Finding the Sources of Missing Heritability Using the Additive Epistatic Interaction Model: A Simulation Study

Yupeng Wang (wyp1125@gmail.com)
BDX Research & Consulting LLC https://orcid.org/0000-0002-3002-8069
Xinyue Liu
BDX Research & Consulting LLC
Paule V. Joseph
National Institutes of Health

Research Article

Keywords: epistatic interaction, missing heritability, simulation, complex diseases, classification, genetic risk score

Posted Date: March 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1323648/v1

Abstract

Objective

Thousands of Genome-Wide Association Studies (GWAS) have been carried out to pinpoint genetic variants associated with complex disease. However, the proportion of phenotypic variance which can be explained by the identified genetic variants is relatively low. Thus, it is desirable to propose new computational models to explain the “missing heritability” problem.

Results

Here, we propose the additive epistatic interaction model, consisting of widespread pure epistatic interactions whose effects are additive and can be summarized by a genetic risk score. Based on a simulated genotype dataset, the additive epistatic interaction model well depicted genetic risks and hereditary patterns of complex diseases. When applied to real genotypic datasets, the additive epistatic interaction model showed potential for accurately classifying human populations from the 1000 Genomes Project, and individuals with and without diabetes from the UK Biobank database. Moreover, the model’s genetic risk score can be replaced by a deep learning model which is more resistant to noises. We suggest that the additive epistatic interaction model might help to explain the “missing heritability” problem. Source code is publicly available at https://github.com/wyp1125/additive_epistasis.

Introduction

Thousands of genome-wide association studies (GWAS) have been carried out to identify genetic variants associated with complex diseases or quantitative traits [1–3]. In a GWAS, several to hundreds of genetic variants may be reported. However, they can explain only a small proportion of heritability, which is often referred to as the “missing heritability” problem [4]. Various theories have been proposed to explain the “missing heritability” problem [5]. One notable theory is epistatic interactions, referring to the phenomenon that genetic variants may show little or moderate individual effects, but strong interaction effects [6, 7]. Other theories such as rare variants [8, 9] and gene-environment interactions [10] have been also widely investigated.

Complex diseases and quantitative traits are often determined by multiple genetic variants [11, 12]. Genetic risk score (GRS), or polygenic score, which adds up the impacts of multiple genetic variants, has become a popular post-GWAS analysis approach [13, 14]. Utilization of GRS has led to summarized disease risks leveraging multiple genetic variants and environmental factors [15–17], which often show higher predictive power than using individual genetic variants. However, previous applications of GRS mainly included genetic variants with marginal effects. Therefore, GRS for epistatic interactions is underexplored. Although aggregated scores of Multifactor Dimensionality Reduction (MDR) were investigated years ago [18, 19], the computational burden of MDR has rendered it impossible to perform genome-wide scan of epistatic interactions. Because genome-wide genotype datasets, especially those

produced by next-generation sequencing technologies, are being increasingly generated, efficient models for summarizing the genetic risks of epistatic interactions are highly desirable.

To date, most computational tools for detecting epistatic interactions actually capture those epistatic interactions with marginal effects. However, in pure epistatic interactions, individual genetic variants have zero marginal effect [20, 21]. Due to computational difficulties and inadequate sample sizes, genetic variants involved in pure epistatic interactions were seldom reported in previous GWAS. However, pure epistatic interactions could be an important genetic component of common traits and diseases, which may substantially account for the "missing heritability" or be associated with causal genetic variants. Here, we propose the additive epistatic interaction model. The model assumes that the effects of multiple pure epistatic interactions are in general linear and additive, and any individual has a genetic risk score which is the number of occurrences of interacting genotypes. We use both simulated and real genotype datasets to show that the additive epistatic interaction model could be useful for assessing the genetic risks of complex diseases.

Methods

See supplementary methods.

Results

Simulation is an effective approach for evaluating new genetic models [22]. We simulated a case-control genotype dataset consisting of 1000 SNP loci and 2000 individuals (see Methods). The 2000 individuals were divided into case (i.e., disease) and control groups of equal size. Genotypes were generated according to the same guiding minor allele frequencies (MAFs) between the case and control groups. In the case group, the 1000 loci composed 500 pairs of interacting loci containing pure epistatic interactions. Therefore, all the 1000 loci should be regarded as association with the disease. The alleles of a pair of interacting loci are denoted by A/a and B/b (**a** and **b** are minor alleles) respectively. We assume that only the **aabb** genotype (i.e., both loci are homozygous for their minor alleles) has interacting effects. Under conventional genetic models, the case group has higher disease allele frequencies than the case group, whereas in this simulated dataset, no locus displays significantly different allele frequencies between the case and control groups, because the minimum *p*-value (χ^2 test) is 7.03×10^{-5} which is higher than the standard GWAS cutoff of 5×10^{-8} as well as the lenient cut-off of 10^{-5} [23] and no locus leads to a significant *p*-value after FDR adjustment. The mean and median relative risks of the **aabb** genotype among all pairs of interacting loci are 1.63 (>1, $p<2.2\times10^{-16}$, *t*-test) and 1.60 respectively, confirming that pure epistatic interactions were successfully generated.

The GRS of an individual is defined as the number of occurrences of the **aabb** genotype among the 500 pairs of interacting loci. We computed the GRS for each case and control individual. Distributions of GRS (Figure 1) clearly show that the case group tends to have higher GRS than the control group ($p<2.2\times10^{-16}$, *t*-test). Although the above comparison is informative, clinical values of GRS depend more on whether the

GRS can distinguish case (i.e., disease) and control (i.e. healthy) statuses. To this end, we computed the area under the curve (AUC), a measure of the discriminative ability ranging from 0 to 1. An AUC of 0.900 was obtained, indicating that the GRS can well depict disease risks.

The exact hereditary mechanisms of complex diseases such as cancer and heart diseases are still little understood. Population and clinical studies suggest that the offspring of healthy parents tend to have less risks of developing a complex disease than the offspring of one parent or both parents with histories of that complex disease [24–26]. We asked whether offspring's risks of developing a complex disease can be depicted by the GRS under the additive epistatic interaction model. Based on the simulated genotype dataset, we derived the genotypes of offspring belonging to three parent groups: 1) case×case, 2) case×control, and 3) control×control (see Methods). We added the GRS distributions of the three offspring groups to Figure 1. The case × case offspring tend to have higher GRS than the other two offspring groups (p=2.47×10^{-7} and p<2.2×10^{-16} respectively, t-test), and the case × control offspring tend to have higher GRS than the control × control offspring (p=1.21×10^{-4}, t-test). This analysis suggests that stronger parental histories of a complex disease do increase offspring's genetic risks of developing that disease. However, the GRS of case × case offspring are greatly reduced relative to the case group (p<2.2×10^{-16}, t-test), suggesting that even with strong parental histories of a complex disease, the genetic risks of developing that disease in offspring are much lower than the same genomes as real patients, because recombination significantly reduces the occurrences of the **aabb** genotype. This observation further indicates that for people with strong family histories of a complex disease, the actual acquirement of that disease may still largely depend on non-parental factors such as accumulation of de novo/somatic mutations and epigenetic modifications. The above analyses collectively suggest that the additive epistatic interaction model and its GRS can well depict the genetic risks and hereditary patterns of complex diseases.

Next, we asked whether the additive epistatic interaction model is applicable to real genotype data. First, we obtained the genotype data of chromosome 20 from the 1000 Genomes Project [27]. The genotype data consisted of 2504 individuals from 5 super populations including African, Ad Mixed American, East Asian, European and South Asian. Here, we treat these super populations as the phenotype, and any two super populations comprise a pair of case-control groups. We inferred that if pure epistatic interactions are widespread in the human genome, the additive epistatic interaction model can be applied to classify super populations. We employed a customized procedure (see Methods) to identify 500 pairs of SNP loci containing pure epistatic interactions for each pair of super populations. Next, we computed AUCs for pairs of super populations based on the GRS of 500 pairs of interacting loci. Note that for each pair of super populations, we exchanged case and control assignments, resulting in two AUC values. These AUCs range from 0.866 to 0.991 (Figure2), suggesting that human super populations might be accurately classified based on the additive epistatic interaction model. Second, we obtained the genotype data of chromosome 20 from the UKBioBank database [28]. We classified 567 individuals with diabetes and 573 individuals without diabetes from the Irish population. We identified 500 pairs of SNP loci containing pure epistatic interactions and computed a GRS for each individual. An AUC of 0.992 was obtained,

suggesting that the additive epistatic interaction model could accurately distinguish disease statuses of genetic disorders.

An important consideration for applying the additive epistatic interaction model is its robustness. In real data applications, especially those utilizing whole-genome sequencing technologies, the numbers of markers are much larger than sample sizes. Thus, most of the identified pairs of interacting loci may not be real. Therefore, we asked whether the additive epistatic interaction model can tolerate some extent of noises and whether advanced machine learning technologies such as deep learning can be integrated to the additive epistatic interaction model. While deep learning has different models, we chose deep neural network (DNN) because different pure epistatic interactions do not contain sequence or interdependency information. We here note that the features of DNN models should be occurrences of the **aabb** genotype, rather than original genotypes because effective feature extraction is still required for deep learning technologies.

We added different ratios of noises (i.e., loci assumed to be interacting but do not contain pure epistatic interactions) to the simulated genotype dataset, and computed AUCs of distinguishing between case and control statuses based on GRS and DNN respectively. Figure 3 shows comparisons of AUCs between GRS and DNN at different noise ratios. In general, the AUCs of both approaches gradually decrease along with increase of noise ratios. When noise ratios are relatively low (e.g., <1), GRS outperforms DNN. When the noise ratio goes up from 1, the AUCs of GRS decrease rapidly, while the AUCs of DNN show resistance to noises, leading to DNN outperforming GRS. This analysis suggests that both GRS and DNN can tolerate a substantial proportion of noises, and DNN can be integrated to the additive epistatic interaction model for dealing with complex datasets.

Discussion

The underlying genetics of complex diseases are very complicated, and a single genetic model may not capture the whole picture. Individual variants with strong marginal effects were most frequently investigated and reported in previous GWAS. Rare variants, which can be reliably assayed using exon sequencing technologies, are increasingly studied in recent years [29, 30]. In contrast, pure epistatic interactions have been seriously underexplored.

Assessing pure epistatic interactions in this study has involved rare variants with 0.01<MAF<0.05. If the sample size is big enough (say >10k), rare variants with even smaller MAFs can be included. Complex diseases may be affected by both common variants and rare variants/mutations. The “missing heritability” should not be a real problem considering that there are still many limitations on both experimental and data analysis sides. Investigation of the interplays among different genetic models/components is highly desirable in future genetic association studies for complex diseases and quantitative traits.

Limitations

The effectiveness of the additive pure epistatic interaction model is contingent on the assumption that real pure epistatic interactions can be identified. Though this genetic scenario can be easily simulated, in real genotypic datasets it is extremely difficult to identify real pure epistatic interactions due to huge numbers of genotype combinations relative to sample sizes. To further validate our models, we repeated our computations on the real genotypic datasets by selecting pure epistatic interactions only from the training datasets, which fully prevented the data leakage problem notable in the machine learning field. However, the AUCs on the testing datasets dropped to ~0.54. This validation suggested that it is essential to build the additive pure epistatic interaction model on top of real pure epistatic interactions, which may be possible for super large study cohorts (e.g. having millions of case-control balanced individuals for a disease status).

Abbreviations

GWAS: Genome-wide association study

GRS: Genetic risk score

MDR: Multifactor dimensionality reduction

DNN: Deep neural network

MAF: Minor allele frequency

AUC: Area under the curve

Declarations

Availability of data and materials

Data of the 1000 Genomes Project were downloaded from the NCBI ftp site (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/). Data of UKBioBank(https://www.ukbiobank.ac.uk)were obtained under Application Number 57780. Source code is publicly available at https://github.com/wyp1125/additive_epistasis.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Funding

PVJ is supported by the National Institute of Nursing Research under award number 1ZIANR000035-01. PVJ is also supported by the Office of Workforce Diversity, National Institutes of Health, and the Rockefeller University Heilbrunn Nurse Scholar Award.

Authors' contributions

Y.W. and P.V.J conceived the study. Y.W. and X.L. made the programs. Y.W. and X.L. performed the analyses. Y.W. and P.V.J wrote the manuscript.

Acknowledgements

We thank Ying Sun for valuable discussion of this research.This research has been conducted using the UK Biobank Resource under Application Number 57780.

References

1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J: **10 Years of GWAS Discovery: Biology, Function, and Translation**. *Am J Hum Genet* 2017, **101**(1):5-22.
2. Visscher PM, Brown MA, McCarthy MI, Yang J: **Five years of GWAS discovery**. *American journal of human genetics* 2012, **90**(1):7-24.
3. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J *et al*: **The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)**. *Nucleic acids research* 2017, **45**(D1):D896-D901.
4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A *et al*: **Finding the missing heritability of complex diseases**. *Nature* 2009, **461**(7265):747-753.
5. Maher B: **Personal genomes: The case of the missing heritability**. *Nature* 2008, **456**(7218):18-21.
6. Cordell HJ: **Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans**. *Human molecular genetics* 2002, **11**(20):2463-2468.
7. Phillips PC: **Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems**. *Nature reviews Genetics* 2008, **9**(11):855-867.
8. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES: **Searching for missing heritability: designing rare variant association studies**. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(4):E455-464.
9. Bandyopadhyay B, Chanda V, Wang Y: **Finding the Sources of Missing Heritability within Rare Variants Through Simulation**. *Bioinformatics and biology insights* 2017, **11**:1177932217735096.

10. Kaprio J: **Twins and the mystery of missing heritability: the contribution of gene-environment interactions**. *Journal of internal medicine* 2012, **272**(5):440-448.
11. Marchini J, Donnelly P, Cardon LR: **Genome-wide strategies for detecting multiple loci that influence complex diseases**. *Nature genetics* 2005, **37**(4):413-417.
12. Yang Q, Khoury MJ, Friedman J, Little J, Flanders WD: **How many genes underlie the occurrence of common complex diseases in the population?***International journal of epidemiology* 2005, **34**(5):1129-1137.
13. Cooke Bailey JN, Igo RP, Jr.: **Genetic Risk Scores**. *Current protocols in human genetics* 2016, **91**:1 29 21-21 29 29.
14. Sugrue LP, Desikan RS: **What Are Polygenic Scores and Why Are They Important?***Jama* 2019, **321**(18):1820-1821.
15. Torkamani A, Wineinger NE, Topol EJ: **The personal and clinical utility of polygenic risk scores**. *Nature reviews Genetics* 2018, **19**(9):581-590.
16. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, Tyrer JP, Chen TH, Wang Q, Bolla MK *et al*: **Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes**. *American journal of human genetics* 2019, **104**(1):21-34.
17. Knowles JW, Ashley EA: **Cardiovascular disease: The rise of the genetic risk score**. *PLoS medicine* 2018, **15**(3):e1002546.
18. Li CF, Luo FT, Zeng YX, Jia WH: **Weighted risk score-based multifactor dimensionality reduction to detect gene-gene interactions in nasopharyngeal carcinoma**. *International journal of molecular sciences* 2014, **15**(6):10724-10737.
19. Dai H, Charnigo RJ, Becker ML, Leeder JS, Motsinger-Reif AA: **Risk score modeling of multiple gene to gene interactions using aggregated-multifactor dimensionality reduction**. *BioData mining* 2013, **6**(1):1.
20. Jiang X, Neapolitan RE: **Mining pure, strict epistatic interactions from high-dimensional datasets: ameliorating the curse of dimensionality**. *PloS one* 2012, **7**(10):e46771.
21. Culverhouse R, Suarez BK, Lin J, Reich T: **A perspective on epistasis: limits of models displaying no main effect**. *American journal of human genetics* 2002, **70**(2):461-471.
22. Hoban S, Bertorelle G, Gaggiotti OE: **Computer simulations: tools for population and evolutionary genetics**. *Nature reviews Genetics* 2012, **13**(2):110-122.
23. Panagiotou OA, Ioannidis JP, Genome-Wide Significance P: **What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations**. *International journal of epidemiology* 2012, **41**(1):273-286.
24. Johns LE, Houlston RS: **A systematic review and meta-analysis of familial prostate cancer risk**. *BJU international* 2003, **91**(9):789-794.
25. Stratton JF, Pharoah P, Smith SK, Easton D, Ponder BA: **A systematic review and meta-analysis of family history and risk of ovarian cancer**. *British journal of obstetrics and gynaecology* 1998,

105(5):493-499.

26. Win AK, Reece JC, Ryan S: **Family history and risk of endometrial cancer: a systematic review and meta-analysis**. *Obstetrics and gynecology* 2015, **125**(1):89-98.
27. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA *et al*: **A global reference for human genetic variation**. *Nature* 2015, **526**(7571):68-74.
28. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J *et al*: **The UK Biobank resource with deep phenotyping and genomic data**. *Nature* 2018, **562**(7726):203-209.
29. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, O'Dushlaine C, Chambert K, Bergen SE, Kahler A *et al*: **A polygenic burden of rare disruptive mutations in schizophrenia**. *Nature* 2014, **506**(7487):185-190.
30. Raghavan NS, Brickman AM, Andrews H, Manly JJ, Schupf N, Lantigua R, Wolock CJ, Kamalakaran S, Petrovski S, Tosto G *et al*: **Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer's disease**. *Annals of clinical and translational neurology* 2018, **5**(7):832-842.

Figures

Figure 1

Comparisons of GRS distributions among case, control and different offspring groups.

Figure 2

AUCs of using the additive epistatic interaction model to classify human super populations.

Figure 3

Comparisons of AUCs between GRS and DNN at different noise ratios.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMethods.docx