Systems Medicine
Volume 2.1, 2019
DOI: 10.1089/sysm.2019.0003

SYSTEMS MEDICINE
Journal of Medical Systems Biology and Network Medicine

Mary Ann Liebert, Inc. publishers

ORIGINAL RESEARCH
Open Access

Epistasis Detection in Genome-Wide Screening for Complex Human Diseases in Structured Populations

Fentaw Abegaz,[1,*] François Van Lishout,[1] Jestinah M. Mahachie John,[1] Kridsadakorn Chiachoompu,[1] Archana Bhardwaj,[1] Elena S. Gusareva,[1] Zhi Wei,[2] Hakon Hakonarson,[3,4] and Kristel Van Steen[1,5];
on behalf of the International IBD Genetics Consortium

Abstract

Over the years, a more prominent role has been given to gene–gene interaction (epistasis) detection, in view of precision medicine and the hunt for novel drug targets and biomarkers for complex diseases. Acknowledging data complexity as embodied by epistasis potentially increases the power of genome-wide association studies (GWAS) and may reveal relevant biological and biochemical pathways previously undetected. Although confounding of GWAS due to shared genetic ancestry has been well recognized, the extent and impact of such confounding in gene–gene interaction association epistasis studies is much less understood. In the same spirit, the role of population substructure in epistasis detection is largely under-investigated, especially outside a regression framework. This is surprising as inadequate handling of such confounding is likely to lead to spurious epistatic associations, hampering replication and the identification of causal loci in synergy. To improve interpretability and replicability of epistasis results, we introduce "MB-MDR for Structured Populations" (Model-Based Multifactor Dimensionality Reduction-Principal Component [MBMDR-PC]). It extends classic MBMDR that was developed for samples sharing the same genetic ancestry, by replacing the original phenotypes with new phenotypes that have been adjusted for population structure as captured by PCs. The method is applied on International Inflammatory Bowel Disease Genetics Consortium Crohn's disease (CD) data from 15 countries. Significant interacting single nucleotide polymorphisms are found within NOD2 and CYLD genes among others that suggest a potential synergetic effect of NOD2 and CYLD on CD. The study highlights the value of examining epistasis effects and indicates the need for further studies to understand the epistasis effects on CD.

Keywords: confounding; epistasis; gene–gene interaction; MBMDR; population structure

Introduction

In the context of genome-wide association studies (GWAS), population stratification refers to genetically distinct subgrouping.[1,2] Several causes exist for population stratification. The basic one being shared genetic ancestry as a result of nonrandom mating between subgroups in a population due to various reasons, which may include social, cultural, or geographical ones. Potential consequences of population stratification are confounding, cryptic relatedness (i.e., unobserved ancestral relationships between individual cases and controls causing them to be nonindependent) and selection bias.[1,2]

Several strategies exist for protecting against population structure in case–control GWA studies, whether due to population stratification or admixture, the most popular one being based on principal component (PC) analysis.[3] The situation is quite less obvious for genome-

[1]GIGA-R, Medical Genomics—BIO3, University of Liege, Liege, Belgium.
[2]Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey.
[3]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania.
[4]Division of Human Genetics, Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania.
[5]WELBIO (Walloon Excellence in Lifesciences and Biotechnology), University of Liege, Liege, Belgium.

*Address correspondence to: Fentaw Abegaz, PhD, GIGA-R, Medical Genomics—BIO3, University of Liege, Liege 4000, Belgium, E-mail: fentawabegaz@gmail.com

wide association interaction studies (GWAIS). One of the main issues is that the increased complexity that comes within reach by performing a GWAIS goes along with a plethora of methodologies (let alone software tools) to detect epistasis or gene–gene interactions. The growing belief in the importance of gene–gene interactions in the development and progression of complex diseases has led to an exponential growth in such tools; to name but a few: generalized linear regression models (GLM), BOOST,[4] Model-Based Multifactor Dimensionality Reduction (MBMDR),[5,6] MDR,[7] Random Forest,[8] PLINK,[9] BiForce,[10] Bayesian Models (e.g., BEAM),[11] and several others. For extensive reviews and appropriate references, please refer to Refs.[12–16]

The literature on epistasis detection in structured populations is very limited, apart from scenarios using a regression framework for association testing. A few exceptions include those strategies that allow submitting population structure corrected phenotypes (i.e., residuals derived from regression models for the original phenotype on selected PCs).[8] Ideally, a general framework and software tool for epistasis detection is available that can offer flexible maneuvering between different measurement scales for phenotypes and genomic predictors. MBMDR offers one such framework and tool.[5,6,17] Interestingly, the MDR-SP method,[18] to our knowledge the only MDR-inspired method that can deal with structured populations, combines MDR[7] with ideas implemented in the EIGENSTRAT software.[3] It adds high/low risk labels to multilocus genotypes after adjusting original phenotypes and genotypes by the first few PCs. For feasibility, the same PCs are used for every multilocus genotype and (pairs, triples, etc.) of loci.

In this article, we incorporate strategy to account for population structure in GWAIS using the MBMDR framework with PCs referred to as MBMDR-PC. 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 fully describe the MBMDR-PC approach and real-life data on inflammatory bowel disease (IBD) from the International IBD Genetics Consortium,[19] are used to further validate the MBMDR-PC approach to correct for confounding due to population structure in epistasis detection. The genome-wide data used for epistasis detection for Crohn's disease (CD) include 15 countries.

CD, a subcategory of IBD, is a complex disorder that can affect any part of the digestive system. Early genome-wide search based on families with multiple affected members with CD reported strong linkage on chromosome 16 and identified the first gene NOD2 as a possible causative locus.[20–22] Genetic studies have identified 163 susceptibility loci for IBD, mostly shared between CD and ulcerative colitis as another major form of IBD.[23] The most significant signals were found in the NOD2 region. This region has been shown to exhibit substantial genetic heterogeneity.[24] Elding et al.[24] demonstrated the independent involvement of a neighboring gene on chromosome 16, called CYLD, also playing a role in the immune system. Moreover, Hrdinka et al.[25] demonstrated that CYLD restricts deposition of Lys63-Ub and Met1-Ub on the Linear Ubiquitin Assembly Complex (LUBAC) substrate RIPK2 to limit NOD2-dependent inflammatory signaling. Since both NOD2 and CYLD are involved in the regulation of immune response, following Elding et al. it is interesting to explore whether the combined mutations of these two genes have a synergistic effect on CD.[24] We will, however, consider epistasis screening, including all available curated genetic markers on chromosome 16.

Our study is important in the light of an increasing number of consortium-based epistasis studies that are marked by heterogeneous sample collections due to complex fine-scale or large-scale population structure.

Materials and Methods

The proposed genome-wide epistasis screening strategy in structured populations is built on the MBMDR.[5,6,26,27] Even though the MBMDR framework can be used for higher level interaction detection and various outcome measurement scales and study designs, in this study we restrict attention to pairwise interactions. A graphical overview of the newly introduced methods is provided in Figure 1 and explained in more detail as follows.

MBMDR-PC: accounting for genomic structure by PCs

Similar to EIGENSTRAT,[3] we use either linear or nonlinear (kernel) PCs to correct for population structure, and coin our strategy MBMDR-PC. PCs as confounders of population structure due to shared genetic ancestry are assumed to relate to both phonotype and genetic markers. In case–control studies to avoid PCs not to reflect case–control differences, PCs can be first computed from within the controls, and then projected onto the cases to generate PCs for the entire data set.[28–30] This approach has been preferred over generating PCs on the pooled case–control data and used in consortium-based IBD GWAS.[19] In MBMDR-PC we

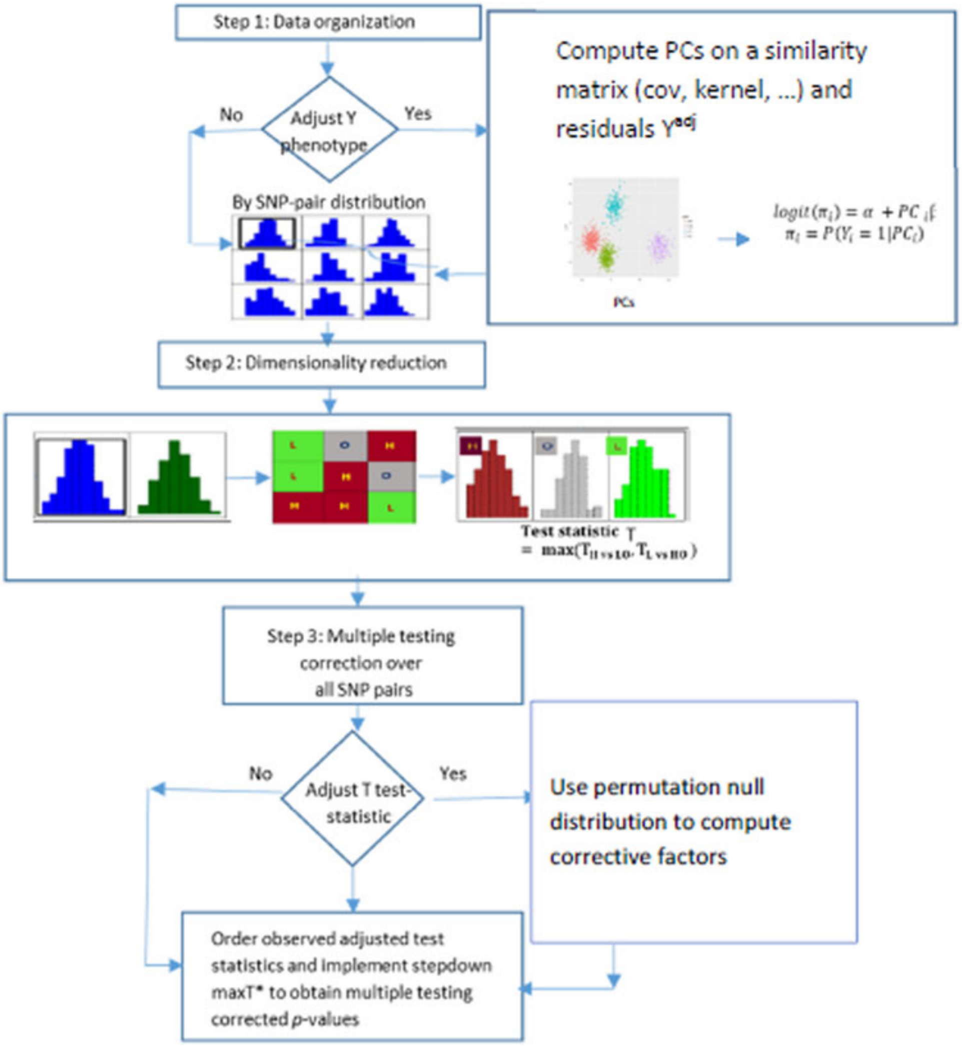
Step 1: Data organization
Adjust Y phenotype
No
Yes
Compute PCs on a similarity matrix (cov, kernel, …) and residuals Y^{adj}
PCs
$logit(\pi_i) = \alpha + PC_i\beta$
$\pi_i = P(Y_i = 1|PC_i)$
By SNP-pair distribution
Step 2: Dimensionality reduction
L
O
H
Test statistic T
$= \max(T_{H\ vs\ LO}, T_{L\ vs\ HO})$
Step 3: Multiple testing correction over all SNP pairs
Adjust T test-statistic
No
Yes
Use permutation null distribution to compute corrective factors
Order observed adjusted test statistics and implement stepdown maxT* to obtain multiple testing corrected *p*-values

FIG. 1. Graphical presentation of the key steps involved in MBMDR for structured populations (Tables 1 and 4 and Fig. 1 in Ritchie et al.[7] describe the implementation of maxT and gammaMaxT in MBMDR). MBMDR, Model-Based Multifactor Dimensionality Reduction.

append the classic data organization step in MBMDR with one that computes new phenotypes, adjusted for population structure. The new phenotypes are taken as input to classic MBMDR, in an attempt to capture genetic interactions that are not spurious due to inadequate handling of population structures. In what follows, we give a detailed outline of MBMDR-PC (Fig. 1).

Step 1: data organization.

As in classic MBMDR, for every two SNP loci j and k, individuals with nonmissing genotype data will exhibit one of nine possible two-locus genotypes {00, 01, 02, 10, 11, 12, 20, 21, 22}, for an individual's genotype G_j at locus j and G_k at locus k taking on 0, 1, 2 as possible values (i.e., number of minor alleles). Specific to MBMDR-PC, the regression framework is adhered to create new phenotypes that are adjusted for population structure captured by a number of PCs. In case–control epistasis studies, where the phenotype Y represents disease status (1 affected, 0 unaffected), and G_j and G_k refer to genetic information at two SNP loci j and k (e.g., using additive encoding), epistasis can be investigated by making inference on the interaction parameter θ in the following logistic regression model:

$$\text{logit}(\pi_i) = \alpha + \beta_1 W_{i1} + \cdots + \beta_r W_{ir} + \gamma_1 G_{ij} + \gamma_2 G_{ik} + \theta G_{ij} G_{ik},$$

where $\pi_i = P\left(Y_i = 1 | W_{i1}, \ldots, W_{ir}, G_{ij}, G_{ik}\right)$ is the probability of disease for subject *i*, possibly conditional on the first *r* PCs $W_{i1}, \ldots, W_{ir}$. The vector $\beta = (\beta_1, \ldots, \beta_r)$ is a vector of regression parameters corresponding to the *r* PCs, α is the intercept term, and γ_1 and γ_2 are the main effects of the two SNPs and θ captures the interaction effect between the two SNPs at loci *j* and *k*. Earlier reports of limitations related to logistic regression for higher-order interaction modeling,[31] including having to make "model assumptions" about mode of inheritance (i.e., related to choosing a particular encoding scheme for genetic exposures) lies at the basis of MBMDR, which is nonparametric in its core. However, when adjustments need to be made for lower-order effects or confounders, the MBMDR paradigm needs to be combined with the regression paradigm. Related to MBMDR-PC, we derive adjusted phenotypes from the aforementioned logistic regression model by subtracting model-fitted values from observed phenotype values:

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

$$Y_i^{\text{adj}} = Y_i - \hat{\pi}_i, \text{where } \hat{\pi}_i = \frac{\exp\left(\hat{\alpha} + \hat{\beta}_1 W_{i1} + \cdots + \hat{\beta}_r W_{ir}\right)}{1 + \exp\left(\hat{\alpha} + \hat{\beta}_1 W_{i1} + \cdots + \hat{\beta}_r W_{ir}\right)}$$

This can be accomplished in R using the package *glm*.[32] The fitted model should be adapted according to the measurement type of the original phenotype, by selecting an appropriate link function linking the linear predictor $\alpha + \beta_1 W_{i1} + \cdots + \beta_r W_{ir}$ to the (adjusted) phenotype. Conditioning on additional confounders (sex, age, etc.) is straightforward by including them in the expression for $\text{logit}(\pi_i)$ earlier.

Step 2: multilocus prioritization and dimensionality reduction. The adjusted phenotype obtained in *Step 1* is subsequently investigated for distributional differences between multilocus genotypes ($G_j \times G_k$). For SNPs at loci *j* and *k*, such investigations are reduced to making inferences about γ as in

$$Y_i^{\text{adj}} = \alpha' + \gamma_m C_{im},$$

for which Y_i^{adj} denotes the adjusted phenotype (*Step 1*) and C_{im} is 1 if the *m*th multilocus genotype derived from G_j and G_k is 1 and 0 otherwise. In practice with MBMDR, a Student's *t*-test is carried out, for each multilocus genotype cell C_{m} at the liberal significance level of 0.10, comparing the mean of cell C_{m} with the mean of the remaining eight cells. If the test is not significant at 0.10 level, the cell is labeled as O (no evidence for risk). If the test is significant at 0.10 the cell is labeled as H (high risk) or L (low risk). The sign of the test statistic is used to distinguish between H and L: a positive (negative) sign refers to risk H (L). The thresholding of 0.10 is motivated in earlier study.[26]

The result is thus a new categorical variable with values H, L, and O, which captures information about the importance of the pair of SNPs with respect to the adjusted phenotype. Subsequently, a final aggregate association test is performed on the new construct and the adjusted phenotype. In particular, we consider the maximum of two *t*-test statistics denoted by $\max(|T_{\text{H}}|, |T_{\text{L}}|)$, again comparing the mean responses of H versus {L, O} categorized individuals and the means of L versus {H, O} labeled individuals, respectively. Contrast tests, comparing H versus L combined multilocus genotypes per SNP pair can be considered as well. These have been evaluated elsewhere[25] but were not considered for the purposes of this article.

Step 3: significance assessment. For every pair of SNPs in the data, we obtain a single test statistic given by

$$T = \max(|T_{\text{H}}|, |T_{\text{L}}|)$$

from Step 2. This maximum test statistic will no longer follow a *t*-distribution, due to the compounding of evidences in Step 2. The significance of *T* per SNP pair is, therefore, assessed through resampling-based strategies. In particular, in this study we generate 999 permutation-based replicates by permuting adjusted trait labels, yet keeping the correlation structure between SNPs intact. Multiple testing is taken care of by a step-down maxT approach as described inadjusted *p*-values,[33] hereby ensuring partial strong control of family-wise error rate at 5%. Strong control properties can be stated under the assumption of the subset pivotality property.[30] For large samples, such as in the real-life data application, we adhered to an approximated step-down maxT adjusted *p*-values, as described in Westfall and Young.[33]

Application on real-life data from International IBD Genetics Consortium (CD)

In particular, we considered data from the International IBD Genetics Consortium, comprising 68,427 samples from 15 countries. This data set underwent rigorous quality control by the IBD consortium.[19] In addition, in view of the intended epistasis analyses on chromosome 16, we furthermore filtered SNPs with linkage disequilibrium (LD) pruning by PLINK[9] using a sliding window of size 50, moving step of 5, and $r^2 = 0.75$ following epistasis protocol recommendations in Gusareva and Van Steen.[14] This resulted in a final data set of 1434 SNPs on chromosome 16 and 52,277 samples of which 18,227 CD cases and 34,050 controls. We used the first five PCs as computed by the IBD consortium, based on Immunochip data with 19,111 SNPs that remained after LD pruning that was performed three times within European controls and after removing SNPs with minor allele frequencies <5%.[19] The PC axes were first generated within the controls, and then projected onto the cases to generate PCs for all samples. These components have proven to be sufficient to control type I error in consortium-based IBD GWAS.[19]

MBMDR was used for the epistasis analysis either without correcting for population structure or with correction using the five aforementioned PCs and MBMDR-PC (codominant main effects correction in the MBMDR software). The resulting multiple testing adjusted p-values indicate statistically significant SNP interactions when <0.05.

Epistasis analysis results obtained by MBMDR-PC and screening chromosome 16 SNPs, with and without correcting for population structure are shown in Figure 2A. These epistasis results are adjusted for potential main effects (codominant encoding scheme in MBMDR). The statistical significance of each pair of SNPs is assessed based on 1000 permutations and adjusting for multiple testing. Without correction for population structure we found 152 significant interacting SNP pairs. With correction for population structure, this number was reduced to 109 SNP pairs (Supplementary Table S1). Of these, 90 interacting SNP pairs were common to both approaches (Fig. 2A). In addition, 19 of the 109 significant interacting SNP pairs, detected after correction for population structure, were not detected by the epistasis analysis without such correction.

Many of the 109 significant SNP pairs resulted from variants from intergenic and long noncoding RNA (lncRNA) regions. In particular, 24 SNPs involved in the interactions are variants in lncRNAs (RP11-327F22.5 and RP11-21B23.2). Details about SNP–SNP interaction counts based on genomic position are shown in Figure 3.

In contrast, 17 unique SNPs involved in 36 pairs of significantly interacting SNPs were from coding regions. Location-wise mapping of these 17 SNPs through the human genome browser at UCSC (http://genome.ucsc.edu/) linked to 7 known genes. The associated statistical epistasis network (obtained through the *igraph* R package) shown in Figure 4A defines an

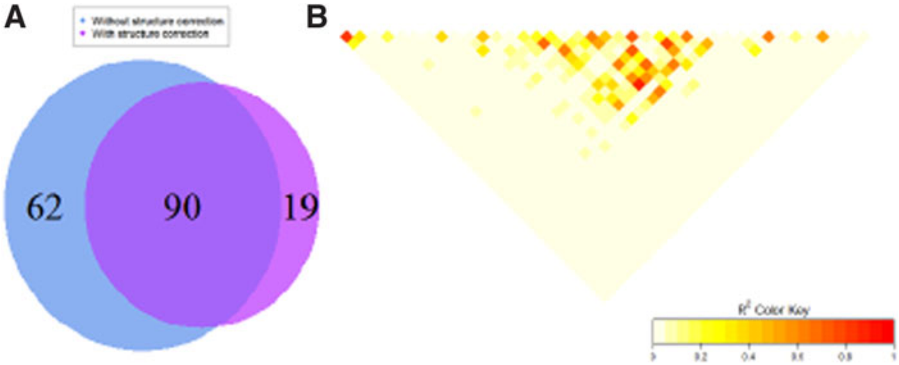

FIG. 2. Pairwise SNPs that were involved in statistically significant interactions with IBD Crohn's disease: **(A)** 109 pairs of significantly interacting SNPs with correction for population structure, 152 pairs without correction for population structure, and 90 pairs common to both approaches, and **(B)** LD plot of SNPs involved in significant interactions. IBD, inflammatory bowel disease; LD, linkage disequilibrium; SNPs, single nucleotide polymorphisms.

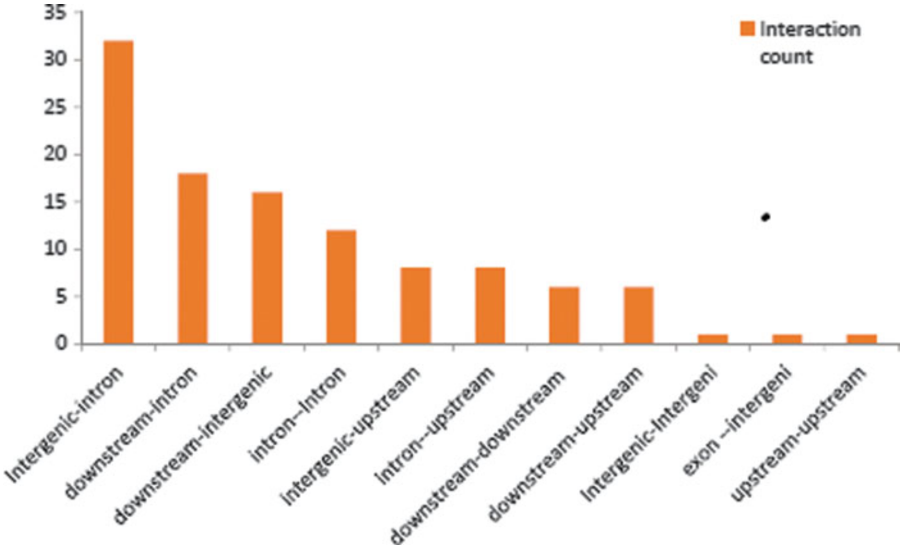

FIG. 3. SNP–SNP interaction counts relate to Crohn's disease based on genomic position. The highest frequency of interaction is found between SNPs in the intergenic and intron regions.

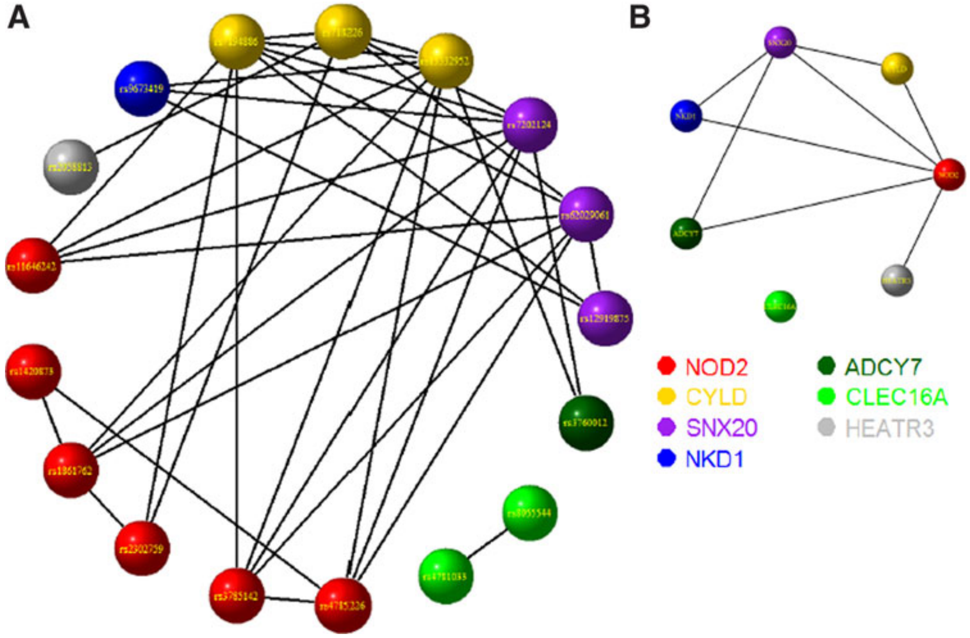

FIG. 4. **(A)** Epistasis network showing the interconnection between pairs of interacting SNPs that have a significant effect on IBD Crohn's disease. The colors show the corresponding genes mapped according to their physical position. **(B)** Gene-level network resulting from mapping the SNPs involved in the epistasis network.

edge between two nodes (2 out of 17 SNPs) if and only if the interacting SNP pair was found statistically associated with CD. The corresponding gene-level network is displayed in Figure 4B and suggests the following interacting gene pairs: NOD2 × CYLD, NOD2 × SNX20, NOD2 × ADCY7, NOD2 × HEATR3, NOD2 × NKD1, SNX20 × NKD1, CYLD × SNX20, and SNX20 × ADCY7. The suggestive interaction between NOD2 and CYLD was derived from rs1332952 × rs4785226 ($r^2 = 0.16$), rs718226 × rs2302759 ($r^2 = 0.18$) and rs7194886 × rs3785142 ($r^2 = 0.32$), among others (Fig. 2A). Notably, NOD2 is connected to all genes except CLEC16A. SNPs occurring in significant SNP × SNP interaction pairs generally did not exhibit strong LD patterns (Fig. 2B).

We next note on the computation time for the real data analysis. In general, the computation time for epistasis analysis using the MBMDR-PC method considered depends on both the number of SNPs and the sample size being analyzed. To analyze all 52,277 individuals and 1374 SNPs from the IBD CD data on a parallel workflow using 96-core computer cluster took 5 min. This computation time is only for epistasis analysis and does not include the prior analysis for extracting PCs.

Discussion

The MBMDR-PC method is a powerful tool to detect epistasis for structured populations when the population structure is adequately captured through the first few PCs. In addition, for this method, population structure can easily be accounted for in regression framework by adding PCs as covariates in the model. Because of the overall good performance (simulation results not shown in this study) of MBMDR-PC we applied it to data from the International IBD Genetics Consortium with the top five PCs, as advocated by the consortium.[19] We found 109 interacting pairs of SNPs of which 36 pairs were physically mapped to 7 pairs of potentially interacting genes, namely NOD2 versus CYLD, NOD2 versus SNX20, NOD2 versus ADCY7, NOD2 versus HEATR3, NOD2 versus NKD1, SNX20 versus NKD1, CYLD versus SNX20, and SNX20 versus ADCY7. Genes such as NOD2, CYLD, HEATR3, and ADCY7 have been reported to be involved in immune system as well as nuclear factor-κB (NF-KappaB) pathways. In particular, independent effects of NOD2 and CYLD have been demonstrated in many GWAS related to CD. In this study we have demonstrated that these genes have a potential synergistic effect on CD. The statistical interaction result between NOD2 and CYLD is supported by a recent study that indicated that CYLD restricts deposition of Lys63-Ub and Met1-Ub on the LUBAC substrate RIPK2 to limit NOD2-dependent inflammatory signaling.[25] Similarly, our finding related to the potential impact of the interaction between HEATR3 and NOD2 on CD is consistent with expression studies of HEATR3 that demonstrated a positive role in NOD2-mediated NF-κB signaling.[34] The NF-κB signaling pathway regulates the expression of hundreds of genes that are involved in diverse and key cellular and organismal processes, including cell proliferation, cell survival, the cellular stress response, innate immunity, and inflammation.[35] It is evident that unregulated activation of NF-κB plays a critical role in the pathogenesis of CD. In a recent investigation of the clinical characteristics and disease outcome of CD patients with varying levels of the NF-κB activation, Han et al.[36] demonstrated that the association of NF-κB activity with specific clinical manifestations in CD patients. However, our statistical evidence for interaction between NKD1 and SNX20 is new; to our knowledge there is no reported evidence for their physical interaction or their coexpression.

Our real-life epistasis analysis for chromosome 16 also indicated the involvement of 24 SNPs in lncRNAs to significantly interact with other SNPs in relation to CD. This links to a recent study that reported the contribution of lncRNAs in IBDs.[37] Moreover, among the significantly interacting SNPs in NOD2, rs5743291 is a nonsynonymous (missense) SNP. This SNP is also involved in changing amino acid V to I within NOD2 (UCSC Genome Browser). In contrast, based on SIFT software prediction, a V to I amino acid change is found to be deleterious possessing a tolerance index score of 0.02 (score <0.05 is predicted to be deleterious), which might result in a potential change in the interaction behavior of NOD2 with other genes for CD as well. Some reported evidence exists that the quantity and the quality of dietary protein consumption and amino acid supplementation may differentially influence the IBD course.[38]

Conclusion

In this study, we have highlighted the importance of detecting and correcting population structure in epistasis studies of complex human diseases. It has been established that not accounting for population structure due to allele frequency differences among populations and subpopulations can result in high false-positive results or reduced power in GWAS. In this regard, the proposed MBMDR-PC approach is a powerful tool to detect epistasis in genome-wide association and interaction studies from heterogeneous populations. Since the performance of the MBMDR-PC approach is highly dependent on how well PCs capture population structure, we recommend to use MBMDR-PC with either linear or nonlinear versions of PC analysis, whenever possible. MBMDR analytics should be seen as part of an entire analysis pipeline that involves making marker selection choices and performing postanalysis steps to validate and replicate findings, as well as to seek biological evidence for flagged interacting regions with MBMDR.[14] Fast implementation for multiple testing correction in exhaustive epistasis screenings[10] makes these MBMDR-based methods efficient tools for GWAIS in structured populations.

Our study 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 is devastating in GWAIS for obtaining meaningful and replicable epistasis results.

Acknowledgments

This study makes use of data generated by the International IBD Genetics Consortium. A full list of investigators who contributed to the data collection is provided at www.ibdgenetics.org/groups.html. We thank Myriam Nemry for suggestions regarding the implementation of ideas in the MBMDR software. 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) [K.V.S.]. We also acknowledge the research opportunities offered by F.N.R.S., including "Foresting in Integromics Inference" (Convention n° T.0180.13) [K.C.], and by the interuniversity research institute Walloon Excellence in Lifesciences and BIOtechnology (WELBIO) [F.A., K.V.S.].

Web Resources

MBMDR-PC is available through the MBMDR software (from version mbmdr-4.4.1 onward), which is downloadable from the BIO3 (University of Liege) website. Main options *—binary–ac number of PCs–d 2D–a CODOMINANT–rc RESIDUALS*

Author Disclosure Statement

No competing financial interests exist.

Supplementary Material

Supplementary Table S1

References

1. Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomark Prev. 2002;11:505–512.
2. 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.
3. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–909.
4. Wan X, Yang C, Yang Q, 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.
5. Cattaert T, Calle ML, Dudek SM, 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.
6. Lishout FV, Gadaleta F, Moore JH, et al. gammaMAXT: a fast multiple-testing correction algorithm. BioData Min 2015;8:36.
7. Ritchie MD, Hahn LW, Roodi N, 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.
8. Zhao Y, Chen F, Zhai R, et al. Correction for population stratification in random forest analysis. Int J Epidemiol. 2012;41:1798–1806.
9. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–575.
10. Gyenesei A, Moody J, Semple CAM, et al. High-throughput analysis of epistasis in genome-wide association studies with BiForce. Bioinformatics. 2012;28:1957–1964.
11. Zhang BY, Zhang J, Liu JS. Block-based Bayesian epistasis association mapping with application to WTCCC type 1 diabetes data. Ann Appl Stat. 2011;5:2052–2077.
12. Shang J, Zhang J, Sun Y, et al. Performance analysis of novel methods for detecting epistasis. BMC Bioinformatics. 2011;12:475.
13. 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.
14. Gusareva ES, Van Steen K. Practical aspects of genome-wide association interaction analysis. Hum Genet. 2014;133:1343–1358.
15. Wei W-H, Hemani G, Haley CS. Detecting epistasis in human complex traits. Nat Rev Genet. 2014;15:722–733.
16. Gola D, Mahachie John JM, van Steen K, et al. A roadmap to multifactor dimensionality reduction methods. Brief Bioinform. 2016;17:293–308.
17. Fouladi R, Bessonov K, Van Lishout F, et al. Model-based multifactor dimensionality reduction for rare variant association analysis. Hum Hered. 2015;79:157–167.
18. 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.
19. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491:119–124.
20. Hugot JP, Laurent-Puig P, Gower-Rousseau C, et al. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. Nature. 1996;379: 821–823.
21. Ohmen JD, Yang H-Y, Yamamoto KK, et al. Susceptibility locus for inflammatory bowel disease on chromosome 16 has a role in Crohn's disease, but not in ulcerative colitis. Hum Mol Genet. 1996;5:1679–1683.
22. Cho JH, Fu Y, Kirschner BS, et al. Confirmation of a susceptibility locus for Crohn's disease on chromosome 16. Inflamm Bowel Dis. 1997;3:186–190.
23. Liu JZ, van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47:979–986.
24. Elding H, Lau W, Swallow DM, et al. Dissecting the genetics of complex inheritance: linkage disequilibrium mapping provides insight into Crohn disease. Am J Hum Genet. 2011;89:798–805.

25. Hrdinka M, Fiil BK, Zucca M, et al. CYLD Limits Lys63- and Met1-linked ubiquitin at receptor complexes to regulate innate immune signaling. Cell Rep. 2016;14:2846–2858.
26. Cattaert T, Calle ML, Dudek SM, 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.
27. 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.
28. Abegaz F, Chaichoompu K, Génin E, et al. Principals about principal components in statistical genetics. Brief Bioinform. 2018 [Epub ahead of print]; DOI: 10.1093/bib/bby081.
29. Clayton D. snpStats: SnpMatrix and XSnpMatrix classes and methods. R Package Version 1280.
30. Clayton D, Leung H-T. An R package for analysis of whole-genome association studies. Hum Hered. 2007;64:45–51.
31. Vermeulen SH, Den Heijer M, Sham P, et al. Application of multi-locus analytical methods to identify interacting loci in case-control studies. Ann Hum Genet. 2007;71:689–700.
32. Team RC. R: A Language and Environment for Statistical Computing. Viena: R Foundation for Statistical Computing. 2019. Vienna, Austria.
33. Westfall PH, Young SS. Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment. John Wiley & Sons. 1993.
34. Zhang W, Hui KY, Gusev A, et al. Extended haplotype association study in Crohn's disease identifies a novel, Ashkenazi Jewish-specific missense mutation in the NF-κB pathway gene, HEATR3. Genes Immun. 2013;14:310–316.
35. Courtois G, Gilmore TD. Mutations in the NF-κB signaling pathway: implications for human disease. Oncogene. 2006;25:6831–6843.
36. Han YM, Koh J, Kim JW, et al. NF-kappa B activation correlates with disease phenotype in Crohn's disease. PLoS One. 2017;12:e0182071.
37. Zacharopoulous E, Gazouli M, Tzouvala M, et al. The contribution of long-coding RNAs in inflammatory bowel diseases. Dig Liver Dis. 2017;49: 1067–1072.
38. Vidal-Lletjós S, Beaumont M, Tomé D, et al. Dietary protein and amino acid supplementation in inflammatory bowel disease course: What impact on the colonic mucosa? Nutrients. 2017;9 [Epub ahead of print]; DOI: 10.3390/nu9030310.

Cite this article as: Abegaz F, Van Lishout F, Mahachie John JM, Chiachoompu K, Bhardwaj A, Gusareva ES, Wei Z, Hakonarson H, Van Steen K; on behalf of the International IBD Genetics Consortium (2019) Epistasis detection in genome-wide screening for complex human diseases in structured populations, *Systems Medicine* 2:1, 19–27, DOI: 10.1089/sysm.2019.0003.

Abbreviations Used

CD	= Crohn's disease
GLM	= generalized linear regression models
GWAS	= genome-wide association studies
GWAIS	= genome-wide association interaction studies
IBD	= inflammatory bowel disease
lncRNA	= long noncoding RNA
LD	= linkage disequilibrium
LUBAC	= Linear Ubiquitin Assembly Complex
MBMDR	= Model-Based Multifactor Dimensionality Reduction
PCs	= principal components
SNPs	= single nucleotide polymorphisms