studies the effect of various degrees of genetic population structure and relatedness on epistasis detection and propose appropriate remedial measures based on linear and nonlinear sample genetic similarity.

Key words: Epistasis, Population Structure, Confounding, GWAS, GWAIS, MB-MDR, Gene-Gene Interaction, Population Stratification, Principal Components

Background

Genome-Wide Association Studies (GWAS) are an effective approach for identifying genetic variants associated to disease risk [1]. In the context of such studies, population stratification refers to systematic ancestry differences between cases and controls [1]. The phenomenon is of particular concern in study designs with unrelated individuals. In contrast, family-based genetic association studies offer protection from population stratification, by using family data as internal controls, although at the expense of some loss of power from genotypic overmatching [2,3]. For case-control genetic association studies, spurious associations are caused by the co-occurrence of two factors: a difference in proportion of individuals from two (or more) subpopulations in cases and controls, and subpopulations having differing allele frequencies at the locus under investigation. This is in fact a special case of Simpson's Paradox [4]. In general, this statistical phenomenon causes a potential bias in data analysis and occurs when a relationship or association between two variables reverses when a third factor, called a confounding variable, is introduced. The paradox also occurs if an association reverses when the data are aggregated over a third variable. Increasing the sample size is usually not a remedy for this issue, but may worsen the problem [5]. Several causes exist for population stratification. The basic one being shared genetic ancestry as a result of non-random mating between subgroups in a population due to various reasons, which may include social, cultural or geographical ones. From an evolutionary point of view, not only population stratification, but also admixture (i.e., inter-mating between genetically distinct groups) is created by human mating patterns. Potential consequences of population stratification are confounding, cryptic relatedness (i.e., unobserved ancestral relationships between individual cases and controls causing them to be non-independent) and selection bias [6,7].

In case/control GWA studies, several strategies have been introduced in the literature for protecting against population structure mainly based on Principal Components Analysis (PCA). In contrast, the extent and impact of confounding due to population structure in gene-gene interaction studies is much less investigated and understood. However, the growing interest in the importance of detecting gene–gene interactions in the development and progression of complex diseases has led to the development of several tools; to name but a few: generalized linear regression models (GLM), BOOST [8], Model Based Multifactor Dimensionality Reduction (MB-MDR) [9,10], Multifactor Dimensionality Reduction (MDR) [11], Random Forest [12], PLINK [13], BiForce [14], Bayesian Models (e.g., BEAM) [15] and several others. For extensive reviews and appropriate references, please refer to [16–20]. However, the literature on epistasis detection in structured populations is very limited, apart from scenarios using a regression framework for association testing. On the other hand, Model-Based Multifactor Dimensionality Reduction (MB-MDR) offers a general framework and software tool for epistasis detection that can offer flexible maneuvering between different measurement scales for phenotypes and genomic predictors [9,10,21]. The MDR-SP method [22] combines MDR [11] with ideas implemented in the EIGENSTRAT software [23], a widely used software in GWAS that detects and corrects for population stratification via PCA.

In this article, we introduce strategies to account for population structure in epistasis studies using the MB-MDR framework. In particular, for the remainder of this article, we restrict attention to case-control study designs (binary original traits) and biallelic Single Nucleotide Polymorphisms (SNPs) as genetic markers. We propose and fully describe three strategies: i) MBMDR-PC, ii) MBMDR-PG and iii) MBMDR-GC. In MBMDR-PC, principal components (PCs) adjusted phenotypes but original genotypes are used to detect epistatic SNP pairs, similar to [23]. In MBMDR-PG adjusted phenotypes are obtained from fitting logistic mixed

(polygenic) models on the original binary trait, hereby allowing to adjust for additional structure such as those arising from family relationships and cryptic relatedness. In MBMDR-GC, we follow principles of Genomic Control correction in GWAS, but allow for multi-locus adaptivity. These methods are evaluated via extensive simulation studies which, to our knowledge, are unique in their kind in that complex nonlinear population structures, in the form of structural epistasis, are considered as well. Here, we let structural epistasis refer to the presence of interacting markers driving population differences or population substructure. All proposed strategies are formally compared to MDR-SP [22] in terms of type I error control and statistical power. Our work is important as it highlights the impact of nonlinear genetic population substructure on epistasis signal detection in GWAIS (Genome-Wide Association Interaction Studies).

Material and Methods

All proposed genome-wide epistasis screening strategies in structured populations are built on the Model-Based Multifactor Dimensionality Reduction method [9,10,24,25], as implemented in version 4.4.1. Even though the MB-MDR framework can be used for higher level interaction detection and various outcome measurement scales and study designs, here we restrict attention to pair-wise interactions with default settings, including lower-order genetic effects correction and multiple testing correction via MaxT [10], unless specified otherwise. We describe the newly introduced methods as follows.

MBMDR-PC: accounting for genomic structure by PCs

In MBMDR-PC, similar to EIGENSTRAT [23] , we use either linear or nonlinear (kernel) PCs to correct for population structure. The popular EIGENSTRAT software to correct for population structure in GWAS contexts uses top linear PCs as covariates in a multiple regression [26]. Many ad hoc procedures and formal statistical tests exist to determine the optimal number of principal components to correct for population structure. It is a common

practice to take between 2-10 principal components for correcting population structure in GWAS involving several countries. Even though linear PCA is most popular and adequate in most cases to capture ancestry genetic background, PCA may fail to capture nonlinear population structure in genetics as shown in [27]. The nonlinear method developed by Alanis-Lobato and colleagues is based on a non-centered Minimum Curvilinear Embedding (ncMCE) kernel. Whereas the latter is able to better capture phylogenetic signals in samples, PCA better seems to reflect geographic dependencies [28]. Alternatively, kernel-based PCA can be adopted to account for nonlinear structures in high dimensional genetics data. In case-control epistasis studies, where the phenotype Y represents disease status (1 affected, 0 unaffected), the new adjusted phenotype Y_i^{adj} can be computed by fitting a logistic regression using the first few (linear or nonlinear) principal components ($W_1, \cdots, W_r$) and subtracting model-fitted values from observed phenotype values:

$$\text{logit}(\pi_i) = \alpha + \beta_1 W_{i1} + \cdots + \beta_r W_{ir},$$

$$Y_i^{adj} = Y_i - \hat{\pi}_i ,$$

$$\text{where} \quad \hat{\pi}_i = \frac{\exp\left(\hat{\alpha} + \hat{\beta}_1 W_{i1} + \cdots + \hat{\beta}_r W_{ir}\right)}{1 + \exp\left(\hat{\alpha} + \hat{\beta}_1 W_{i1} + \cdots + \hat{\beta}_r W_{ir}\right)}.$$

The new adjusted phenotype Y^{adj} is taken as input to classic MB-MDR, in an attempt to capture genetic interactions that are not spurious due to inadequate handling of population structures. Detail of the MBMDR-PC approach is outlined in [29,30].

MBMDR-PG: accounting for genomic structure due to families and cryptic relatedness via the extended Polygenic Model

Family structure or cryptic relatedness may induce phenotypic similarity between individuals and may confound gene-phenotype associations in GWAS when not properly accounted for. Whereas PCs have proven useful in GWAs and structured populations due to shared genetic ancestry, they are not suitable to adequately protect for the effects of familial or cryptic

relatedness on GWAS [1]. With the recent developments of computationally efficient algorithms, mixed models have become feasible in the context of GWAS as well as GWAIS, in structured populations, whether this structure presents population stratification, known or unknown relatedness. For quite some time, GWAS for binary traits have been analyzed with linear mixed models, assuming that little harm is done when sample sizes are in the thousands as is often the case with consortium data [31]. However, Chen et al. [32] showed that linear mixed models are really inappropriate for analyzing binary traits when population stratification induces violation of the constant residual variance assumption in linear mixed models. Therefore, these authors developed a computationally efficient logistic mixed model for binary trait GWAS in the presence of population structure as well as familial and cryptic relatedness. In the same spirit of the logistic regression models adopted before, a logistic mixed model that includes interaction effect between two SNPs can be defined as

$$\text{logit}(\pi_i) = \alpha + \gamma_1 G_{ij} + \gamma_2 G_{ik} + \theta G_{ij} G_{ik} + \vartheta_i + \varepsilon ,$$

$$\vartheta \sim N\left(0, \sigma_g^2 \Omega\right), \text{ and } \varepsilon \sim N(0, \sigma_e^2),$$

where $\pi_i = P\left(Y_i = 1 | G_{ij}, G_{ik}, \vartheta\right)$ is the probability of disease for subject i, conditional on SNPs G_{ij}, G_{ik} and random effects ϑ_i. Here, ϑ is a $N \times 1$ vector of random effects assumed to follow a multivariate Gaussian distribution, σ_g^2 is the additive genetic variance, and Ω is the genetic similarity matrix between all pairs of individuals (dimension $N \times N$) such that Ω_{il} represents the similarity between individuals i and l. An estimate of the genetic similarity matrix, Ω, is required which can be obtained from a large number of genetic variants [33]. Fitting the model involves integrating over the random effects vector ϑ with respect to the Gaussian distribution so that the likelihood is maximized with respect to the parameters $\left\{\alpha, \gamma_1, \gamma_2, \theta, \sigma_g^2, \sigma_e^2\right\}$ [34]. In MBMDR-PG we obtain the adjusted phenotype from the residuals of fitting the logistic random effect model using the R package *GMMAT* (Generalized Linear Mixed Model Association

Test) [32]. Then, similar to MBMDR-PC we use the adjusted phenotype as input for interaction analysis with MBMDR.

MBMDR-GC: accounting for genomic structure via genomic control

The genomic control method introduced in [35] is computationally simple and fast to control for population structure in case-control association studies. The key idea is to divide the observed association test statistic by a single factor, λ_{GC} , which measures the overall inflation in association test statistic due to population stratification. The factor λ_{GC} can be estimated by dividing the mediums of the observed association test statistics across a set of markers by the theoretical median of the association test statistic. Notably, corrective factors computed in this sense may turn out to be less than 1 and may actually inflate observed test values rather than deflating them. Although genomic control have proven useful in a variety of contexts, Price et al. [23] pointed out that the common deflation factor applied to all SNPs where some SNPs differ in their allele frequencies across ancestral populations more than others could lead to loss of power. As a solution, [36] considered test specific genomic control. MBMBDR-GC also employs test-specific genomic control, adapted to the MB-MDR testing framework.

In MBMBDR-GC principles of classic GC in GWAS for structured populations are adopted [23]. Large differences between several multi-locus genotype frequencies across populations may lead to power loss when a single corrective inflation factor GC is used. Therefore, in MBMDR-GC the definition of GC is adapted and the permutation null data generated in MBMDR (*step 3*, Figure 1) is exploited to estimate a test-specific GC factor, similar to [36]. In particular, for the jth SNP-SNP interaction pair, the corrective factor $\lambda_{GC,j}$ is estimated as

$$\lambda_{GC,j} = \frac{The\ median\ of\ observed\ interaction\ test\ statistics\ across\ all\ pairwise\ interactions}{The\ \ expected\ median\ of\ the\ jth\ SNP-SNP\ interaction\ test\ statistic\ T_j}, j = 1, \dots, J,$$

where J is the total number of pairwise interactions and for which the expected median of the j^{th} SNP-SNP interaction test statistic is computed from 1000 permutations under a null distribution constructed by randomly permuting the phenotype values. Then the adjusted test statistic for the j^{th} interaction pair becomes $T_j/\lambda_{GC,j}$, which serve as input to the MB-MDR multiple testing routines instead of their unadjusted counterparts.

Application on synthetic HapMap data

To assess type I error control and power performance of MBMDR-PC, MBMDR-PG and MBMDR-GC, and to compare it to MDR-SP [22], we set up a series of simulation settings, involving either unrelated or related individuals, as depicted in Figure 1. Two main strategies were adopted for our simulation study with unrelated individuals involving: 1) HapMap data as a template with and without structural epistasis, 2) newly generated discrete populations without a reference template, with and without structural epistasis (referred to as model-based data). More detailed explanations are given next.

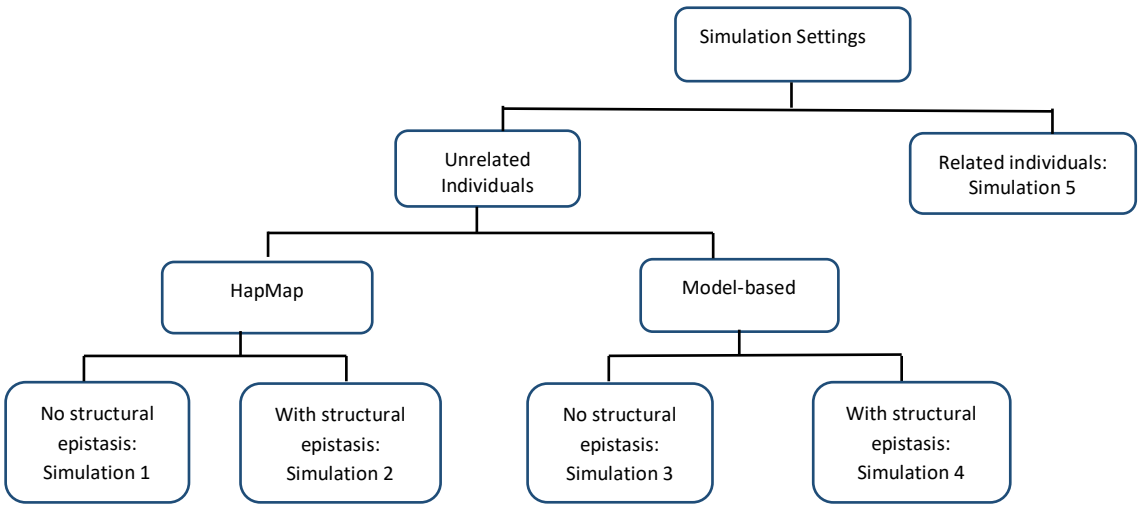

Figure 1. Flow of considered simulation settings.

In summarizing the simulation results, type I error rates were obtained as the proportion of the number of simulated datasets for which a pair of SNPs were found significant at the 5% level after correcting for multiple testing. Similarly, the power was obtained as the proportion of the number of simulated datasets for which only the functional pair of SNPs were found significant at the 5% level after correcting for multiple testing.

Simulation setting 1: *Synthetic* data derived from HapMap in the absence of structural epistasis

For each simulated set, we considered 200, 300 or 400 individuals, half of them cases and half of them controls, 80% controls and 40% cases as European (CEU) and 20% controls and 60% cases as African (YRI). We followed the simulation strategy adopted for MDR-SP [22] to generate genotype data with unlinked SNPs. In particular, $L \in \{200, 400, 800\}$, independent SNPs were randomly selected from the total number of SNPs from the pooled HapMap3 (CEU and YRI) data with quality control that includes (only founders, HWE *p*-value threshold of 0.001, individual and genotype missing rates of 5% and 2%, respectively, minor allele frequency $MAF > 0.05$ and LD pruning threshold of $r^2 = 0.75$) (http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html), and minor allele frequencies were extracted for these two populations. Genotypes were then generated under the assumption of Hardy-Weinberg equilibrium (HWE). The genotypes for the L unlinked SNPs were subsequently used to compute principal components and the first 10 principal were retained to capture population substructure.

Since the aim of this study is not to evaluate multiple testing strategies but to evaluate approaches for population structure control in epistasis, SNPs screened for epistasis were generated as follows. A total of only 10 candidate SNPs were selected at random from the available CEU and YRI SNP panel, with the restriction that the minor allele frequency

difference between CEU and YRI was larger than $d \in \{0.1, 0.3\}$. Genotypes for unlinked null loci were generated as before. A total of 1000 replicates of null data (i.e., no association between SNPs and trait) were created by repeating the process above 1000 times and by randomly assigning individuals to disease. To be able to assess power, each genetic replicate was appended with 2 functional SNPs. Disease status generation was based on 6 pure epistasis models (Supplementary material – Table S1). These models are heavily used in the epistasis field, for instance to evaluate MDR [37], MDR-PDT [38], MBMDR [24] and MDR-SP [22]. They involve equal MAFs for functional SNP pairs, with $\text{MAFs} \in \{0.50, 0.25, 0.10\}$ and no main effects. We randomly selected 2 SNPs from the pooled CEU and YRI HapMap data ensuring that the MAF in the CEU population at each of the 2 SNP was within ±0.02 of the given MAF in the chosen pure epistasis disease model. Two-locus genotypes for the functional SNP pair were then generated, conditional on fixed and equal numbers of cases and controls (100, 150 and 200) each. This process was repeated 1000 times.

Simulation setting 2: Synthetic data derived from HapMap with structural epistasis

To introduce structural epistasis into our simulation study for GWAIS, we considered four HapMap populations: 2 closely related populations CHB and JPT (F_{ST} = 0.007) and 2 distant populations CEU and YRI (F_{ST}=0.153). Then, to detect epistasis via adjusting nonlinear structural differences between these populations we applied the aforementioned MB-MDR methods for structured populations. We subsequently identified all significant SNP-SNP interaction pairs, adjusted for main effects and corrected for multiple testing with default options. Based on these results, several approaches were taken to generate genotypes in the absence or presence of epistatic differences between populations. Approach *1*: we generated 10,200 unlinked random genotypes including a) 10,000 SNPs randomly generated from the pooled CEU, YRI, CHB and JPT data, without association to disease and population

structure, similar to simulation set 1 and b) 100 pairs of SNPs randomly selected from the aforementioned significant pairs of SNPs related to population structure comparing CHB to JPT, and CEU to YRI. From these 100 pairs, we extracted the empirical proportion of corresponding 9 two-locus genotype combinations. The associated penetrance functions were used to generate the additional 200 unlinked genotypes, by conditioning on fixed sample sizes of $\{100, 250\}$ from each of the four populations. Approach 2: we simulated 110 candidate random genotypes including a) 100 without association to disease with population structure similar to Approach 1 -b) and 5 significantly interacting SNP pairs with population structure similar to Approach 1 -b). Approach 3: Two functional genotypes were randomly selected from the significant pairs that were found to be associated to population structure in such a way that the MAFs in the CEU and CHB populations at each SNP was within ±0.1 of the given MAFs in the disease model (Supplementary Table S1). A total of 1000 replicates were generated for total samples sizes of $\{400, 1000\}$ and proportions of cases and controls according to 60:40.

Unlinked genotypes obtained via Approach 1 were used to extract principal components to control for population structure. The first 10 principal components were used to capture population structure in epistasis analyses. Candidate genotypes generated via Approach 2 were used to evaluate type I error rates of proposed population correction strategies in GWAIS, whereas functional genotypes as in Approach 3 were used in methods power analyses. Various ways of computing principal components were implemented to capture synthetic data underlying population structure. In particular, we considered linear principal component analysis (linear PCA), as applied to genetic data in [23] , kernel PCA with a radial basis kernel, as implemented in the R package *kernlab* (Kernel-Based Machine Learning Lab), and ncMCE (non-centered multicurvilinearity embedding) kernel-based PCA introduced in [27] as an alternative to capture nonlinear genetic differences between populations.

Simulation setting 3: Model-based discrete populations in the absence of structural epistasis

Here, we simulated a large number of biallellic genotype frequencies for each individual in subpopulations, using Balding-Nicholas models [39], similar to [35,36]. First, an ancestral allele frequency p_a was randomly sampled from the uniform distribution in the interval [0.05, 0.95]. Second, Wright's coefficient of inbreeding F_{ST} was specified for the subpopulations $F_r \in \{0.01,\ 0.03\},\ r = 1, 2$. Third, the allele frequency $p_{ij}^{(r)}$ of individual i for genotype j in subpopulation r was simulated from a beta distribution with parameters $p_a\left(\frac{F_r}{1-F_r}\right)$ and $(1-p_a)\left(\frac{F_r}{1-F_r}\right),\ r = 1, 2,\ i = 1, \ldots, N \text{ and } j = 1, \ldots, M$. Then, genotype values $\{0, 1, 2\}$ were simulated from a multinomial distribution with probabilities - computed without assuming HWE - by $\left\{ F_t\, p_{ij}^{(r)} + (1-F_t)\left(p_{ij}^{(r)}\right)^2,\ 2(1-F_r)\, p_{ij}^{(r)}\left(1-p_{ij}^{(r)}\right), \mathrm{F}_r\left(1-p_{ij}^{(r)}\right) + (1-F_r)\left(1-p_{ij}^{(r)}\right)^2, \right\}$ (see [36] and references therein). We thus generated 1000 unlinked SNPs to calculate principal components similar to [23] . In addition, 100 SNPs were generated in a similar way as the unlinked SNPs to evaluate type I error rates. For power comparison two functional SNPs were generated taking into account the six genetic disease models presented as supplementary information (Table S1). This procedure was repeated 1000 times with total samples sizes of $\{500, 1000\}$ and proportions of cases and controls according to 60:40.

Simulation setting 4: Model-based discrete populations in the presence of structural epistasis

Instead of relying on a data-driven empirical penetrance table for structural epistasis as before (Simulation setting 2), we considered a checker-board type of model as in Table 1,

which describes epistatic genetic difference between the populations using the XOR model. In Table 1, the parameter β_0 was taken to be the average penetrance (in the absence of any genetic effect), whereas β_1 captured the increase in penetrance when having the specific 2-locus genotype. In our simulations we assumed $\beta_0 = 0$ and β_1 = 0.35 and 0.20 for populations 1 and 2, respectively. Then, we generated 1000 unlinked random genotypes including a) 800 SNPs randomly generated similar to *Simulation setting 3* using F_{ST} in the two subpopulations $F_r \in \{0.001,\ 0.001\},\ r = 1, 2$; b) 100 pairs from each population similar to *Simulation setting 2 (1b)* using the population-specific penetrance values given in Table 3 with $\beta_{j,1} = \beta_1 + \varepsilon$, $j = 1, \ldots, 50$, where ε is randomly drawn from $uniform\ (0, 0.05)$. To assess type I error rate, 120 SNPs are generated of which 100 similar to (a) and 10 pairs similar to (b). A total of 1000 replicates were generated for total samples sizes of $\{200, 500, 1000\}$ and proportions of cases and controls according to two scenarios 60:40 and 80:20. This dataset was used to construct principal components.

Table 1. Checkerboard stratification penetrance models for structural epistasis

Population 1				Population 2			
	BB	Bb	Bb		BB	Bb	bb
AA	β_0	$\beta_0 + \beta_1$	β_0	AA	$\beta_0 + \beta_1$	β_0	$\beta_0 + \beta_1$
Aa	$\beta_0 + \beta_1$	β_0	$\beta_0 + \beta_1$	Aa	β_0	$\beta_0 + \beta_1$	β_0
Aa	β_0	$\beta_0 + \beta_1$	β_0	aa	$\beta_0 + \beta_1$	β_0	$\beta_0 + \beta_1$

Simulation setting 5: Simulating genotypes for related individuals

Inspired by [40], we simulated 1000 replicate datasets consisting of 250 nuclear families, with the number of children drawn from a multinomial distribution with probabilities 1/4 to have one child, 1/2 to have two children, and 1/4 to have three children. On average, this gave rise to 1000 individuals. To generate parental genotypes, we generated 10 biallelic markers in linkage equilibrium, and assuming Hardy-Weinberg equilibrium. The allele frequencies of the functional SNP pair (SNP_1, SNP_2) were taken to be equal, and varied as $(p_1, p_2) = (p, p),\ p \in$

$(0.1,0.25,0.5)$. The allele frequencies of the 8 remaining non-functional SNPs were fixed at $p_j = 0.1 + (j-3)0.05,\ j = 3, \ldots 10$. Children's genotypes were assumed to follow Mendelian inheritance patterns. Disease penetrance for parents and children was based on Model M170, as discussed in [41]. This epistasis model is similar to Model 1 in Table S1 (supplementary material). However, we fixed the total heritability h^2 and the proportion of the total variance explained by the two-locus model variance at 0.5 and 0.05, respectively. As family relationships may induce phenotype similarity, this simulation setting was used to evaluate the performance of MBMDR-PG.

Results

Simulation setting 1. Type I error estimates obtained for simulation setting 1 via application of MBMDR-PC, MBMDR-PG, MBMDR-GC and MDR-SP to 1000 replicated samples are presented in Table 2. In case of a single homogeneous population (CEU only) none of the estimated type I errors are significantly different from the nominal 0.05 FWER level, with a 95% confidence interval of (0.036, 0.064) [22]. This is the case, for all considered combinations of population structure correction methods, sample sizes and number of SNPs. In case of structured samples (in particular consisting of CEU and YRI), MBMDR-PC estimated type I errors presented in Table 2 always follow Bradley's liberal criterion. In addition, it can be seen from Table 2 that all the estimated type I error rates for MBMDR-PC are within the 95% confidence interval but it is not the case for MDR-SP. However, many type I error rate estimates based on MBMDR-GC do not fall within the 95% interval. The results of MBMDR-PG are similar to MBMDR-PC (results not shown).

Table 2. Estimates of Type I error for MBMDR-PC, MBMDR-GC and MDR-SP, with a nominal 0.05 FWER level.

			b = 40%	
Method	Sample sizes	Markers	*d* = 0.1	*d* = 0.3
MBMDR-PC	200	200	0.050	0.046
		400	0.051	0.051
		800	0.046	0.048
	300	200	0.046	0.047
		400	0.047	0.049
		800	0.049	0.047
	400	200	0.050	0.053
		400	0.051	0.054
		800	0.048	0.046
MDR-SP	200	200	0.054	0.054
		400	0.062	0.055
		800	0.062	0.050
	300	200	0.051	0.056
		400	0.055	0.051
		800	0.044	0.046
	400	200	0.044	0.050
		400	0.046	0.052
		800	0.044	0.065
MBMDR-GC	200	200	0.059	0.058
		400	0.061	0.065
		800	0.064	0.067
	300	200	0.060	0.065
		400	0.062	0.071
		800	0.063	0.068
	400	200	0.066	0.069
		400	0.061	0.071
		800	0.063	0.066

Note: *d* denotes the difference of candidate allele frequencies in the two subpopulations, and *b* denotes the percentage of cases from the European subpopulation

In Figure 2, for allele frequencies difference $d = 0.1$ and percentage of European population $b = 40\%$, MBMDR-PC is significantly more powerful than MDR-SP under all models considered in particular for small sample sizes. Moreover, MBMDR-PC outperforms MDR-SP even for large sample sizes in the models 5 and 6. Notably, these epistasis models are the toughest of the 6 considered Ritchie models in that they involve functional SNP pairs with the lowest MAFs (0.10). As the sample size increases the power of both methods

increases. We also included the power results of MBMDR-PG, which are almost similar to MBMDR-PC. Similar results follow when $d = 0.3$ and $b = 40\%$. In addition, the results of power based on varying number of unlinked markers are included in Figure S1 (Supplementary material) that suggest there is not much difference in the power of MBMDR-PC using 200, 400 and 800 unlinked markers for computing principal components to control population structure in our data simulation.

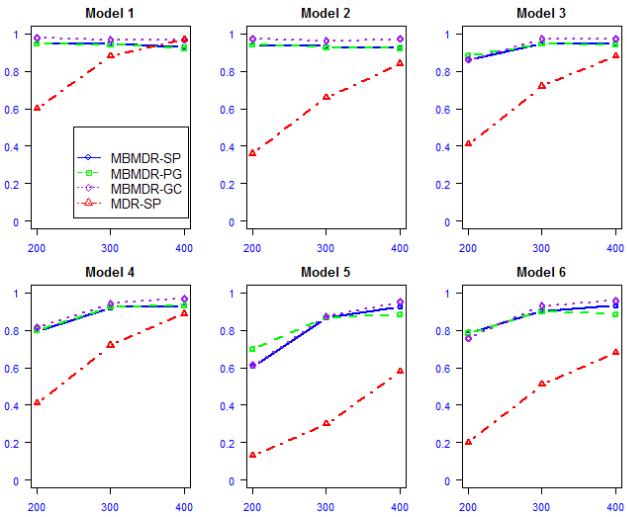

Figure 2. Power estimates for MBMDR-PC (blue/solid line), MBMDR-PG (green/dashed line) and MDR-SP (red/dotted line) under the six disease models based on simulated data on CEU and YRI populations with a difference of 0.3 minor allele frequency between the two populations. Percentage of cases and control from the CEU are 40% and 80%, respectively. The power (y-axis) is computed using 10 candidate SNPs. PCs are computed from 200 unlinked SNPs.

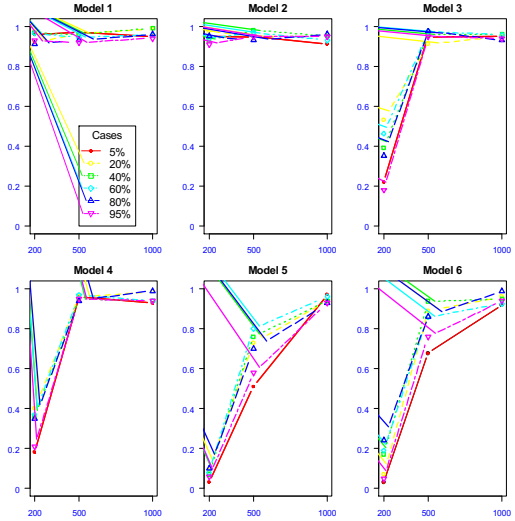

Figure 3. Power estimates according to varying proportions of cases and controls in six disease epistasis models and variable sample sizes (200, 500, 1000). The percentage of cases in one of the two populations are shown.

Simulation setting 2. The estimated power of MBMDR-PC under the discrete population simulation setting are shown in Figure 3. The results show that MBMDR-PC has high power in all scenarios of varying case-control proportions in all disease models with large samples. The power of MBMDR-PC is low for small sample sizes (100 samples from each of the two populations) for disease models with moderate and small minor allele frequencies. In general, the simulation results of varying case-control proportions have no considerable impact on the

power of MBMDR-PC method for large samples of 1000 or more. The results of estimated Type I error rates for varying proportions of cases and controls with and without main effect and principal component corrections are displayed in Figure S2 (Supplementary material). From this figure we see that MBMDR-PC performs well in controlling type I error rate at the nominal 0.05 FWER level with and without main effects correction in all scenarios of case-control proportions (Figure S2 A and B). Use of the original MBMDR without population and main effect corrections in case of structured population leads to inflated type I error rates (Figure S2 D and C) in case of small samples and large difference in case-control proportions.

Simulation setting 3. To evaluate the performance of MBMDR-PC in multiple subpopulations we evaluate three principal component extraction methods: linear, kernel and ncMCE. Pairwise PC-plots for the first three principal components computed from the unlinked null SNPs are shown in Figure 4. The plot of the first and the second PCs obtained from linear PCA method (Figure 4 A1) fails to separate CHB and JPT populations. Similar result were reported in [27]. However, the plot of the second and third linear PCs (Figure 4 A3) differentiate all four populations. In case of kernel-based PCs, the four populations are clearly separable in any of the pairwise PC plots of the first three kernel PCs (Figure 4 B1-B3). On the other hand, the plot of the first versus the second ncMCE based PCs (Figure 4 C1) was able to reveal the hierarchical structure of the four populations, reflecting the phylogenetics of these populations, as discussed in Alanis-Lobato and colleagues [27].

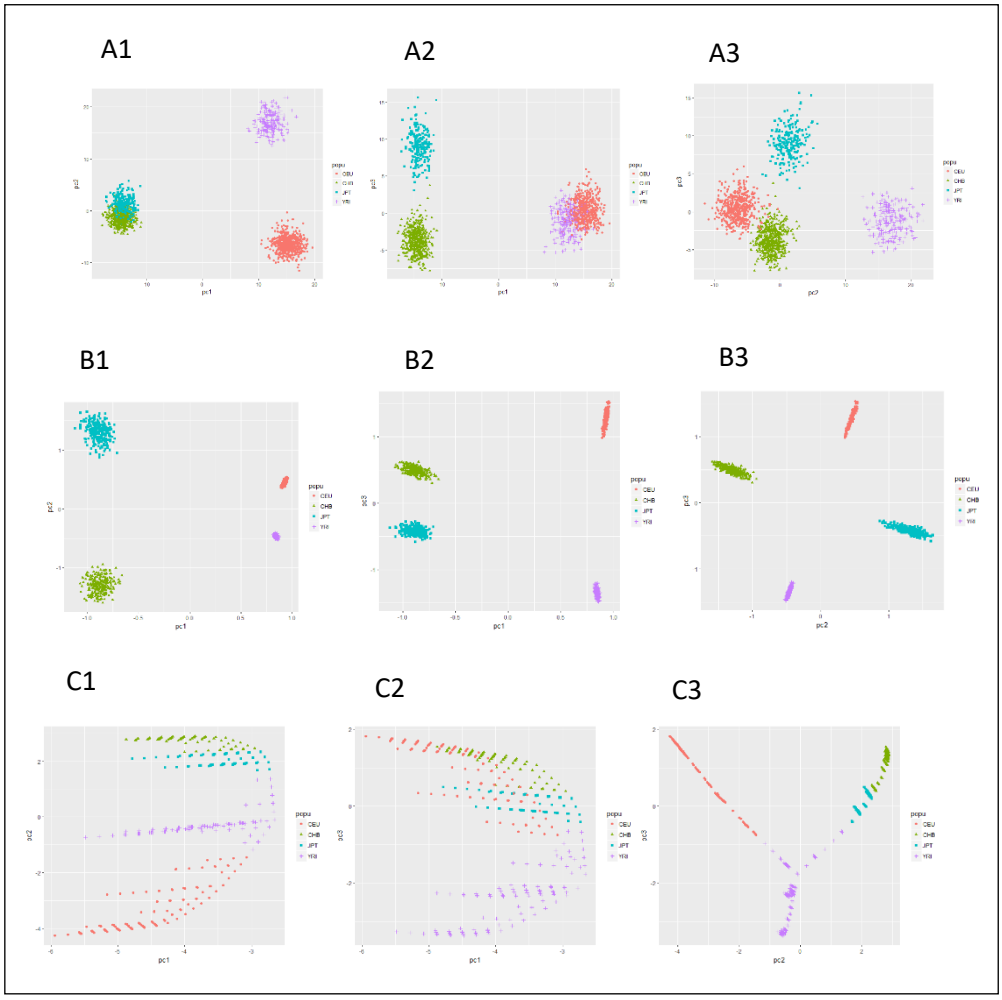

Figure 4. Pairwise plots of the first three principal components computed using linear (A1-A3), kernel (B1-B3) and ncMCE (C1-C3) PCA methods.

As can be seen from Table 3, none of the considered simulation scenario's show marked differences regarding type I error control or power to detect epistasis, when using the first 10 PCs computed via linear, kernel or ncMCE PCA methods with MBMDR-PC. In contrast, type I error estimates are somewhat inflated with MBMDR-GC. However, MBMDR-PC and MBMDR-GC give comparable power estimates, except for epistasis models with low frequency causal variants (Models 5 and 6).

Table 3. Estimates of power and type I error rates of MBMDR-PC with population structure captured by linear, kernel and ncMCE principal components and MBMDR-GC, with a nominal 0.05 FWER level.

	MBMDR-GC		MBMDR-PC					
			Linear PCA		Kernel PCA		ncMCE	
Sample sizes	200	500	200	500	200	500	200	500
Type I Error Rates	0.077	0.085	0.052	0.054	0.054	0.050	0.047	0.055
Power								
Model 1	0.945	0.735*	0.927	0.730*	1.000	0.740*	0.929	0.770*
Model 2	0.805	0.131*	0.846	0.404*	0.864	0.560*	0.895	0.556*
Model 3	0.653	0.954	0.658	1.000	0.731	0.970	0.669	0.968
Model 4	0.483	0.956	0.481	0.960	0.476	0.945	0.490	0.960
Model 5	0.074	0.784	0.259	0.950	0.238	0.950	0.258	0.958
Model 6	0.161	0.918	0.447	0.970	0.421	0.970	0.446	0.956

*The values are close to 1 when the power is calculated based on all significant interactions that include the true signal.

Simulation setting 4*.* The scatter plot on the first 2 linear and kernel principal components for a single simulated dataset (see Methods section) is shown in Figure 6. Linear PCA indicates a nonlinear genetic background structure (Figure 5 A). This is confirmed by kernel based PCA, which clearly separates the two subpopulations (Figure 5 B).

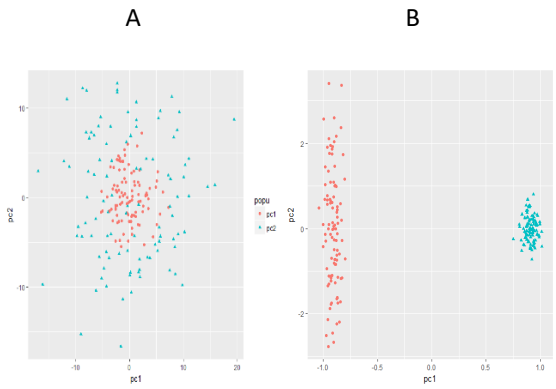

Figure 5. Plots of the first two principal components (A) linear PCA and (B) kernel PCA.

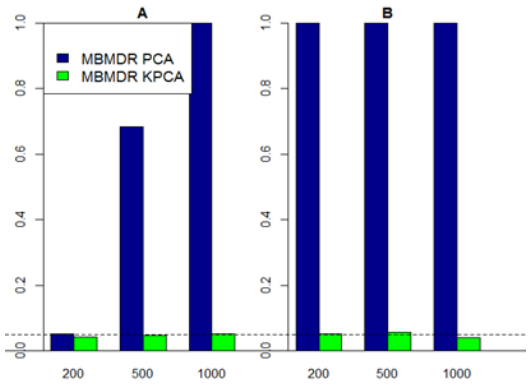

Figure 6. Estimated type I error rates for MBMDR-PC with case-control ratios (A) 60:40 and (B) 80:20. PC approaches considered: linear PCA (blue bars), kernel PCA (green bars).

The estimated results of type I error rates of MBMDR-PC using linear and kernel principal components are presented in Figure 6. In the presence of phenotypic and structural epistasis, linear PCA based MBMDR-PC highly inflates the type I error which is substantially higher than the nominal 0.05 FWER level. For example, for a total sample size of 500 (cases and controls jointly) and case-control ratio is 60:40 and 80:20, the type I error rates of linear MBMDR-PC are 0.7 and 1.0 for the nominal level 0.05 (Figure 6 A and 6 B, respectively). Type I error rates of linear MBMDR-PC increases as the sample size increases. Furthermore, type I errors estimates get worse for linear PCA based MBMDR-PC with increasing levels of unbalancedness (Figure 6 B, 80:20). In comparison, the estimated type I error rates of kernel based MBMDR-PC are not significantly different from the nominal level 0.05 in all the scenarios considered.

Simulation setting 5*.* The estimated type I error rate for simulation setting based on related samples obtained from 1000 replicates (as explained in the Methods section) is 0.051., which is close to the nominal 0.05 level. Power estimates for epistasis model M170 (see Methods) increase with increasing minor allele frequencies for the causal epistasis SNP pair (0.45, 0.885 and 0.911 for MAFs of 0.1, 0.25 and 0.5, respectively).

Discussion

In this work, we have highlighted the importance of detecting and correcting for population structure in epistasis studies. Using extensive simulations we have shown that in genome-wide epistasis studies failure to account for complex population structure results in inflated false positives or low power to detect true signals of epistasis. When evaluating the impact of ignored or inadequately captured population structure in GWAIS, we not only considered epistatic ancestry informative markers, but also paid special attention to the idea of nonlinearity in population genetics [27]. In addition, we considered the influence of unequal sample sizes. As epistasis analytic tool, we relied on MB-MDR analytics, which should be seen as part of an entire analysis pipeline that involves making marker selection choices and performing post-analysis steps to validate and replicate findings, as well as seeking biological evidence for flagged interacting regions [18].

That ignoring population structure due to allele frequency differences among populations and subpopulations can result in high numbers of false positives or reduced power in GWAS is not new. In GWAS, Structured Association (SA) [42–45], Genomic Control (GC) [35,36], Principal Component Analysis (PCA) [23] and Mixed Modeling (MM) [46] are the main 4 strategies to deal with confounding associations due to shared genetic ancestry. The basic idea of SA is to infer the underlying population structure and then to incorporate this information in subsequent testing for genetic associations of interest. In contrast, the basic idea of GC is to correct the null distribution of genetic association tests for the effects of the unspecified population structure [35]. The statistical advantage of SA methods depends on the degree of information provided by the available marker data to make inferences about

the true structure. Classic GC methods rely on adjusting all marker-trait associations in the same way, which ignores the strength of the relationship between the genealogy of the genetic marker under study and the (hidden) pedigree structure, and thus also dependencies between markers. The basic idea of MM for controlling population structure is to account for pairwise relatedness between individuals, for example, using a kinship matrix. It is an approach that naturally accommodates familial and cryptic relatedness in the data. Mixed models have long been computationally expensive; it took until the development of more efficient algorithms for them to gain popularity in population structure control. Some of the algorithm improvements are incorporated in the following approaches: compressed-MLM [46], EMMA (Efficient Mixed-Model Association) [47], EMMAX (EMMA eXpedited) [48], GEMMA (Genome-Wide Efficient Mixed-Model Association) [49], LRLMM (low rank linear mixed model [33], FaST-LMM (Factored Spectrally Transformed Linear Mixed Model) [50], FaST-LMM-Set [51], GRAMMAR-Gamma (fast variance components-based two-step method) [52], and FarmCPU (Fixed and random model Circulating Probability Unification) [53]. PCA allows data transformation to a new coordinate system such that the projection of the data along the first new coordinate has the largest variance, the second principal component has the second largest variance, and so on. The relative straightforwardness of PCA, its ease of use, the availability of efficient algorithms and its ability to detect individuals with unusual or differential ancestry [28,54–58] has made PCA among the most heavily used strategies in the context of genetic association studies in structured populations. Once principal components are obtained, several choices can be made to use these for the purpose of confounding correction in GWAS. Assuming that the GWAS is performed within a regression framework, the most straightforward approach is to include the first few principal components, capturing genetic ancestry of each individual, as fixed effects in a (generalized) linear model. Alternatively, instead of directly including the principal components in a

regression model, both phenotype and genotypes can be adjusted by top PCs as in EIGENSTRAT [23]. The adjusted phenotype is defined as the residual of fitting an appropriate generalized linear regression model of phenotype on a number of principal components. A similar model fitting is performed to obtain adjusted genotypes [12].

The aforementioned methods naturally extend to epistasis detection frameworks, in particular those that allow for a regression model component in their methodology. One such a framework is MB-MDR (32), which adds a model-based component to Multifactor Dimensionality Reduction, hereby enabling adjusting for lower order genetic effects or confounders (46). Our proposed methods for detecting epistasis in the presence of population structure, MBMDR-PC, MBMDR-PG and MBMDR-GC, build on MB-MDR. MBMDR-PC and MBMDR-PG involve first deriving adjusted phenotypes (residuals) obtained from fitting appropriate generalized linear models with the first few principal components (linear or nonlinear) as covariates, and generalized linear mixed models with a kinship matrix to capture the covariance structure of random effects, respectively. MBMDR-GC involves computing a SNP-pair specific correction factor for each MB-MDR observed test value. This was inspired by earlier observations that the distribution of MB-MDR test statistics may largely vary from one SNP-pair to another due to a combination of disease prevalence and minor allele frequencies of SNPs under testing (results not shown). The generated null data under the hypothesis of no trait associations are used twice with MBMDR-GC: first to estimate the expected MB-MDR test value for each SNP pair *j* under this null, and second to assess the statistical significance of observed MB-MDR test values that are adjusted by $\lambda_{GC,j}$. With equal MB-MDR test null distributions across SNP pairs, no genetic associations with the trait and no population structure, the expected $\lambda_{GC,j}$ should approximate 1. With unequal MB-MDR test null distributions, observed test values will receive higher chances to become significant with

higher expected SNP-pair related median test values, computed in the absence of genetic and confounder associations with the trait.

In general, our simulation results showed that in the presence of population structure MBMDR-PC and MBMDR-PG consistently control type I error rate at the nominal level compared to MBMDR-GC which had a slightly inflated type I error rate. Also, our three methods of population structure correction were more powerful than MDR-SP. Thus, MBMDR-PC and MBMDR-PG for GWAIS adjusted for confounding by (non)linear population structure give promising results and are to be preferred over MDR-SP in the considered simulation settings. Our results also suggested that there is no need to compute population controlling PCs for every SNP pair separately. For related samples, MBMDR-PG based on a generalized linear mixed model should be used. In many instances of mild population structure, MB-MDR with codominant correction exhibits a comparable performance to MBMDR-PC. All analyses can easily accommodate covariates using similar principles as in MBMBDR-PC and MBMDR-GC.

The outperformance of MBMDR-PC clearly depends on the ability of the principal components to capture population structure well. We chose the checker board stratification model in order to inject strong nonlinear genetic differences between two populations; more work is needed to investigate a variety of complex nonlinear stratification models and to assess their occurrence in real-life. Overall, widely used linear PCA fails to properly differentiate such complex populations Kernel-based strategies offer an interesting alternative, especially when additional efficient computational tools are developed to extract non-linear PCs from large genetic datasets as those collected within disease-specific

consortiums. Our simulation results that compare MBMDR-PC with linear and kernel PCs showed that MBMDR-PC with linear PCs gives inflated type I errors, which becomes worse as the ratio of case-control becomes increasingly unbalanced. On the contrary, MBMDR-PC based on kernel PCs effectively controlled for both linear and non-linear population structure and maintained the type I error rates at the required nominal levels.

In conclusion, MBMDR-PC is a generally well performing approach, compared to the computationally intensive MBMDR-PG and MBMDR-GC approaches, although its performance is highly dependent on how well PCs capture population structure. We therefore recommend to use both linear and nonlinear versions of PCA, whenever possible. Fast implementation for multiple testing correction in exhaustive epistasis screenings [10] makes our proposed MB-MDR based methods efficient tools for GWAIS in structured populations. Our work is important in view of ongoing initiatives of epistasis detection in large-scale heterogeneous consortium data, as we have shown that inadequate capturing of population structure may severely jeopardize obtaining meaningful and replicable epistasis findings.

Supplementary Information

Supplementary Table S1, Figure S0, Figure S1 and Figure S2.

Abbreviations

GC: Genomic Control

GWAS: Genome-wide Association Study

GWAIS: Genome-wide Association Interaction Studies

MAF: Minor Allele Frequency

MB-MDR: Model-Based Multifactor Dimensionality Reduction

MDR: Multifactor Dimensionality Reduction

MDR-SP: Multifactor Dimensionality Reduction for Structured Population

PC: Principal Components

PCA: Principal component analysis

PG: Polygenic

SA: Structured Association

SNP: Single Nucleotide Polymorphism

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgement

We thank Myriam Nemry for suggestions regarding the implementation of ideas in the MBMDR software.

Funding

This research was in part funded by the Fonds de la Recherche Scientifique (F.N.R.S.), in particular "Integrated complex traits epistasis kit" (Convention n° 2.4609.11) [KVS]. We also acknowledge research opportunities offered by F.N.R.S., including “Foresting in Integromics Inference” (Convention n° T.0180.13) [KC], and by the interuniversity research institute Walloon Excellence in Lifesciences and BIOtechnology (WELBIO) [FA, KVS].

Authors' contributions

FA, KVS and FVL conceived and designed the study. FA, KVS and FVL carried out the implementation. FA, KVS, FVL, KC and DD analyzed and interpreted the results. FA, KVS, FVL, JMJ, KC, AB, DD, ESG, ZW and HH wrote the manuscript.

Authors details

[1] GIGA-R, Medical Genomics – BIO3, University of Liège, Liege, Belgium
[2] Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA
[3] Center for Applied Genomics, The Children’s Hospital of Philadelphia, Philadelphia, PA, USA

[4] Division of Human Genetics, Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[5] WELBIO (Walloon Excellence in Lifesciences and Biotechnology), University of Liege, Liege, Belgium

Availability of data and materials

Codes to implement MBMDR-PC, MBMDR-PG and MBMDR-GC are available via the MBMDR software (from version mbmdr-4.4.1 onwards), which is downloadable from http://bio3.giga.ulg.ac.be/index.php/software/mb-mdr/ . Main options

MBMDR-PC: *--binary –ac number of PCs –d 2D –a CODOMINANT –rc RESIDUALS*
MBMDR_PG: *--continuous* –d 2D –a CODOMINANT*
MMBMDR_GC: *--binary –d 2D –mt STRAT3*
**: residuals obtained using the R package GMMAT*

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

References

1. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11: 459–463. doi:10.1038/nrg2813

2. Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet. 1996;59: 983–989.

3. Horvath S, Xu X, Laird NM. The family based association test method: strategies for studying general genotype--phenotype associations. Eur J Hum Genet EJHG. 2001;9: 301–306. doi:10.1038/sj.ejhg.5200625

4. Simpson EH. The Interpretation of Interaction in Contingency Tables. J R Stat Soc Ser B Methodol. 1951;13: 238–241.

5. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet. 2004;36: 512–517. doi:10.1038/ng1337

6. Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2002;11: 505–512.

7. Wacholder S, Rothman N, Caporaso N. Counterpoint: Bias from Population Stratification Is Not a Major Threat to the Validity of Conclusions from Epidemiological Studies of Common Polymorphisms and Cancer. Cancer Epidemiol Prev Biomark. 2002;11: 513–520.

8. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NLS, et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet. 2010;87: 325–340. doi:10.1016/j.ajhg.2010.07.021

9. Cattaert T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, Urrea V, et al. Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise. Ann Hum Genet. 2011;75: 78–89. doi:10.1111/j.1469-1809.2010.00604.x

10. Lishout FV, Gadaleta F, Moore JH, Wehenkel L, Steen KV. gammaMAXT: a fast multiple-testing correction algorithm. BioData Min. 2015;8. doi:10.1186/s13040-015-0069-x

11. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet. 2001;69: 138–147. doi:10.1086/321276

12. Zhao Y, Chen F, Zhai R, Lin X, Wang Z, Su L, et al. Correction for population stratification in random forest analysis. Int J Epidemiol. 2012; 1798–1806. doi:10.1093/ije/dys183

13. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81: 559–575.

14. Gyenesei A, Moody J, Semple CAM, Haley CS, Wei W-H. High-throughput analysis of epistasis in genome-wide association studies with BiForce. Bioinformatics. 2012;28: 1957–1964. doi:10.1093/bioinformatics/bts304

15. Zhang BY, Zhang J, Liu JS. Block-based Bayesian Epistasis Association Mapping with Appliation to WTCCC Type 1 Diabetes Data. Ann Appl Stat. 2011;5: 2052–2077. doi:10.1214/11-AOAS469

16. Shang J, Zhang J, Sun Y, Liu D, Ye D, Yin Y. Performance analysis of novel methods for detecting epistasis. BMC Bioinformatics. 2011;12: 475. doi:10.1186/1471-2105-12-475

17. Li M, Lou X-Y, Lu Q. On epistasis: a methodological review for detecting gene-gene interactions underlying various types of phenotypic traits. Recent Pat Biotechnol. 2012;6: 230–236.

18. Gusareva ES, Van Steen K. Practical aspects of genome-wide association interaction analysis. Hum Genet. 2014;133: 1343–1358. doi:10.1007/s00439-014-1480-y

19. Wei W-H, Hemani G, Haley CS. Detecting epistasis in human complex traits. Nat Rev Genet. 2014;15: 722–733. doi:10.1038/nrg3747

20. Gola D, John M, M J, van Steen K, König IR. A roadmap to multifactor dimensionality reduction methods. Brief Bioinform. 2016;17: 293–308. doi:10.1093/bib/bbv038

21. Fouladi R, Bessonov K, Van Lishout F, Van Steen K. Model-Based Multifactor Dimensionality Reduction for Rare Variant Association Analysis. Hum Hered. 2015;79: 157–167. doi:10.1159/000381286

22. Niu A, Zhang S, Sha Q. A Novel Method to Detect Gene–Gene Interactions in Structured Populations: MDR-SP. Ann Hum Genet. 2011;75: 742–754. doi:10.1111/j.1469-1809.2011.00681.x

23. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38: 904–909. doi:10.1038/ng1847

24. Cattaert T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, Urrea V, et al. A detailed view on Model-Based Multifactor Dimensionality Reduction for detecting gene-gene interactions in case-control data in the absence and presence of noise. Ann Hum Genet. 2011;75: 78–89. doi:10.1111/j.1469-1809.2010.00604.x

25. Mahachie John JM, Van Lishout F, Van Steen K. Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data. Eur J Hum Genet. 2011;19: 696–703. doi:10.1038/ejhg.2011.17

26. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38: 904–909. doi:10.1038/ng1847

27. Alanis-Lobato G, Cannistraci CV, Eriksson A, Manica A, Ravasi T. Highlighting nonlinear patterns in population genetics datasets. Sci Rep. 2015;5: 8140. doi:10.1038/srep08140

28. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 2008;40: 646–649. doi:10.1038/ng.139

29. Abegaz F, Van Lishout F, Mahachie John JM, Chiachoompu K, Bhardwaj A, Gusareva ES, et al. Epistasis Detection in Genome-Wide Screening for Complex Human Diseases in Structured Populations. Syst Med. 2019;2: 19–27. doi:10.1089/sysm.2019.0003

30. Abegaz F, Lishout FV, John JMM, Chiachoompu K, Bhardwaj A, Gusareva ES, et al. Epistasis Detection using Model Based Multifactor Dimensionality Reduction in Structured Populations. bioRxiv. 2019; 541946. doi:10.1101/541946

31. Astle W, Balding DJ. Population Structure and Cryptic Relatedness in Genetic Association Studies. Stat Sci. 2009;24: 451–471. doi:10.1214/09-STS307

32. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. Am J Hum Genet. 2016;98: 653–666. doi:10.1016/j.ajhg.2016.02.012

33. Hoffman GE. Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. PLOS ONE. 2013;8: e75707. doi:10.1371/journal.pone.0075707

34. Eu-ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SMB, Blackwell JM, Cordell HJ. Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data. PLoS Genet. 2014;10: e1004445. doi:10.1371/journal.pgen.1004445

35. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55: 997–1004.

36. Wang K. Testing for genetic association in the presence of population stratification in genome-wide association studies. Genet Epidemiol. 2009;33: 637–645. doi:10.1002/gepi.20415

37. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol. 2003;24: 150–157. doi:10.1002/gepi.10218

38. Martin E r., Ritchie M d., Hahn L, Kang S, Moore J h. A novel method to identify gene–gene effects in nuclear families: the MDR-PDT. Genet Epidemiol. 2006;30: 111–123. doi:10.1002/gepi.20128

39. Balding DJ, Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica. 1995;96: 3–12. doi:10.1007/BF01441146

40. Cattaert T, Urrea V, Naj AC, Lobel LD, Wit VD, Fu M, et al. FAM-MDR: A Flexible Family-Based Multifactor Dimensionality Reduction Technique to Detect Epistasis Using Related Individuals. PLOS ONE. 2010;5: e10304. doi:10.1371/journal.pone.0010304

41. Li W, Reich J. A complete enumeration and classification of two-locus disease models. Hum Hered. 2000;50: 334–349. doi:22939

42. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association Mapping in Structured Populations. Am J Hum Genet. 2000;67: 170–181.

43. Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 2000;155: 945–959.

44. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic Structure of Human Populations. Science. 2002;298: 2381–2385. doi:10.1126/science.1078311

45. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664. doi:10.1101/gr.094052.109

46. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010;42: 355–360. doi:10.1038/ng.546

47. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178: 1709–1723. doi:10.1534/genetics.107.080101

48. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42: 348–354. doi:10.1038/ng.548

49. Zhou X, Stephens M. Genome-wide Efficient Mixed Model Analysis for Association Studies. Nat Genet. 2012;44: 821–824. doi:10.1038/ng.2310

50. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011;8: 833–835. doi:10.1038/nmeth.1681

51. Listgarten J, Lippert C, Kang EY, Xiang J, Kadie CM, Heckerman D. A powerful and efficient set test for genetic markers that handles confounders. Bioinformatics. 2013;29: 1526–1533. doi:10.1093/bioinformatics/btt177

52. Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS. Rapid variance components-based method for whole-genome association analysis. Nat Genet. 2012;44: 1166–1170.

53. Liu X, Huang M, Fan B, Buckler ES, Zhang Z. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLOS Genet. 2016;12: e1005767. doi:10.1371/journal.pgen.1005767

54. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLOS Genet. 2006;2: e190. doi:10.1371/journal.pgen.0020190

55. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 2007;3: 1672–1686. doi:10.1371/journal.pgen.0030160

56. Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, et al. Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet EJHG. 2008;16: 1413–1429. doi:10.1038/ejhg.2008.210

57. Reich D, Price AL, Patterson N. Principal component analysis of genetic data. Nat Genet. 2008;40: 491–492. doi:10.1038/ng0508-491

58. Novembre J, Peter BM. Recent advances in the study of fine-scale population structure in humans. Curr Opin Genet Dev. 2016;41: 98–105. doi:10.1016/j.gde.2016.08.007